ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Multidisciplinary Innovation in Engineering Science & Technology

2020

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
(Special Issue)

Editor-in-Chief

Prof. Passerini Kazmerski
Pritzker School of Molecular Engineering, University of Chicago, USA

Guest Editors

Prof. Wang Xiu Ying
Chongqing University, China

Prof. Yu Xiao Yan
Chongqing Normal University, China

Prof. María Jesús Espinosa Trujillo
Universidad Tecnológica Metropolitana, Mexico

Prof. Ahmad Yusairi Bani Hashim
Universiti Teknikal Malaysia Melaka, Malaysia

Prof. Mohamed Abdelaziz Hassan Eleiwa
University of Hail, KSA

Prof. Nicolae Tudoroiu
John Abbott College, Canada

Editorial

The Special Issue on *Multidisciplinary Innovation in Engineering Science & Technology (2020)* in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* presents a compelling collection of research that reflects the growing importance of innovation driven by multidisciplinary collaboration. The year 2020 marked a period in which the convergence of diverse scientific and engineering domains became increasingly essential for addressing complex and rapidly evolving global challenges. This special issue captures that shift by bringing together contributions that integrate knowledge, methodologies, and technologies from multiple disciplines to advance both research and practical applications.

A key theme across the papers in this issue is the role of innovation at the intersection of engineering science and emerging technologies. The contributions highlight how advancements in computational methods, artificial intelligence, and data analytics are being combined with traditional engineering principles to create more efficient, adaptive, and intelligent systems. These developments are evident in applications such as smart infrastructure, advanced manufacturing, healthcare technologies, and environmental systems, where multidisciplinary approaches enable more comprehensive and effective solutions.

The issue also emphasizes the significance of collaborative research in fostering technological breakthroughs. By integrating expertise from fields such as mechanical, electrical, civil, and computer engineering, along with applied sciences, researchers are able to address complex problems that cannot be solved within a single domain. The studies included demonstrate how such collaboration leads to the development of innovative models, improved system designs, and enhanced analytical frameworks. This multidisciplinary synergy not only accelerates the pace of innovation but also broadens the scope of potential applications.

Another important aspect highlighted in this collection is the focus on sustainability and resilience. The challenges faced in 2020 underscored the need for engineering solutions that are robust, flexible, and environmentally responsible. The research presented explores strategies for optimizing resource utilization, integrating renewable energy systems, and designing infrastructure that can adapt to changing conditions. These contributions reflect a growing commitment to aligning technological advancement with sustainable development goals and societal needs.

Advances in digital technologies and their integration into engineering systems are also prominently featured. The use of simulation tools, digital twins, and real-time monitoring systems has transformed the way engineering processes are designed and managed. The papers in this issue illustrate how these technologies enable more accurate predictions, improved decision-making, and enhanced system performance. This digital transformation is a key driver of innovation, enabling the development of interconnected and intelligent systems across various sectors.

The editorial team expresses its sincere gratitude to all authors, reviewers, and contributors whose dedication and expertise have made this special issue possible. Their efforts have resulted in a high-quality compilation of research that advances multidisciplinary innovation in engineering science and technology. The works presented in this issue highlight the increasing importance of integrated approaches in driving technological progress, offering valuable insights and inspiration for future research and development in an increasingly interconnected world.

Guest Editor

Prof. María Jesús Espinosa Trujillo

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **April 2020**

CONTENTS

Mitigating Congestion in Restructured Power System using FACTS Allocation by Sensitivity Factors and Parameter Optimized by GWO
by Anubha Gautam, Parshram Sharma and Yogendra Kumar

Offline Signature Recognition and Verification Using ORB Key Point Matching Techniques
by Aravinda Chickmaglore Venkataramu, Atsumi Masahiko, Akshaya, Amar Prabhu Gurupura, Udaya Kumar Reddy Kyasambally and Rajashekar

Qualitative Properties of a Cell Proliferating Model with Multi-phase Transition and Age Structure
by Youssef El Alaoui and Larbi Alaoui

Standardized UCI-EGO Dataset for Evaluating 3D Hand Pose Estimation on the Point Cloud
by Sinh-Huy Nguyen and Van-Hung Le

A Novel Representative k-NN Sampling-based Clustering Approach for an Effective Dimensionality Reduction-based Visualization of Dynamic Data
by Dharamsotu Bheekya, Kanakapodi Swarupa Rani, Salman Abdul Moiz and Chillarige Raghavendra Rao

A Study on Intelligent Dialogue Agent for Older Adults' Preventive Care – Towards Development of a Comprehensive Preventive Care System
by Sho Hirose, Daisuke Kitakoshi, Akihiro Yamashita, Kentarou Suzuki and Masato Suzuki

Synthesis and Characterization of Graphene Oxide Under Different Conditions, and a Preliminary Study on its Efficacy to Adsorb Cu2+
by Olayinka Oluwaseun Oluwasina, Surjyakanta Rana, Sreekantha Babu Jonnalagadda and Bice Susan Martincigh

Case Study to Determine the Causes of Fire in Agriculture
by Marianna Tomašková, Darina Matisková and Michaela Balážiková

Automatic Comprehension and Summarisation of Legal Contracts
by Sibusiso Kubeka and Abejide Ade-Ibijola

Supervised Learning Techniques for Stress Detection in Car Drivers
by Pamela Zontone, Antonio Affanni, Riccardo Bernardini, Leonida Del Linz, Alessandro Piras and Roberto Rinaldo

Analysis of Green Building Effect on Micro grid Based on Potential Energy Savings and BIM
by Ihsan Mizher Baht, Petre Marian Nicolae, Ileana Diana and Nameer Baht

Effective Segmented Face Recognition (SFR) for IoT

by Fei Gao and Jiangjiang Liu

Using Leader Election and Blockchain in E-Health
by Basem Assiri

Organizational, Social and Individual Aspect on Acceptance of Computerized Audit in Financial Audit Work
by Bambang Leo Handoko, Nada Ayuanda and Ari Tihar Marpaung

Fast Stream Cipher based Chaos Neural Network for Data Security in CAN Bus
by Zhongda Liu, Takeshi Murakami, Satoshi Kawamura and Hitoaki Yoshida

Low Power Bulk Driven Series Parallel OTA for Low Frequency Applications
by Sushma Padubidri Shivaprasad and Sreemannarayanay Kulkarni

Automated Extraction of Heavyweight and Lightweight Models of Urban Features from LiDAR Point Clouds by Specialized Web-Software
by Sergiy Kostrikov, Rostyslav Pudlo, Dmytro Bubnov, Vladimir Vasiliev and Yury Fedyay

Power Loss Minimization using the Integration of DGs and Reconfiguration of Distribution System: Applied on Real Distribution Feeder of Urbain Areas of Kenitra City in Morroco
by Ismail Moufid, Soukaina Naciri, Hassan EL Moussaoui, Tijani Lamhamdi and Hassane El Markhi

Method of Modelling Prices for R&D Products in the Case of their Transfer from Engineering Universities to the Business
by Oleksandra Mrykhina, Lidiya Lisovska, Ihor Novakivskyj, Terebukh and Valentyna Zhukovska

Maximum Power-Point Tracking and Stall Control with Eddy Current Brake System on Small-Scaled Wind Turbines and its Application on Agricultural Harvesting
by Anupa Koswatta, Faramarz Alsharif, Yasushi Shiroma, Shiro Tamaki and Junji Tamura

Transient Analysis of a Line-Start Synchronous Reluctance Motor with Symmetrical Distributed Brass Rotor Bars
by Mbika Muteba

Using TOST in Teaching Operating Systems and Concurrent Programming Concepts
by Tzanko Golemanov and Emilia Golemanova

Personality Measurement Design for Ontology Based Platform using Social Media Text
by Andry Alamsyah, Sri Widiyanesti, Rizqy Dwi Putra and Puspita Kencana Sari

Experimental Studies of the Silicon Photomultiplier Readout Electronics Based on the Array Chip MH2XA030
by Oleg Dvornikov, Vladimir Tchekhovski, Yaroslav Galkin, Alexei Kunz and Nikolay Prokopenko

Optimization of Dual Motion Mechanism with Double Grooved Cams for High-voltage Gas Circuit Breaker
by Masanao Terada, Yuki Nakai, Hiroaki Hashimoto, Daisuke Ebisawa, Hajime Urai and Yasunobu Yokomizu

The Impact Assessment of the Errors in Determining the Mass and Zero Lift-Drag Coefficient on the Aircraft's Performance Data
by Klyagin Viktor Anatolievich and Laushin Dmitry Andreevich

Design of an EEG Acquisition System for Embedded Edge Computing
by Kanishk Rai, Keshav Kumar Thakur, Preethi K Mane and Narayan Panigrahi

Infrared Uplink Implementation for Software Defined Visible Light Communication Systems
by Oswaldo René Banda-Sayco

Downlink Indoor Coverage Performance of Unmanned Aerial Vehicle LTE Base Stations
by Mahmut Demirtas, Kerem C¸agdas¸ Durmus¸, Gülçín Tanıs¸, Caner Arslan and Metin Balcı

Composition of Methods to Ensure Iris Liveness and Authenticity
by Ali Al-Rashid

Balancing Exploration-Exploitation in the Set Covering Problem Resolution with a Self-adaptive Intelligent Water Drops Algorithm
by Broderick Crawford, Ricardo Soto, Gino Astorga, José Lemus-Romani, Sanjay Misra, Mauricio Castillo, Felipe Cisternas-Caneo, Diego Tapia and Marcelo Becerra-Rozas

Performance Analysis of Joint Precoding and Equalization Design with Shared Redundancy for Imperfect CSI MIMO Systems
by Bui Quoc Doanh, Ta Chi Hieu, Truong Sy Nam, Pham Thi Phuong Anh and Pham Thanh Hiep

Electro-tactile Stimulation for Augmenting Finger Motoric Learning
by Daniel Sutopo Pamungkas and Arjon Turnip

Applications of TCAD Simulation Software for Fabrication and study of Process Variation Effects on Threshold Voltage in 180nm Floating-Gate Device
by Thinh Dang Cong, Toi Le Thanh, Hao Mai Tri, Phuc Ton That Bao and Trang Hoang

Aviation MRO: Impact of Physical Environment Factors on Job Performance in Aircraft Maintenance Organization
by Kamal Jaiswal, Serdar Dalkilic, Evangelos Papageorgiou and Balgopal Singh

Design of Purebred Dog Recommendation System Using MCDM Approach
by Phie Chyan

Four-Dimensional Sparse Data Structures for Representing Text Data
by Martin Marinov and Alexander Efremov

Nearest Neighbour Search in k-dSLst Tree

by Meenakshi Hooda and Sumeet Gill

Research of the Effect of Rotation and Low-Frequency Vibration on the Robotic Assembly Process
by Mikhail Vladimirovich Vartanov, Trung Ta Tran and Van Dung Nguyen

Modelling and Simulation of Aerodynamic Parameters of an Airship
by Anoop Sasidharan Pillai and Venkata Ramana Murthy Oruganti

Industry 4.0 Operators: Core Knowledge and Skills
by Olayan Alharbi

Contextualization of the Augmented Reality Quality Model through Social Media Analytics
by Jim Scheibmeir and Yashwant Malaiya

A Toolkit for the Automatic Analysis of Human Behavior in HCI Applications in the Wild
by Andrea Generosi, Silvia Ceccacci, Samuele Faggiano, Luca Giraldi and Maura Mengoni

Real-Time Identification and Classification of Driving Maneuvers using Smartphone
by Munaf Salim Najim Al-Din

Trace-Driven Simulation of LoRaWAN Air Channel Propagation in an Urban Scenario
by Eugen Harinda, Hadi Larijani and Ryan M. Gibson

Multi-Directional Light Sensing Using A Rotating Sensor
by Hoang Anh Dung, Nguyen Manh Cuong and Nguyen Phan Kien

Dynamic Decision-Making Process in the Opportunistic Spectrum Access
by Mahmoud Almasri, Ali Mansour, Christophe Moy, Ammar Assoum, Denis Lejeune and Christophe Osswald

Applicability of Generalized Metropolis-Hastings Algorithm to Estimating Aggregate Functions in Wireless Sensor Networks
by Martin Kenyeres and Jozef Kenyeres

Sensorless Control and Corrected Error Commutation of the Brushless DC Motor
by Anatoly Stepanov and Vitaly Enin

Assessing Heutagogical Elements in Learning of Engineering Education: Instrument Validation
by Mimi Mohaffyza Mohamad, Alias Masek, Jailani Md Yunos, Maizam Alias, Nor Hidayah Hamdan and Andika Bagus Nur Rahma Putra

IT GRC Smart Adviser: Process Driven Architecture Applying an Integrated Framework
by Meriyem Chergui and Aziza Chakir

Improved Nonlinear Fuzzy Robust PCA for Anomaly-based Intrusion Detection
by Amal Hadri, Khalid Chougdali and Raja Touahni

Differential Evolution based Hyperparameters Tuned Deep Learning Models for Disease Diagnosis and Classification
by Sathyabama Kaliyapillai and Saruladha Krishnamurthy

Multi Operated Virtual Power Plant in Smart Grid
by Yevhen Fediv, Olha Sivakova and Mykhailo Korchak

Enhanced Collaborative Constellation for Visible Light Communication System
by Manh Le Tran and Sunghwan Kim

A Novel Demand Side Management by Minimizing Cost Deviation
by Vikas Anand Vatul, Arputha Aravinth, Narayanan K, Gulshan Sharma and Tomonobu Senjyu

Trajectory Tracking Control of a DC Motor Exposed to a Replay-Attack
by Reda El Abbadi and Hicham Jamouli

A Comprehensive Review of Traditional Video Processing
by Helen Kottarathil Joy and Manjunath Ramachandra Kounte

Overmind: A Collaborative Decentralized Machine Learning Framework
by Puttakul Sakul-Ung, Amornvit Vatcharaphrueksadee, Pitiporn Ruchanawet, Kanin Kearpimy, Hathairat Ketmaneechairat and Maleerat Maliyaem

Based on Reconfiguring the Supercomputers Runtime Environment New Security Methods
by Andrey Molyakov

Cluster Centroid-Based Energy Efficient Routing Protocol for WSN-Assisted IoT
by Nalluri Prophess Raj Kumar and Josemin Bala Gnanadhas

Complex Order PIα+jβDγ+jθ Design for Surface Roughness Control in Machining CNT Al-Mg Hybrid Composites
by Ravi Sekhar, Tejinder Paul Singh and Pritesh Shah

Economic and Environmental Analysis of Life Expectancy in China and India: A Data Driven Approach
by Nittaya Kerdprasop, Kittisak Kerdprasop and Paradee Chuaybamroong

Deep Learning Model for A Driver Assistance System to Increase Visibility on A Foggy Road
by Samir Allach, Mohamed Ben Ahmed and Anouar Abdelhakim Boudhir

ISR Data Processing in Military Operations
by Ladislav Burita and Ales Novak

A Hybrid Approach for Intrusion Detection using Integrated K-Means based ANN with PSO Optimization
by Jesuretnam Josemila Baby and James Rose Jeba

Logistics Solutions in the Preparation Phase for the Appearance of Disasters

by Erika Barojas-Payán, Diana Sánchez-Partida, Miguel-Josué Heredia-Roldan, Victorino Juárez-Rivera and Jesús Medina-Cervantes

Multi-Model Security and Social Media Analytics of the Digital Twin
by Jim Scheibmeir and Yashwant Malaiya

Distributed Microphone Arrays, Emerging Speech and Audio Signal Processing Platforms: A Review
by Shahab Pasha, Jan Lundgren, Christian Ritz and Yuexian Zou

The Contribution of Wind Energy Capacity on Generation Systems Adequacy Reliability using Differential Evolution Optimization Algorithm
by Athraa Ali Kadhem, Noor Izzri Abdul Wahab and Ahmed Abdalla

Computational Intelligence and Statistical Learning Performances on Predicting Dengue Incidence using Remote Sensing Data
by Nittaya Kerdprasop, Kittisak Kerdprasop and Paradee Chuaybamroong

DC-DC Buck Converter Driver with Variable Off-Time Peak Current Mode Control
by Osvaldo Gasparri, Paolo Del Croce and Andrea Baschirotto

Design and Optimization of a Three Stage Electromechanical Power Unit using Numerical Methods
by Yashwant Kolluru, Rolando Doelling and Lars Hedrich

Blueprint Model: An Agile-Oriented Methodology for Tackling Global Software Development Challenges
by Andre Figliuolo da Cruz, Cristiano Pereira Godoy, Lanier Menezes dos Santos, Lucas Frota Marinho, Marco Santarelle Jardim, Elisangela Paiva da Silva, C´ıcero Augusto Pahins, Paulo Fonseca and Felipe Taliar Giuntini

Bilateral Communication Device for Deaf-Mute and Normal People
by Raven Carlos Tabiongan
Semiclassical Theory for Bacteria Motility Under External Electric Fields and Interactions with Nanodevices
by Huber Nieto-Chaupis

Comparison of Gaze Points Among Viewing Conditions (Resolution, Display Size, Viewer Position) During Video Viewing
by Miho Shinohara, Yusuke Nosaka, Reiko Koyama, Riko Nakagawa, Takuya Sarugaku and Mitsuho Yamada

Environmental Entrepreneurship as an Innovation Catalyst for Social Change: A Systematic Review as a Basis for Future Research
by Carol Dineo Diale, Mukondeleli Grace Kanakana-Katumba and Rendani Wilson Maladzhi

Enhancing Decision Making Capabilities in Humanitarian Logistics by Integrating Serious Gaming and Computer Modelling
by Za'aba Bin Abdul Rahim, Giuseppe Timperio, Robert de Souza and Linda William

To the Question of Multi-Criteria Optimization of Aircraft Components in Order to Optimize its Life Cycle
by Sergey Alekseevich Serebryansky and Alexander Vladimirovich Barabanov

A Framework for Measuring Workforce Agility: Fuzzy Logic Approach Applied in a Moroccan Manufacturing Company
by Fadoua Tamtam and Amina Tourabi

A Hybrid Mod
by Chalinee Partanapat, Chuleerat Jaruskulchai and Chanankorn Jandaeng

A Simulation Based Proactive Approach for Smart Capacity Estimation in the Context of Dynamic Positions and Events
by Naeem Ahmed Haq Nawaz, Hamid Raza Malik, Ahmed Jaber Alshaor and Kamran Abid

Effective Learning of Tax Regulations using Different Chatbot Techniques
by Rafael Mellado-Silva, Antonio Faúndez-Ugalde and María Blanco-Lobos

Newton-Raphson Algorithm as a Power Utility Tool for Network Stability
by Lambe Mutalub Adesina, Ademola Abdulkareem, James Katende and Olaosebikan Fakolujo

Electronic Warfare Methods Combatting UAVs
by Miroslav Kratky, Vaclav Minarik, Michal Sustr and Jan Ivan

A Solution Applying the Law on Road Traffic into A Set of Constraints to Establish A Motion Trajectory for Autonomous Vehicle
by Quach Hai Tho, Huynh Cong Phap and Pham Anh Phuong

Bayes Classification and Entropy Discretization of Large Datasets using Multi-Resolution Data Aggregation
by Safaa Alwajidi and Li Yang

Short-Term Dynamic Exchange Rate Model: IFEER Concept Development
by Anton Kuzmin

Extending the Classifier Algorithms in Machine Learning to Improve the Performance in Spoken Language Understanding Systems Under Deficient Training Data
by Sheetal Jagdale and Milind Shah

Spot Toyota: Design and Development of a Mobile Application for Toyota's Promotion Actions to the Young Audience
by Nuno Martins and Joel Enes

The Sound of Trust: Towards Modelling Computational Trust using Voice-only Cues at Zero-Acquaintance
by Deborah Ooi Yee Hui, Syaheerah Lebai Lutfi, Syibrah Naim, Zahid Akhtar, Ahmad Sufril Azlan Mohamed and Kamran Siddique

Reliability Improvement of Radial Distribution System by Reconfiguration
by Srividhya. P, Mounika. K, Kirithikaa. S, Narayanan. K, Gulshan Sharma, Girish Ganesan. R and Tomonobu Senjyu

American Sign Language Recognition Based on MobileNetV2
by Kin Yun Lum, Yeh Huann Goh and Yi Bin Lee

Neural Network-Based Fault Diagnosis of Joints in High Voltage Electrical Lines
by Marco Bindi, Igor Aizenberg, Riccardo Belardi, Francesco Grasso, Antonio Luchetta, Stefano Manetti and Maria Cristina Piccirilli

Design of Interactive Aids for Children's Teeth Cleaning Habits
by Cheng Chuko, Fang-Lin Chao and Hsin-Yu Tsai

An Economic Theory Perspective for the Fight Against Poverty in the Peruvian Andes
by Robert Antonio Romero-Flores

Remote Control of Garden Plantation Water Pumps using Arduino and GSM Mobile
by Beza Negash Getu, Mohamed Abdulkadir and Michael Tous

Decision Making System for Improving Firewall Rule Anomaly Based on Evidence and Behavior
by Suchart Khummanee, Phatthanaphong Chomphuwiset and Potchara Pruksasri

Growth Models and Age Estimation of Rice using Multitemporal Vegetation Index on Landsat 8 Imagery
by Abdi Sukmono, Arief Laila Nugraha, Arsyad Nur Ariwahid and Nida Shabrina

Genetic Organization and Evolution of Electromechanical Objects with Adaptive Geometry of Active Zone
by Vasyl Shynkarenko, Ali Makki, Viktoriia Kotliarova, Anna Shymanska and Pavlo Krasovskyi

IoT Based Human Activity Recognition System Using Smart Sensors
by Deepti Sehrawat and Nasib Singh Gill

The Capacity Factor of Renewable Energy Power Plants During Electric Power Peak Times in Jeju Island
by Zulmandakh Otgongerel, Gaemyoung Lee and Ankhzaya Baatarbileg

Efficient Discretization Approaches for Machine Learning Techniques to Improve Disease Classification on Gut Microbiome Composition Data
by Hai Thanh Nguyen, Nhi Yen Kim Phan, Huong Hoang Luong, Trung Phuoc Le and Nghi Cong Tran

The Newtonian Model of the Smolensk Catastrophe
by Józef Pawelec

Application of EARLYBREAK for Line Segment Hausdorff Distance for Face Recognition
by Chau Dang-Nguyen and Tuan Do-Hong

A Fuzzy-PID Controller Combined with PSO Algorithm for the Resistance Furnace
by Trinh Luong Mien, Vo Van An and Bui Thanh Tam

Warehouse Relocation of a Company in the Automotive Industry Using P-median
by Zarate-Zapata, Aldo Cesar, Garzón-Garnica, Eduardo Arturo, Cante-Mota, Román, Olmos-Álvarez, Fernando, Martinez-Flores, José Luis, Sánchez-Partida and Diana

Solutions for Building a System to Support Motion Control for Autonomous Vehicle
by Quach Hai Tho, Huynh Cong Phap and Pham Anh Phuong

Review of Pedagogical Principles of Cyber Security Exercises
by Mika Karjalainen and Tero Kokkonen

Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques
by Moulay Smail Bouzakraoui, Abdelalim Sadiq and Abdessamad Youssfi Alaoui

Real-Time Traffic Sign Detection and Recognition System for Assistive Driving
by Adonis Santos, Patricia Angela Abu, Carlos Oppus and Rosula Reyes

A Typological Study of Portuguese Mortality from Non-communicable Diseases
by Ana Paula Nascimento, Cristina Prudêncio, Mónica Vieira, Rui Pimenta and Helena Bacelar-Nicolau

The Design Process in the Improvement of The Experience Between a Brand and its Target Audience Through a Digital Product: The Lexus Portugal's used Car Website Case Study
by Nuno Martins, Juan-Ramon Martin-Sanroman and Fernando Suárez-Carballo

Finding Association Patterns of Disease Co-occurrence by using Closed Association Rule Generation
by Panida Songram, Phattanaphong Chompowiset and Chatklaw Jareanpon

Automatic Stochastic Dithering Techniques on GPU: Image Quality and Processing Time Improved
by Giorgia Franchini, Roberto Cavicchioli and Jia Cheng Hu

Social-Interactive Learning Concept Used for Electronic Educational Resource "Post-Graduate Foreign Language" and the Obtained Learning Curve
by Natalya Chernova, Victor Chernov, Margarita Emelianova, Raisa Akhunzianova and Danil Sukhopluev

Correlation-Based Incremental Learning Network for Gas Sensors Drift Compensation Classification
by Panida Lorwongtrakool and Phayung Meesad

Improved System Based on ANFIS for Determining the Degree of Polymerization

by Marcel Nicola, Marian Duță, Maria-Cristina Nițu, Ancuța-Mihaela Aciu and Claudiu-Ionel Nicola

Investigation of Dielectric Properties of Indigenous Blended Ester Oil for Electric System Applications
by D.M. Srinivasa, Usha Surendra and V.V. Pattanshetti

Fuzzy Recognition by Logic-Predicate Network
by Tatiana Kosovskaya

Hand-Based Biometric Recognition Technique – Survey
by Katerina Prihodova and Miloslav Hub

Optimal Hydrokinetic Turbine Array Placement in Asymmetric Quasigeostrophic Flows
by Victoria Monica Miglietta and Manhar Dhanak

An Innovative Angle of Attack Virtual Sensor for Physical-Analytical Redundant Measurement System Applicable to Commercial Aircraft
by Antonio Vitale, Federico Corraro, Nicola Genito, Luca Garbarino and Leopoldo Verde

Newton-Euler Based Dynamic Modeling and Control Simulation for Dual-Axis Parallel Mechanism Solar Tracker
by Sarot Srang, Sopagna Ath and Masaki Yamakita

Adaptive Identification Method of Vehicle Model for Autonomous Driving Robust to Environmental Disturbances
by Yohei Yamauchi and Mitsuyuki Saito

Eliminating Target Anopheles Proteins to Non-Target Organisms based on Posterior Probability Algorithm
by Marion Olubunmi Adebiyi and Oludayo Olufolorunsho Olugbara

Johnson Noise and Optical Characteristics of Polymer Nanocomposites Based on Colloidal Quantum Dots and In-Situ Nanoparticles Formation
by Fatin Hana Naning, S. Malik, Lee Feng Koo, Tze Jin Wong and Pang Hung Yiu

Transient Response & Electromagnetic Behaviour of Flexible Bow-Tie Shaped Chip-less RFID Tag for General IoT Applications
by Muhammad Usman Ali Khan, Raad Raad and Javad Foroughi

Optimization of Multi-user Face Identification Systems in Big Data Environments
by Majdouline Meddad, Chouaib Moujahdi, Mounia Mikram and Mohammed Rziza

Surge Protection Device for Ex Application
by Teik Hua Kuan, Kuew Wai Chew and Kein Huat Chua

A Method for Detecting Human Presence and Movement Using Impulse Radar
by Young-Jin Park and Hui-Sup Cho

Using Classic Networks for Classifying Remote Sensing Images: Comparative Study
by Khalid A. AlAfandy, Hicham Omara, Mohamed Lazaar and Mohammed Al Achhab

Ethics as a Motivation Indicator in Second Language Vocational Digital Teaching
by Natalya Viktorovna Matveeva and Ludmila Vladimirovna Makar

A Novel Quantum No-Key Protocol for Many Bits Transfer with Error Correction Codes
by Duc Manh Nguyen and Sunghwan Kim

Development of a Wireless Displacement Estimation System Using IMU-based Device
by Tri Nhut Do, Quang Minh Pham, Hoa Binh Le-Nguyen, Cao Tri Nguyen and Hai Minh Nguyen-Tran

Optimal Irrigation Strategy using Economic Model Predictive Control
by Luisella Balbis

Design and Implementation of Quad-Site Testing on FPGA Platform
by Basavaraj Rabakavi and Saroja V Siddamal

Non-Ridership Presumption Toward New Bus Rapid Transit Purwokerto-Purbalingga Execution
by Fauzan Romadlon

Mentoring Model in an Active Learning Culture for Undergraduate Projects
by Wongpanya Nuankaew, Kanakarn Phanniphong, Sittichai Bussaman, Direk Teeraputon and Pratya Nuankaew

Innovative Solution for Parking-Sharing of Private Institutions Using Various Occupancy Tracking Methods
by Adrian Florea, Valentin Fleaca and Simona Daniela Marcu

Ferromagnetic Core Reactor Modeling and Design Optimization
by Subash Pokharel and Aleksandar Dimitrovski

Performance Analysis of Grid-Connected PV Rooftop, at Sakon Nakhon Province, Thailand
by Supalak Sathiracheewin, Patamaporn Sripadungtham and Settakorn Kamuang

Innovative Course Delivery using Analyze – Group – Design – Optimize (AGDO) Methodology: Case Study of Entity-Relationship Model
by Aparna Sharma, Rishabh Singh, Prathamesh Churi and Mahesh Mali

Fine Tuning the Performance of Parallel Codes
by Sanaz Gheibi, Tania Banerjee, Sanjay Ranka and Sartaj Sahni

Deaf Chat: A Speech-to-Text Communication Aid for Hearing Deficiency
by Mandlenkosi Shezi and Abejide Ade-Ibijola

Effect of Heat Retention Time and Pouring Temperature on Graphite Shape and Mechanical Properties of Gray Cast Iron
by Hong-Nga Thi Pham

An Enhanced Conceptual Security Model for Autonomous Vehicles
by Abdulla Obaid Al Zaabi, Chan Yeob Yeun, Ernesto Damiani and Gaemyoung

Low Power Fast Settling Switched Capacitor PTAT Current Reference Circuit for Low Frequency Applications
by Muhammed Mansoor C. B., Hanumantha Rao G. and Rekha S.

Blockchain Application in Higher Education Diploma Management and Results Analysis
by Fernando Richter Vidal, Feliz Gouveia and Christophe Soares

Evaluating the Effectiveness of Query-Document Clustering Using the QDSM Measure
by Claudio Gutierrez-Soto, Marco Palomino, Arturo Curiel, Hector Riquelme Cerda and Fernando Bejar Rain

Method of Analysis and Classification of Acoustic Emission Signals to Identify Pre-Seismic Anomalies
by Marapulets Yury, Senkevich Yury, Lukovenkova Olga and Solodchuk Alexandra

Need of E-Recruitment System for Universities: Case of Pulchowk Campus, Nepal
by Vijay Yadav, Ujjwal Gewali, Suman Khatri, Shree Ram Rauniyar and Aman Shakya

A Fast Adaptive Time-delay-estimation Sliding Mode Controller for Robot Manipulators
by Dang Xuan Ba

Investigation of the Effect of FBG Profiles, Temperature and Transmission Distance for Environmental Sensing & Monitoring
by Muhammad Arif Riza and Yun Ii Go

Procrustes Dynamic Time Wrapping Analysis for Automated Surgical Skill Evaluation
by Safaa Albasri, Mihail Popescu, Salman Ahmad and James Keller

Effect of Cover Number on Distilled Water Production of Distillers with a Novel Water Feeding
by Mirmanto Mirmanto, Made Wirawan, I Made Adi Sayoga, Abdullah Abdullah and Muhamad Faisal

Barriers and Supports in Engineering Career Development: An Exploration of First-Year Students
by Rosmery Ramos-Sandoval and Jano Ramos-Diaz

Prediction of Vessel Dynamic Model Parameters using Computational Fluid Dynamics Simulation
by Nu'man Amri Maliky, Nanda Pratama Putra, Mochamad Teguh Subarkah and Syarif Hidayat

Design and Implementation of DFT Technique to Verify LBIST at RTL Level
by Nagaraj Vannal and Saroja V Siddamal

Dense SIFT–Flow based Architecture for Recognizing Hand Gestures
by Suni S S and K Gopakumar

Scattering of H-Wave by a Moving Dispersive Conducting Complex Object
by Esmail Mohamed Abuhdima, Gurcan Comert, Ahmed El Qaouaq and Ashleigh Nicole Reeves

The Impact of Innovation on the Performance of Manufacturing Enterprises in Vietnam
by Thi Anh Van Nguyen and Khac Hieu Nguyen

Matrix-based Minimal Cut Method and Applications to System Reliability
by Emad Kareem Mutar

Renewable Electric Power from the Equine Treadmill: An Evaluation of the Potential
by Faizan Dastgeer and Hasan Erteza Gelani

Control of Soft Robotic Artificial Muscle with Hand Gesture Using Leap Motion Sensor
by Victoria Oguntosin and Akindele Ayoola E

Dissection of Quantitative Trait Loci (QTL), Annotation of Single Nucleotide Polymorphism (SNP), and Identification of Candidate Genes for Grain Yield in Triticum turgidum L. var durum
by Issame Farouk, Ahmad Alsaleh, Jihan, Fatima Gaboun, Bouchra Belkadi, Abdelkarim Filali Maltouf, Zakaria Kehel, Ismahane Elouafi, Nasserelhaq Nsarellah, Dimah Habash and M. Miloudi Nachit

Simulated IoT Based Sustainable Power System for Smart Agriculture Environments
by Shahenaz S. Abou Emira, Khaled Y. Youssef and Mohamed Abouelatta

VLSI Architecture for OMP to Reconstruct Compressive Sensing Image
by Santosh Bujari and Saroja V Siddamal

Essential Features/Issues of a Multi-Phase Switching Synchronous Buck Regulator
by Hani Ahmad-Assi, Nour Sultan Gammoh and Mariana Awni Al Bader

Review of Different Methods for Optimal Placement of Phasor Measurement Unit on the Power System Network
by Ademola Abdulkareem, Divine Ogbe and Tobiloba Somefun

Synthesis of CVs Using a Context-free Grammar
by Darren Tafadzwa Semusemu and Abejide Ade-Ibijola

Nature Inspired and Transform Based Image Encryption Techniques: A Comparative Study
by Bhagyashri Pandurangi R, Chaitra Bhat and Meenakshi R. Patil

Human Emotion Recognition Based on EEG Signal Using Fast Fourier Transform and K-Nearest Neighbor
by Anton Yudhana, Akbar Muslim, Dewi Eko Wati, Intan Puspitasari, Ahmad Azhari and Murein Miksa Mardhia

5G, Vehicle to Everything Communication: Opportunities, Constraints and Future Directions
by Boughanja Manale and Tomader Mazri

Posture Recognition Method for Caregivers during Postural Change of a Patient on a Bed using Wearable Sensors
by Kodai Kitagawa, Koji Matsumoto, Kensuke Iwanaga, Siti Anom Ahmad, Takayuki Nagasaki, Sota Nakano, Mitsumasa Hida, Shogo Okamatsu and Chikamune Wada

Development of Secondary Processing Data Methods under Single Point Thunderstorm Activity Monitoring
by Anatoly Panyukov and Alexander Bogushov

Tolerance of Characteristics and Attributes in Developing Student's Academic Achievements
by Wongpanya Nuankaew and Pratya Nuankaew

Feature Extractors Evaluation Based V-SLAM for Autonomous Vehicles
by Mounir Amraoui, Rachid Latif, Abdelhafid El Ouardi and Abdelouahed Tajer

Effects of Resting Actions Using Smart Toys During Break Times on Concentration in E-learning
by Takashi Ito, Kenichi Takahashi and Tomoko Kajiyama

A Proactive Mobile Edge Cache Policy Based on the Prediction by Partial Matching
by Lincan Li, Chiew Foong Kwong and Qianyu Liu

Improvised E-Rickshaws for Indian Roads by Effective Battery-Ultracapacitor Hybridization
by Shimin Vayal Veetil, Varsha Shah and Makarand Lokhande

PV Integrated Recursive Least Mean Square Estimation Based Shunt Active Power Filter
by Ragam Rajagopal, K. Palanisamy and S. Paramasivam

Advanced Control Strategies on Nonlinear Testbench Dynamometer System for Simulating the Fuel Consumption
by Marika Fanesi and David Scaradozzi

Defeating Anti-Debugging Techniques for Malware Analysis Using a Debugger
by Jong-Wouk Kim, Jiwon Bang and Mi-Jung Choi

Synthesis of SQL Queries from South African Local Language Narrations
by George Obaido, Abejide Ade-Ibijola and Hima Vadapalli

Development of a Technology and Digital Transformation Adoption Framework of the Postal Industry in Southern Africa: From Critical Literature Review to a Theoretical Framework
by Kgabo Mokgohloa, Grace Kanakana-Katumba and Rendani Maladzhi

Examination of a Skill Sampling Method of an Athlete Using the Athlete's Movement and Eye Movement for the Development of an AI Coach
by Takuya Sarugaku, Jun Lee, Yasuaki Matsumoto and Mitsuho Yamada

Evolution of Teaching Approaches for Science, Engineering and Technology within an Online Environment: A Review
by Rendani Wilson Maladzhi and Grace Mukondeleli Kanakana-Katumba

Development and Performance Analysis of HRPL Using 6LoWPAN CC2538 Module for IoT Ecosystem
by Nin Hayati Mohd Yusoff and Nurul Azma Zakaria

Priority-based Scheduling Algorithm for NOMA-integrated V2X
by Ala Din Trabelsi, Hend Marouane and Faouzi Zarai

Design and Development of Electronic Sensor and Monitoring System of Smart Low-cost Phototherapy Light System for Non-Invasive Monitoring and Treatment of Neonatal Jaundice
by Paul Cabacungan, Carlos Oppus, Gregory Tangonan, Nerissa Cabacungan, John Paul Mamaradlo and Neil Angelo Mercado
Agricultural Data Fusion for SmartAgro Telemetry System
by Ioana Marcu, Ana-Maria Drăgulinescu, Carmen Florea, Cristina Bălăceanu, Marius Alexandru Dobrea and George Suciu

Laser Deprocessing Technique and its Application to Physical Failure Analysis
by Yanlin Pan, Jia Rui Thong, Pik Kee Tan, Siong Luong Ting and Chang Qing Chen

An Overview on CryptDb and Word2vec Approaches
by Hana Yousuf, Asma Qassem Al-Hamad and Said Salloum

A Circular Invariant Convolution Model-Based Mapping for Multimodal Change Detection
by Redha Touati, Max Mignotte and Mohamed Dahmane

Empirical Probability Distributions with Unknown Number of Components
by Marcin Kuropatwi´nski and Leonard Sikorski

C-Band FMCW Radar Design and Implementation for Breathing Rate Estimation
by Mohammad Mohammad Abdul-Atty, Ahmed Sayed Ismail Amar and Mohamed Mabrouk

Basic Study of 3-D Non-Invasive Measurement of Temperature Distribution using Ultrasound Images during HIFU Heating
by Ryosuke Sakakibara, Yasuhiro Shindo, Kazuo Kato, Pak Kon Choi and Akira Takeuchi

Wideband Active Switch for Electronic Warfare System Applications
by Mahadev Sarkar, Gaurav Anand and Sivakumar Ramadoss

Evaluating the Impact of Semantic Gaps on Estimating the Similarity using Arabic Wordnet
by Mamoun Abu Helou

Evaluation of Simple Space Interpolation Methods for the Depth of Precipitation: Application for Boyacá, Colombia
by Pedro Mauricio Acosta Castellanos, Alejandra Castro Ortegon and Hugo Fernando Guerrero Sierra

SH-CNN: Shearlet Convolutional Neural Network for Gender Classification
by Chaymae Ziani and Abdelalim Sadiq

Handling Priority Data in Smart Transportation System by using Support Vector Machine Algorithm
by Sara Ftaimi and Tomader Mazri

Fabrication and Optimization of High Frequency ZnO Transducers for Both Longitudinal and Shear Emission: Application of Viscosity Measurement using Ultrasound
by Hatem Dahmani, Ibrahim Zaaroura, Abbas Salhab, Pierre Campistron, Julien Carlier, Malika Toubal, Souad Harmand, Vincent Thomy, Marc Neyens and Bertrand Nongaillard
Parameter Estimation for Industrial Robot Manipulators Using an Improved Particle Swarm Optimization Algorithm with Gaussian Mutation and Archived Elite Learning
by Abubakar Umar, Zhanqun Shi, Lin Zheng, Alhadi Khlil and Zulfiqar Ibrahim Bibi Farouk

A Software-Defined Network Approach for The Best Hospital Localization Against Coronavirus (COVID-19)
by Bilal Babayigit and Eda Nur Hascokadar

NemoSuite: Web-based Network Motif Analytic Suite
by Wooyoung Kim, Yi-Hsin Hsu, Zican Li, Preston Mar and Yangxiao Wang

On Design of IoT-based Power Quality Oriented Grids for Industrial Sector
by Nesma N. Gomaa, Khaled Y. Youssef and Mohamed Abouelatta

A Novel Way to Design ADS-B using UML and TLA+ with Security as a Focus
by Pranay Bhardwaj, Carla Purdy and Nawar Obeidat

Real-time Gradient-Aware Indigenous AQI Estimation IoT Platform
by Hasan Tariq, Abderrazak Abdaoui, Farid Touati, Mohammad Abdullah Al Hitmi, Damiano Crescini and Adel Ben Mnaouer

Japanese Abstractive Text Summarization using BERT
by Yuuki Iwasaki, Akihiro Yamashita, Yoko Konno and Katsushi Matsubayashi

Efficient and Scalable Ant Colony Optimization based WSN Routing Protocol for IoT
by Afsah Sharmin, Farhat Anwar and S M A Motakabber

Design of Power Efficient Routing Protocol for Smart Livestock Farm Applications
by Shahenda S. Abou Emira, Khaled Y. Youssef and Mohamed Abouelatta

New Properties of Crimes in Virtual Environments
by Roman Dremliuga, Natalia Prisekina and Andrei Yakovenko

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 01-10 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Mitigating Congestion in Restructured Power System using FACTS Allocation by Sensitivity Factors and Parameter Optimized by GWO

Anubha Gautam[1,*], Parshram Sharma[1], Yogendra Kumar[2]

[1]*Department of Electrical Engineering, J C Bose UST, YMCA, Faridabad, 121006, India*

[2]*Department of Electrical Engineering, MANIT, Bhopal, 462003, India*

ARTICLE INFO

Article history:
Received: 07 June, 2020
Accepted: 11 August, 2020
Online: 08 September, 2020

Keywords:
Deregulation
Congestion
ATC
ACPTDF
FACTS
GWO

ABSTRACT

In modern deregulated power industry, private sector has invested a lot to supply for extended power demand using the preexisting power system framework. This resulted into increased loading of transmission lines which has to work now to hit their thermal limits. The overloading of transmission line resulted in congestion and hence increase in loss of power in the system. One of the efficient ways to reduce congestion is by enhancing the available transfer capacity (ATC) of the power system. ATC enhancement can be achieved by application of FACTS devices. This paper presents an innovative method to mitigate congestion by locating TCSC in the IEEE 30 bus system. The allocation of TCSC is done by using ACPTDF sensitivity factors while the parameter setting is done by applying Grey Wolf Optimization (GWO) method. The effective application of GWO is demonstrated in this paper to reduce active power loss, enhancement of ATC value with reduction of reactive power loss and to optimize TCSC size through a multi objective function. The suitability of algorithm is established through concerned figures and tables.

1. Introduction

This paper is an extension of work originally presented in 3rd International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE) [1].

With deregulation act in 2003, the reliability of the power system is enhanced in terms of availability and economics. The private sector intervened in the power generation and used the preexisting transmission system for distribution through pools. This resulted in overloading of lines to work under congestion, reaching their thermal and voltage limits [2]. The congestion resulted in huge amount of power losses thus effecting the economy of power transmission. There are two ways to relive the congested system. The cost-free method and the non-cost-free method. The cost-free method is one with no enhancement of operational cost. This is effectively achieved by incorporating Facts devices [3]. Power system is unevenly loaded. This results in inefficient output of the circuits. With uneven sharing of load through the lines, some lines become overloaded while others turn out to be under loaded. This distorts the voltage profile of the interconnected system [4]. FACTS being optimized for their respective parameters such as voltage angles, circuit reactance and voltage magnitudes, can be successfully incorporated in power system to modify the line parameters. This results in establishing a preferred bus and generator voltage profiles [5]. The system efficiency in terms of enhance loadability can be improved by suitably designing the controller of FACTS devices [6]. Maximum load on power system is industrial and domestic inductive load. Thus, there is significant voltage drop at these loads resulting in uneven voltage profile of system. Hence to reduce system inductive voltage drop, the inductive reactance has to be reduced in order to increase the power transfer capacity (PTC) of the system. The series FACTS device such as TCSC plays a vital role in achieving the reactance regulation [7]. To utilize the system at its maximum capacity together with power transmission economics, the transfer capacity of system must be enhanced to maximum value. System loadability improvement by increasing ATC value was achieved by an optimal power-flow-based model, for maximum power transfer by incorporating optimized FACTS control in the system [8]. Properly tuning

*Corresponding Author: Anubha Gautam., gautamanubha.12@gmail.com

FACT controller and its optimal location can reduce the system transmission losses and in return increases the power trans mitting capacity [9]. Different approaches were applied to optimize the location of fact devices. A number of heuristic methods like GA & BA were applied for optimal tuning of FACTS controller to enhance system ATC [10].

Further sensitivity index-based method for optimal allocation of FACTS devices such as SVC and TCSC was applied to enhance system power transfer capacity and hence ATC improvement [11]. The above-mentioned techniques resulted in ATC enhancement but this does not suffice the same improvements in other system parameters such as active power loss, reactive power loss and voltage profile regulation. The main objective of this paper is to develop a method with the incorporation of sensitivity index, ACPTDF together with an heuristic algorithm, Grey Wolf Optimization, (GWO) for minimizing the multi-objective function considered. This paper extends the area of implementation from a single objective of congestion management by ATC enhancement to a multi-objective of reducing active and reactive power losses as well as regulating the voltage profile of the system.

ATC of a power system is the power transfer capacity available above the maximum demand of the system, to be utilized for commercial activities between power suppliers and consumers. ATC is the back bone of any power system as it is directly influencing the power markets technically as well as economically [12]. The ATC of a power system can be enhanced by different methods and FACTS are very efficient in the same. The calculation of ATC can be done by different methods. A method was proposed to first calculate the reactive power flow and then by using PTDF as sensitivity factor ATC was calculated [13]. For enhancing the ATC value, generator terminal voltage and the output power can be worked within the defined security limits. ATC calculation and its values in power system database are therefore very crucial for the power market participants so as it can be used economically for industrial back up [14].

2. Related works

ATC for any system is the basis of restructuring of power system. Power system capability and its strength depends upon it ATC value [15].

The power system is interconnected, hence for enhancing the ATC of a particular system it is required that the power flow during the process must be technically very controlled. In other words, due to interconnection between different areas the ATC enhancement may result in change in power flow levels, resulting in unstable system. FACTS devices are quite a handful solution for this problem. Different types of FACTS devices use their specific properties to dynamically control the voltage magnitude, voltage angles and impedance of the lines while ATC enhancement is worked out. [16], [17].The FACTS devices are operated with the help of controllers. These controllers are either thyristor-controlled switch based or voltage source converter (VCS) based. These controllers helps to enhance the ATC while compensating for reactive power flow and reducing the active power losses [18]. A number of heuristic methods such as GA and PSO have been applied to program the controllers so as the FACTS device can compensate for reactive power and reduce active power loss while regulating the voltage profile of the system [19], [20].

Out of different FACTS devices TCSC is one of the most widely used device. ATC enhancement was done by applying continuation power flow (CPF) method taking thermal limits and voltage profile into account [21] .Other FACTS devices such as SSSC, STATCOM and UPFC have been modeled using heuristic methods like PSO and sensitivity index PTDF for ATC enhancement [22]. The controller of SSSC, UPFC and STATCOM have also been modeled by novel current based modeling for enhancing ATC instead of laying down new transmission line or rescheduling of generator [23]. ATC enhancement have also been done and compared by some more heuristic methods such as GA, PSO & FA under different contingency conditions [24].

3. Calculation of ATC

ATC in a power system can be calculated in numerous ways. Some methods are CPFM (continuous power flow method), linear approximation method etc. In this paper power sensitivity indices method is applied to calculate the ATC for standard IEEE 30 Bus system. The sensitivity factor applied here is Power Transfer Distribution Factor (PTDF). The PTDF can be of two types, DCPTDF and ACPTDF. Here we are applying ACPTDF for the calculation of ATC. ACPTDF determines the change in system power flow with the change in power transaction in some other line at steady state as well as under contingency conditions [25]. For a bilateral transaction between bus m and bus n. Here bus m is considered to sell power to bus n . The PTDF measures the change in real power flow of line i-j due to bilateral transaction between m & n [26].

Mathematically,

$$ACPTDF_{ij,mn} = \frac{\Delta P_{ij}}{P_{mn}} \quad (1)$$

where

P_{mn} is the power transaction between the bus m and n. Here bus m is considered to sell power to bus n. Bus n is considered to buy power in energy pooling system.

ΔP_{ij} is the change in power flow between bus i and j due to the bilateral transaction between bus m and n. This can be calculated as:

$$\Delta P_{ij} = \left[\frac{\partial P_{ij}}{\partial V_i}\right]\Delta V_i + \left[\frac{\partial P_{ij}}{\partial V_i}\right]\Delta V_j + \left[\frac{\partial P_{ij}}{\partial \delta_i}\right]\Delta \delta_i + \left[\frac{\partial P_{ij}}{\partial \delta_i}\right]\Delta \delta_j \quad (2)$$

$$\text{or, } \Delta P_{ij} = \left[\frac{\partial P_{ij}}{\partial \delta_2} \ldots \frac{\partial P_{ij}}{\partial \delta_n} \frac{\partial P_{ij}}{\partial V_2} \ldots \frac{\partial P_{ij}}{\partial V_n}\right] * \begin{bmatrix} \partial\delta_2 \\ . \\ . \\ \partial\delta_n \\ \partial V_2 \\ . \\ . \\ \partial V_n \end{bmatrix} \quad (3)$$

$$\text{or, } \Delta P_{ij} = \left[\frac{\partial P_{ij}}{\partial \delta_2} \ldots \frac{\partial P_{ij}}{\partial \delta_n} \frac{\partial P_{ij}}{\partial V_2} \ldots \frac{\partial P_{ij}}{\partial V_n}\right] * \begin{bmatrix} J1 J2 \\ J3 J4 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ . \\ . \\ +Pt \\ 0 \\ . \\ . \\ -Pt \\ 0 \end{bmatrix} \quad (4)$$

Now the power transfer in the line between buses i and j can be calculated for different values of PTDF as in equation (5)

$$T_{ij,mn} = \begin{Bmatrix} \dfrac{(P_{ij}^{\max} - P_{ij}^{0})}{PTDF_{ij,mn}}; & PTDF_{ij,mn} > 0 \\ \alpha(\inf inite)\ ; & PTDF_{ij,mn} = 0 \\ \dfrac{(-P_{ij}^{\max} - P_{ij}^{0})}{PTDF_{ij,mn}}; & PTDF_{ij,mn} < 0 \end{Bmatrix} \quad (5)$$

In equation (5)

P_{ij}^{max} is active power flow limit of line *i-j*.

P_{ij}^{0} is base case power flow in line *i-j*.

$PTDF_{ij,\,mn}$ is the Power Transfer Distribution Factor for line i-j due to the exchange of active power between bus m and n.
Then ATC can be calculated as:

$$ATC_{mn} = \min\left\{T_{ij,mn}\right\}, ij \in N_L \quad (6)$$

N_L is the total number of lines.

4. Reactive power flow

A fundamental aspect of the compensation and control of reactive power is its balance. The shunt capacitance of the transmission line yields reactive power proportional to the square of the voltage. Since the voltage must be kept within ± 5% of the rated voltage, the output or consumption of reactive power is relatively constant. The series inductance of the transmission line consumes reactive power proportional to the square of the current. Since the current varies from the duration of maximum demand to the duration of minimum demand, the reactive power consumption is also modified by the transmission line.

Figure number (1) shows the reactive power flow in the transmission system.

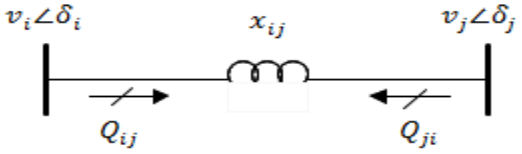


Figure 1: Reactive power flow in the Power system

The reactive power losses in the transmission line can be mathematically expressed as:

$$Q_p = Q_{ij} + Q_{ji} \quad (7)$$

where , Q_{ij} can be written as:

$$Q_{ij} = \frac{v_i}{x_{ij}}\left(v_i - v_j \cos\delta\right) \quad (8)$$

from equation (7) and equation (8) the expression for reactive power loss can be calculated as:

$$Q_p = \frac{v_i^2}{x_{ij}} + \frac{v_j^2}{x_{ij}} - \frac{2v_i v_j}{x_{ij}}\cos\delta \quad (9)$$

On simplification

$$Q_L = min\ \{\textstyle\sum_{k=1}^{nl} g_k(V_i^2 + V_j^2 - 2V_i V_j\, sin(\delta_i - \delta_j))\} \quad (10)$$

5. Modeling of TCSC

During the steady state, the compensator can freely change between reactance values according to its control. To avoid over-compensation of the line, limits are recommended for the TCSC reactance, given by the equation (7).

$$-0{,}8\ X_L \le X_{TCSC} \le 0{,}2\ X_L\ p.u. \quad (11)$$

The proposed limit vary among different research works, but a tendency of capacitive factor greater than 50% of reactance of the line and inductive factor less than 25% of the inductance of the line is maintained.

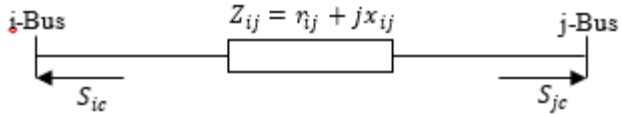


Figure 2: TCSC Power Injection Model [27]

In order to be used in power flow, the TCSC is modelled in form of line impedance with the built-in transformer reactance, in order to obtain the compensation scheme as shown in Figure (2)
The change in admittance is governed by the equation:

$$\Delta y_{ij} = y'_j - y_{ij} = \left(G'_{ij} + jB'_{ij}\right) - \left(G_{ij} + jB_{ij}\right) \quad (12)$$

$$G_{i,j} + jB_{i,j} = \frac{1}{Z_{i,j}} \tag{13}$$

$$\left.\begin{aligned} G_{i,j} &= \frac{r_{i,j}}{r_{i,j}^2 + x_{i,j}^2} \\ B_{i,j} &= \frac{-x_{i,j}}{r_{i,j}^2 + x_{i,j}^2} \end{aligned}\right\} \tag{14}$$

$$G'_{i,j} = \frac{r_{ij}}{r_{i,j}^2 + (x_{i,j} + x_{TCSC})^2} \tag{15}$$

$$B'_{i,j} = \frac{-(x_{i,j} + x_{TCSC})}{r_{ij}^2 + (x_{i,j} + x_{TCSC})^2} \tag{16}$$

where,

G_{ij} is the conductance of line ij before applying TCSC

G'_{ij} is the conductance of line ij after applying TCSC

B_{ij} is the suceptance of line ij before applying TCSC

B'_{ij} is the suceptance of line ij after applying TCSC

Equation (8) depicts that there is a variation of admittances by the application of the TCSC, so the admittances matrix of the system will be modified as indicated (13).

$$Y'_{BUS} = Y_{BUS} + \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \Delta y_{i,j} & 0 & \cdots & 0 & -\Delta y_{i,j} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\Delta y_{i,j} & 0 & \cdots & 0 & \Delta y_{i,j} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ & Col-i & & \ldots & Col-j & & \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ Line-i \\ Line-j \end{matrix} \tag{17}$$

The active and reactive power flow, including the admittance variation implemented by the TCSC can be explained by equation (14) & (15):

$$P_{ijrCSC} = V_i^2 G'_{ij} - V_i V_j \left[G'_{ij} Cos(\delta_{ij}) - B'_{ij} Sin(\delta_{ij})\right] \tag{18}$$

6. Objective Function

The objective function includes:

a) ATC maximization:
$ATC_{mn} = \min\{T_{ij,mn}\}$

b) Reactive power loss minimization:
$Q_L = min\{\sum_{k=1}^{nl} g_k(V_i^2 + V_j^2 - 2V_iV_j\, s\, in(\delta_i - \delta_j))\}$

Thus, the multi-objective function can be written as:

$$f(x) = w_1 \times (\max ATC) + w_2 \times \min P_L) \tag{19}$$

7. Constraints and limits

While performing optimal power flow (OPF) on a system, there are certain parameters which are to be implemented with constraints and pre-defined limits as stated in the standard IEEE bus system appendix. Basically, there are two types of parameters. One parameter which involves expenses such as generation of power at generating end, P_i^G. The other parameter which does not involve expenses are magnitude of voltages at generators, V_i^G and the transformer taps, t_{ij}.

The constraints in the system can be represented as summarized in Table number (1) below:

Table 1: Constraints incorporated in the system

Constraints	Equations
Power Balance (MW)	$P_{gn} - P_{dn} - P_n(V_n - \delta_n) = 0$
Power Balance (MVAr)	$Q_{gn} - Q_{dn} - Q_n(V_n - \delta_n) = 0$
Generated Power (MW)	$P_n^{min} \le P_n \le P_n^{max}$
Generated Power (MVAr)	$Q_n^{min} \le Q_n \le Q_n^{max}$
Bus Voltage Limits	$0.95pu \le Vi \le 1.05pu$
Generator Voltage Limits	$0.95pu \le Vg \le 1.50pu$
TCSC Reactance Limits	$-0.8X_L \le X_{tcsc} \le 0.2X_L$

Where,

P_{gn} & Q_{gn} represents real and reactive power generations at *nth* bus

P_{dn} & Q_{dn} represents real and reactive power demand at *nth* bus

P_n & Q_n represents real and reactive power injected at *nth* bus

V_n & δ_n are voltage and resultant angle at *nth* bus

8. Grey Wolf Optimization

The heuristic technique implemented here to optimize the TCSC parameter is Grey Wolf Optimization (GWO). This technique is based on the well-organized social hierarchies of a pack of grey wolf. Grey wolf have a very typical and well defined hunting action. The hunting of prey is led by the most powerful alpha (a) wolf. The next hierarchy is taken by beta (b) wolf and the next one is taken by gamma(g) wolf. Rest of the wolfs in the pack are omega wolf. All the wolfs are guided by the alpha wolf. Hence the alpha wolf position in solution space is considered as the best solution, beta wolf, the next best and gamma wolf position the third best solution. The omega wolfs always follow the three best wolves throughout searching. GWO is divided into three processes, encircling, hunting and attacking the prey. Mathematically, the circling of prey can be symbolized as below [28].

$$\bar{D} = \left| \bar{C} * \bar{X}_{(prey)(t)} - \bar{X}_{GW(t)} \right| \tag{20}$$

$$\bar{X}_{(GW)(t+1)} = \left| \bar{X}_{(prey)(t)} - D \right| \tag{21}$$

where t is the current time, $\bar{X}$ is the vector representing location of the prey, $\bar{X}$ _GW is a vector representing location of grey wolf, C & A are coefficient vectors and mathematically presented as:.

$$\vec{A} = 2\vec{a} * \vec{r}_1 - \vec{a} \tag{22}$$

$$\vec{C} = 2 * \vec{r}_2 \quad (23)$$

where "a", is the error that is introduced in the system so as to avoid premature convergence of the algorithm. Its value is decreased from 2 to 0 through a series of iteration.

$\vec{r}_1$ & $\vec{r}_2$ represents arbitrary values between 0 and 1.

As the power system equations are highly non-linear and the solution can't be realized by traditional methods so Grey wolf algorithm is simulated mathematically to locate the position of prey (solution). First three positions Alpha, Beta & gamma of wolf are best fitness values and position of omega wolves are updated with respect to the position of alpha, beta & gamma wolves and are mathematically be represented as:

$$\vec{X}_1 = \vec{X}_{-a}(t) - \vec{A}_1 * \vec{D}_{-}a$$
$$\vec{X}_2 = \vec{X}_{-b}(t) - \vec{A}_1 * \vec{D}_{-}b$$
$$\vec{X}_3 = \vec{X}_{-g}(t) - \vec{A}_1 * \vec{D}_{-}g$$

where $\vec{D}_{-}a, \vec{D}_{-}b, \vec{D}_{-}g$ are defined as

$$\vec{D}_{-}a = |C * \vec{X}_{-a}(t) - \vec{X}_{GW}(t)| \quad (24)$$
$$\vec{D}_{-}b = |C * \vec{X}_{-b}(t) - \vec{X}_{GW}(t)| \quad (25)$$
$$\vec{D}_{-}g = |C * \vec{X}_{g}(t) - \vec{X}_{GW}(t)| \quad (26)$$

The algebraic sum of three locations of wolves is averaged which gives the best location of grey wolf

$$X_{GW}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (27)$$

This algorithm is applied in following steps:

1. The population of grey wolf is initialized with the initialization of initial parameters as:
   - Size of the search space (defined by problem constraints)
   - Number of search agents (here taken as 100)
   - Vectors a, A & C
   - Maximum number of iterations (100)
2. The wolves are randomly distributed in the pre-defined search space
3. The fitness value of each search agent is calculated and then indexed to get the best three fitness values. [from equation (24), (25) and (26)]
4. The three best positions are considered as best fitness values.
5. With respect to these three positions the fitness value of other wolfs are calculated.
6. Again the fitness values are sorted to get the updated positions of the wolfs [using equation (27)]
7. The fitness values are indexed and again first three values are considered as best fitness values.
8. The iterations are carried out till maximum iterations have reached or the fitness value become constant for defined number of iterations (in this paper equals to 50)

8.1 Pseudo code for GWO

Result: Bset value of fitness function

generate (X) / / Initialize the population of grey wolves

initialize Parameter (a, A, C);

evaluate X (0);

select new (Alpha, Beta, Delta, X (0));

for *e = 1 to EVALMAX* to do

  for *every Wolf l in Omega* do

    for *i = 0 to DIM* do

update Position (l, i); // Update current position

. **end for**

adjust parameters (a, A, c); // adjust the algorithm parameters

evaluate X (e + 1);

. select new (Alpha, Beta, Delta, X (e + 1));

. e = e + 1;

  end for

end for

9. Methodology adopted

9.1 For ATC maximization (Without TCSC)

Figure (3) presents the sequence to calculate the ATC value without the application of TCSC. Here GWO is used only in NR with OPF only. While section 9.2 shows the steps followed to calculate ATC value using GWO optimized TCSC size and ACPTDF sorted location of TCSC. In the second case all the calculations are done once TCSC is located at a suitable position.

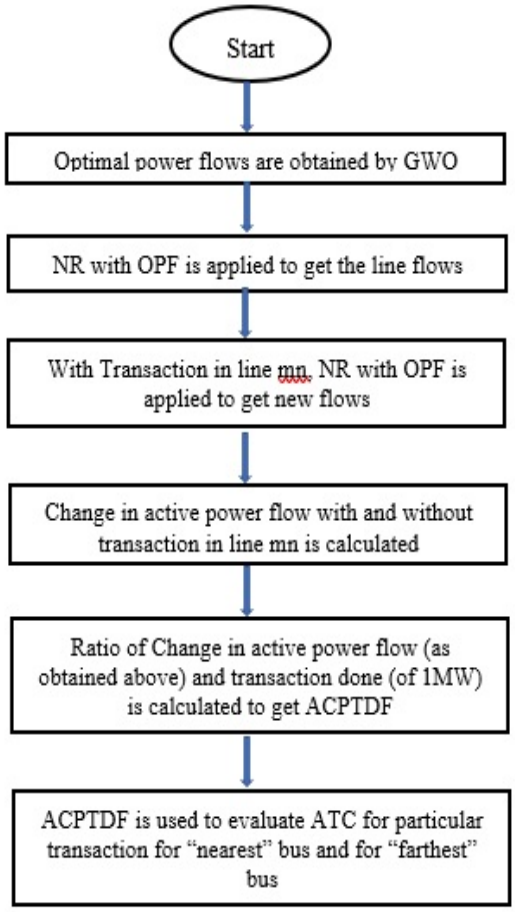


Figure 3: Process to calculate ATC without TCSC

9.2 For Minimization of Reactive power loss (TQL)

For mminimization of power loss same two methods i.e. with and without TCSC are applied but with an objective to minimize reactive power loss only. With decreased value of reactive power loss, it can be observed that the ATC value is also decreased.

Here IEEE 30 bus system with six generators at bus number 1, 2, 5, 8, 11 and 13 is utilized for the application of the selected methodology. ACPTDF values are calculated for the transaction at a particular bus and its effect on power flow at all the other buses. Figure (4) gives the presentation of ACPTDF values when transactions are done between bus number 2 to 5 ad bus 2 to 26.

A number of evolutionary programming methods have been applied enhancement of ATC for a system but GWO turned to be one of the most suitable method to give optimized result for the chosen objective. Firefly algorithm have been used for calculation of ATC with different FACTS devices [29] .GWO when used under similar circumstances with TCSC gave better results. Figure (5) gives a detailed presentation of the values of ATC and active power losses with both the methods are applied.

From fig (5) it can be depicted that while no FACTS device is connected in the system, the ATC value with GWO comes to be 12.18 MW significantly higher with that obtained for FA which comes to be 7.47 MW.

Similarly, when the methods are applied for Total reactive power loss (TQL) reduction the power loss in case of GWO is 4.89 MW which is lesser then 5.01 MW obtained by applying FA. Moreover, the ATC value in case of TPL minimization is 10.54MW which is higher than 6.235 MW obtained by FA.

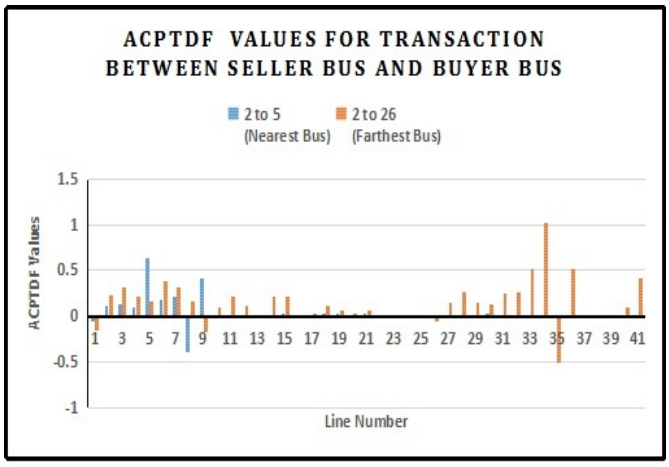


Figure 4: The ACPTDF values calculated for transaction between bus 2-5 and 2-26

10. Results and Analysis

The proposed optimization method is verified on 41 line, six generators, standard IEEE 30 BUS system [30] as shown in figure (6). Here except bus number 2, 5, 8, 11 and 13 where generators are connected, all the buses are considered as load bus. Also, all the generator buses are considered as seller buses while all other buses are the buyer buses.

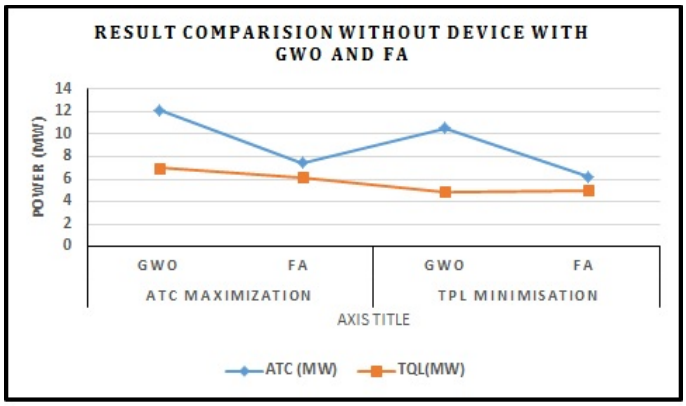


Figure 5: Result comparison between GWO and FA without device

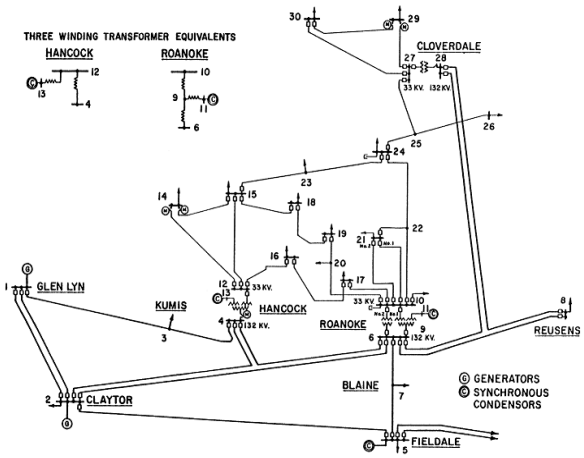


Figure.6: Standard IEEE 30 bus system [30]

10.1 ATC Maximization

The algorithm firstly applied for the objective of enhancement of ATC to reduce the congestion in the given system. The first step in this process is to calculate ATC with the help of ACPTDF. Figure (7) represents the effect of Generator at bus number 2 on ATC values of different parts of the system considered.

It can be seen that bus number 5 is nearest to generator and hence it has maximum value of ATC which equals 116.65 MW. On the other hand, bus number 26 is farthest from bus 2 so have a minimum value of ATC equal to 12.18 MW.

Figure number (8) represents the ATC distribution in the system due to generator at bus number 5 for all transactions. It can be well depicted that ATC value is largest, 184.56 MW for transaction between 5 to 2 and minimum, 12.16 MW for transaction between 5 to 26. The effect of generator at bus number 8 can be seen from figure number (9). This figure shows the distribution of ATC throughout the system for all transactions. Here it can be seen that maximum value of ATC, 70.41 MW is obtained at the line between bus 8-12 as bus 12 is nearest to bus 8. While the minimum value of ATC, 12.86 MW is obtained at line between 8-26.

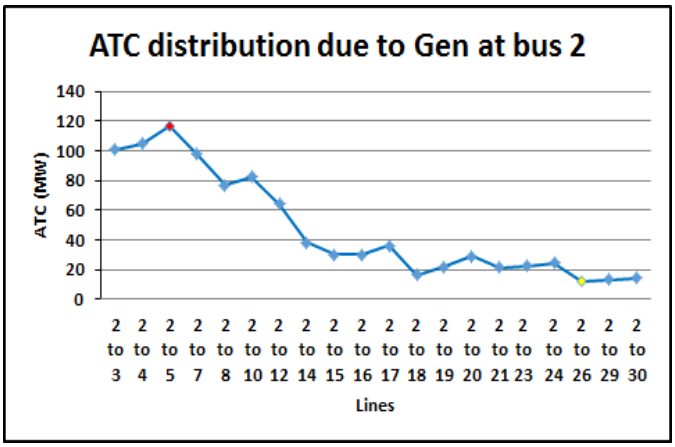


Figure 7: ATC distribution in the system due to generator connected at bus number 2

The effect of generator at bus number 11 on the distribution of ATC in the system for all transactions is shown in figure (10). It is clear from the figure that maximum value of ATC, 7.77 MW,

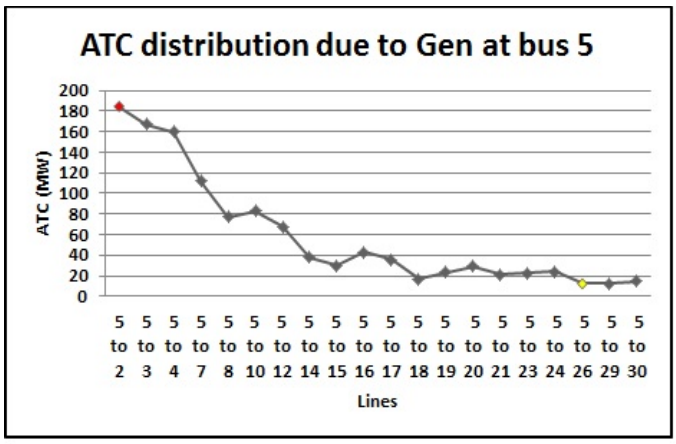


Figure 8: ATC distribution in the system due to generator connected at bus number 5

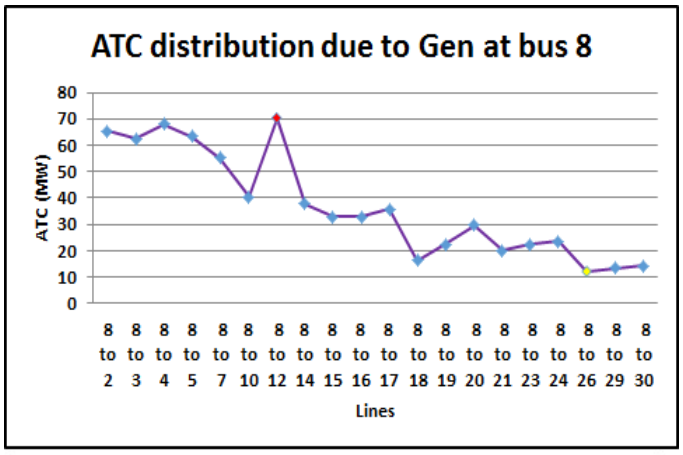


Figure 9: ATC distribution in the system due to generator connected at bus number 8

during all transactions is obtained at the line between seller bus 11 and the buyer bus 16. On the other hand, the minimum ATC value is obtained at the line between bus 11 and 26.

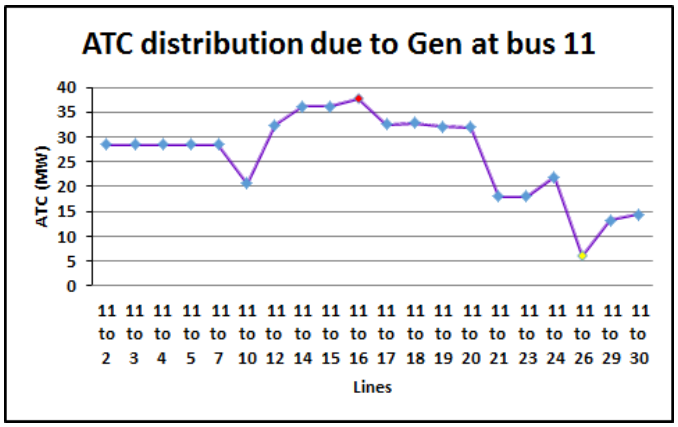


Figure 10: ATC distribution in the system due to generator connected at bus number 11

Effect of generator at bus number 13 is detailed in figure number (11). It is clear from the figure that maximum value of ATC, is obtained at line 13-7 and the minimum value of ATC, 8.89 MW is between bus 13 and 26 which is the farthest bus.

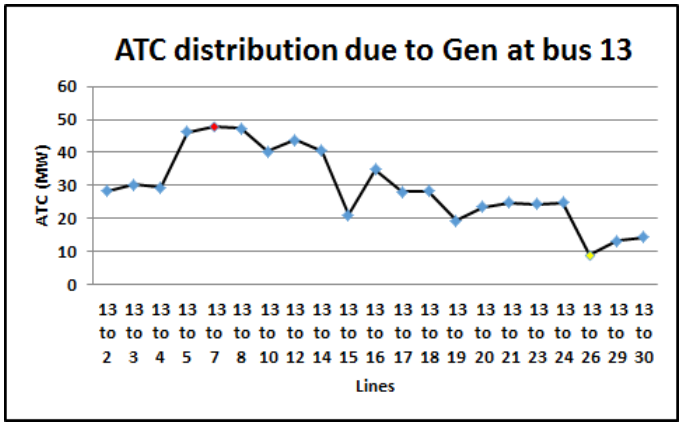


Figure 11: ATC distribution in the system due to generator connected at bus number 13

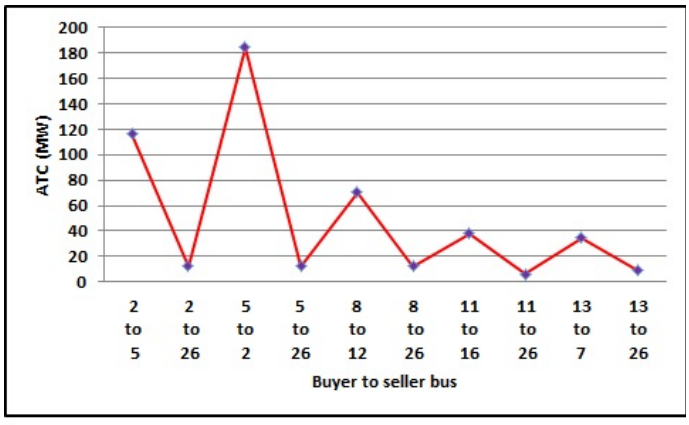


Figure 12: Variation of ATC values for all transactions due to generator at different buses

The maximum and the minimum ATC values can be represented in figure (12) for all generators.

It can be seen that the bus nearer to the generator bus are having higher ATC values. The lines 2-5, 5-2, 8-12, 11-16 and 13-7 are having higher ATC values as compared to their corresponding

farther buses.After ATC being calculated for all generators and all buses, now the ATC value is to be calculated by incorporating TCSC. The position of TCSC is searched by ACPTDF where the value of ATC is maximum after being sorted by GWO for its optimized parameter. NR with OPF is performed to get the enhanced values of ATC after incorporating TCSC the result are being summarized in table number 2. From table no 2 the results for change in ATC and reactive power loss (TPL) can be predicted. The calculations were done by giving weightage to ATC maximization in equation number (19).. Together with the increase of ATC the power loss is also increased.

Table 2: ATC maximization consolidated results

Bus		ATC (MW)		TQL (MVAr)	
SB	BB	Without TCSC	With TCSC	Without TCSC	With TCSC
2	5	116.65	120.75	82.33	30.5
	26	12.18	13.18	28.15	30.3
5	2	184.56	215.13	29.78	30.1
	26	12.26	13.45	28.48	30.7
8	12	70.41	74.056	28.26	30.4
	26	12.18	16.57	28.52	30.8
11	16	37.77	38.08	28.39	30.6
	26	6.09	8.01	29.24	31.3
13	7	34.67	36.89	27.56	29.6
	26	8.89	11.46	28.95	30.5

This increment in ATC value is higher when, percentage increase is compared from ATC enhancement by other optimization methods. In this case the percentage increase in ATP and TQL is shown in figure (13).

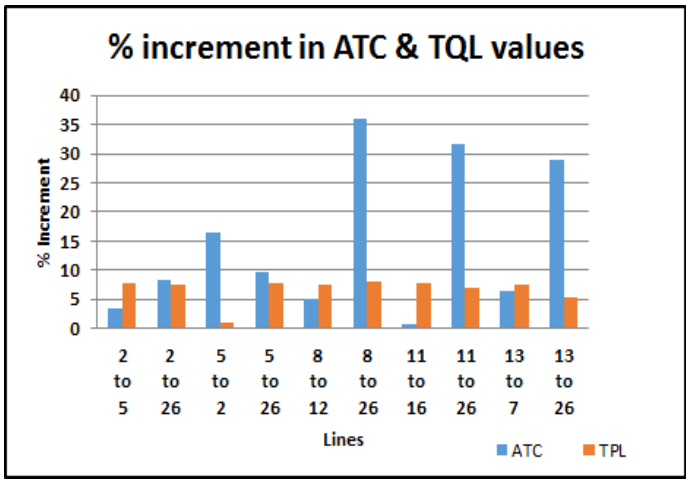


Figure 13: Percentage increment in ATC & TQL values with the incorporation of

TCSC

10.2 Reactive Power loss minimization (TQL)

The second objective to minimize reactive power loss in the system is obtained by giving weightage to the TQL term in the objective function in equation (19).

Reactive power loss minimization results are depicted in table no.3. The table bears consolidated results when to reduce the reactive power losses is the objective. The table shows that reactive power losses are significantly reduced.

Table 3: TQL minimization consolidated results

Bus		ATC (MW)		TQL (MVAr)	
SB	BB	Without TCSC	With TCSC	Without TCSC	With TCSC
2	5	116.65	90.67	28.33	25.36
	26	12.18	8.34	28.15	25.36
5	2	184.56	165.67	29.78	25.36
	26	12.26	9.45	28.48	25.36
8	12	70.41	45.89	28.26	25.36
	26	12.18	7.54	28.52	25.36
11	16	37.77	32.78	28.39	25.36
	26	6.09	6.03	29.24	25.36
13	7	34.67	32.43	27.56	25.36
	26	8.89	9.54	28.95	25.36

Together with reduction in reactive power losses, ATC value is also reduced. Again, in this case, because of the optimized TCSC location, reactive power loss is minimized. The percentage reduction in reactive power loss is higher as compared to other evolutionary programming such as FA. Moreover, the reduction in ATC value with TCSC in line is less as compared to when other FACTS devices are used. Figure (14) shows the percentage reduction in TQL and ATC with GWO optimized TCSC.

Table no 4: OPF result validation for ATC values in IEEE 30 bus system for far end bus

Bilateral Transactions Between Buses		ATC Values (MW)	
From	To	FA [29]	GWO
2	26	7.6697	13.18
5	26	7.67154	13.45
8	26	7.81371	16.57
11	26	7.70317	8.01
13	26	7.66233	11.46

The OPF results for ATC maximization are validated with those obtained by Firefly Algorithm [29] and are presented in Table 4. The ATC value at the bus connected at the far end is low due to the transmission losses. The ATC value obtained by applying GWO is significantly higher as compared to that

obtained by FA when far end values are compared. This validates GWO algorithm.

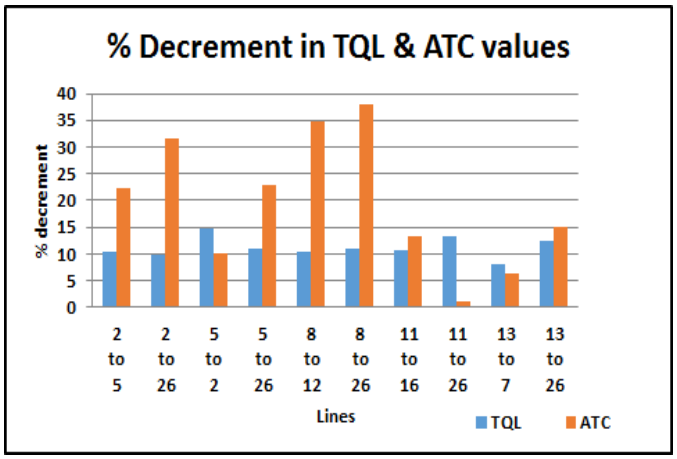


Fig no 14: Percentage decrement in TQL & ATC values with the incorporation of TCSC

11. Conclusion

In the current deregulated version of power system, where the expansion of the power system framework, due to economical and geographical constraints, can't be expanded. Thus, the private market participants supplying their generation through the exiting lines, forcing the lines to work at their voltage and thermal limits hence creating congestion. To overcome this congestion, keeping the operational cost same, a versatile FACTS device, TCSC is tested here in the standard IEEE 30 bus system. TCSC parameters are optimised with the help of very versatile Grey Wolf Optimizer and the location of TCSC is optimized with the help of power system sensitivity factors, ACPTDF. The objectives of enhancing ATC value to reduce congestion in the system and to minimise reactive power loss are achieved. The objectives are achieved here within the prescribed voltage and thermal limit constraints. GWO being guided by the best fitness value (alpha wolf) as well as two more near to best fitness values i.e. beta and gamma wolves. This ensures the speed of convergence. This optimiser uses three more variables a, A and C which ensures that GWO must not get stuck in the local minima and premature convergence. The ATC values for farther buses are quit low but this method optimised the TCSC parameter and its location such that the ATC value at far end should have higher values for greater utilization of lines and hence reduced the congestion. This paper utilised the pre-existing transmission lines for applying GWO and placing TCSC in the system to achieve the objectives of maximising ATC and reduction of reactive power loss to manage congestion in the system.

References

[1] A. Gautam, P.R. Sharma, Y. Kumar, "Placement of TCSC applying Grey Wolf Optimization," 313–318, 2019, doi:10.1109/RDCAPE47089.2019.8979016.

[2] K.S. Verma, "Enhancement of Available Transfer Capability by the use of UPFC in Open Power Market," 463–467, 2002.

[3] M.B. K. Selvi, N. Ramaraj, "Total Transfer Capability Evaluation using Genetic Algorithms," IE(I) Journal, **88**, 60–64, 2007.

[4] Y.H.S. and T.J. L. Gyugyi, Flexible AC Transmission Systems (FACTS), The Institutions of Electrical Engineers, Padstow, 1999.

[5] N.G. Hingorani. and L. Gyugyi, Understanding FACTS, concepts and technology of flexible ac transmission systems, Wiley-IEEE Press, 2000.

[6] S. Gerbex, R. Cherkaoui, A.J. Germond, "Optimal location of multi-type FACTS devices in a power system by means of genetic algorithms," IEEE Transactions on Power Systems, **16**(3), 537–544, 2001, doi:10.1109/59.932292.

[7] A. Arzani, M. Jazaeri, Y. Alinejad-Beromi, "Available transfer capability enhancement using series FACTS devices in a designed multi-machine power system," Proceedings of the Universities Power Engineering Conference, 4–9, 2008, doi:10.1109/UPEC.2008.4651434.

[8] Y. Xiao, Y.H. Song, C. Liu, Y.Z. Sun, "Using FACTS Devices," IEEE Transactions on Power Systems, **18**(1), 305–312, 2003.

[9] W. Ongsakul, P. Jirapong, "Optimal allocation of facts devices to enhance total transfer capability using evolutionary programming," Proceedings - IEEE International Symposium on Circuits and Systems, 4175–4178, 2005, doi:10.1109/ISCAS.2005.1465551.

[10] R. Mohamad Idris, A. Khairuddin, M.W. Mustafa, "A multi-objective bees algorithm for optimum allocation of FACTS devices for restructured power system," IEEE Region 10 Annual International Conference, Proceedings/TENCON, 1–6, 2009, doi:10.1109/TENCON.2009.5395826.

[11] J. Kumar, A. Kumar, W.Z. Gao, S. Yan, J. Wang, S. Gu, "ATC enhancement with SVC and TCSC using PTDF based approach in deregulated electricity markets," Advanced Materials Research, **516**–**517**, 1337–1341, 2012, doi: 10.4028/www.scientific.net/AMR.516-517.1337.

[12] K. Mushfiq-Ur-Rahman, M. Saiduzzaman, M.N. Mahmood, M.R. Khan, "Calculation of available transfer capability (ATC) of Bangladesh power system network," 2013 IEEE Innovative Smart Grid Technologies - Asia, ISGT Asia 2013, 1–5, 2013, doi:10.1109/ISGT-Asia.2013.6698716.

[13] Ibraheem, N.K. Yadav, "Implementation of FACTS Device for Enhancement of ATC Using PTDF," International Journal of Computer and Electrical Engineering, **3**(3), 343–348, 2011, doi:10.7763/ijcee. 2011.v3.338.

[14] P.W. Sauer, "Technical challenges of computing available transfer capability (ATC) in electric power systems," Proceedings of the Hawaii International Conference on System Sciences, **5**(2), 589–593, 1997, doi:10.1109/HICSS.1997.663220.

[15] M.A. Khaburi, M.R. Haghifam, "A probabilistic modeling-based approach for Total Transfer Capability enhancement using FACTS devices," International Journal of Electrical Power and Energy Systems, **32**(1), 12–16, 2010, doi:10.1016/j.ijepes.2009.06.015.

[16] S. Ahmad, F.M. Albatsh, S. Mekhilef, H. Mokhlis, "An approach to improve active power flow capability by using dynamic unified power flow controller," 2014 IEEE Innovative Smart Grid Technologies - Asia, ISGT ASIA 2014, 249–254, 2014, doi:10.1109/ISGT-Asia.2014.6873798.

[17] G. Hamoud, "Assessment of Available Transfer Capability of transmission systems," IEEE Transactions on Power Systems, **15**(1), 27–32, 2000, doi:10.1109/59.852097.

[18] X. Jiang, X. Fang, J.H. Chow, A.A. Edris, E. Uzunovic, M. Parisi, L. Hopkins, "A novel approach for modeling voltage-sourced converter-based FACTS controllers," IEEE Transactions on Power Delivery, **23**(4), 2591–2598, 2008, doi:10.1109/TPWRD.2008.923535.

[19] S. Panda, N.P. Padhy, "Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design," Applied Soft Computing Journal, **8**(4), 1418–1427, 2008, doi:10.1016/j.asoc.2007.10.009.

[20] R. Eberhart, J. Kennedy, "New optimizer using particle swarm theory," Proceedings of the International Symposium on Micro Machine and Human Science, 39–43, 1995, doi:10.1109/mhs.1995.494215.

[21] T. Nireekshana, G. Kesava Rao, S. Siva Naga Raju, "Enhancement of ATC with FACTS devices using Real-code Genetic Algorithm," International Journal of Electrical Power and Energy Systems, **43**(1), 1276–1284, 2012, doi:10.1016/j.ijepes.2012.06.041.

[22] K. Bavithra, S.C. Raja, P. Venkatesh, "Optimal setting of FACTS devices using particle swarm optimization for ATC enhancement in deregulated power system," IFAC-Papers On Line, **49**(1), 450–455, 2016, doi:10.1016/j.ifacol.2016.03.095.

[23] S. Rahimzadeh, M.T. Bina, "Looking for optimal number and placement of FACTS devices to manage the transmission congestion," Energy Conversion and Management, **52**(1), 437–446, 2011, doi:10.1016/j.enconman.2010.07.019.

[24] A. Gupta, A. Kumar, "ATC Determination with Heuristic techniques and Comparison with Sensitivity based Methods and GAMS," Procedia Computer Science, **125**(2017), 389–397, 2018, doi:10.1016/j.procs.2017.12.051.

[25] A. Kumar, J. Kumar, "ATC determination with FACTS devices using PTDFs approach for multi-transactions in competitive electricity markets," International Journal of Electrical Power and Energy Systems, **44**(1), 308–

317, 2013, doi:10.1016/j.ijepes.2012.07.050.
[26] Shweta, V.K. Nair, V.A. Prakash, S. Kuruseelan, C. Vaithilingam, "ATC Evaluation in A Deregulated Power System," Energy Procedia, **117**, 216–223, 2017, doi:10.1016/j.egypro.2017.05.125.
[27] G.A.N.S.B.S.S. N.M. Tabatabaei, "Optimal Location Of Facts Devices Using Adaptive Particle Swarm Optimization Mixed With Simulated Annealing," International Journal on "Technical and Physical Problems of Engineering, **3**(7), 60–70, 2011.
[28] S. Mirjalili, S.M. Mirjalili, A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, **69**, 46–61, 2014, doi:10.1016/j.advengsoft.2013.12.007.
[29] M. Venkateswara Rao, S. Sivanagaraju, C. V. Suresh, "Available transfer capability evaluation and enhancement using various FACTS controllers: Special focus on system security," Ain Shams Engineering Journal, **7**(1), 191–207, 2016, doi:10.1016/j.asej.2015.11.006.
[30] IEEE 30 Bus System. http://www.ee.washington.edu/research/pstca/pf30/pg_tca30bus.htm

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 01-07 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Offline Signature Recognition and Verification Using ORB Key Point Matching Techniques

Aravinda Chickmaglore Venkataramu[*,1], Atsumi Masahiko[2], Akshaya[1], Amar Prabhu Gurupura[1], Udaya Kumar Reddy Kyasambally Rajashekar[1]

[1]*Department of Computer Science and Engineering, N.M.A.M Institute of Technology, Nitte 574110, India*

[2]*Department of Electronic and Computer Engineering, Ritsumeikan University, Kusatsu, 5258577, Japan*

ARTICLE INFO

Article history:
Received: 08 May, 2020
Accepted: 17 June, 2020
Online: 06 July, 2020

Keywords:
Image matching
Robust independent elementary features (BRIEF)
Oriented FAST, Rotated BRIEF (ORB).

ABSTRACT

An extensive work has been carried out in the field of human transcribe-verification and transcribe-recognition by extensive scholars across the globe from past decades. In order to demeanour immense experiments for considering the performance of the newly intended models and to substantiate the efficacy of the proposed model which is moderately required. This paper monologue the problem of signature-verification and recognition using diverse ORB key points and Convolution Neural Network. This method reckons on descriptors for detection and matching. The systematic approach is tested more on an few real and few fake signatures. Many features and combination of features were proposed for signature substantiation and acknowledgement. Numerous experiments are conducted to determine the capability of the proposed models in selective genuine and forgery signature. In this context, a large signature corpus comprising of 29950 offline signatures from 605 persons is created during the course of the research work. Finally the achievement was achieved about 90% of correct accuracy of the given original signature.

1 Introduction

In modern society there is a greater demand for reliable methods of personal authentication because authentication is a crucial aspect of our daily life activities such as document authentication, having access to computing facilities or information, seeking entry to secured zone, etc. These activities require fast and effective automatic personal authentication systems. Existent classes in handwritten signatures: 1) Online/dynamic signatures 2) Offline/static signatures. A specially designed electronic device like digitizing tablet is used to record online-signature along with features like position, pressure, elevation etc., whereas, an offline signature is a digitized signature image taken using electronic devices like scanner, camera etc.[1]-[3] During the present day situation the digital images play a very basic role in many applications and have become a important source of information. As we all knew that feature detection and feature matching will plays the vital role in image processing applications, like image retrieval, object detection and many more. The offline signature and verification are related to probing the components of the signature into integral elements for extracting features by matching techniques is as shown in the Figure 1. The SURF technique is a feature finder which estimate of SIFT and performs faster than SIFT without reducing the quality of identified points[4]-[5]. These two techniques are thus based on descriptor and a detector. One more well-known technique BRIEF is an alternative for SIFT which requires less complexity with similar matching performance.Rublee et al. proposed Oriented FAST and Rotated BRIEF (ORB) as another efficient alternative for SIFT and SURF. In this paper the main concentration was on the ORB feature rather than SIFT and SURF. The experiment was carried out for the robustness of the signature key point techniques against rotation, scaling, and deformity due to horizontal or vertical shears.

2 Related work carried out so far

As we all knew that signature verification and recognition are considered to be most complex and challenging problem in pattern recognition area by virtue of large intraclass divergence between individual and less interclass deviations between signatures of individuals in toady's' global field.

[*]Aravinda C.V, N.M.A.M Institute of Technology Nitte INDIA , +91-9742260265 & araivnda.cv@nitte.edu.in,0077ev@ed.ritsumei.ac.jp

2.1 Template Matching

This is one of the oldest analogous artistry used for pattern-matching. This technique is used for a signature class denoted by a template and is being compared with the original to establish its authentication. As referred in the "Dynamic time warping", (DTW) is the materialistic technique used to establish a match between a test signature and a template[6].

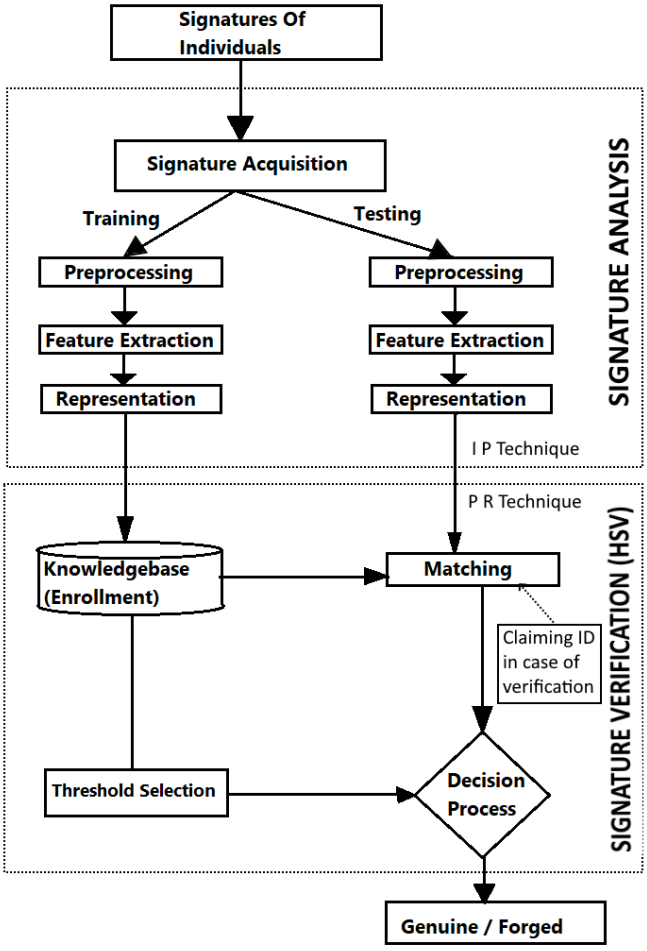


Figure 1: Depiction of HSVR system

2.2 Distance Classifiers

By using this technique such as Euclidean distance experiments were carried out for tracing the solutions and considered this technique by taking comprehensive features such as "heights", "width", "slant angles", "maximum horizontal projection shift as measure of similarity". Fang et al. used Mahalanobis distance for computing similarity among the local and global feature vectors. This work used a data set which was consisting of 800 signatures from 20 individuals. In this method the nearest neighbor and a threshold classifier are used[7].

2.3 Bayesian Classifiers

As per the author considered the pertinence of the refitted version of Bayesian network for offline. This kind of network is different from the traditional Bayesian. This network node are splits into two classes namely common and alternative hypothesis, to build the network in tree structure. This captures the conditional probability of every node and topological relations between the components with the nodes of the network. this method uncertainties in structure description between various components which are encoded. The FRR of 20% and the FAR of 14% were reported on a small database of 160 signatures from eight individuals [8].

2.4 Support Vector Machines

As per the author measured a Human Signature Verification and Recognition system as 2 class pattern-recognition problem. By this method "DTW technique", "SVM classifier", "PCA" and "Bayes, concepts are used. First the test signature is combined with referred signature using DTW to find authenticity of test signatures. This feature is then classified as authentic or forgery. For this case linear classifier in combination with PCA, Bayes, SVM classifier are used for classification purposes. The FRR of 1.64% and the FAR of 3.85% are obtained on a dataset of 306 signatures.

3 Overview of Feature matching techniques

3.1 SIFT

The SIFT mainly projected for solving the image rotation, transformations and opinion point in matching features, as per the author Lowe. The main key functionality is as follows (a) Evaluating the scale space using Gaussian techniques (b) point localization (c) orientation (d) generator.

3.2 SURF

The SURF is nearly alike to DoG features. The replacement of Gaussian average, the squares are used for approximation. It uses Binary Large Object detector based on Hessian matrix to find the point of interest and wavelet response for feature description as mentioned [9]-[10].

3.3 ORB

The ORB is an amalgamation of FAST key point and BRIEF descriptor with minor changes. The main technique of FAST is to identify the key points along with Harris corner to find top N points. This will not figure out the orientation and rotation variant but computes the weighted centroid of the located point. The resultant gives the corner point to centroid and this leads to orientation. The ORB rotation matrix is calculated using the orientation of patch and BRIEF descriptors are controlled to orientation[11].

3.4 Feature matching and representation

Even though there has been widely published papers and solved problem based on this offline signature, still there is lot of research problems which is still being carried out for analysis purpose. The

vital issue of this analysis is to resolve the most pertinent features. Special features like characterization of signatures based on feature and formulate the corresponding between the insinuation and test scribbles are the significant issue in authentication[12]-[13]. This study mainly relies on the effectual representation of signatures and learning similar features. These features are derived from geometrical properties for offline signatures as shown in the Figure 2.

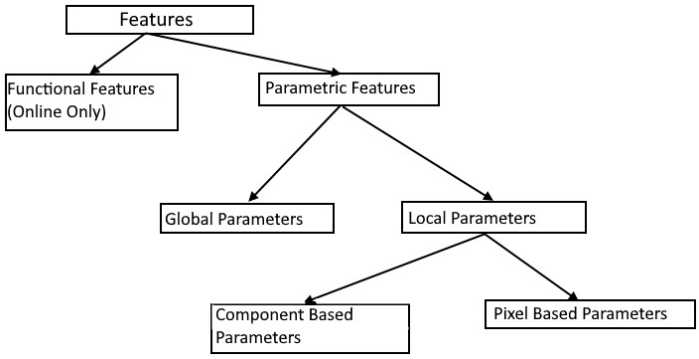


Figure 2: Organization of Offline and Online signature

4 Proposed Technique

In the proposed method the process of verification and recognition is carried on a huge database which consists of both genuine and forgery signatures. These signature samples features are extracted from a few samples of original specimen and rest original samples and forgery samples are used for verification process. The flow of the process is as shown in the Figure 3 The steps carried is as follows

1. CNN classifier is used to recognize which person the signature belongs to
2. If the recognition threshold is >80% feature matching technique is used, else it will be rejected.
3. Extracting the signature part from signature database and resize the same
4. Calculating the keypoints using ORB technique.
5. Comparing the key points with original sample within database.
6. If the number of matches between the given sample and the original image is > than the average number of matches between the original signatures it will be accepted else rejected.
7. Carried out the experiment for matching and classifying the genuine and forgery.

5 Feature Detection And Matching Using ORB

The most relevant for problem solving the task relating to application as mentioned in the paper [14]. The offline signature samples specific structures like point, edges were considered. The resultant neighbourhood operation is applied on the sample. The ORB featured was measured for the experimental purpose, whereas the SIFT performed better detection on the same problem. This builds well-known "Features from accelerated segment test", "Features from accelerated segment test", key-point detector and "Binary Robust Independent Elementary Features descriptor".

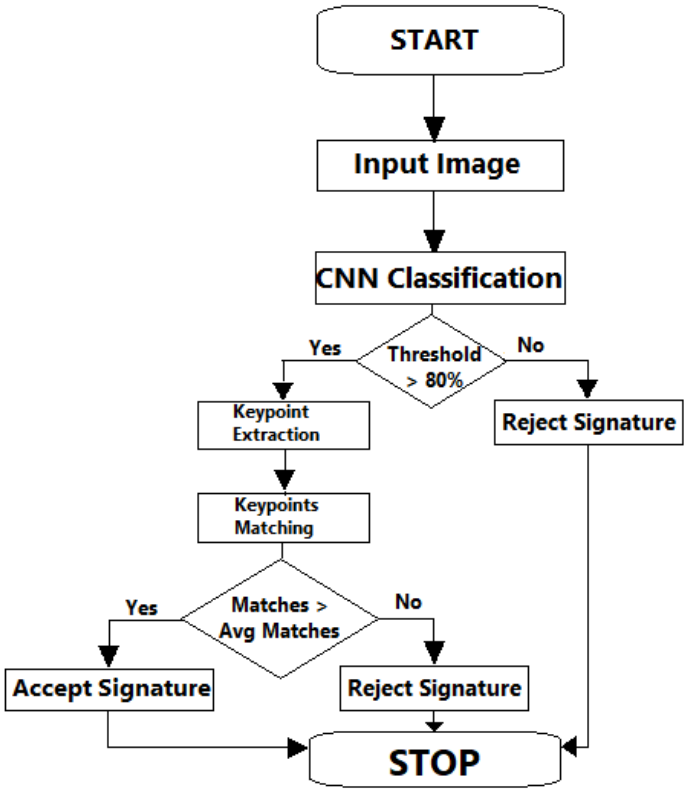


Figure 3: Architecture flow of Human Signature Verification Recognition

"Features from accelerated segment test", is a corner detection method, which could be used to extract feature points is as shown in Figure 5. The "Features from accelerated segment test", Algorithm applied on the signature sample to find the key points is as follows

1. Chosen a pixel p on the signature sample which is to identified as key point or not a key point. This intensity be I_p.
2. Applied the appropriate threshold value t.
3. After this the pixel p was considered as a corner set of n contiguous pixels in the circle and then all the brighter pixels than I_p +t was considered.
4. To improve the accuracy of the algorithm, first the intensity of the pixel was compared 1,5,9,13 of the circle with I_p. Finally all least three pixels out of these four pixels were considered because of satisfying the threshold criterion.
5. The least three of the four pixel values I1, I5, I9, I13 were not found above nor below the I_p + t, then p is not an interest point was discarded.

6. These steps were repeated to all the pixels in the signatures The BRIEF performance is used to find the key points which was recognized by the prior FAST algorithm where this method was considered for processing and converting this into binary feature vector to epitomize the image of the given sample [15]-[17]. This consist of 0 and 1 were every key-point is designated by a feature vector which is 128-512-bit string and the selected random pair of pixels by default was identified as neighbourhood key point. The discern neighbourhood around pixel is known as a patch, which is a square of some pixel width and height. The first pixel in the random pair was drawn from Gaussian and centred around the key point with standard deviation. This creates a vector for each key-point in an signature sample is as shown in the Figure 7.

Algorithm 1: The rBRIEF() algorithm for ORB is as follows

Result: Greedy Search
Run each test against all training patches obtained from the sample;
Order the tests by their distance from a mean of 0.5, forming the vector T;
Greedy search;
while *R > threshold* **do**
  Put the first test into the result vector R and remove it from T;
  Take next test from T and compare with R;
  if *discard it* **then**
    If its absolute correlation of R ¿ threshold;
    else;
  else
    add it to R;
  end
  if *If length(R) < 256* **then**
    Repeat until length of R = 256 ;
  else
    raise threshold and try again;
  end
end

ORB feature matching with Brute Force Technique To attain the fast and accurate key-point matching BFT was instigated. The gain of using this descriptor was one feature in first set and it is matched with all residual features and the second using some distance calculation. Finally, the adjoining one is returned.

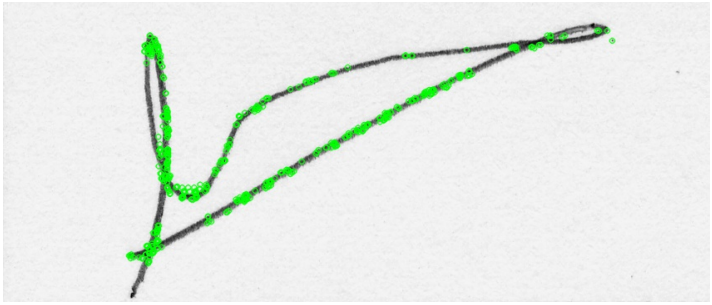


Figure 4: Detected keypoints

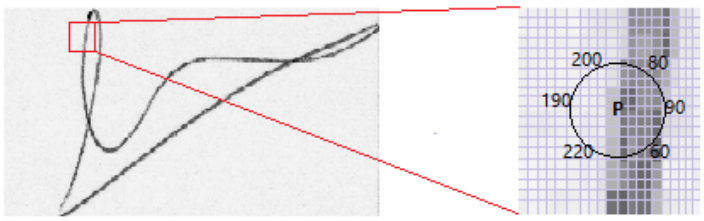


Figure 5: FAST Algorithm work flow

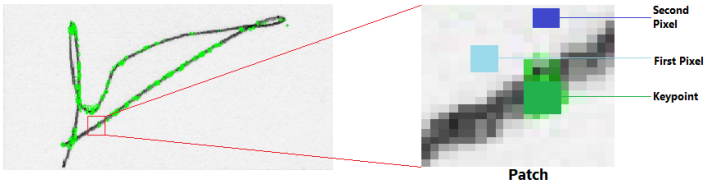


Figure 6: BRIEF Algorithm work flow

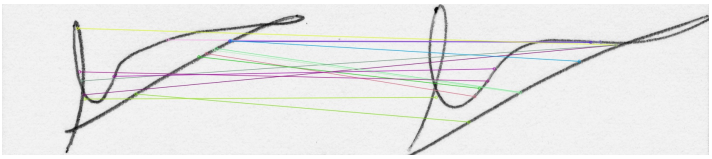


Figure 7: Good matches of keypoints from 2 signature samples

6 Experimental Results Conducted

F.A.R Choosing the counterfeit signature sample and discerning it is a genuine signature. It is as given by the equation 1

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\left(\frac{s4}{V(x)^2}\right)-3\frac{(n-1)^2}{(n-2)(n-3)} \tag{1}$$

False Acceptance Rate=Number of Forgery Signature Accepted/Number of Forgery Tested

False Rejection Rate (F.R.R) Snubbing the signature though it is frank sampled signature discerning it is a forgery sample. This is given by the equation 2

$$Skewness = \frac{n\sqrt{n-1}}{n-2}\frac{\sum_{i=1}^{n}\left(X_i - {X_{avg}}^3\right)}{n-2\left(\sum_{i=1}^{n} left(X_i - {X_{avg}}^3\right)^3/2\right.} \tag{2}$$

Skewness can also be defined as ratio of the 3rd moment and 2nd moment raised to the power of 3/2 ratio of the 3rd moment and the standard deviation as mentioned in the equation.
3 and 4

$$Skewness = \frac{m_3}{{m_2}^3/2} \tag{3}$$

$$Skewness = \frac{m_3}{\sigma^3} \tag{4}$$

F.R.R=Number of Genuine Signature Rejected/Number of Geniune Tested

Table 1: Online signature dataset for research purpose

Datasets	No. of Signers	No. of Genuine	No. of Forgeries	total no. of signatures
MCYT -75[151]	75	75*15	75*15	2250
GDPS -signature	160	160*24	160*30	8640

Table 2: Summary of DB-605 datase

Datasets	No. of Individuals	No. of Genuine	No. of Forgeries	total no. of signatures
DB -200	200	200*28	200*28	11200
MCYT -300	330	330*25	330*25	16500
DB -605	605	14975	14975	29950



Figure 8: Images of MCYT-75: (a) Accurate signatures (b) Corresponding fraudulence signatures

Table 3: Representation of C.N.N Architecture

Layer(type)	Output-Shape	Parameter
convl2d_1 (Convl2D)	(None, 173, 173, 32)	320
convl2d_2 (Convl2D)	(None, 169, 169, 64)	51264
batch_normalization_1	(Batch (None, 169, 169, 64)	256
activation_1 (Activation)	(None, 169, 169, 64)	0
max_pooling2d_1	(MaxPooling2 (None, 84, 84, 64)	0
global_average_pooling2d_1	((None, 64)	0
dense_1 (Dense)	(None, 70)	4550

Total-parameters: 56,390
Trainable-parameters: 56,262
Non-trainable-parameters: 128

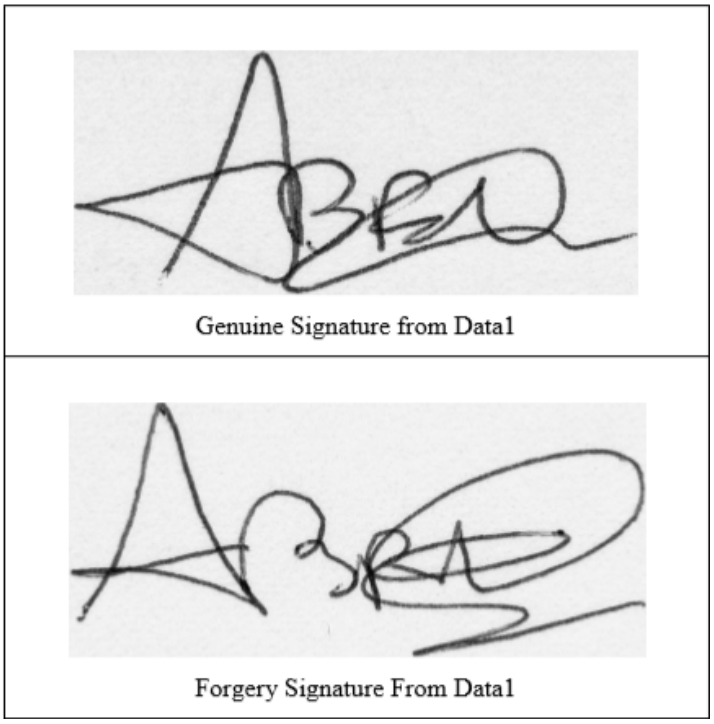


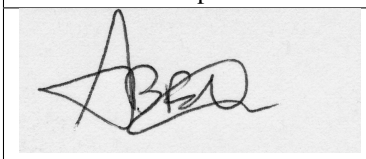
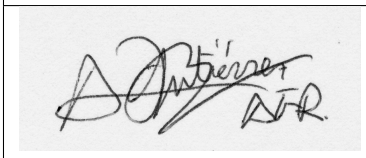
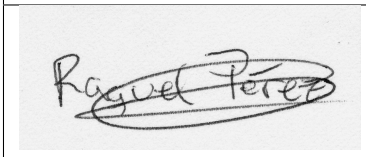
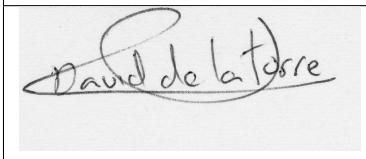
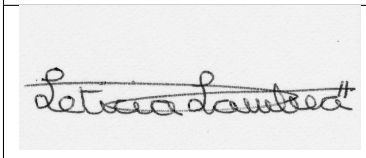
Figure 9: Sample of Feature extraction

6.1 Neural Network setup for experiment

Initialize position: This is the first stage, where NN is put up by describing the Network architecture (i/p and o/ps) . Further, the additive the number of patterns which will be used to train the network has to be considered along with the number of units in output layers are decided **Generate Training Set:**The quality of cognitive process is deliberated as important to the neural network; this gives an ability to recognize any changes in the signature images of the same person. Many individual learning algorithms have been developed for the verification/recognition purposes.

Create Neural-Network Set: In the antecedent period, the signature samples of every persons are divided into training and testing set. The training set are fed into a multilayer network. Multilayer NN is nothing but network with multiple hidden layers. In this work, data flow in the forward direction that is from input layer to output layer.

Table 4: Summary of datasets used

Sample	FAR	FER	Accuracy
	0	0	100
	0	20	90
	0	20	90
	0	40	80
	0	0	100
	0	20	90

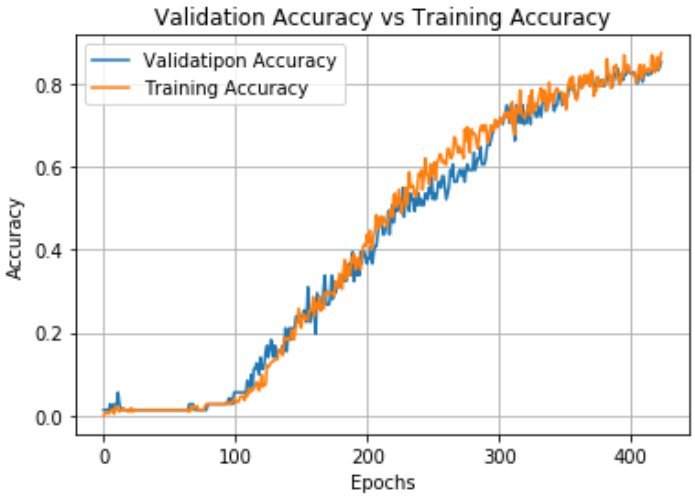


Figure 10: Training Accuracy and Validation Accuracy

In this phase, weights are stated for every construction using the simple weighting equation which is preferred to be forwarded that is, data flows from the input layer to the output layer. This stage specifies the weights which are available on every connection using the simple weighting equation. Weight$(t+1)$ = $weight\,(t) + \eta\,(\omega weight\,(t))$ the weight t+1 is equal to the weight at time t plus the learning rate times the calculated weight change.

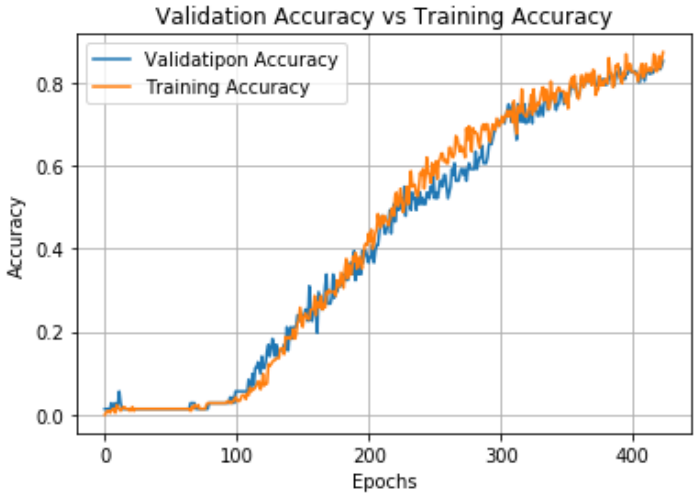


Figure 11: Training Accuracy and Validation Accuracy



Figure 12: Training Loss and Validation Loss

7 Conclusion

In the paper Contour based technique was used to classify signature into 2 classes (genuine and fake) the accuracy achieved was around 81% though this method used was fast and robust, the main disadvantage posed by the method was it required 2 classes of each signature and about 15 samples from each class. By applying the Multiple classifiers technique using Hu and Hog features, it was able to achieve an accuracy of around 90%. In this experiment, we set out to find various methods which proved effective for signature classification. Though the method was fast, it posed the same disadvantage of having to use 2 classes per person (genuine and fake) and consider about 15 samples each. This was far from a practical approach as collecting 15 samples for each person was not possible. In this paper, we sought out to find a practical approach

to classify signatures into real and fake considering practicality of the problem. Hence as a practical approach, we collected 3 genuine signatures from each person and used a combination of deep learning and machine learning to classify signatures. The accuracy in this case was found to be around 92% but this method gets rid of the disadvantages faced in the previous 2 cases and makes the classification task suit the real world problem.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We would like to thank to N.M.A.M Institute of Technology, Nitte, INDIA and Ritsumeikhan University Japan,for there support towards this paper.

References

[1] T.S. Wilkinson, T.S. Pender, J.W. Goodman, "Use of synthetic discriminant functions for hand-written signature verification". Appl. Opt. **30** (23), 33453353, 1991.

[2] J.E. Ward, R.E. Wendell, "A new norm for measuring distance which yields linear location problems". Operation Reasearch, (28) 836844, 1980.

[3] Xiao, "Signature verification using a modified Bayesian network", Article in Pattern Recognition **35**(5) 983-995, May 2002. 10.1016/S0031-32030100088-7

[4] M.B. Yilmaz, B. Yanikoglu, C. Tirkaz, "Offline signature verification using classifier combination of HOG and LBP features- 2011 international joint 2011.DOI: 10.1109/IJCB.2011.6117473

[5] David G. lowe, "Distinctive Image Features from Scale-Invariant Key-Points", International Journal of Computer Vision, **60**(2), pp. 91-110, 2009.

[6] M. Ramakrishna, S.S. Shylaja, "Is ORB efficient Over Surf for Object Recognition?", International Journal of Advanced Research in Science Engineering and Technology, 2014, **3**(8), pp. 2783-2788, ISSN: 2278-1323

[7] J. Li, X. Li, N. Yang, Xingming, "Segmentation Based Image Copy-Move Forgery Detection Scheme", IEEE Transactions on Information Forensics and Security, 2015, **10**(1), pp.507-518, 2015. DOI: 10.1109/TIFS.2014.2381872

[8] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Key-Points", International Journal of Computer Vision, **60**(2), pp.91-110, 2009. DOI https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[9] B. Moghaddam, C. Nastar and A. Pentland, "A Bayesian similarity measure for deformable image matching", Image and Vision Computing, **19**(5), pp. 235-244, 2001. https://doi.org/10.1016/S0262-8856(00)00059-7

[10] M. Gzel, "A Hybrid Feature Extractor using Fast Hessian Detector and SIFT", Technologies, **3**(2), pp. 103-110, 2015. https://doi.org/10.3390/technologies3020103

[11] X. Bo, W. Junwen, L. Guangie, D. Yuewei, "Image Copy-Move forgery Detection Based on SURF", International Conference on Multi-media Information Networking and Security, 2010, pp. 889-89. DOI: 10.1109/MINES.2010.189

[12] E.R.W. Garage, M. Park, "ORB: An efficient alternative to SIFT or SURF", IEEE International Conference on Computer Vision, 2011, pp. 2564-2571. DOI: 10.1109/MINES.2010.189.

[13] T. Mohmood, "A Survey on Block Based Copy-Move Image Forgery Detection Techniques", International Conference, pp.1-6, 2015. DOI: 10.1109/ICET.2015.7389169

[14] P. Sykora, P. Kamencay and R. Hudec, Comparison of SIFT and SURF Methods for Use on Hand Gesture Recognition based on Depth Map, AASRI Procedia, vol. 9, pp. 19-24, 2014. https://doi.org/10.1016/j.aasri.2014.09.005.

[15] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "ORB: and efficient alternative to SIFT or SURF", IEEE International Conference on Computer Vision, 2011. DOI: 10.1109/ICCV.2011.6126544.

[16] C.V. Aravinda, L. Meng, Akshaya, U. K. Reddy, "Signature Recognition and Verification Using Multiple Classifiers Combination of Hu's and HOG Features", 2019 International Conference on Advanced Mechatronic Systems (ICAMechS) 2019. DOI: 10.1109/icamechs.2019.8861686 ISBN: 9781728134802

[17] C.V. Aravinda, L. Meng, Akshaya, U.K. Reddy, "An approach for signature recognition using contours based technique", 2019 International Conference on Advanced Mechatronic Systems (ICAMechS), 2019. DOI: 10.1109/icamechs.2019.8861686 ISBN: 9781728134802

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 01-08 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Qualitative Properties of a Cell Proliferating Model with Multi-phase Transition and Age Structure

Youssef El Alaoui*,[1], Larbi Alaoui[2]

[1]*Faculty of Sciences, University Mohammed V - Agdal, Rabat, 10010, Morocco*

[2]*International University of Rabat, Sala Al-Jadida, 11100, Morocco*

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 18 October, 2020
Online: 08 November, 2020

Keywords:
Semigroup
Dual semigroup
Asynchronous exponential growth

ABSTRACT

In this paper we study a cell division cycle modeled by a system of partial differential equations with an age structure. This model translates the many regulatory mechanisms within the cell cycle where it introduces the notion of phases. The individual cell can be either in I phases where the transition between theses phases are ordered and unidirectional. The model is related to the suns and stars caluculus via the dual semigroups of operators that are considered as solution of an abstract integral equation equivalent to a Volterra type equation of the form $w(t) = \Phi(w_t)$. We will determine the core operator Φ and prove that the semigroup solution of the model possesses the asynchronous exponential property. The model permits different types of controls where the provided framework allows better control on the model parameters and yields the characterization of the intrinsic rate of natural increase through properties of the core operator Φ. Finally, we demonstrate that the asymptotic behavior of the model is governed by the simple dominant eigenvalue and its associated eigenvector, that leads to the dispersion of the cell structure through the future generations.

1 Introduction

This paper is considered as an extension of the work originally presented in 2019 International Conference of Computer Science and Renewable Energies ICCSRE [1], that represents more refined work where we investigated the long time behavior of a cell cycle model using a well founded mathematical theory.
The cell cycle is an ordered set of molecular events that a cell undergoes until it divides into two daughter cells. It has been recognized that the cell cycle can be divided into two major phase: The interphase and the mitosis phase denoted by M. The interphase is commonly composed of three phases : G_1, S, and G_2. It is the longest part of the cell cycle and is devoted to duplication of the cell's component. During the S phase, the chromosomes are duplicated. It is known that throughout the M phase, the cell split, producing two daughter cells [2]. In each phase the progression in the division cycle is reglemented by a serial of molecular regulation events which are ordered, directional and punctuated by checkpoints. These latter make use of some protein dephosphorylation and degradation where they are under control of the molecular circadian clocks. In fact, several animal experiments showed that the role of these clocks are essential to exert a rhythmic regulating control of each step [3], [4]. These steps are intracellular cascades which ensures the smooth conduct of the proliferation. The extracellular factors such growth factors play an important role in the controling strategy which ensures the environmental requirements are adapted to the proliferation of the cell. Thus, the regulation of the cell cycle is a crucial process to the survival of the cell. It is known that there exists two key classes of regulatory molecules: cyclins (regulatory subunits) and cyclin dependent kinases (catalytic subunits) denoted by CDKs where through phosphorylation activates or inactivates target proteins to permit entry in the next phase of the cell cycle.

Through the cell cycle, there exists different restriction points. The first one is called start checkpoint. It regulates the progression in the G1 phase. The transition from G_1 to S is regulated by CDK4/cyclin D and CDK6/cyclin D where CDK is expressed in cells and cyclin D is synthesized due to the growth factors. Various cyclins are produced at each phase: S, G_2 and M, resulting in the periodic formation of distinct combinations of cyclin-CDK complexes that trigger the cell cycle events [2, 5].

These various mechanisms of regulation are showed to be inefficient in cancer cells. In addition, the disruption of circadian clocks is showed to enhance the tumour growth [6]–[8]. In fact, the circadian rhythms play an important role in the cell cycle control where they contribute to the cell cycle regulation by the synchronisation of the

*Corresponding Author: Youssef EL ALAOUI, is.youssefelalaoui@gmail.com

control system. However, these dynamics have still to be understood and investigated. Further, various speculation tried to explain this phenomenon which give rise to the concept of synchronisation, that is, to the speculation that less synchronised phases are within cell population, the faster is proliferation. It is known that circadian proteins control CDKs and theses latter control phase transitions. In the following, we want to investigate the question if circadian clocks control the proliferation of the cell population.

The paper is concerned with a system of partial differential equations (PDE) describing the case of a multi transition phases cycle [9] in which cells can transit from one phase to the next one. The cell cycle is assumed to be divided into I different phases where the individual cell can be distinguished from one another according to their age a. These cells can growth, die and transit to the next phase where the rates describing the associated physiological process depend only on the cell's age. Then, the model read as follows (see figure 1)

$$\begin{aligned} &\frac{\partial}{\partial t}p_i(t,a)+\frac{\partial}{\partial a}p_i(t,a)=-\mu_i(a)p_i(t,a)-\mathcal{K}_{i\to i+1}(a)p_i(t,a)\\ &p_i(t,0)=\int_0^\infty \mathcal{K}_{i-1\to i}(a)p_{i-1}(t,a)da \quad \text{for } 2\le i\le I \\ &p_1(t,0)=2\int_0^\infty \mathcal{K}_{I\to 1}(a)p_I(t,a)da \end{aligned} \tag{1}$$

Here $p_i(t,a)$ is the density of the population of cells at the ith phase of age a at the instant t, that is, $\int_{a_1}^{a_2} p_i(t,a)da$ is the number of cells at the ith phase with an age in the interval (a_1,a_2) at the instant t. The transition from the state i to the state j is given by the rate $\mathcal{K}_{i\to j}(a)$, that is $\mathcal{K}_{i\to j}(a)dt$ is the probability that a cell having an age a at instant t in the phase i transit in the time interval $(t,t+dt)$ to the phase j, in which they start with age 0, the last phase when $i=I$ refer to mitosis as it is expressed by the boundary conditions in (1). At the ith state, the cells go to apoptosis at the rate $\mu_i(a)$, that is $\mu_i(a)dt$ is the probability that a cell having age a at instant t die in the time interval $(t,t+dt)$ in the phase i. Since the effective separation of the mother cell into two daughter cells, that is, since the cell birth, the individual cell can quantitated by a continuous variable corresponding to cell age.

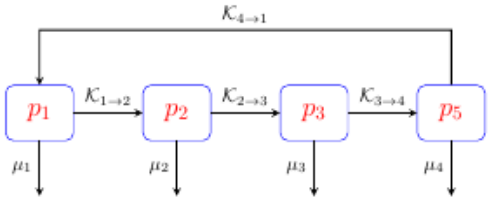


Figure 1: Flow chart of the model describing the evolution of cells along each phase of the cycle and the transition to the next phase between the cell densities. Here I=4, this corresponds to G1-S-G2-M.

It is known that cells transit through the cell cycle with variability in intermitotic times which correspond to different biological processes such the proteic (G_1 phase) or DNA (S phase) or microtuble (M phase) synthesis . As contrary to the model given in [9], the progression speed is assumed to be constant, that is one does not take into account the external growth factors. Indeed, the progression speed can be enhanced by the external growth factors in the early of the G_1 phase.

The authors in [9]–[16] considered a partial differential equations modeling a proliferating cell population with taking into consideration their external controls which can be hormonal, circadian, pharmacological . . . that leads to investigate the consequences of a periodic control on these models. Their results assert that the effect of circadian control on tumour proliferation is indirect.

Here, the death rates are supposed only age dependent. However, the case time-periodic death rates studied in [10, 11], using the Floquet theory, had shown a higher first eigenvalue and thus enhanced proliferation of the population.

For the one phase cell transition model [16, 17], it has been proved that if we influence the cell cycle with some control of a periodic function, we obtain that cells with cell cycle duration slightly lower than the control function period are selectively more advantageous, in addition, they hypothetize that the effect of disruption of circadian rhythms on tumour growth enhancement is indirect. The model for I phases was first introduced in [9], where the authors studied the long time behaviour of this model with the effect of a circadian control using an entropy method. Their results show that it imposes a circadian rythm to the cell cycle.

Mathematical cell cycle models have been treated since the 1960s by authors such as Fredrickson, Bell, Sinko and Streifer. These models were traditionally formulated as partial differential equations for the age and/or size distribution of the cell population. Diekmann et al. [18] used semigroup methods to give conditions for when there is, and also when there is not, asynchronous exponential growth of the population in a constant environment. More recently, Diekmann and coworkers have argued in favour of using (abstract) integral equations for the birth rate instead of PDE, also, and in particular, for the nonlinear models resulting from interaction via the environment.

The purpose of this paper is to create a mathematical framework for the analysis of this type of models by using the theoretically well established semigroup method based on the suns and stars calulcus using the perturbed dual semigroups of operators on the spaces of the form $\mathcal{E}:=L^1((-h,0),\mathcal{F})$. These semigroups are solutions of the integral equation [19]–[21] of the type :

$$p(t)=P_0(t)f+\int_0^t P_0^{\odot *}(t-\tau)\Phi(p(\tau))d\tau, \quad \text{for } t\ge 0 \tag{2}$$

where $h>0$, $\mathcal{F}$ is a Banach space and $\Phi:\mathcal{E}\to\mathcal{Z}$ and $\mathcal{Z}$ is some bigger Banach space. It has been proved that this method is well suited for applications, in particular in the field of cell biology models [22, 23].

In the following, the main result is to provide necessary and sufficient conditions on the parameters building the cell division cycle model (1) in order to yield the property of asynchronous exponential growth.The asynchronous exponential property is explained mathematically by the fact that a simple dominant eigenvalue govern the spectrum of the infinitesimal generator of the semigroup solution of the model. In the context of the semigroup theory, this operator will be the associated infinitesimal generator. We can say that the asynchronous exponential growth occurs when the following estimate $\|e^{-\lambda_0 t}p(t,\cdot)-\mathcal{P}p_0\|$ decreases to 0 exponentially as t tends to infinity, where the vector population is defined by $p(t,a)=(p_i(t,a))_{1\le i\le I}^{tr}$ and $p_0(a)$ is the initial given distribution of the population. This latter

means that when the cell population is multiplied by an exponential factor, know an asymptotic stabilization around a unidimensional projection. It is noted that this projection is only dependent on the initial structure of the concerned population.
We will use the theory developed in the works [19]–[21] to exhibit the link between the solutions of the model and the perturbed dual semigroups. First, we determine the operator Φ. Secondly, we derive different results, only based on the properties of the operator Φ ranging from existence to qualitative properties and particulary give a characterization of the Malthusian parameter as a spectral radius of some bounded operator.
Throughout this paper, the two parameters $\mathcal{K}_{i\to j}(a)$ and $\mu_i(a)$ are supposed to satisfy the assumption

$\left(\mathbf{H}_{\mathcal{K},\mu}\right)$: (1) The rates $\mathcal{K}_{i\to j}, \mu_i$ defined from $(0,\bar{a})$ to $\mathbb{R}$ are continuous.
(2) $\mathcal{K}, \mu \in L^\infty((0,\bar{a}),\mathbb{R})$ such that there exists $0 < \bar{\mathcal{K}} < \infty$ where $0 < \mathcal{K}_{i\to j}(a) \le \bar{\mathcal{K}}$ and$\mathcal{K} \not\equiv 0$ in addition there exists $\bar{\mu} > 0$ such that $0 \le \mu_i \le \bar{\mu}$

2 Well posedness

In this section, the mathematical analysis of the multi phases proliferating cell cycle model will be provided using the theory developed for the perturbed dual semigroups [24]–[27] associated to the equation of Volterra type :

$$p(t) = \phi(p_t) \tag{3}$$

where the Hale notation $p_t(s) := p(t+s)$ is used (see [28]–[35]), and the operator $\phi : \mathcal{E} := L^1((-h,0),\mathcal{F}) \to \mathcal{F}$. Our construction will permit to link the solutions of the model (1) to the solutions of the integral equation

$$p(t) = P_0(t)f + j^{-1}\left(\int_0^t P_0^{\odot *}(t-s)\Phi(p(s))ds\right) \tag{4}$$

where P_0 satisfies the properties of a semigroup of bounded linear operators on $\mathcal{E}$ with $\mathcal{E}^\odot \subset \mathcal{E}^*(\subset$ stands as the linear subspace) on which P_0^* yields the strong continuity, that is,

$$\mathcal{E}^\odot = \{f^* \in \mathcal{E}^* : \lim_{t\downarrow 0}\|P^*(t)f^* - f^*\| = 0\}$$

It can be shown that this subspace is closed in the norm topology of $\mathcal{E}^*$ and weak* dense subspace on $\mathcal{E}^*$ [24]. In addition, the semigroup on $\mathcal{E}^\odot$ is denoted by $P_0^\odot := P_0^*|_{\mathcal{E}^\odot}$. It is obvious that this restriction build a semigroup that is strongly continuous on the subspace $\mathcal{E}^\odot$. Since we consider an element of $\mathcal{E}$ as a bounded linear functional on $\mathcal{E}^\odot$. We can embed $\mathcal{E}$ into $\mathcal{E}^{\odot *}$ via the map q of $\mathcal{E}$ into $\mathcal{E}^{\odot *}$ defined by :

$$< f^\odot, qf >=< f, f^\odot >, \qquad \text{for } f \in \mathcal{E}, f^\odot \in \mathcal{E}^\odot$$

It is obvious that $q(\mathcal{E}) \subset \mathcal{E}^{\odot\odot}$. When $q(\mathcal{E}) = \mathcal{E}^{\odot\odot}$ then $\mathcal{E}$ is said to be $\odot$-reflexive with respect to $P(t)$.

Go back to the study of the model (1). The state space considered is $\mathcal{E} := L^1([0,\tilde{a}],\mathbb{R}^I)$. The parameter $\tilde{a} > 0$ is the maximal age that an individual cell can attain. This space is endowed with norm

$$\|f\|_{\mathcal{E}} = \int_0^{\tilde{a}} \|f(a)\|_{\mathbb{R}^I} da, \qquad f \in \mathcal{E}$$

Next, we introduce the operator $\chi(a)$ that yields the equation

$$\begin{aligned} \chi'(a) &= -\chi(a)\mathcal{M}(a) \\ \chi(0) &= Id(\text{ stands for identity}) \end{aligned} \tag{5}$$

with

$$\mathcal{M}(a) = \mathrm{Dg}\,(m_{i,i}(a))_i \qquad 1 \le i \le I \tag{6}$$

where Dg stands for the diagonal matrix and the entries of the matrix $\mathcal{M}$ hold

$$m_{i,j}(a) := \begin{cases} \mu_i(a) + \psi_{i\to i+1}(a) & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{7}$$

The aim of the introduction of this transformation is to demonstrate that $\{P(t)\}_{t\ge 0}$ the semigroup solution of the model (1) is equivalent to $(P_\phi(t))_{t\ge 0}$ the semigroup solution of (4). Thus, via this machinery and under the properties of the operator ϕ, we will establish the asynchronous exponential property only by handling with the equivalent semigroup and go back to the original one by using the cited transformation.
Before going to give the main results, we need some settings. First, let set $p(t,a) := (p_i(t,a))_{1\le i\le I}^{tr}$. So via this vector, our initial model (1) can be reformulated as an vector type Mckendrick Von Foerster model

$$\frac{\partial}{\partial t}p(t,a) + \frac{\partial}{\partial a}p(t,a) = -\mathcal{M}(a)p(t,a) \tag{8}$$

with the boundary condition at age 0

$$p(t,0) = \int_0^{\tilde{a}} \mathcal{B}(a)p(t,a)da \tag{9}$$

The matrix $\mathcal{M}$ is given above by (6) and (7) and the matrix $\mathcal{B}$ is given by

$$B(a) = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 2K_{I\to 1} \\ K_{1\to 2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & K_{2\to 3} & 0 & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & K_{I-1\to I} & 0 \end{pmatrix}$$

After we gave a compact form to the original model (1). Now, we consider the transformation

$$m(t,.) := \Gamma p(t,.) \tag{10}$$

where $(\Gamma g)(a) := \chi(a)g(a)$ with $a \in (0,\bar{a})$ and the operator χ satisfies the first order differential equation (5). Via this latter, we arrive at the equivalent system where it is more easily to handle than the original one (8). The model is formulated as given

$$\begin{aligned} \frac{\partial}{\partial t}m(t,a) + \frac{\partial}{\partial a}m(t,a) &= 0 \\ m(t,0) &= \int_0^{\tilde{a}} \theta(a)m(t,a)da \end{aligned} \tag{11}$$

The kernel θ is given by

$$\theta(a) = \mathcal{B}(a)\chi^{-1}(a), \quad \text{for } a \in (0,\tilde{a}) \tag{12}$$

The diagonal matrix χ^{-1} is given by

$$\chi^{-1}(a) = \mathrm{Dg}\left(\tilde{z}_{i,i}(a)\right)$$
$$\tilde{z}_{i,j} := \begin{cases} e^{-\int_0^a m_{i,i}(\tau)d\tau} & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{13}$$

Definition 2.1 *The strongly continuous semigroup $\{P(t)\}$ is said to be a translation semigroup on $L^1((-r,0),\mathcal{F})$ with $r > 0$ and $\mathcal{F}$ is a Banach space if the following $(P(t)f)(x) = f(x+t)$ is satisfied when $x + t < 0$.*

Diekmann et al. have demonstrated in [20] that the space $\mathcal{E}$ is $\odot$-reflexive with respect to P_0, where the operator P_0 is defined as the translation semigroup to the right and 0 on $\mathbb{R}^+$. Its generator $\mathcal{A}_0$ is given by

$$\mathcal{D}(\mathcal{A}_0) = \{f \in \mathcal{E} : f \in AC \text{ and } f(0) = 0\}$$
$$\mathcal{A}_0 f = -f'$$

where AC stands for the absolutely continuous space. We have the following identification (see [19] and [20])

$$L^1([0,\bar{a}],\mathbb{R}^I)^{\odot} \to C_0((-\tilde{a},0],\mathbb{R}^I)$$
$$L^1([0,\bar{a}],\mathbb{R}^I)^{\odot *} \to \mathrm{NBV}([0,\tilde{a});\mathbb{R}^I)$$

The semigroup $P_0^{\odot *}$ is the translation to the right and 0 on $\mathbb{R}^+$. Its generator $\mathcal{A}_0^{\odot *}$ is given by

$$\mathcal{D}(\mathcal{A}_0^{\odot *}) = \{f^{\odot *} \in \mathcal{E}^{\odot *} : f^{\odot *}(a) = \int_0^a g(s)ds \text{ and } g \in \mathcal{E}^{\odot *}\}$$
$$\mathcal{A}_0^{\odot *} f^{\odot *} = -f^{\odot *} = g$$

Using integration and the natural embedding q from $\mathcal{E}$ into $\mathcal{E}^{\odot *}$ and taking into account the boundary condition, we arrive at

$$m(t) = P_0(t)g + q^{-1}\int_0^t P_0^{\odot *}(t-s)\Phi(m(s))ds \tag{14}$$

where the core operator Φ defined from $\mathcal{E}$ to $\mathcal{E}^{\odot *}$ is given as a perturbation with finite dimensional range

$$\Phi(g) = \phi_1(g)\mathcal{H}_1 + \phi_2(g)\mathcal{H}_2 + \ldots + \phi_I(g)\mathcal{H}_I \tag{15}$$

The Heaviside function $\mathcal{H}_i$ is given by $\mathcal{H}_i(a) = e_i$ if $a \in (0,\tilde{a})$ else $\mathcal{H}_i(a) = 0$ where $\{e_1, e_2, \ldots, e_I\}$ is the canonical basis of $\mathbb{R}^I$. The canonical injection q is defined by $(qx)(a) = \int_0^a x(s)ds,\ s \in [0,\tilde{a})$ where the inverse is clearly differentiation.
The operator given in (11) is defined from $\mathcal{E}$ to $\mathbb{R}^I$ and it is given by the integral formula

$$\Phi g = \int_0^{\tilde{a}} \theta(s)g(s)ds \tag{16}$$

the kernel θ is given by (12). In the following, we adopt the following Hale notation

$$m_t(a) := m(t-a), \quad \text{for all } a \in [0,\tilde{a}]$$

with

$$m_0(a) = g(a) \qquad a \in [0,\tilde{a}] \tag{17}$$

In fact, the solutions of the formulation given by the integral equation (14) is proven to be equivalent to the solutions of the renewal equation [20]

$$\begin{aligned} m(t) &= \Phi(m_t), \quad \text{for } t > 0 \\ m_0(a) &= g(a), \quad a \in [0,\tilde{a}] \end{aligned} \tag{18}$$

where the latter generates a translation semigroup [28] given by

$$\mathcal{T}_\Phi g(a) = \begin{cases} g(a-t) & \text{if } a - t > 0 \\ \Phi(\mathcal{T}_\Phi(t-a)g) & \text{if } a - t \le 0 \end{cases}$$

The following result gives us the existence and uniqueness of the perturbed semigroup solution of the abstract integral equation (14)

Theorem 2.2 *Since the assumption $(\mathbf{H}_{\mathcal{K},\mu})$ holds. We obtain that $\phi \in \mathcal{L}(\mathcal{E},\mathcal{F})$. Then for the initial data $g \in \mathcal{E}$, the equation (14) has unique solution $m(t)$ on $[0,T)$ for some positive real number T where the operator $\mathcal{A}_\phi$ characterized as the following*

$$\begin{aligned} \mathcal{D}(\mathcal{A}_\phi) &= \{\varphi \in W^{1,1}((0,\bar{a}),\mathbb{R}^I), \varphi(0) = \phi\varphi\} \\ \mathcal{A}_\phi\varphi &= -\varphi', \qquad \varphi \in \mathcal{D}(\mathcal{A}_\phi) \end{aligned} \tag{19}$$

is the inifinitesimal generator of the semigroup P_ϕ on $\mathcal{E}$.

Proof. The hypothesis given by $(\mathbf{H}_{\mathcal{K},\mu})$ allows to us to obtain the first assertion about the boundedness of ϕ. Then, by using regular arguments on contraction mapping theory, we obtain the existence of the perturbed semigroup $(P_\phi(t))_{t\ge 0}$ solution of (14) on $\mathcal{E}$. □

3 Compactness and irreducibility

It is known that the eventual compactness result is often used to determine the asymptotic behavior of semigroups. In fact, it implies that the growth bound of the semigroup denoted by $\omega(T(t))$ equals the spectral bound denoted by $s(A)$, where

$$\omega(T) = \inf\{w : \|T(t)\| \le Me^{wt}, M \ge 0\}$$

and

$$s(A) = \begin{cases} \sup\{\Re e\lambda : \lambda \in \sigma(A)\} & \text{if } \sigma(A) \ne \emptyset \\ -\infty & \text{if } \sigma(A) = \emptyset \end{cases}$$

Definition 3.1 *We say that the semigroup $\{P(t)\}_{t\ge 0}$ is eventually compact if for some $t_0 \ge 0$ the operators $P(t)$ are compact for all $t \ge t_0$.*

We will give a criterion to obtain the compactness of the semigroup.

Lemma 3.2 *Since $\Phi \in \mathcal{L}(\mathcal{X},\mathbb{R}^I)$ is with finite Range. The semigroup $\{P(t)\}_{t\ge 0}$ given by (14) is eventually compact.*

Proof. The core operator Φ is spanned by the linearly independent set $\{\mathcal{H}_i\}_{1\le i\le I}$ in $\mathcal{X}^{\odot *}$, that is, Φ is with Range$(\Phi) < \infty$, where the core operator Φ takes the form given by (15). We know that there exists $\Phi_i^* \in L^\infty([-\tilde{a},0],\mathbb{R}^I)$ such that :

$$\Phi_i(\zeta) = \int_0^{\tilde{a}} < \zeta(a), \Phi_i^*(-a) > da, \qquad \forall \zeta \in \mathcal{X}$$

We conclude that :

$$\Phi(\zeta) = \sum_{i=1}^{n} \int_0^{\tilde{a}} < \zeta(a), \Phi_i^*(-a) > \mathcal{H}_i da, \qquad \forall \zeta \in \mathcal{X}$$

Let $t > 2\tilde{a}$, we have $\forall \tau, \tau' \in [0, \tilde{a}],\ \tau' < \tau$ for all f in the unit ball of $\mathcal{X}$. Then,

$$\begin{aligned}
&\|[P(t)\zeta](\tau) - [P(t)\zeta](\tau')\|_{\mathbb{R}^n} \\
&= \|q^{-1} \int_0^t P_0^{\odot *}(t-s) \left\{ (\Phi(P(s)\zeta)ds)(\tau) - (\Phi(P(s)\zeta)ds)(\tau') \right\} \|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(- \int_{t-\max\{t,\tau\}}^{t} \Phi(P(s)\zeta)ds + \int_{t-\max\{t,\tau'\}}^{t} \Phi(P(s)\zeta)ds \right) \|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(\int_{t-\max\{t,\tau'\}}^{t-\max\{t,\tau\}} \Phi(P(s)\zeta)ds \right) \|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(\sum_{i=1}^{n} \int_0^{\tilde{a}} \int_{t-\max\{t,\tau'\}}^{t-\max\{t,\tau\}} < P(s)\zeta(a), \Phi_i^*(-a) > \mathcal{H}_i ds\, da \right) \|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(\sum_{i=1}^{n} \int_0^{\tilde{a}} \int_{t-\max\{t,a+\tau'\}}^{t-\max\{t,a+\tau\}} < \Phi(P(s)\zeta), \Phi_i^*(-a) > \mathcal{H}_i ds\, da \right) \|_{\mathbb{R}^n}
\end{aligned}$$

Let $\alpha_1 = t - \tau$ and $\alpha_2 = t - \tau'$

$$\begin{aligned}
&\|[P(t)\zeta](\tau) - [P(t)\zeta](\tau')\|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(\sum_{i=1}^{n} \int_{\alpha_1}^{\tilde{a}+\alpha_1} \int_0^a < \Phi(P(s)\zeta), \Phi_i^*(\alpha_1 - a) > \mathcal{H}_i ds\, da \right. \\
&\left. - \sum_{i=1}^{n} \int_{\alpha_2}^{\tilde{a}+\alpha_2} \int_0^a < \Phi(P(s)\zeta), \Phi_i^*(\alpha_2 - a) > \mathcal{H}_i ds\, da \right) \|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(\sum_{i,j=1}^{n} \int_{\alpha_1}^{\tilde{a}+\alpha_1} \int_0^a < \Phi_j(P(s)\zeta)\mathcal{H}_j, \Phi_i^*(\alpha_1 - a) > \mathcal{H}_i ds\, da \right. \\
&\left. - \sum_{i,j=1}^{n} \int_{\alpha_2}^{\tilde{a}+\alpha_2} \int_0^a < \Phi_j(P(s)\zeta)\mathcal{H}_j, \Phi_i^*(\alpha_2 - a) > \mathcal{H}_i ds\, da \right) \|_{\mathbb{R}^n} \\
&= \|q^{-1} \left(\sum_{i,j=1}^{n} \int_0^a \left\{ \int_{\alpha_2}^{\tilde{a}+\alpha_2} < \Phi_j(P(s)\zeta)\mathcal{H}_j, \Phi_i^*(\alpha_1 - a) - \Phi_i^*(\alpha_2 - a) > \right. \right. \\
&\mathcal{H}_i da + \int_{\alpha_1}^{\alpha_2} < \Phi_j(P(s)\zeta)\mathcal{H}_j, \Phi_i^*(\alpha_2 - a) > \mathcal{H}_i da \\
&\left. \left. + \int_{\tilde{a}+\alpha_1}^{\tilde{a}+\alpha_2} < \Phi_j(P(s)\zeta)\mathcal{H}_j, \Phi_i^*(\alpha_2 - a) > \mathcal{H}_i da \right\} ds \right) \|_{\mathbb{R}^n} \\
&\leq \sup_{1 \leq i,j \leq n} \|\mathcal{H}_i\| \|\mathcal{H}_j\| \sum_{i,j=1}^{n} \left(\sup_{0 \leq s \leq 2\tilde{a}} |\Phi_j(P(s))\zeta)| \times \right. \\
&\left\{ \int_{\alpha_2}^{\alpha_2+\tilde{a}} |\Phi_j^*(\alpha_1 - a) - \Phi_j^*(\alpha_2 - a)| da + \int_{\alpha_1}^{\alpha_2} |\Phi_i^*(\alpha_2 - a)| da \right. \\
&\left. \left. + \int_{\tilde{a}+\alpha_1}^{\tilde{a}+\alpha_2} |\Phi_i^*(\alpha_2 - a)| da \right\} \right)
\end{aligned}$$

Hence, we obtain that $\{P(t)\zeta\}_{t\geq 0}$ is equicontinuous since $|\alpha_2 - \alpha_1|$ tends to zero uniformly for ζ in the unit ball of $C([0, \tilde{a}], \mathbb{R}^n)$. The rest of the proof is due to applying a famous result of Arzela-Ascoli, we obtain then that $\{P(t)\}_{t\geq 0}$ is eventually compact. □

Proposition 3.3 *Let assume that* $(\mathbf{H}_{\mathcal{K},\mu})$ *is satisfied. Then the strongly continuous semigroup P and* $\mathcal{T}_\Phi$ *are eventually compact.*

Proof. The core operator Φ given by the formula (15) is a perturbation of finite rank. Then by using the following result [35, Proposition 4], we obtain the eventual compactness of the semigroup solution $\{T_\Phi(t)\}_{t\geq 0}$. The perturbed semigroup $\{W(t)\}_{t\geq 0}$ is eventually compact by applying the result in Lemma3.2. □
In the follwing we define the family $(\tilde{\Psi}_\lambda)_{\lambda\in\mathbb{C}}$ of operators such that their entries are defined by

$$\tilde{\Psi}_{\lambda,i,j} := < \nabla_{\lambda,j}, \nabla_i^* >, \quad \lambda \in \rho(\mathcal{A}_0) \tag{20}$$

with

$$\nabla_{\lambda,i} = q^{-1} R(\lambda, \mathcal{A}_0^{\odot *}) \nabla_i^{\odot *}, \quad \lambda \in \rho(\mathcal{A}_0)$$
$$\{\nabla_i^{\odot *}\}_{1\leq i\leq I} \text{ is a linearly independent set in } \mathcal{X}^{\odot *}$$

If we adopt this formulation to the integral equation given by (14) such that

$$\begin{aligned}
\nabla_i^{\odot *} &= \mathcal{H}_i \\
\nabla_{\lambda,j} &= \left(q^{-1} R(\lambda, \mathcal{A}_0^{\odot *}) \mathcal{H}_j \right)(s) = e^{\lambda s} e_j \\
\nabla_i^* &= \Theta_{i,j}
\end{aligned}$$

where Θ is given by

$$\begin{pmatrix}
0 & 0 & \cdots & 0 & 2\psi_{I\to 1}\bar{z}_{I,I} \\
\psi_{1\to 2}\bar{z}_{1,1} & 0 & \cdots & \cdots & 0 \\
0 & \psi_{2\to 3}\bar{z}_{2,2} & \ddots & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \psi_{I-1\to I}\bar{z}_{I-1,I} & 0
\end{pmatrix}$$

Therefore we obtain

$$\tilde{\Psi}_\lambda x = \Phi(e^{\lambda \cdot} \otimes x) \quad \text{for } x \in \mathbb{R}^I$$

where the relation $\otimes$ is defined by

$$(e^{\lambda \cdot} \otimes x)(s) := e^{\lambda s} x \quad \text{for } s \in \mathbb{R}.$$

Next we will give a characterization of the spectrum of the infinitesimal generator $\mathcal{A}_\Phi$ associated to the perturbed semigroup W_Φ by using the family of operators defined above. This characterization permits to show that the spectrum is a point spectrum and there exists a characteristic equation, the roots of which are the eigenvalue of the infinitesimal generator of the associated perturbed semigroup. Furthemore, we will give a criterion which assure the simplicity of these eigenvalues.

Proposition 3.4 *λ belongs to $\sigma(\mathcal{A}_\Phi)$ if and only if λ belongs to $\sigma_p(\mathcal{A}_\Phi)$ if and only if*

$$\prod_{i=1}^{I} \int_0^{\tilde{a}} \exp\left(- \int_0^x \lambda + m_i(\tau) d\tau \right) \mathcal{K}_{i\to i+1}(x) dx = \frac{1}{2} \tag{21}$$

Proof. By transforming the integral equation (14) using the Laplace transformation, we obtain the following resolvent equation

$$(I - q^{-1} R(\lambda, \mathcal{A}_0^{\odot *})\Phi) R(\lambda, \mathcal{A}) = R(\lambda, \mathcal{A}_0), \text{ for } \lambda \in \rho(\mathcal{A}_0)$$

Then it is clear that λ belongs to the spectrum of $\mathcal{A}_\Phi$ if and only if the operator $I - q^{-1} R(\lambda, \mathcal{A}_0^{\odot *})\Phi$ is not invertible. By straightforward computation it is turn out that it is satisfied if and only if $\det(\mathbb{I} - \tilde{\Psi}_\lambda) = 0$ that is equivalent to the assertion (21) since the term $q^{-1} R(\lambda, \mathcal{A}_0^{\odot *})\Phi$ is bounded and $\sigma(\mathcal{A}_0) = \emptyset$. By the eventual compactness of the semigroup the spectrum of $\mathcal{A}_\Phi$ is ponctual. □

Proposition 3.5 *The spectral bound of $\mathcal{A}_\Phi$ equal to the growth bound of W_Φ, that is,*

$$s(\mathcal{A}_\Phi) = \omega(W_\Phi)$$

such that there exists a real dominant eigenvalue $\tilde{\lambda}$ of $\mathcal{A}_\Phi$ solution of

$$\tilde{\lambda} = s(\mathcal{A}_\Phi)$$
$$\prod_{i=1}^{I} \int_0^{\tilde{a}} \exp\left(-\int_0^x \lambda + m_i(\tau)d\tau\right)\mathcal{K}_{i\to i+1}(x)dx = \frac{1}{2} \qquad (22)$$

Proof. Since our semigroup solution of (14) is eventually compact, the result given in [20, Theorem 2.7] provides to us that $s(\mathcal{A}_\phi) = \omega(W_\phi)$. The map given by the spectral radius of the operator $\tilde{\Psi}_\lambda$ is continuous. In addition, the operator $\tilde{\Psi}$ decreases and it satisfies

$$r(\tilde{\Psi}_\lambda) \text{ tends to } +\infty \quad \text{as } \lambda \text{ tends } -\infty$$
$$r(\tilde{\Psi}_\lambda) \text{ tends to } 0 \quad \text{as } \lambda \text{ tends } +\infty$$

Then there exists a unique real root $\tilde{\lambda}$ that holds (22) where $\tilde{\lambda}$ is an eigenvalue of $\mathcal{A}_\phi$. Using the fact that the spectral radius of $\tilde{\Psi}$ is monotone, we have the result

$$\tilde{\lambda} = \sup\{\Re e\lambda : \lambda \in \sigma(\mathcal{A}_\phi)\} = s(\mathcal{A}_\phi)$$

Finally, since the semigroup is eventually compact, the real eigenvalue $\tilde{\lambda}$ solution of (22) is dominant.□

We can say from the above result that less control on the transition and mortality rates, the higher the growth rate, that is the proliferation is faster within cells with less synchronised phases.

Proposition 3.6 *The operator $\tilde{\Psi}_\lambda$ is irreducible such that $\tilde{\lambda}$ is a pole of order 1.*

Proof. The irreducibility is a strong positivity property for the semigroup that provides to us a convenient criterion about the simplicity of the poles. Let $g, g^* \in \mathbb{R}^I$ we have

$$< \tilde{\Psi}_\lambda g, g^* >= \int_0^{\tilde{a}} \exp(-\lambda a) \sum_{i=1}^{I-1} \mathcal{K}_{i\to i+1} g_{i+1}^* \bar{z}_{i,i} g_i + 2\mathcal{K}_{I\to 1} g_1^* \bar{z}_{I,I} g_I da > 0$$

□

Next, we investigate the behavior in the long term of the perturbed semigroup solution of (14), where we show that the property of asynchronous exponential property is fufilled

Theorem 3.7 *The solution semigroup $\{P_\phi(t)\}_{t\geq 0}$ of* (14) *holds*

$$\|\exp(-\lambda_0 t)P_\phi(t) - Q\| \leq \beta e^{-\gamma t} \text{ for } \gamma > 0, \beta \geq 1, \forall t \geq 0$$

such that Q given by

$$Qg = \alpha(g)(e^{\lambda_0 \cdot} \otimes \varphi_{\lambda_0})$$

and

$$\alpha(g) = \frac{< \psi_{\lambda_0}^*, \phi(\theta \mapsto \int_0^\theta \exp(\lambda_0(\theta - s))g(s)ds >}{< \psi_{\lambda_0}^*, \phi(\theta \mapsto \theta \exp(-\lambda_0\theta) \otimes \psi_{\lambda_0} >}, \quad \psi \in \mathcal{E} \qquad (23)$$

where $\psi_{\lambda_0} > 0$ ($\psi_{\lambda_0}^* > 0$) is the eigenvector of $\tilde{\Psi}_{\lambda_0}$ (respectively of $\tilde{\Psi}_{\lambda_0}^*$) such that $< \psi_{\lambda_0}^*, \psi_{\lambda_0} >= 1$.

Proof. Our state space $\mathcal{X}$ can be decomposed as the following

$$\mathcal{X} = N_{\tilde{\lambda}} \oplus R_{\tilde{\lambda}}$$

where

$$\begin{cases} N_{\tilde{\lambda}} = \ker(\mathcal{A}_\Phi - \tilde{\lambda}I) = e^{\tilde{\lambda}\cdot} \otimes \ker(I - \tilde{\Psi}_{\tilde{\lambda}}) \\ \quad = \{\alpha e^{\tilde{\lambda}\cdot} \otimes \psi_{\tilde{\lambda}} : \alpha \in \mathbb{R}\} \\ R_{\tilde{\lambda}} = \text{Range}(\mathcal{A}_\Phi - \tilde{\lambda}\mathbb{I}) \end{cases}$$

where $\varphi_{\tilde{\lambda}}$ is the positive eigenvector of $\tilde{\Psi}_{\tilde{\lambda}}$ associated to the eigenvalue $\tilde{\lambda}$ solution of (22). The projection of g on the subspaces $N_{\tilde{\lambda}}$ and $R_{\tilde{\lambda}}$ gives us

$$g = \alpha(g)\exp(-\tilde{\lambda}.) \otimes \psi_{\tilde{\lambda}} + \rho$$

Then our semigroup solution of (14) can be rewritten as

$$P_\Phi(t)g = \alpha(g)\exp(-\tilde{\lambda}(t - .))\psi_{\tilde{\lambda}} + (P_\Phi(t)|_R)(\rho)$$

Using the fact that $\omega(P_\Phi(t)|_R) < \tilde{\lambda}$. Then we obtain that

$$P_\Phi(t)g = \alpha(g)\exp(-\tilde{\lambda}(t - .))\psi_{\tilde{\lambda}} + o(\exp(\tilde{\lambda}t))$$

Then we obtain

$$\|\exp(-\tilde{\lambda}t)P_\Phi(t)g - Qg\|_{\mathcal{E}} \leq \beta\exp(-\delta t)\|g\|_{\mathcal{E}} \quad \text{for } \gamma > 0, \beta \geq 1, g \in \mathcal{E}$$

where $Qg = \alpha(g)\exp(-\tilde{\lambda}.)\psi_{\tilde{\lambda}}$. Then

$$\alpha(g) = \frac{< \psi_{\tilde{\lambda}}^*, \phi(\theta \mapsto \int_0^\theta \exp(\tilde{\lambda}(\theta - s))g(s)ds >}{< \psi_{\tilde{\lambda}}^*, \phi(\theta \mapsto \theta\exp(-\tilde{\lambda}\theta) \otimes \psi_{\tilde{\lambda}} >}, \quad g \in \mathcal{E}$$

such that $\psi_{\tilde{\lambda}}^*$ is a positive eigenvector of $\tilde{\Psi}_{\tilde{\lambda}}^*$ that yields the condition $< \psi_{\tilde{\lambda}}^*, \psi_{\tilde{\lambda}} >= 1$. □

Proposition 3.8 *The semigroup solution of* (1) *possesses the asynchronous exponential property.*

Proof. Using the relation $(\Gamma f)(s) = \chi(s)f(s), \forall s \in (0, \tilde{a})$ we obtain the similar semigroup

$$P(t)f = \Gamma^{-1}(P_\Phi(t)\Gamma f), \quad \forall f \in \mathcal{E}$$

Then by considering the formula $\tilde{Q}f = \Gamma^{-1}(Q(\Gamma f))$ we obtain the estimate

$$\begin{aligned} &\|\exp(-\tilde{\lambda}t)P(t)f - \tilde{Q}f\|_{\mathcal{E}} = \|\Gamma(\exp(-\tilde{\lambda}t)P(t)f - Qf)\|_{\mathcal{E}} \\ &\leq \|\exp(-\tilde{\lambda}t)(P_\Phi(t)\Gamma f) - (\alpha(\Gamma f)\exp(-\tilde{\lambda}.)\psi_{\tilde{\lambda}})\|_{\mathcal{E}} \\ &\leq \beta\exp(-\gamma t)\|f\|_{\mathcal{E}} \end{aligned}$$

We can conclude that the solution of (1) yields the asynchronous exponential property. □

As generations follow one another, the synchronisation between phases are less expressed within the evolving cell population, that is we observe a variability in the structure when a cell transition from the ith phase to the $(i+1)^{th}$ phase, and becomes fully desynchronised. In the following a threshold phenomenon appear and it can be formulated as the spectral radius of some integral operator

Proposition 3.9 *Let $r(\tilde{\Psi}_\lambda)$ be the spectral radius of $\tilde{\Psi}_\lambda$ defined by* (20)*. Then the following holds :*

1. *If $\int_0^{\tilde{a}} \exp\left(-\int_0^x m_i(\tau)d\tau\right)\mathcal{K}_{i\to i+1}(x)dx < \frac{1}{2}$ the trivial equilibrium state is asymptotically stable.*
2. *If $\int_0^{\tilde{a}} \exp\left(-\int_0^x m_i(\tau)d\tau\right)\mathcal{K}_{i\to i+1}(x)dx > \frac{1}{2}$ the trivial equilibrium state is unstable.*

Proof. We have shown that $\tilde{\lambda} = s(\mathcal{A}_\phi)$. Due to the monotonicity of the operator $\tilde{\Psi}_\lambda$ we have that the spectral bound $s(\mathcal{A}_\phi) < 0$ if

$$r(\tilde{\Psi}_0) = \int_0^{\tilde{a}} \exp\left(-\int_0^x m_i(\tau)d\tau\right)\mathcal{K}_{i\to i+1}(x)dx < \frac{1}{2}$$

which corresponds to the trivial equilibrium state. Then, since from

$$\|P(t)f\|_\mathcal{E} \leq e^{\tilde{\lambda}t}\|f\|_\mathcal{E}$$

we have $\lim_{t\to\infty} P(t)f = 0$ and it is unstable when

$$\int_0^{\tilde{a}} \exp\left(-\int_0^x m_i(\tau)d\tau\right)\mathcal{K}_{i\to i+1}(x)dx > 1/2 \quad \square$$

4 Conclusion

The study of the cell population dynamics by taking into account the proliferation process and generations overlapping take an important part in the mathematical modeling and analysis. Various analytical and probabilist methods were performed to this aim. In particular, the semigroup method, that demonstrate it is theoretically well established combining between different theories such as operator theory, functional analysis, spectral theory ...

In this paper, the dual semigroup theory provide a useful framework to examine these processes. In particular, the context of the control of the cell cycle where many complex regulatory processes are involved where the structuration by age permits to us to track the individual cell position through the multiple generations. We proved the asynchronous exponential property that consists as an important trait within the cell populations. In fact, the synchronisation or desynchronisation is essential in tumour therapy in which drugs can synchronize cell phases and others can destroy cells in the same phases.

Here, we investigated the role of the circadian clock in the control of the proliferation within the cell population. Mathematically, we can conclude that the disruption of the circadian clock, in particular the desynchronisation between the transition from a phase to another one has a major effect on the proliferation of cells. In addition, we demonstrate that as time evolves, the cells are more desynchronised than the younger generations. In this context, natural questions arise about the differences in synchronisation between healthy and cancerous cells. This question is essential in order to understand the cell cycle and to use such knowledge to make optimisation methods for cancer therapy.

In this paper, we consider only age dependent parameter, in perspective we want to examine the effect of time dependent and different periodic controls on the cell proliferation and to compare theses controls between them.

References

[1] Y. El Alaoui, L. Alaoui, "Mathematical analysis of an age structured multi phases cell cycle model with checkpoints." in 2019 IEEE International Conference of Computer Science and Renewable Energies (ICCSRE-IEEE), 2019, doi:10.1109/iccsre.2019.8807639 .

[2] H. Lodish, A. Berk, C.A. Kaiser, M. Krieger, M.P. Scott, A. Bretscher, H. Ploegh, P.T. Matsudaira, Molecular cell biology, W.H. Freeman, New York, 2007.

[3] S. M. Reppert, D. R. Weaver, Coordination of circadian timing in mammals, Nature, **418**(6901):935-41,2002, doi: 10.1038/nature00965.

[4] G. Filipski, P.F. Innominato, M.W. Wu, X. M. Li, S. Iacobelli, L.J. Xian, F. Levi, Effect of light and food schedules on liver and tumour molecular clocks in mice, Journal of the National Cancer Institute, **97**, 507–517, 2005, doi:10.1093/jnci/dji083.

[5] D. Morgan, The cell cycle: Principles of control, Primers in Biology series, Oxford University Press, 2006.

[6] E. Filipski, V.M. King, X.M. Li, T.G. Granda, M. Mormont, X.H. Lin, B. Claustrat, M.H. Hastings, F. Lévi, Host circadian clock as a control point in tumor progression, Journal of the National Cancer Institute **94**(9): 690-7, 2002, doi: 10.1093/jnci/94.9.690.

[7] E. Filipski, F. Delaunay, V.M. King, M. -W. Wu, B. Claustrat, A. Gréchez-Cassiau, C. Guettier, M.H. Hastings, F. Lévi, Effects of chronic jet lag on tumor progression in mice, Cancer Research **64**, 7879-7885, 2004, doi:10.1158/0008-5472.CAN-04-0674.

[8] E. Filipski, P.F. Innominato, M.W. Wu, X.M. Li, S. Iacobelli, L.J. Xian, F. Lévi, Effects of light and food schedules on liver and tumor molecular clocks in mice, Journal of the National Cancer Institute **97** (7), 507–517, 2005, doi:10.1093/jnci/dji083.

[9] J. Clairambault, B. Laroche, S. Michler and B. Perthame, A mathematical model of the cell cycle and its control, Research Report, **4892**, INRIA. 2003, URL :https://hal.inria.fr/inria-00071690.

[10] J. Clairambault, P. Michel, B. Perthame, Circadian rhythm and tumour growth, Comptes Rendus de l'Académie des Sciences, Série I, Mathématique, Elsevier, **342**, 17–22, 2006, URL: https://hal.archives-ouvertes.fr/hal-00113511.

[11] J. Clairambault, P. Michel, B. Perthame, A mathematical model of the cell cycle and its circadian control, in: I.A. Deutsch, L. Brusch, H. Byrne, G. de Vries, H.P. Herzel, Mathematical Modeling of Biological Systems, **I**, Birkhäuser Boston, 247–259, 2007, doi:10.1007/978-0-8176-4558-8_21

[12] J. Clairambault, S. Gaubert, B. Perthame, An inequality for the Perron and Floquet eigenvalues of monotone differential systems and age structured equations, Comptes Rendus Mathematique, **345**(10):549–554, 2007, doi: 10.1016/j.crma.2007.10.001.

[13] F. Bekkal Brikci, J. Clairambault, B. Perthame, Analysis of a molecular structured population model with possible polynomial growth for the cell division cycle, Mathematical and Computer Modelling, **47**(7–8):699–713, 2008, doi:10.1016/j.mcm.2007.06.008.

[14] F. Bekkal Brikci, J. Clairambault, B. Ribba, B. Perthame, An age-and-cyclin-structured cell population model for healthy and tumoral tissues,Journal of Mathematical Biology, **57**(1):91–110, 2008, doi:10.1007/s00285-007-0147-x.

[15] M. Doumic, Analysis of a population model structured by the cells molecular content, Mathematical Modelling of Natural Phenomena,**2**(3):121–152,2007, doi:10.1051/mmnp:2007006.

[16] J. Clairambault, S. Gaubert and T. Lepoutre, Comparison of Perron and Floquet eigenvalues in age-structured cell division models, Mathematical Modelling of Natural Phenomena, **4**:183-209,2009, doi: 10.1051/mmnp/20094308.

[17] J. Clairambault, S. Gaubert and T. Lepoutre, Circadian rhythm and cell population growth, Mathematical and Computer Modelling, **53**(7-8):1558-1567, 2011, doi:10.1016/j.mcm.2010.05.034.

[18] O. Diekmann, H.J.A.M. Heijmans, and H.R. Thieme, On the stability of the cell size, Journal of Mathematical Biology **19**: 227-248, 1984, doi: 10.1007/BF00277748.

[19] O. Diekmann, S.A. Van Gils, S.M. Verduyn Lunel, AND H.-O. Walther, Delay Equations, Functional, Complex and Nonlinear Analysis, Springer-Verlag, New York, 1995.

[20] O. Diekmann , P. Getto and M. Gyllenberg, Stability and bifurcation analysis of volterra functional equations in the light of suns and stars, SIAM Journal on Mathematical Analysis, **39**(4): 1023–1069, 2007, doi: 10.1137/060659211.

[21] O. Diekmann, M. Gyllenberg, Equations with infinite delay: blending the abstract and the concrete, Journal of Differential Equations, **252**(2):819–851, 2012, doi: 10.1016/j.jde.2011.09.038.

[22] A. M. de Roos, O. Diekmann, P. Getto, M.A. Kirkilionis, Numerical equilibrium analysis for structured consumer resource models, Bulletin of Mathematical Biology **72**(2):259–297, 2010, doi: 10.1007/s11538-009-9445-3.

[23] O. Diekmann, M. Gyllenberg, J.A.J. Metz, S. Nakaoka and A.M. de Roos, Daphnia revisited: local stability and bifurcation theory for physiologically structured population models explained by way of an example, Journal of Mathematical Biology, **61**(2):277–318, 2010, doi: 10.1007/s00285-009-0299-y.

[24] Ph. Clement, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, Perturbation theory for dual semigroups. I. The sun-reflexive case, Mathematische Annalen, **277**:709–725, 1987, doi:10.1007/BF01457866.

[25] Ph. Clement, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, Perturbation theory for dual semigroups. II. Time-dependent perturbations in the sun-reflexive case, Proceedings of the Royal Society of Edinburgh Section A: Mathematics, **109**: 145–172, 1988, doi: 10.1017/S0308210500026731

[26] Ph. Clement, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, Perturbation theory for dual semigroups. III. Nonlinear Lipschitz continuous perturbations in the sun-reflexive case, In Proceedings of Volterra Integrodifferential Equations in Banach Spaces and Applications 1987, 67–89, 1989.

[27] Ph. Clement, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, Perturbation theory for dual semigroups. IV. The intertwining formula and the canonical pairing, in Trends in Semigroup Theory and Applications Trieste-Italy, **116**:95–116, 1989.

[28] Y. El Alaoui and L. Alaoui, Asymptotic Behavior in a cell proliferation Model with Unequal division and Random Transition using translation semigroup,Indian Journal of Sciences and Technology, **10**(28):1-8, 2017, doi: 10.17485/ijst/2017/v10i28/101042.

[29] L. Alaoui and Y. El Alaoui, AEG property of a cell cycle model with quiescence in the light of translation semigroups, International Journal of Mathematical Analysis, **9**(51): 2513-2528, 2015, doi:10.12988/ijma.2015.59221.

[30] L. Alaoui, Nonlinear homogeneous retarded differential equations and population dynamics via translation semigroups, Semigroup Forum, **63**:330–356, 2001, doi: 10.1007/s002330010083.

[31] L. Alaoui, Age-dependent population dynamics and translation semigroups, Semigroup Forum, **57**: 186-207, 1998, doi: 10.1007/PL00005973.

[32] L. Alaoui, A cell cycle model and translation semigroups, Semigroup Forum, **54**, 1997, doi: 10.1007/BF02676597.

[33] L. Alaoui, Generators of translation semigroups and asymptotic behavior of the Sharpe-Lotka model, Differential Integral Equations, **9**:343-362, 1996.

[34] L. Alaoui, Population dynamics and translation semigroups, dissertation, University of Tübingen, 1995.

[35] L. Alaoui and O. Arino, Compactness and spectral properties for positive translation semigroups associated with models of population dynamics, Differential Integral Equation, **6**:459-480, 1993.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 1, 01-09 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Standardized UCI-EGO dataset for evaluating 3D hand pose estimation on the point cloud

Sinh-Huy Nguyen[1], Van-Hung Le[*,2]

[1]*Institute of Information Technology, MIST, 100000, Vietnam*

[2]*Tan Trao University, Tuyen Quang, 22000, Vietnam*

ARTICLE INFO

Article history:
Received: 14 September, 2020
Accepted: 23 December, 2020
Online: 10 January, 2021

Keywords:
UCI-EGO dataset
3D CNNs
3D hand pose estimation
Point cloud data

ABSTRACT

To evaluate and compare methods in computer vision, scientists must use a benchmark dataset and unified sets of measurements. The UCI-EGO dataset is a standard benchmark dataset for evaluating Hand Pose Estimation (HPE) on depth images. To build robotic arms that perform complex operations such as human hands, the poses of the human hand need to be accurately estimated and restored in 3D space. In this paper, we standardized the UCI-EGO dataset to evaluate 3D HPE from point cloud data of the complex scenes. We also propose a method for fine-tuning a set parameter to train the estimation model and evaluating 3D HPE from point cloud data based on 3D Convolutional Neural Networks (CNNs). The CNNs that we use to evaluated currently the most accurate in 3D HPE. The results of the 3D HPE from the point cloud data were evaluated in two branches: using the hand data segment and not using the hand data segment. The results show that the average of 3D joint errors of the 3D HPE is large on the UCI-EGO dataset (87.52mm) and that the error without using the hand data segment is many times higher than the estimated results when using the hand data segment (0.35ms). Besides, we also present the challenges of estimating 3D hand pose and the origin of the challenge when estimating real image dataset.

1 Introduction

In computer vision when evaluating and comparing the methods, the scientists must use a benchmark dataset and unified sets of measurements. The benchmark dataset usually includes training sets and testing sets/ validation set [1], this ratio is defined in the cross-validation parameter [2]. And these sets just include the annotation data of each sample. The UCI-EGO dataset has been published in [3] [1] and evaluated in many studies of HPE [4]–[6]. However, these ratings are evaluated in 2D space on the depth image. The UCI-EGO dataset provided the annotation data, each key point is represented by the structure (x, y, z), (x, y) are the coordinates on the depth image, z is the depth value of pixel which has the coordinates (x, y). However, the actual depth values of the annotation data are different from the depth data on the depth image. They are shown as Figure 2.

Nowadays, building robotic arms with hands that can perform many complex actions like human hands is an issue that needs research [7]. Since the human hands have many degrees of freedom (DOF), the complex actions can be performed. In order to build a robotic hand that can perform complex operations (Figure 3), first of all, it is necessary to restore and estimate the hand poses in the 3D space. Therefore, we continue to perform research on estimating human hand pose in the 3D space. In particular, estimating the hand pose on the data obtained from the EGOcentric VIsion (EGO-VI) sensor, contains many challenges such as missing, data loss, or obscuring.

Therefore, this paper includes the main contributions as follows:

- Standardizing the annotation data of the UCI-EGO dataset based on the depth value on the depth image. That means replacing the depth value of each point in the annotation data with the depth value of the corresponding point at that coordinate on the depth image, as illustrated in Figure 4. From this, using 3D hand pose annotation data to train the hand pose estimation model in the 3D space.
- Fine-tuning a set of parameters to train the hand pose estimation model in the 3D space based on the (V2V - V2V-PoseNet [8]) and evaluating the 3D HPE based on the most accurate 3D CNN (V2V). The estimation results are presented and

[*]Corresponding Author: Van-Hung Le, Tan Trao University, & Lehung231187@gmail.com
[1]`https://github.com/hassony2/inria-research-wiki/wiki/uci-ego-dataset`.

evaluated on the point cloud data. This is the same data as the real world.

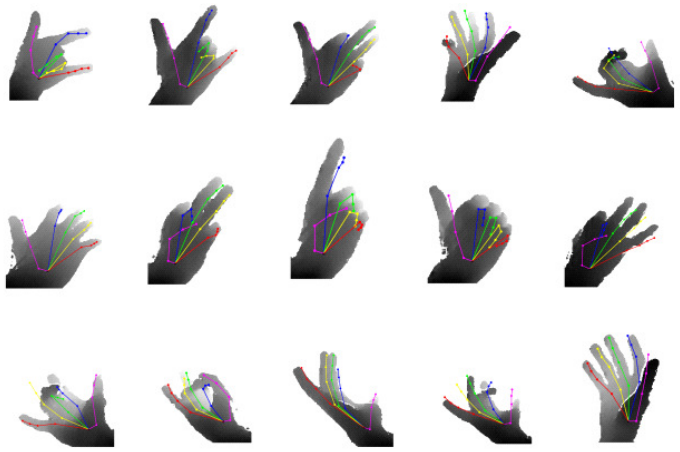


Figure 1: Illustration of 2D HPE results on the UCI-EGO dataset [6].

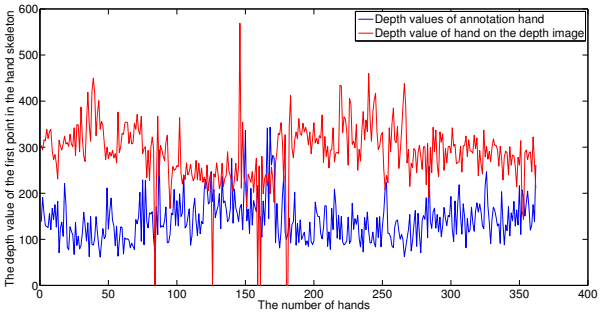


Figure 2: The depth value of the first point in the hand skeleton of annotation data and depth image data.

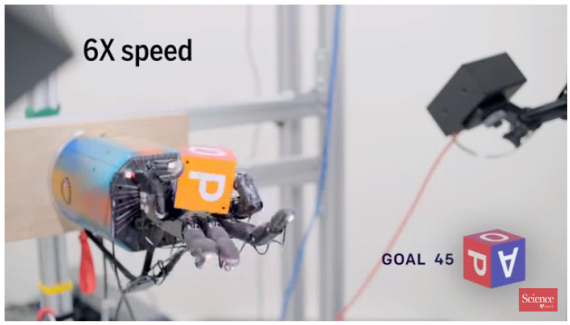


Figure 3: Robot arm illustration follows the operation of human hands [7].

- Presenting and comparing some 3D HPE results on the full hand dataset and the dataset obtained from the EGO-VI sensor. Presenting some challenges of estimating hand pose in the 3D space when estimating on the data obtained from EGO-VI sensors.

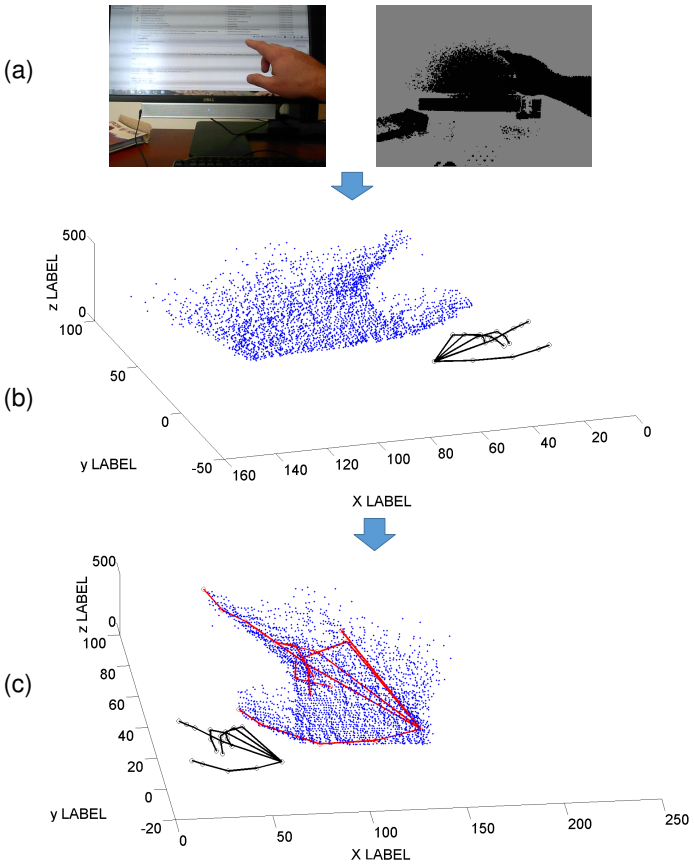


Figure 4: Illustrating of standardized the UCI-EGO annotation data process. (a) is the RGB-D images; (b) is the annotation data of UCI-EGO (the blue points are the point cloud of hand, the black skeleton is the annotation of UCI-EGO dataset); (c) is the standardized the annotation of UCI-EGO.

During the research on recognizing the daily activities of the human hand based on data collected from the EGO-VI sensor. We only research within a limited range as follows: We are interested in the dataset obtained from the EGO-VI sensor, namely the UCI-EGO dataset studied in this paper; We are also interested in 3D CNN that use point cloud as the input because the point cloud data is real data similar to the real environment.

The organization of the paper is shown as follows: Section 1 first introduce the benchmark dataset, the existence problem of the UCI-EGO [3] dataset, and the 3D HPE problem, we also introduce the application of 3D HPE to build robotic arms. Section 2 presents studies on the benchmark datasets to evaluate 3D HPE and some results. Section 3 presents the standardization of the UCI-EGO dataset and 3D HPE in the point cloud data. Section 4 presents the results and some discussions of 3D HPE. Finally, there are some conclusions and the next research direction of the paper (Section 5).

2 Related Works

Evaluating on the benchmark datasets is an important step to confirm the correctness of the detection, recognition, and estimation model of computer vision. Currently, there are many datasets for evaluating 3D HPE. The datasets are listed in Tab. 6 of [9]. In this paper, we only reintroduce some of the datasets used to evaluate 3D

HPE and some results based on typical CNNs.

In [10], published MSRA dataset, [2]. It includes 76k depth images of nine subjects of the right hands are captured using Intel's Creative Interactive Gesture Camera. Each subject include 17 gestures captured and include about 500 frames with 21 3D annotation hand joints for each frame: wrist, index mcp(metacarpal bone), index pip(proximal phalanges), index dip(distal phalanges), index tip, middle mcp(metacarpal bone), middle pip(proximal phalanges), middle dip(distal phalanges), middle tip, ring mcp (metacarpal bone), ring pip(proximal phalanges), ring dip(distal phalanges), ring tip, little mcp(metacarpal bone), little pip(proximal phalanges), a little dip(distal phalanges), little tip, thumb mcp(metacarpal bone), thumb pip(proximal phalanges), thumb dip(distal phalanges), and thumb tip. The size of the captured image is 320×240 pixels. The camera's intrinsic parameters are also provided, i.e. principal point of the image is (160, 120) and the focal length is 241.42. This dataset only has depth images, especially the hand data that is segmented with environmental data. This is a benchmark dataset for the evaluation of 3D HPE, the results of the CNNs are shown in table 2 of [11].

In [12] [3], the author includes 72,757 frames of the training set captured from a single person and 8,252 frames of the testing set captured two different persons from three MS Kinect v1, i.e. a frontal view and two side views. Each frame is a couple of RGB-D images. This dataset provided 25-joints in the annotation data with 42 DOF. The authors used the Randomized Decision Forest (RDF) to train a binary classification model by this dataset. And then this classification segments each pixel that belongs to a hand or background in the depth image. 3D HPE results of the CNNs are shown in table 2 of [11].

In [13] [4], the author includes 22K frames for training and 1.6K frames for testing, they captured from the Intel's Creative Interactive Gesture Camera with 10 subjects to take 26 different poses. It also provides 3D annotation data with 16 hand joints: palm, thumb root, thumb mid, thumb tip, index root, index mid, index tip, middle root, middle mid, middle tip, ring root, ring mid, ring tip, pinky root, pinky mid, and pinky tip.

The above are the datasets collected from a fixed number of perspectives of the image sensors. In many real applications, the image sensors are mounted on the body to collect data from the environment. These datasets are named the "Egocentric" dataset. In [14], the author published the **UCI-EGO** dataset, in [15] the author published the **Graz16** dataset, in [16] the author published the **Dexter+Object** dataset, in [17] the author published the **First-Person Hand Action** (FHAD) dataset, in [18] the author published the **UCI-EGO-Syn** dataset, in [15] the author published the **CVAR** dataset. Most EGO-VI datasets contain challenges for 3D HPE as follows: The frames do not contain hands, and the hands are obscured by objects in the scene; Data of the fingers is obscured; Data of visible hand only data of palm. The percent of challenges in CVRA [15] and UCI-EGO-Syn [18] datasets are shown in Figure 5.

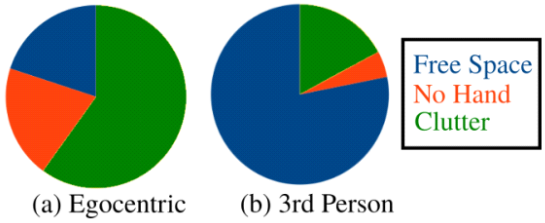


Figure 5: The percent of challenges in CVRA [15] and UCI-EGO-Syn [18] datasets.

In the past 5 years, many CNNs and related studies have been published for HPE. However, most of these studies were evaluated on the MSRA, NYU, ICVL datasets, the results are shown in table 2 of [11]. In this paper, we are interested in estimating 3D hand pose from the 3D annotation of hand skeleton data. The method we use is V2V, it has the input data of 3D annotation data and the point cloud data of hand. In [19], the author proposed 3D CNN for 3D HPE. This network projects 3D points of the hand following: x-direction, y-direction, z-direction. Synthesized in these three directions is encoded as 3D volumes storing the projective Directional Truncated Signed Distance Function (D-TSDF). Special, it only uses three 3D convolutional layers and three fully-connected layers to train the model. The estimated results have an average error of 9.58mm on the MSRA dataset.

In [20], the author proposed a deep regression network (SHPR-Net) for 3D HPE. This network consists of two components: A semantic segmentation network (SegNet) and the hand pose regression network (RegNet). The first component is used to segment the joints, the parts of the hand. That is, each part of the hand is segmented and labeled, RegNet is used to predict the 3D coordinates of the match corresponding to the segmented hand data areas. The estimated results have an average error of 10.78mm on the NYU dataset. In [21], the author proposed Hand PointNet to estimate 3D hand pose from the segmented hand on a depth image by using random decision forest [13], then convert to point cloud data using the Eq. 1. This method has improved the basic PointNet by using a hierarchical PointNet to generate the hierarchical feature extraction. Specifically, it uses three point set abstraction levels. Besides, In [22] and [23], the authors proposed the 3D DenseNet, Point-to-Point Net, respectively. The estimated error of result on the above methods is usually less than 10mm.

3 Standardized UCI-EGO Dataset and 3D HPE by V2V

3.1 Standardized UCI-EGO Dataset

As shown in Figure 4(b), the annotation data of the UCI-EGO dataset needs to be calibrated to meet the evaluation of 3D HPE in 3D space / on the point cloud data. Deviation in 3D annotation data of the hand pose is due to the 3D annotation data generated from the semi-automatic labeling tool. This issue is covered in Sec. 4.1 by UCI-EGO dataset introduction. This process is done as follows: The

[2] https://www.dropbox.com/s/c91xvevra867m6t/cvpr15_MSRAHandGestureDB.zip?dl=0.
[3] https://jonathantompson.github.io/NYU_Hand_Pose_Dataset.htm.
[4] https://labicvl.github.io/hand.html

coordinates of each keypoint $K(x, y, z)$ in the UCI-EGO annotation data, where (x, y) is the coordinate of K on the depth image, z is the depth value of K in the depth image [3]. In this paper, we replace the depth value of the K point in UCI-EGO [3] with the depth value of the point with coordinates (x, y) on the depth image. However, there are many cases of data loss or missing in the depth images, especially on the depth image sensors collected on previous depth sensors such as the Microsoft Kinect Version 1 [24]. We solve this problem by using the mean depth value of the k-neighbors of the K on the depth image, where $k = 3$, $d_K = mean(d_{K1}, d_{K2}, d_{K3})$.

After that, the coordinates (x, y) of the keypoint on the image combined with the depth value to generate one point (x_a, y_a, z_a) in 3D space / point cloud data [25] according to Eq. 1.

$$\begin{aligned} x_a &= \frac{(x - cx_d) * z}{fx_d} \\ y_a &= \frac{(y - cy_d) * z}{fy_d} \\ z_a &= z \end{aligned} \tag{1}$$

where fx_d, fy_d, cx_d, and cy_d the intrinsics of the depth camera.

The results of the standardized annotation data are illustrated on the point cloud data of the hand as shown in Figure 4(c).

3.2 3D HPE by V2V

Based on the results of the 3D HPE of the 3D CNN on ICVL, NYU, MSRA datasets. The V2V [8] network has the best estimation result (average of 3D joints error is 6.28mm, 8.42mm, 7.49mm, respectively). Therefore, we used the V2V network to evaluate HPE on the UCI-EGO dataset. The execution process is shown below.

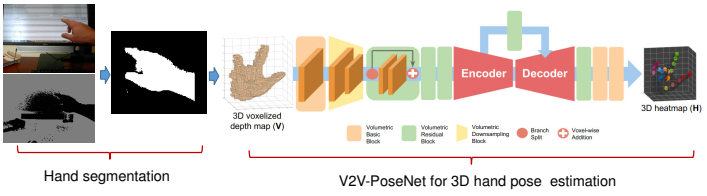


Figure 6: The hand data segmentation and 3D HPE based on 3D CNN (V2V) [8].

As shown in Figure 6 (left), the hand is in the complex scene, where the input data of V2V is the point cloud data of the segmented hand. Therefore, we propose a preprocessing step to segment the hand data with the environment and other objects.

We based on the annotation data on the keypoints of the hand skeleton frame on the depth image to crop a region container on the depth image with a rectangle that is bounding box of the keypoints on the depth image. We then find the maximum depth value of the keypoints M_d^h on the depth image. We rely on the context of capturing data from the EGO-VI sensor, the hand that is usually closest to the sensor. Therefore, the data near the hand that is large M_d^h is not part of the hand data.

As shown in Figure 6 (right), the input of the V2V method is 3D voxelized data. Thus, it reprojects each pixel of the depth map to the 3D space. After that, this space is discretized based on the pre-defined voxel size. V2V-PoseNet is based on the hourglass model [26] and is designed to be divided into four volumetric blocks. The first volumetric basic block includes a volumetric convolution, volumetric batch normalization, and the activation function. The location of the first volumetric basic block is in the first and last parts of the network. The second volumetric residual block extended from the 2D residual block in [27]. The third volumetric downsampling block is similar to the volumetric max-pooling layer. The last block is the volumetric upsampling block, which consists of a volumetric deconvolution layer, volumetric batch normalization layer, and the activation function.

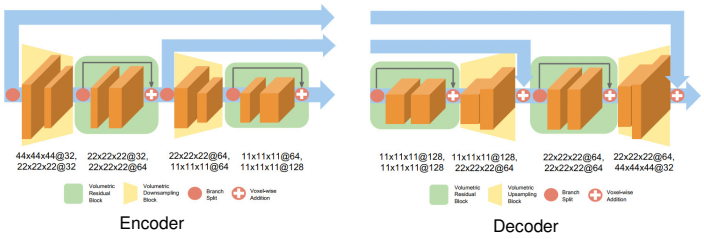


Figure 7: Encoder and decoder phase in the network architecture of V2V-PoseNet [8].

Each phase of the V2V-PoseNet method in Figure 6 is shown in Figure 7. Each phase consists of four blocks as shown in Figure 7. Therein, the volumetric downsampling block reduces the feature map space while the volumetric residual bock increases the number of channels in the encoder phase. Otherwise, the volumetric upsampling block enlarges the feature map space. When upsampling, to compress the extracted features the network reduce the number of channels. To predict each keypoint of the hand in 3D space through two stages: encoder, decoder. They are connected by the voxel-wise. To supervise the per-voxel likelihood in the estimating process, V2V generates a 3D heatmap, wherein the mean of the Gaussian peak is positioned at the ground-truth joint location in Eq. 2.

$$D_n^*(i, j, k) = exp\left(- \frac{(i - i_n)^2 + (j - j_n)^2 + (k - k_n)^2}{2\sigma^2} \right) \tag{2}$$

where n^{th} keypoint whose ground-truth 3D heatmap is denoted D_n^*, (i_n, j_n, k_n) is the ground-truth voxel coordinate of nth, and $\sigma = 1.7$ is the standard deviation of the Gaussian peak [8]. V2V also uses the mean square error as a loss function L in Eq. 3.

$$L = \sum_{n=1}^{N} \sum_{i,j,k} \| D_n^*(i, j, k) - D_n(i, j, k) \|^2 \tag{3}$$

where D_n^* and D_n are the ground-truth and estimated results for n^{th} keypoint, respectively, and the number of keypoints is denoted N.

4 Experimental Results

4.1 Dataset

In this paper, we are trained and tested on the UCI-EGO [14] dataset. It provides about 400 frames prepared the 3D annotation. 3D annotations of keypoints with 26 joint points are also provided. To

annotate this dataset for evaluating 3D HPE and hand tracking the authors developed a semi-automatic labeling tool. It can annotate the accurate partially occluded hands and fingers in the 3D space by using the techniques: A few 2D joints are first manually labeled in the image and have used to select the closest synthetic samples in the training set; After that, a full hand pose is generated combining the manual labeling and the selected 3D sample; This pose is manually refined, resulting to the selection of a new sample, and the creation of a new pose; This process is repeated until acceptable labeling is achieved. This dataset captured from a chest-mounted Intel Senz3D RGB-D camera/EGO-VI and capture 4 sequences, 2 for each subject (1 male and 1 female), as illustrated in Figure 8. The authors labeled the keypoints of any visible hand in both RGB and Depth images every 10 frames, as illustrated in Figure 9.

We perform experiments on PC with Core i5 processor - RAM 8G, 4GB GPU. Pre-processing steps were performed on Matlab, fine-tuning, and development process in Python language on Ubuntu 18.04.

Before performing 3D HPE from the point cloud data, we propose a pre-processing step to segment the hand from the complex scene data, as shown in Figure 14. The depth image contains the depth value of the hand data is the closest (the hand is closest to the sensor). From there we use a threshold d_{thres} which is the maximum depth value of the hand to segment the data of the hand and other data in the complex scene.

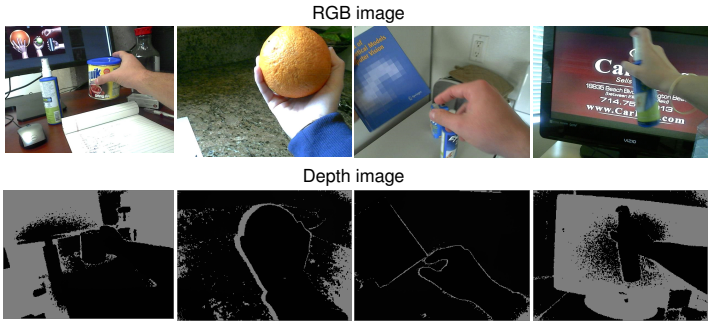


Figure 8: Illustration of hand grasping object in the UCI-EGO dataset [14].

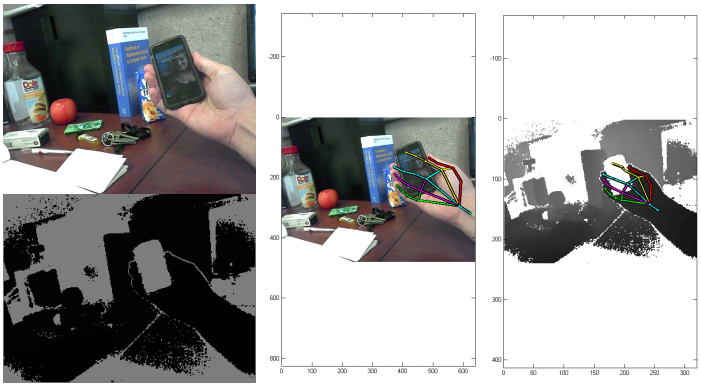


Figure 9: Illustration of 2D hand pose ground truth on the RGB and depth images.

[5] https://github.com/dragonbook/V2V-PoseNet-pytorch.

4.2 V2V Parameters

Like in the original study of V2V-PoseNet [8], this deep network is developed in the PyTorch framework. The zero-mean Gaussian distribution with $\sigma = 0.001$ is initialized to all weights. The learning rate is set 0.00025 and batch size is set 4. This is the maximum value that V2V can train the model on our computer. The size of input is $88 \times 88 \times 88$. This deep network also uses the optimizer method of Adam [28]. To standardize data for training, V2V rotates [-40, 40] degrees in XY space, scale [0.8, 1.2] in 3D space, and translate with the size of voxels [-8, 8] in 3D space. We trained the model for 15 epochs. Implementation details of V2V are shown in the link [5].

In the original V2V study [8], the model trained only for 10 epochs. Although, the depth data value on the MSRA dataset is from 0.3m-0.7m, while the depth data value on the UCI-EGO dataset is from 0.3-0.45m. In each epoch, we computed the total loss function of batch sizes by Eq. 4.

$$T_L = \sum_{j=1}^{N_p} L_j \tag{4}$$

where $N_p = N_s/batch\_size$ is the batch size number of the training data, N_s is the sample number of the training data. L_j is the total loss function j^{th} of each batch size.

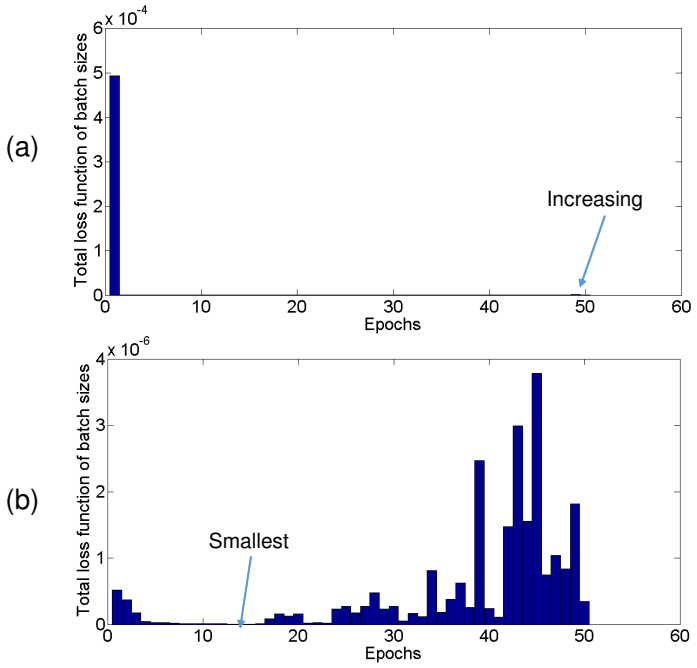


Figure 10: The total loss function of batch sizes at each epoch. (a) is the total loss function of batch sizes of training data when hand data is segmented on the depth image. (b) is the total loss function of batch sizes of validation data when hand data is segmented on the depth image.

We then compare the total loss function at each epoch. In a model with the smallest total loss function, that model is the best model for estimating the 3D hand pose. During training, we found that the total value of loss function up to the epoch 15th does not

decrease any more. We have trained the estimation model for 50 epochs, the total loss function of batch sizes at each epoch are shown in Figure 10. It can be seen that the value of the loss function decreases drastically and is about 10^{-11} at the epoch 10th. Specifically, on the validation data, there is the value of the smallest loss function at the epoch 10th as Figure 10(b). Therefore in this paper, we train only 15 epochs.

4.3 Evaluation Measure

As the evaluations of the previous 3D HPE method, we used the average 3D distance error (as shown in Eq. 5) to evaluate the results of the 3D HPE on the dataset.

$$\widehat{Err_a} = \frac{1}{Num_s} \sum_{n=1}^{Num_s} \frac{1}{21} \sum_{k=1}^{21} DIS(p_g, p_e) \quad (5)$$

where $DIS(p_g, p_e)$ is the distance between a ground truth joint p_g and an estimated joint p_e; Num_s is the number of testing frames. In this paper, we evaluated the 21 joints of hand pose, illustrated in Figure 11.

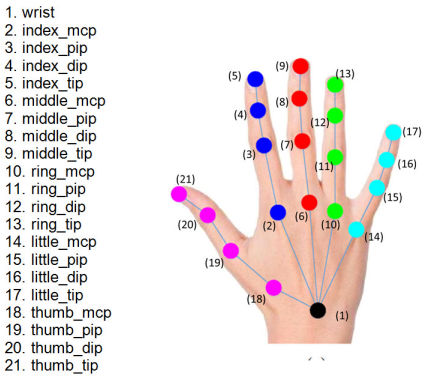


Figure 11: Illustrating the hand joints of the UCI-EGO [14] dataset.

In this paper, we use the rate at 1:5, which means 80% for training and 20% for testing. This ratio is based on the division of [15](using 5-fold cross-validation for testing and training). This means the UCI-EGO dataset uses 283 samples for training and 71 samples for testing. Although, the UCI-EGO dataset provides 26 annotation points. However, we only use 21 annotation data points to evaluate the 3D HPE. The order of points is shown in Figure 11.

4.4 Results and Discussions

As the evaluation of 3D HPE results shown in Tab. 2 of [11], also use the 3D distance error (mm) to evaluate the estimation results on the UCI-EGO dataset. The average 3D distance error is shown in Table 1. The processing time of training process and 3D HPE process is shown in Figure 12. As figure 12, the processing time to train for 50 epochs is $1.472h$ and $0.442h$, it is calculated by $(1m46s = 106s) * 50epochs = 1.472h$ and $(1m46s = 106s) * 15epochs = 0.442h$, respectively. The processing time for testing is shown in Table 2. It is calculated by $128ms/362samples$ and $137ms/362samples$, respectively.

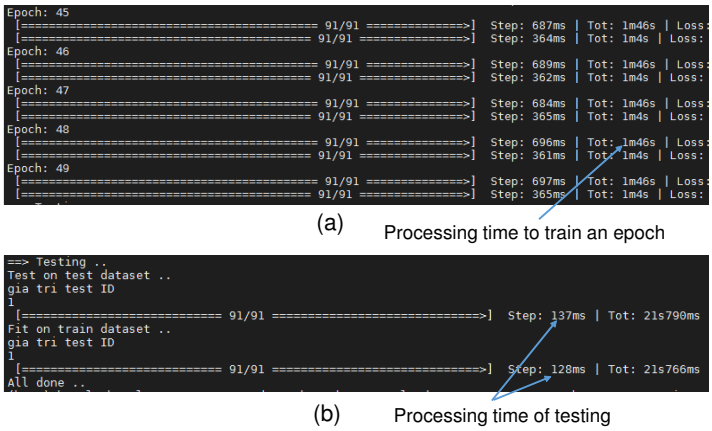


Figure 12: Illustrating the processing time to train the estimation model and testing 3D HPE. (a) is the processing time to train an epoch. (b) is the processing time for testing.

Table 1: The average 3D distance error of the V2V on the UCI-EGO dataset for 3D HPE when trained through 15 epochs and 50 epochs.

Training	Measuremet/ Method		V2V
15 epochs	**Average of 3D joints error Err_a(mm)**	**Hand segmentation**	**87.52**
		No hand segmentation	95.49
50 epochs	**Average of 3D joints error Err_a(mm)**	**Hand segmentation**	**87.07**
		No hand segmentation	88.73

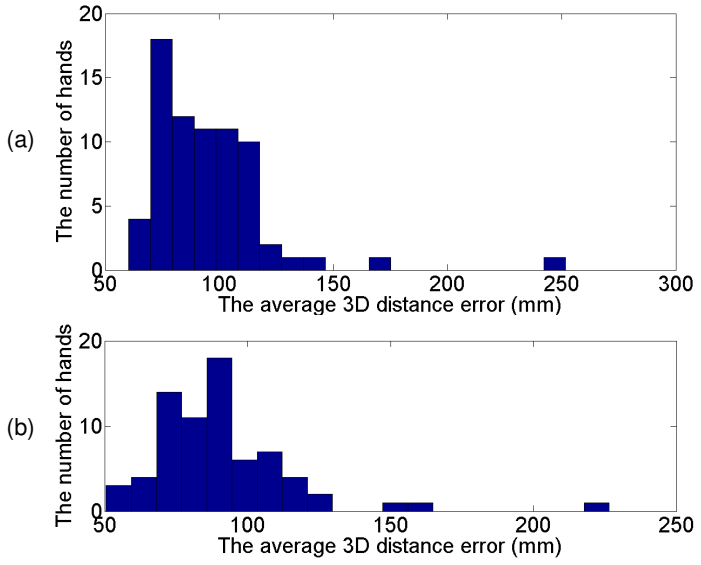


Figure 13: The distribution of 3D joints error for 3D HPE based on the UCI-EGO dataset by V2V-PoseNet [8] when trained through 15 epochs. (a) The distribution of 3D joints error when using hand segmentation on the depth image; (b) The distribution of 3D joints error when do not use hand segmentation on the depth image.

Table 2: The average of processing time of the V2V on the UCI-EGO dataset to estimate a 3D hand pose.

Measurement/ Method		V2V
Processing time (ms)/ hand	Hand segmentation	0.35
	No hand segmentation	0.38

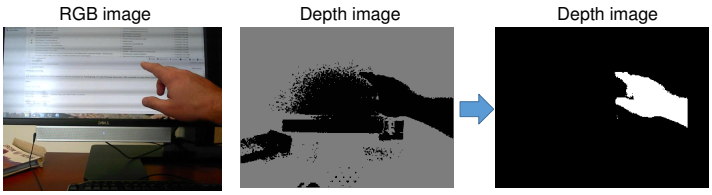


Figure 14: Illustrating the data well segmented of hand in the complex scene.

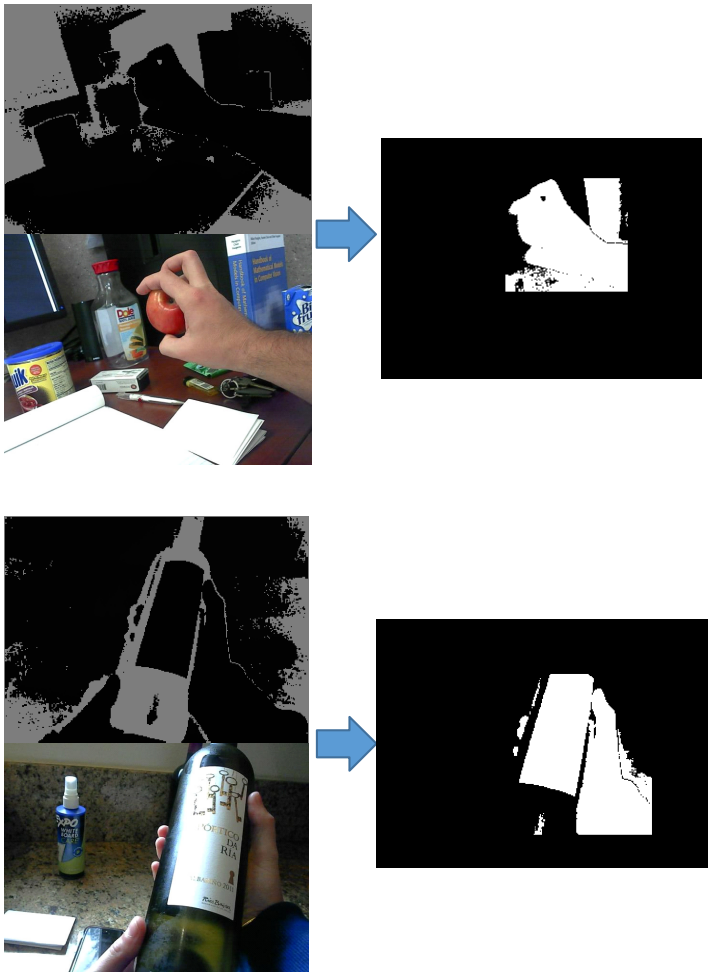


Figure 15: Illustrating of the hand data and the objects in the UCI-EGO dataset.

From table 2 of research in [11] and Table 1, the error results of estimating the 3D hand pose on the EGO-VI dataset are much larger than those estimated on the obtained datasets from a fixed number of perspectives. The time of estimated hand joints when using V2V is enormous, as shown in Table 2. This high estimate time due to carrying CNN using the input data is the point cloud data that is not reduced by number of points.

As Table 1 and Figure 13, the 3D HPE results when using the hand data segment are better when not using the hand data segment in the complex scenes. The error distribution when using the hand data segment concentrated in bins closer to 0.

Although when segmenting the hand data, the estimation results were better than when the hand was not segmented. However, the estimated results have not improved much. Since in the UCI-EGO dataset only about 9% of the hands are well segmented with other subject's data, as illustrated in Figure 14. The remaining about 90% of the hands are grasping objects like phone, book, spray bottle, bottle, ball, etc. Therefore, the hand data gets stuck with the data of the objects being handled, as illustrated in Figure 15.

Figure 16 shows the results of estimating 3D hand pose on the point cloud data in two methods: The segmented hand data (a) and no segmented hand data (b).

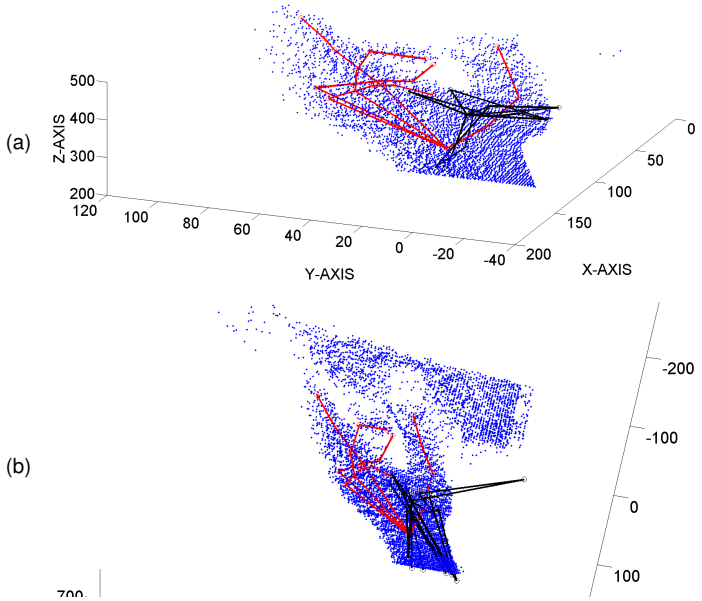


Figure 16: The results of the estimated 3D hand pose on the point cloud. (a) is the result of estimating the 3D hand pose in the 3D space on the segmented hand data; (b) is the result of estimating the 3D hand pose in the 3D space on hand data is not segmented. The blue points are the point cloud of hand and others object. The red skeleton is the ground truth of 3D hand pose, the black skeleton is the estimated 3D hand pose.

Based on the reading paper of [29], we find that the error of estimating the 3D hand pose on the UCI-EGO-Syn [18] dataset is high with: Hough [30], RDF [31], Deep Prior [32], PXC [33], Cascader [14], EGO.WS. [3]. The error distribution is from 35 to 100mm, as illustrated in Figure 18.

The UCI-EGO dataset has the hand data that performs grasping objects and is collected from an EGO-VI mounted on the person, the hand data is obscured, as shown in Figure 15. In particular, the data of the fingers is obscured. In this paper, we used V2V for estimating 21 joints of hand (3D hand pose), the input data of V2V is the coordinate of 21 joints in the 3D space of ground truth data, the output is also 21 joints, as illustrated in Figure 6. Therefore, the fields of the obscured fingers, the hand joints can still be estimated.

However, the point cloud data of the hand is much missing, the estimation results have large errors, as illustrated in Figure 17.

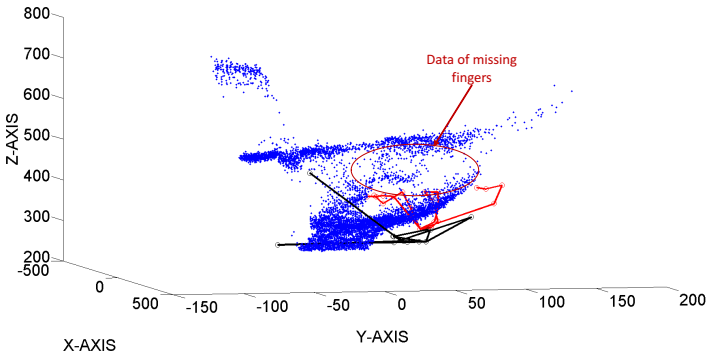


Figure 17: Illustrating of estimation results on the data with obscured fingers of Figure 9. The red skeleton is the ground truth of the 3D hand pose, the black skeleton is the estimated 3D hand pose.

During the research on 3D HPE, we found the challenges as follows:

- ***The high degree of freedom***: In realistic/3D space human hand models have between 15 and 24 degrees of freedom [34]. From 21 joints there are about 63 coordinates of in the 3D space. To train such a large dimension vector requires a very strong model and very large learning data to train model.
- ***Data obscured***: As shown above, the hand data is obscured making the hand's point cloud data missing. This makes the 3D HPE result from a high error value. This problem can be seen when comparing the results estimated on datasets with complete hand data (MSRA, ICVL, NYU) (table 2 of [11]) with the estimation results on the UCI-EGO-Syn [18] dataset (Figure 18). The quality of the depth images is also an issue affecting the 3D data / point cloud data. The depth images can be collected from a stereo camera or ToF (Time of Flight), this data still contains error noise.
- ***Hand size in the space***: When moving in the 3D space to perform operations, the size of the hand is constantly changing. To train hand estimation models with different sizes, a large number of samples and strong models are required.
- ***3D hand pose annotation***: The quality of the 3D HPE model depends on the training data. To prepare the training data requires an expensive system like in the FPHA (First-Person Hand Action) dataset [17], or use the estimated data through an estimation model like in the Ho-3D dataset [35].

5 Conclusion

To perform complex operations such as human hands then robot hand operations need to be built based on human hand operation. To do this, the poses of the human hand need to be accurately estimated and restored in 3D space. In this paper, we perform the standardization of the UCI-EGO dataset to evaluate 3D HPE. Simultaneously, we retrained, evaluated 3D HPE, and presented the results on the point cloud data based on the V2V-PoseNet. The estimation results on the UCI-EGO dataset have been a large error and many challenging. The estimation results above also show that estimating 3D hand pose on the EGO-VI dataset is challenging and needs to be studied to improve the accuracy of 3D HPE.

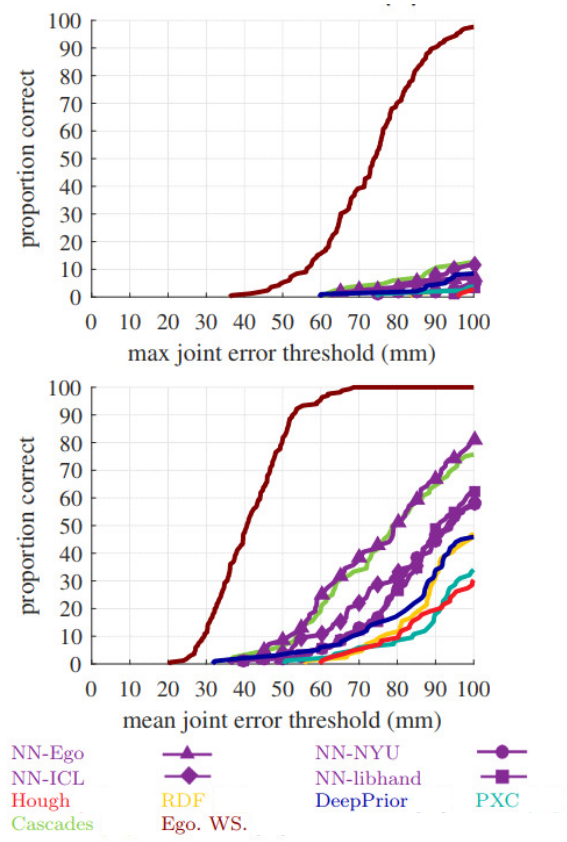


Figure 18: Distribution of 3D joints error estimation [29] on the UCI-EGO-Syn [18] dataset.

Acknowledgement

This research is funded by Tan Trao University in Tuyen Quang, Vietnam.

References

[1] S. Anasua, Y. Yang, V. Mauno, "Variation benchmark datasets: update, criteria, quality and applications," Database, Volume 2020, 2020, baz117, https://doi.org/10.1093/database/baz117, 2020.

[2] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," `https://machinelearningmastery.com/k-fold-cross-validation/`, [Accessed 1 September 2020].

[3] G. Rogez, J. S. Supanvcivc, D. Ramanan, "First-person pose recognition using egocentric workspaces," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 07-12-June, 4325–4333, 2015, doi:10.1109/CVPR.2015.7299061.

[4] R. Gregory, S. S. James, R. Deva, "Egocentric Pose Recognition in Four Lines of Code," https://arxiv.org/pdf/1412.0060.pdf, 2014.

[5] S. S. James, R. Grégory, Y. Yi, S. Jamie, R. Deva, "Depth-based hand pose estimation: data, methods, and challenges," International Journal of Computer Vision, **126**, 1180–1198, 2015.

[6] S. Yuan, Q. Ye, B. Stenger, S. Jain, T. K. Kim, "BigHand2.2M benchmark: Hand pose dataset and state of the art analysis," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, 2605–2613, 2017, doi:10.1109/CVPR.2017.279.

[7] M. Hutson, "Watch a robot hand learn to manipulate objects just like a human hand," https://www.sciencemag.org/news/2018/07/watch-robot-hand-learn-manipulate-objects-just-human-hand, [Accessed 1 September 2020].

[8] G. Moon, J. Y. Chang, K. M. Lee, "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 5079–5088, 20118.

[9] R. Li, Z. Liu, J. Tan, "A survey on 3D hand pose estimation: Cameras, methods, and datasets," Pattern Recognition, **93**, 251–272, 2019, doi:10.1016/j.patcog.2019.04.026.

[10] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition," in Proc. ICCV, 264–272, 2015, doi:10.1109/CVPR.2018.00035.

[11] C.-h. Yoo, S.-w. Kim, S.-w. Ji, Y.-g. Shin, S.-j. Ko, "Capturing Hand Articulations using Recurrent Neural Network for 3D Hand Pose Estimation," https://arxiv.org/abs/1911.07424, 2019.

[12] J. Tompson, M. Stein, Y. Lecun, K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," ACM Transactions on Graphics, **33**(5), 2014.

[13] D. Tang, H. J. Chang, A. Tejani, T. K. Kim, "Latent regression forest: Structured estimation of 3D hand poses," IEEE Transactions on Pattern Analysis and Machine Intelligence, **39**(7), 1374–1387, 2017, doi:10.1109/TPAMI.2016.2599170.

[14] G. Rogez, M. Khademi, J. S. Supanvcivc, J. M. Montiel, D. Ramanan, "3D hand pose detection in egocentric RGB-D images," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 8925, 356–371, 2015, doi:10.1007/978-3-319-16178-5_25.

[15] M. Oberweger, G. Riegler, P. Wohlhart, V. Lepetit, "Efficiently creating 3D training data for fine hand pose estimation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, 4957–4965, 2016, doi:10.1109/CVPR.2016.536.

[16] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, C. Theobalt, "Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input," in Proceedings of European Conference on Computer Vision (ECCV), 2016.

[17] G. Garcia-Hernando, S. Yuan, S. Baek, T.-K. Kim, "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations," .

[18] S. S. James, R. Gregory, Y. Yi, S. Jamie, R. Deva, "Depth-based hand pose estimation: methods, data, and challenges," International Journal of Computer Vision, Springer Verlag, 2018, **126**(11), 1180-1198, 2018. doi: 10.1007/s11263-018-1081-7, 2018.

[19] L. Ge, H. Liang, J. Yuan, D. Thalmann, "3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi:10.4324/9781315556611.

[20] X. Chen, G. Wang, S. Member, C. Zhang, K. I. M. Member, X. Ji, "SHPR-Net : Deep Semantic Hand Pose Regression From Point Clouds," IEEE Access, **PP**(c), 1, 2018, doi:10.1109/ACCESS.2018.2863540.

[21] L. Ge, Y. Cai, J. Weng, J. Yuan, "Hand PointNet : 3D Hand Pose Estimation using Point Sets," Cvpr, 3–5, 2018.

[22] L. Ge, H. Liang, J. Yuan, S. Member, D. Thalmann, "Real-time 3D Hand Pose Estimation with 3D Convolutional Neural Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, **8828**(c), 2018, doi:10.1109/TPAMI.2018.2827052.

[23] L. Ge, Z. Ren, J. Yuan, "Point-to-point regression pointnet for 3D hand pose estimation," in European Conference on Computer Vision, volume 11217 LNCS, 489–505, 2018, doi:10.1007/978-3-030-01261-8_29.

[24] D. Abdul, H. Ammar, "Recovering Missing Depth Information from Microsoft Kinect," pdfs.semanticscholar.org, 2011.

[25] N. Burrus, "Kinect Calibration," http://nicolas.burrus.name/index.php/Research/KinectCalibration, [Accessed 25 July 2020].

[26] N. A., Y. K., D. J., "Stacked hourglass networks for human pose estimation," in In European Conference on Computer Vision, 2016.

[27] H. Kaiming, Z. Xiangyu, R. Shaoqing, S. Jian, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[28] P. K. Diederik, B. Jimmy, "Adam: A Method for Stochastic Optimization," in In ICLR, 2015.

[29] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, D. Ramanan, "Depth-Based Hand Pose Estimation: Methods, Data, and Challenges," International Journal of Computer Vision, **126**(11), 1180–1198, 2018, doi:10.1007/s11263-018-1081-7.

[30] X. C., C. L., "Efficient Hand Pose Estimation from a Single Depth Image." in International Conference on Computer Vision (ICCV), 2013.

[31] K. C., Kırac, K. F., Y. E., A. L., "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in International Conference on Computer Vision (ICCV), 2012.

[32] M. Oberweger, P. Wohlhart, V. Lepetit, "Hands Deep in Deep Learning for Hand Pose Estimation," in Computer Vision Winter Workshop, 2015.

[33] Intel, "Perceptual computing SDK," 2013.

[34] S. Cobos, M. Ferre, R. Aracil, "Simplified Human Hand Models Based On Principal Component Analysis," in IFIP Conference on Human-Computer Interaction, 610–615, 2010.

[35] S. Hampali, M. Rad, M. Oberweger, V. Lepetit, "HOnnotate: A method for 3D Annotation of Hand and Objects Poses," https://arxiv.org/abs/1907.01481, 2019.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 09-21 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

A Study on Intelligent Dialogue Agent for Older Adults' Preventive Care - Towards Development of a Comprehensive Preventive Care System -

Sho Hirose[*,1], Daisuke Kitakoshi[1], Akihiro Yamashita[1], Kentarou Suzuki[2], Masato Suzuki[1]

[1]*National Institute of Technology, Tokyo College, Hachioji, Tokyo, 193-0997, Japan*

[2]*Kyorin University, Mitaka, Tokyo, 181-8612, Japan*

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 18 October, 2020
Online: 08 November, 2020

Keywords:
Intelligent Dialogue Agent
Comprehensive Preventive Care System
Reinforcement Learning

ABSTRACT

Preventive care approaches have attracted much attention in Japan, which is one of the world's most super-aged societies. These approaches aim to decrease the number of people who require nursing care or other human support. Our research group has developed several kinds of preventive care systems, including a fall prevention system, a cognitive training system, and an Intelligent Dialogue Agent (IDA). In this paper, we focus on familiarizing older adults with the IDA – a conversational agent that encourages older adults to engage it in natural conversations, while monitoring their health in the process – and we introduce a Speech Content Coordinating Function (SCCF) to the IDA to further improve older adult's familiarity with and interest in the agent. We also have a plan to develop a Comprehensive Preventive Care System (CPCS) which can encourage proactive conversation and provide effective monitoring of older adults as well as habitual cognitive training. To evaluate the basic characteristics and impressions of the CPCS, a prototype version that consists of the IDA and the Mechanism for Cognitive Training (MCT) is developed. The results of the experiments indicated that the SCCF can adequately adjust speech content depending on the user's circumstances. We also confirmed that the prototype CPCS achieves synergistic effects (e.g., more detailed monitoring of older adult's health condition, increased frequency of cognitive training) when its components are used together.

1 Introduction

This paper is an extension of work originally presented at the 2019 International Conference on Technologies and Applications of Artificial Intelligence [1].

Japan is currently known as the most advanced super-aged society. In 2018, Japanese people aged 65 and older accounted for 28.1% of the total population [2]. As the aging demographic increases, the number of people requiring long-term care also increases, as does the number of people who need nursing care or other human support, and that number in Japan has reached 6.41 million, as certified by the Long-Term Care Insurance Act [3]. Preventive care approaches, which aim to decrease the number of people who require nursing care or other human support by reducing the decline in individuals' mental and/or physical functions, have attracted a lot of attention. These approaches also aim to control increasing medical costs and the associated burden because the number of older adults is increasing while the working population is expected to decline due to declining total fertility rates. In recent years, many researchers in various fields have focused on the prevention not only of dementia but also of fall, one of the causes of the need for nursing care, and have shown a certain degree of effectiveness in preventive care [4]– [12]. However, some older adults experience physical and/or psychological distress due to preventive care activities, and therefore have trouble engaging in these activities on a regular basis.

At the same time, in the past decade, various voice-interaction-based devices and softwares (Dialogue Agents) such as smart speakers (e.g., Google Home and Amazon Echo) and voice assistants (e.g., Siri and Cortana) [13] have been developed and employed to provide useful information and a variety of services in response to users' prompts. In addition, there has been much research into the use of Dialogue Agents for older adults with dementia and mild cognitive impairment [14]– [17]. However, it should be noted that these devices and the Dialogue Agents generally do not talk to users

[*]Corresponding Author: Sho Hirose, Email: s20613@tokyo.kosen-ac.jp

without a specific request or prompt. However, we believe that introducing a function that makes it possible for Dialogue Agents to speak to users spontaneously in order to elicit a reaction will play an important role in endowing the agents with a sense of humanity and familiarity.

Against these backgrounds, previous studies examined Intelligent Dialogue Agents (IDAs) that allow natural and flexible communication with and monitoring of older adults by using a Smart Speaker, a friendly interface, to investigate older adults' impressions of the Smart Speaker and the basic characteristics of the IDA [18]. The IDA employed a Spontaneous Talking Function (STF) that allows the IDA to talk to older adults without users' requests to encourage them to actively dialogue. Our research group has shown that the STF is effective in some ways, particularly in providing increased convenience and familiarity to older adults. Our research group is also developing a cognitive training system based on a memory game on a tablet device [19] and a Fall Prevention System for regular preventive care exercise [3] to prevent dementia and fall, one of the main factors leading to a need for nursing care. Additionally, our research group is conducting a study to estimate the behavior of older adults who live alone from data gathered by supervision sensors installed in their homes [20]. In this paper, we introduce a Speech Content Coordinating Function (SCCF) into the IDA in order to further improve the user's familiarity with and interest in the IDA. The SCCF determines the content of speech by using a reinforcement learning algorithm according to the user's preferences and condition. In order to investigate the basic characteristics of the SCCF and users' impression of it, and to evaluate whether the SCCF was able to acquire appropriate policy depending on its users, several experiments were conducted. In the experiments, we developed a text-based agent in the form of a chat-bot and introduced the SCCF into it.

In addition, this paper considers the feasibility of a Comprehensive Preventive Care System (CPCS). Although we aim to develop the CPCS by integrating our preventive care systems, it is currently in the conceptual phase. Once fully developed, the CPCS will contribute to the encouragement of active conversation and the provision of not only effective monitoring of the health condition of older adults but also regular cognitive training and fall prevention exercise. As the first step to examine the adequacy and impressions of the CPCS, we conducted the experiment with a prototype version of the CPCS using both the IDA with a sensor and the Mechanism for Cognitive Training (MCT) which is at present equivalent to the cognitive training system.

At present, neither the IDA nor the prototype CPCS has been tested by older adults, because their basic characteristics and impressions have not been evaluated. Therefore, we enrolled students as participants in the experiments as the first phase to confirm the future applicability of the IDA and CPCS for older adults.

2 Preliminaries (Reinforcement Learning)

Before describing our study, we will first introduce the concept of Reinforcement Learning (RL). RL is a machine learning method [21]. An RL agent attempts to acquire an appropriate policy (a state-to-action map) based on observations and trial-and-error interactions with its environment. A policy characterizes an agent's behavior and is described using the function:

$$\pi : \mathcal{S} \times \mathcal{A} \rightarrow \boldsymbol{R}, \tag{1}$$

where $\mathcal{S}$ and $\mathcal{A}$ denote sets of states and actions, respectively. The pair $(s, a)(\forall s \in \mathcal{S}, \forall a \in \mathcal{A})$ is referred to as a rule, which states that "if an agent observes a state s, it outputs an action a". At each time step t, an agent observes state s_t, and selects action a_t using π.

RL algorithms are roughly classified into two approaches, exploration-oriented and exploitation-oriented approaches. The former approaches, which include Q-leaning, Sarsa, and so on, aim to optimize the total discounted reward typically in environments modeled by Markov Decision Process (MDP). The latter ones, such as Bucket Brigade and REINFORCE, are based on the notion of credit assignment and aim to increase learning speed and to acquire reasonable policies even in non-MDP environments although it is not ensured to acquire the optimal policy by using them. Because the environments we treat in this study seem to be non-MDP ones, the exploitation oriented RL algorithms are applied. Brief explanations about REINFORCE and Bucket Brigade, which are employed in the IDA and MCT (described later), respectively, are presented below.

2.1 REINFORCE

REINFORCE [22] is an exploitation-oriented approach based on a policy-gradient reinforcement learning algorithm. Policy-gradient reinforcement learning algorithms such as natural policy gradients [23], and stochastic gradient methods [24] learn policies directly by updating parameters θ that characterize a policy π_θ, as opposed to many conventional reinforcement learning algorithms such as Q-learning [25] or Sarsa [26]. The values of θ are related to conditional probabilities $p(a|s)$. In other words, the policy-gradient reinforcement learning algorithms aim to acquire a probability distribution that selects the optimal action to take in each state so as to maximize the sum of the rewards. The goal of these algorithms is thus to obtain the parameters θ, which characterize the probabilistic policy π_θ that maximizes the sum of rewards, updated according to the following formula (2):

$$\theta \leftarrow \theta + \eta \Delta J(\theta), \tag{2}$$

where η is the learning rate. REINFORCE updates θ by $\Delta J(\theta)$ shown below:

$$\Delta J(\theta) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \theta} \left(r_{m,t} - b\left(\theta\right) \right) \ln \pi_\theta(a_{m,t}|s_{m,t}), \tag{3}$$

where $s_{m,t}$, $a_{m,t}$, and $r_{m,t}$ are state, action and reward at the t-th time step in the m-th episode ($t = 1, ..., T; m = 1, ..., M$), respectively. In the above formula, an episode denotes the sequence of rules selected between rewards, and one time step is counted when the agent selects a rule. The term $b(\theta)$ reduces the variance of the estimate and serves to stabilize the policy learning process, called the baseline ($b(\theta) = \Sigma_{m=1}^{M} \Sigma_{t=1}^{T} r_{m,t}$), and π_θ denotes the agent's policy. In the present study, the probability that the IDA will select its action a given a state s is computed by the softmax function (4) with the

inverse temperature parameter β, because both the state and action spaces are discrete.

$$\pi_\theta(a|s)(= p(a|s)) = \frac{\exp(\beta\theta_{(a|s)})}{\sum_{a'}\exp(\beta\theta_{(a|s)})} \tag{4}$$

REINFORCE is used to coordinate the IDA's speech content depending on the user's circumstances in this study.

2.2 Bucket Brigade

In the Bucket Brigade algorithm [27], a policy is characterized by rule-weights $w(s_t, a_t)$, which are used as an agent's policy π. When an agent acquires reward R after it selects rule (s_t, a_t) at time step t, the weights of rule $w(s_t, a_t)$ and $w(s_{t-1}, a_{t-1})$ are reinforced as follows:

$$\begin{aligned} w(s_t, a_t) &= w(s_t, a_t) - C_{bid} \times w(s_t, a_t) - \\ &\quad C_{tax} \times w(s_t, a_t) + R, \qquad (5) \\ w(s_{t-1}, a_{t-1}) &= w(s_{t-1}, a_{t-1}) + \\ &\quad C_{bid} \times w(s_t, a_t), \qquad (6) \end{aligned}$$

where C_{bid} and C_{tax} denote a "bid" parameter specifying a degree of propagating the weight $w(s_t, a_t)$ to the weight $w(s_{t-1}, a_{t-1})$, in which the rule (s_{t-1}, a_{t-1}) plays a role of "trigger" to select the next rule (s_t, a_t), and a "tax" parameter which determines a discount rate for the weight of the rule selected at step t, respectively.

As described earlier, the Bucket Brigade is one of the algorithms categorized as an exploitation-oriented approach. The above formulae mean that this algorithm updates a rule weight at each step, even when no reward is given, and also that this algorithm is relatively easy to implement compared to other conventional RL methods. Moreover, the agent with the Bucket Brigade can acquire reasonable policies in many cases even if the agents do not have a sufficient number of trials because it actively employ rewarded past experience. This study applies the algorithm to the function for adjusting the difficulty level of cognitive training for respective users in the mechanism providing the users with cognitive training, which is described in the next section.

3 Preventive Care Systems

In this section, we present the preventive care systems that are currently under development or about to start development. We first overview the Intelligent Dialogue Agent (IDA), which our research group is currently developing, then introduce the Mechanism for Cognitive Training (MCT) as another preventive care framework, and finally provide brief outline of the Comprehensive Preventive Care System (CPCS), which will consist of the above two components and will be developed as our final goal.

3.1 Intelligent Dialogue Agent (IDA)

We have developed the Intelligent Dialogue Agent (IDA) with the aim of encouraging natural and flexible conversation with its users (older adults) and of monitoring older adult's health condition based on their behavior and dialogue history. The IDA talks spontaneously, using a variety of contents such as notification of daily routines (e.g., taking medicine, walking), trivia, news, weather and traffic information, and food information. The sensor mounted on the IDA also can monitor older adults within its detection area.

3.1.1 Architecture of the IDA

Figure 1 shows the final design of the IDA. The IDA consists primarily of an input/output (I/O) unit, a learning unit, and a database.

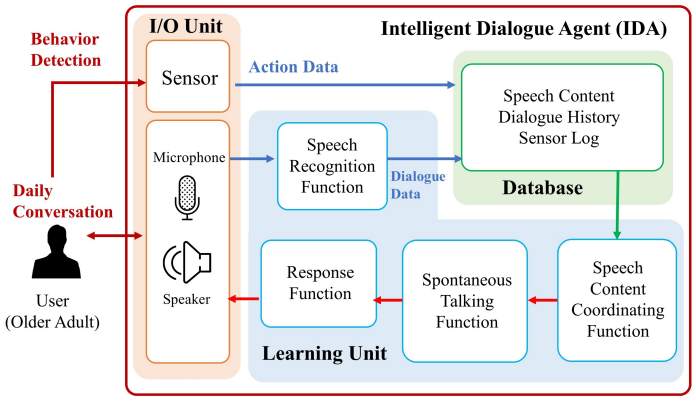


Figure 1: Final design of the IDA

(1) I/O Unit

The I/O unit is composed of a microphone, a speaker and a sensor. The user's motion is recorded by the sensor in order to track his or her characteristic behavior, such as lifestyle activities (e.g., waking and sleeping times, regular exercise). The pyroelectric sensor we adopted in this study, and which is embedded in the IDA, is both inexpensive and easy to use. Moreover, our research group has already employed a similar type of sensor and confirmed that it worked effectively to capture the motions of individuals [20]. This sensor can detect objects separate from the user, as well as even the slightest movement of people separate from the user. Although the IDA currently employs a microphone and speaker which are implemented by the included smart speaker (Google Home Mini) as its I/O unit, these components have certain limitations; for example, the smart speaker cannot acquire the user's response and dialogue history in real time. We thus plan to develop and implement an IDA that includes the highly desirable speech recognition and synthesis functions without using a smart speaker.

(2) Database

The database stores logs of the user's activity as detected by the sensors, and the history of dialogue between the IDA and the user, which is converted from speech information into text data by the Speech Recognition Function, a component of the Learning Unit. The data stored in the database are used to estimate the user's condition and to adjust the behavior of the IDA. We adopted a NoSQL cloud-hosted database known as the Firebase Realtime Database (https://firebase.google.com) for this component because it is compatible with Google Home Mini, which we employed in the IDA.

(3) Learning Unit

The learning unit has four functions: the Spontaneous Talking Function (STF), the Speech Recognition Function, the Response Function, and the Speech Content Coordinating Function (SCCF). A brief explanation of the four functions is as follows. The IDA can output its speech content without prompting by the user at the appropriate time by using the STF. The STF triggers speech to the user based on the user's daily routines and behavioral logs stored in the database. The Speech Recognition Function converts voice information obtained from the microphone, such as questions from the user or the user's responses to the IDA's speech, into textual information. The Response Function transforms the next speech content into voice information and transfers it to the speaker. In addition, the SCCF provides appropriate speech content depending on the user's preferences and condition. We expect that introducing the SCCF into the IDA will improve the convenience and familiarity of the IDA and hence the user's willingness to continue using the IDA. In order to adjust the content, timing, and frequency of speech, the following history information should be taken into account:

1. The behavioral history of the user (i.e., the frequency and number of times the user is detected) obtained by sensors on the I/O unit.
2. The history and frequency of dialogue between the user and the IDA.
3. What kinds of speech content were previously produced by the IDA and at which time slots.
4. Basic information about the user (e.g., age, gender, address, preferred topics for conversation)

3.1.2 The Speech Content Coordinating Function (SCCF)

The results of the experiments and questionnaires conducted in our previous study confirmed that the provision of appropriate information by the IDA increases the willingness of users to continue participating in dialogue with the IDA [1]. In the present study, we introduced the SCCF into the IDA and investigated its basic characteristics and performance.

In order to acquire the adequate behavior (policy) to determine natural responses depending on the particular user's preferences and/or circumstances, we apply the REINFORCE algorithm described in Section 2.1 to the SCCF as the policy learning algorithm. To conduct policy learning, we defined state, action, and reward for the IDA. The IDA's action is defined as the speech category $c \in \boldsymbol{C}$ that the agent employs to talk to the user ($\boldsymbol{C}$: A set of speech categories). There are 13 different speech categories: $\boldsymbol{C}$ = {time signal, greeting, today's weather, weekly weather forecast, today's trivia, domestic news, international news, economic news, technology news, science news, entertainment news, sports news, and each user's daily routine work}. The state of the IDA is defined as the pair of speech categories (c_1, c_2) that constituted the agent's output in the last two time steps. This state definition is based on the assumption that there is some relationship between past speech content and current content in human-to-human conversation ($c_1 \in \boldsymbol{C}, c_2 \in \boldsymbol{C}$). The user's feedback on the speech content spoken by the IDA is employed as the reward. All feedback is classified into one of three categories: positive (e.g., good, useful, beneficial), negative (e.g., bad, not useful, detrimental), or uninteresting, and the reward values obtained by the IDA are determined according to the feedback categories. Before establishing the reward setting, we asked a speech therapist at a long-term care insurance facility for older adults to comment on "how older adults come to have a good impression of speech content". She responded as follows:
[Comments of speech therapist]

"Even if older adults express negative responses to the information provided by the agent's speech, they may not necessarily have a negative evaluation of the corresponding speech (i.e., any user's reaction to the speech is evidence of being interested in the speech content)."

Based on this information from the speech therapist, we set up our SCCF to have two reward settings, one employing the speech therapist's comments and the other based on another assumption under which "the users' interest in speech directly corresponds to interest in the content of the speech." When the IDA acquires a reward, the weight parameter θ for each state and action are updated, and the probabilities for selecting actions are changed.

We substituted the softmax function (4) into the REINFORCE updating equation ((2), (3)) and the number of time steps T was set at one because we defined one episode as ending with one instance of speech. The following formula derived from (2), (3), and (4) is applied to the policy learning (parameter update) in the SCCF.

$$\theta \leftarrow \theta + \eta' \frac{1}{M} \sum_{m=1}^{M} (r_{m,t} - b(\theta))(1 - \pi_\theta(a|s)) \tag{7}$$

3.2 The Mechanism for Cognitive Training (MCT)

Our research group has developed the Mechanism for Cognitive Training (originally proposed as the cognitive training system) that encourages older adults to engage in cognitive training to help prevent dementia [19]. As shown in Figure 2, the MCT provides older adults with a brain-twister-based memory game as cognitive training, provided in such a way that they can regularly perform this memory game through repetitive interactions with a software agent on a tablet device. The software agent also serves as both an opponent who plays the game with the older adult users and a conversation partner; the goal of the latter function is to increase the familiarity of older adults with the MCT. Our system is expected to prevent dementia by preserving and improving cognitive functions such as short-term memory and encouraging and improving the user's motivations to use them. Furthermore, the MCT has a difficulty adjustment function that can set the appropriate game difficulty level according to differences between users and each individual user's current circumstances in order to create an environment that makes it easier for users to continue cognitive training for long periods of time, since they neither get bored because the game is too easy nor give up because it is too difficult. The database stores information on each user's game-playing conditions, such as memory time, response time, accuracy rate, and difficulty, and uses these data to adjust the difficulty level and provide feedback on the memory game results to the users.

The difficulty adjustment function employs the Bucket Brigade algorithm described in Section 2.2 in order to coordinate the game difficulty level. The Bucket Brigade algorithm is known to have two

characteristics: (i) advancing policy learning even when no reward is given; and (ii) additionally acquiring reasonable policies in many cases, even if the agents do not have a sufficient number of trials, by actively employing rewarded past experiences.

The software agent in the MCT is based on the concept of Human-Agent Interaction (HAI), which aims to design an appropriate framework of interactions between humans and agents. The MCT currently has a function to coordinate speech content depending on the game-playing frequency and/or time interval between bouts of gameplay so that a user can play the game (engage in the cognitive training exercise) comfortably. However, since it is intended that the software agent will be implemented on a tablet device, its functional and performance limitations may prevent it from fully achieving the desired functionality. Therefore, both functions for coordinating speech content and sensors for detecting the user's motions will be implemented in the IDA. In addition, the functions currently embedded in the software agent on the MCT will be implemented in the CPCS as functions of the IDA.

As a result of several experiments, we have confirmed that older adults were able to increase the difficulty level and the accuracy of the cognitive training by using the MCT. Furthermore, introducing the difficulty adjustment function made it possible to appropriately coordinate the training difficulty depending on the circumstances of the older adult playing the game. Although the use of MCT clearly had a positive effect on maintaining/improving the short term memory and cognitive function of older adults, we also confirmed that experiments over a longer period with a much large number of participants will be needed in order to make a detailed assessment of the MCT's effectiveness in dementia prevention.

3.3 Comprehensive Preventive Care System (CPCS)

The Comprehensive Preventive Care System (CPCS) aims to allow older adults to enjoy engaging in cognitive training while being monitored by their family and friends through active conversation. The CPCS consists of an IDA, which can carry out natural conversation with users, and an MCT (detailed in section 3.3) which provides users with a memory game on a tablet device as a kind of cognitive training (Figure 2).

Our research group is carrying out research on and development of each component with a view towards their eventual integration into the CPCS. Because the integration of the IDA and the MCT will allow the CPCS to employ the data stored in both in parallel, this system will be able to produce a synergistic effect between the cognitive training and monitoring functions. For example, the cognitive training status obtained by the MCT will be combined with the dialogue/behavior history acquired by the IDA, making possible a more detailed understanding of the user's condition (e.g., the user's degree of fatigue, whether he or she has taken a break while engaging in cognitive training, and so on). One of the final purposes of the CPCS is to notify family and friends of the circumstances of an older adult user obtained by the system, which also aims to accomplish the following three goals:

1. To maintain and improve the cognitive function of older adults;
2. To promote active conversation between older adults and their family, friends and the IDA; and
3. To create an effective environment where older adults and their families and/or community groups can be mutually involved to better monitor the older adult's health condition, and promote their care and dementia prevention.

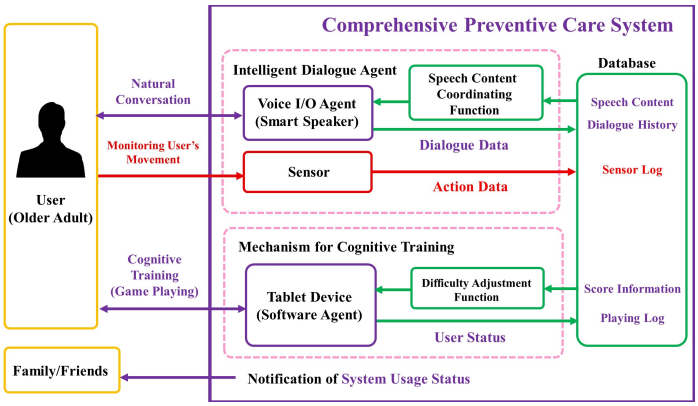


Figure 2: the Diagram of the CPCS

4 Experiments

As described in Section 1, we have not yet had an opportunity to employ older adults in the evaluation of the SCCF and the CPCS. In addition, many aged households lack a Wi-Fi environment which would be needed in order to use and evaluate the CPCS at home. For these reasons, we conducted several experiments with students to confirm that the SCCF can appropriately coordinate the speech content depending on the user's circumstances and to evaluate the basic characteristics and impressions of the CPCS. Moreover, we used the prototype version of the CPCS (described in a later section) in these experiments, because at present we have not evaluated the full version of the CPCS, but only its respective components (the IDA and the MCT).

Before conducting the experiment, a simple written questionnaire was administered to older adults in order to confirm their awareness of smartphones, the internet environment at their home, and the degree of interest/expectation in regard to spontaneous talking by the IDA (smart speaker). The questionnaire results showed that (i) older adults' degree of recognition of smart speakers was low; (ii) the majority of them did not have an internet (Wi-Fi) connection that would permit the use of a smart speaker at home; however, (iii) they had a high level of interest about conversation with the IDA (smart speaker), particularly its spontaneous talking function. Regarding item (iii), the older adults had on average 2.3 topics that they wanted the IDA to talk about, and no older adults answered that the talking function was "unnecessary" or that they had "no topics which they wanted to make the smart speaker talk about". The detailed results of the questionnaire are presented in Appendix (A).

4.1 Evaluation of the SCCF

In our first experiment, we aimed to evaluate the basic characteristics of the SCCF by employing a prototype of the IDA (henceforth referred to simply as the IDA) developed as a chat-bot using the Slack message application. The chat-bot periodically talks to provide information and learns from its users' feedback. The experiment was conducted for six days and our subjects were 10 students (average age: 18.9).

As the speech content of the chat-bot, six out of 13 categories were randomly selected before starting the experiment. The total number of states was $6 \times 6 = 36$ because the chat-bot had up to two previous categories (c_1, c_2) as pairs. From the set of selected speech categories $\boldsymbol{C'}$ ($|\boldsymbol{C'}| = 6$), the chat-bot selects speech category c according to the current policy. The initial value of the parameters θ which characterize a policy π_θ equals θ_0. Therefore, the chat-bot chooses a speech category with uniform probability at the first stage of the experimental period. Each participant responded to the speech content provided by the IDA with one of three kinds of feedback (P/N: positive/negative impression; U: uninterested). Reward values were determined by the collective feedback of users. The following two types of reward settings were utilized based on comments received from the speech therapist.

[Setting R_A]

Under this setting, in accordance with the speech therapist's assumption that "older adults would be interested in a topic if their evaluation was not U," a positive reward (+1) was assigned if the participant's feedback was P or N, and no reward was given otherwise.

[Setting R_B]

Under this setting, positive (+1) reward, negative (-1) reward, or no reward was given if the user's evaluation of a topic was P, N, or U, respectively.

The participants gave feedback for approximately 300 speech contents for each setting. To avoid any influence of the order of the participants' evaluations in settings R_A and R_B, the participants were grouped into two equally sized groups, and the order was changed in different groups. The parameters used in the policy learning and action selection are shown in Table 1.

Table 1: Parameters used in reinforcement learning

Parameter	Value
η'	0.22
M	3
θ_0	10
β	15

In addition, subjective evaluations was conducted after the end of the experimental period (three days) for each reward setting and after the end of the experiment to examine the participants' impressions of the SCCF. We also administered the questionnaire regarding the level of interest in each category at the end of the experiment because we did not want it to affect the participants' evaluation of the chat-bot's speech.

Which speech categories provided by the IDA were most preferred varied among users. The favorite speech categories and/or favored order of talking could also vary within the same user depending on the current circumstances (e.g., on holiday, at work, while eating, before bedtime, and so on). For these reasons, it was difficult to acquire a policy for providing optimal speech contents in the experiments; however, we can say that we were able to acquire a policy better expressing the user's preference by the learning in the SCCF if the speech categories preferred by the user were easy to be provided from the IDA compared to the other categories. The experiment thus confirmed the probability of the IDA generating speech for each category after learning the preferences of users while investigating the favorite categories of users via the questionnaire. By comparing the above results, we discuss the adequateness of the SCCF by evaluating whether the SCCF can learn the preference of users and how accurately the learning result matches the actual preference of users.

4.2 Preliminary Experiments with the CPCS

Several experiments were conducted to investigate the synergistic effect of using the MCT in conjunction with the IDA. As part of our preparation for conducting the experiment with older adults, we asked 10 students (the same as the participants in section 3.1) to participate in our preliminary experiments. However, after excluding one participant (participant F) due to an equipment malfunction that resulted in incomplete sensor data, a total of nine participants were included in these experiments.

A prototype version of the CPCS was employed for these experiments. The system consists primarily of the MCT and the IDA, but does not have functions that allow the two mechanisms to communicate or that allow either of the components to use data collected by the other. Using both mechanisms in parallel would contribute to the combined use of the log data obtained by each. We conducted two kinds of experiments to investigate differences in impressions of the CPCS when the IDA talked spontaneously and when it did not. This experiment was conducted in two periods, a speech period and a nonspeech period, of three days each for a total experimental duration of six days. However, to ensure that the order of the two periods did not affect the subjects' impressions of the CPCS, the order was randomized for each participant. The IDA employed in the experiment had a simple STF consisting of a smart speaker (Google Home Mini) and a pyroelectric sensor without introducing the SCCF, and provided information (speech content $c \in \boldsymbol{C''} = \boldsymbol{C} \backslash$ {each user's daily routine work}) selected randomly every 15 min between 9:00 and 23:00 during the speech period. Additionally, we asked participants to listen to the IDA's speech at least eight times a day. Sensors detected the participants' behavior throughout the experimental period (sampling interval: 1 s; detection range: 115° within a 2-m view range). The difficulty adjustment function was included in the MCT. We asked the participants to play a 10-min memory game every day near the location where the sensor was installed.

Written questionnaires were also administered after the first half of the experiment and at the end of the experiment to inquire about the number of times participants heard the IDA speaking, whether the difficulty level of the memory game was appropriate, and their

Table 2: The actual speech probability of the chat-bot's speech at the end of the experiment for each participant.

Reward setting	R_A						R_B					
Participant / ranking	1st place	2nd place	3rd place	4th place	5th place	6th place	1st place	2nd place	3rd place	4th place	5th place	6th place
A	**0.680**	0.040	0.050	0.060	0.080	0.090	**0.300**	**0.280**	**0.180**	0.090	0.040	0.110
B	**0.505**	**0.257**	0.040	0.099	0.059	0.030	**0.310**	**0.380**	0.140	0.030	0.070	0.070
C	0.030	0.000	0.050	**0.416**	0.040	**0.455**	**0.300**	**0.180**	**0.210**	0.150	0.120	0.040
D	**0.198**	0.139	**0.208**	**0.238**	0.079	0.129	**0.416**	0.069	**0.238**	0.079	0.079	0.109
E	0.108	**0.206**	**0.235**	0.157	0.118	**0.167**	**0.168**	**0.218**	0.129	0.139	0.139	**0.208**
F	0.109	0.059	**0.327**	0.119	**0.218**	**0.168**	**0.168**	0.149	**0.208**	0.099	**0.267**	0.099
G	**0.300**	**0.260**	0.100	**0.180**	0.080	0.080	0.139	**0.287**	0.119	**0.257**	0.119	0.069
H	0.000	0.000	0.000	0.020	**0.980**	0.000	0.149	0.129	**0.228**	**0.277**	0.099	0.119
I	**0.190**	**0.170**	0.120	0.100	**0.250**	**0.170**	**0.304**	0.127	0.147	**0.196**	0.157	0.069
J	**0.240**	**0.256**	0.112	0.096	0.088	**0.208**	0.160	**0.260**	0.080	0.070	**0.200**	**0.230**
Average	**0.236**	0.139	0.124	0.148	**0.199**	0.150	**0.241**	**0.208**	**0.168**	0.139	0.129	0.112

willingness to continue to use the CPCS.

Similarly to the case of IDA, how efficiently the CPCS prototype can monitor a user's behavior would be influenced by its installation site and the daily habits of the user. However, synergistic effects are also expected through the concomitant use of the IDA and the MCT, and could have a positive impact on the frequency with which users engage in cognitive training. To clarify these issues, we evaluated and discussed the impressions of users, the effectiveness of the system at monitoring users, and the relationships between dialogue with the IDA and the frequency of cognitive training by examining data such as the sensor log, the result of the questionnaire/interview with the participants, and the training frequency in parallel.

5 Results and Discussion

This section describes our experimental results and discusses our investigation of the basic characteristics and performance of the SCCF as well as the effectiveness of the CPCS prototype presented in Section 4.

5.1 Performance Evaluation of the SCCF

Let us first discuss the performance and adequacy of the SCCF in providing appropriate information according to the user's preferences and condition. Table 2 lists the probability of the chat-bot talking about each category over the last 100 times in the three-day experimental period. This table shows the results for all participants. In this table, "ranking" corresponds to the participants' interest levels for each speech category obtained by the questionnaire administered after the experiment. This table revealed that the probability of talking acquired by the learning tends to be high when the rank of the interest level is high, especially under the reward setting R_B. The bold digits in the table indicate areas in which probability increased from its initial value (16.66...%). Table 3 represents the results of the questionnaire concerning the participants' interest levels for each category on a five-point scale (the other questionnaire results are shown in Appendix (B)). Digits in boldface indicate areas with an interest level greater than 3. Table 2 shows that, for reward setting R_B, the probabilities of the top three speech categories tend to be larger than the initial probability value while the probabilities of the bottom three categories tend to be smaller. Additionally, the results presented in Tables 2 and 3 show that, under reward setting R_B, there is a positive correlation between the probabilities of speech from the chat-bot at the end of the experiment and interest in the actual content of the speech under reward setting R_B. In other words, the probability that the chat-bot provides the participants with topics of interest increased throughout the experiment, while the probability that it provided topics for which many participants' feedback was U (no interest) decreased. Therefore, we believe that the chat-bot with reward setting R_B was able to provide information that reflected the users' thoughts and states.

On the other hand, the chat-bot with reward setting R_A failed to acquire the appropriate policy. An important feature in the chat-bot with setting R_A was that the probability of a particular category being provided was too large (up to 98%) when targeting a specific participant compared to the results obtained with setting R_B (up to 41%). Critically, reward setting R_A does not include negative rewards; regardless of whether the user's feedback is positive or negative, a positive reward is given. Thus, it is much more likely that the chat-bot with the reward setting R_A will be given a positive reward than the chat-bot with reward setting R_B. However, it seems that the speech probability was changed too drastically for the chat-bot with reward setting R_A under the current parameter settings.

Table 3: Degree of interest in participants' speech categories (on a 5-point scale).

Participant / ranking	1st place	2nd place	3rd place	4th place	5th place	6th place
A	**5**	**4**	**4**	3	2	2
B	**5**	**4**	3	3	3	2
C	**4**	**4**	**4**	3	1	2
D	**5**	**5**	**4**	1	2	1
E	**5**	3	3	**4**	3	2
F	**5**	**5**	**4**	3	2	1
G	**4**	**4**	**4**	**4**	3	2
H	**4**	**4**	3	2	1	1
I	**5**	**4**	**4**	2	1	1
J	**4**	**4**	**5**	**4**	3	3
Average	**4.6**	**4.1**	**3.8**	2.9	2.1	1.7

5.2 Basic Characteristics of the CPCS

This section discusses whether the prototype version of the CPCS can monitor and/or understand a user's circumstances in detail using the log data collected by the IDA and MCT, and evaluates users' impressions of the system as well. The sensor data of participant F were incomplete due to an equipment malfunction and are therefore not included in these results.

5.2.1 *Differences in the Participants' Amounts of Activity*

Let us consider the amount of activity detected by the pyroelectric sensor on the CPCS. A participant's amount of activity Act^x is defined as the number of times the sensor detects his/her motion for x min. The maximum and minimum values of Act^x equal $60x$ and 0, respectively. Since the sensor may also detect objects other than the user, the value of Act^x must be corrected depending on the experimental circumstances. Table 4 shows the maximum amount of activity per 30 min for each participant over the course of the experiment. As shown in the table, Participant B had the greatest amount of activity (number in red) while Participant A had the lowest amount (blue). Figure 3 also shows sketches of the rooms usually used by Participants A and B, and the location and orientation of the sensor in those rooms.

Installation of the sensor in a location where it can detect the participant's motion appropriately is essential in order for behavior to be estimated accurately. Since Participant B answered in the interview after the experimental period that she often reads for long periods of time at her desk, we believe that this caused her amount of activity to be greater than that of the other participants because she was active within the detection range of the sensor. These results suggest that environmental factors, such as the location and direction of the sensor, may influence the recorded amount of activity (some noise and error in sensing may also influence the results).

Therefore, in the case that the CPCS in its current form would be employed in practice by older adults, it is necessary to consider in advance where the sensors should be installed to effectively track the older adults' amount of activity. At the same time, if the CPCS could be installed at the appropriate location, we can expect to properly understand their behavior.

Table 4: Maximum amount of activity per 30 min for each participant.

Participant	Act^{30}	Participant	Act^{30}
A	1210	G	1301
B	1800	H	1353
C	1312	I	1711
D	1386	J	1503
E	1672		

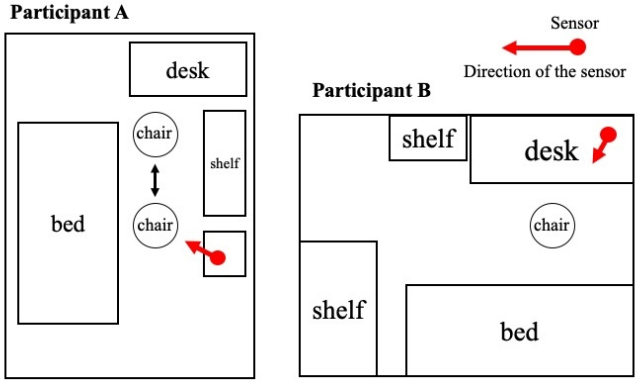


Figure 3: Diagram of the rooms where the experiment was conducted for Participants A and B

5.2.2 *Sensor Logs and Time Spent on the Memory Game*

Figure 4 depicts the average number of cognitive training memory games played every 30 min on weekdays (bar graph) and activity patterns (i.e., transition in the amount of activity) recorded by the sensor (line graph). The results shown in this figure are the average values for seven participants who showed similar behavioral trends. Participant I was excluded because his academic year and life patterns were different from those of the others. Here, interval (b) in Figure 4 is characterized by low activity because the participants were at school and not near the CPCS installation site. In addition, as can be seen at interval (c), there was an overall tendency for the participants to play the memory game more often at night, increasing their activity level at the same time. In contrast, interval (a) shows a larger amount of activity compared to interval (b), but almost no gameplay history except for Participant E. This result indicates that the participants were preparing to go to school during the corresponding period.

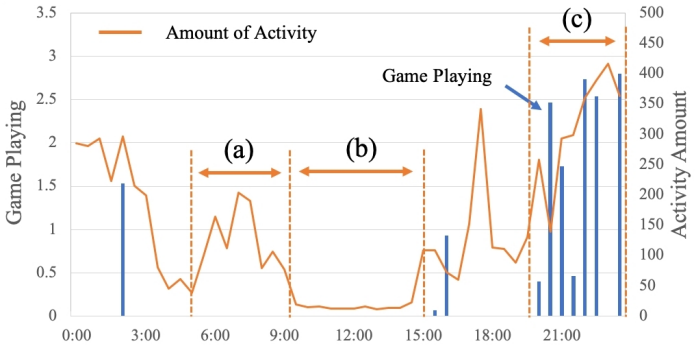


Figure 4: Average amount of game playing and activity patterns on weekdays.

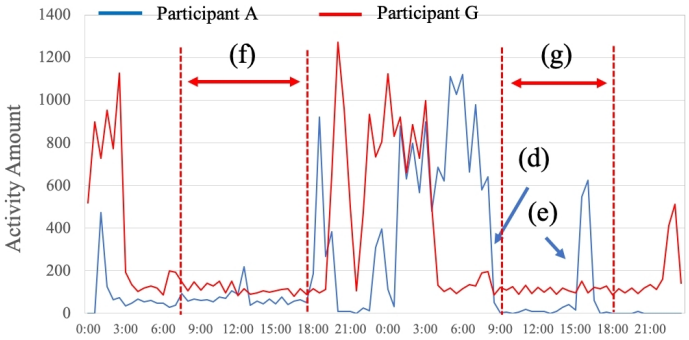


Figure 5: Weekend activity patterns for Participants A and G

At the same time, we can also see significant differences between the participants' behavior on weekday vs. weekends in terms of their life rhythms, such as sleeping hours and daily habits. Figure 5 shows the weekend activity patterns for Participants A and G, which were particularly distinctive. These patterns allow us to estimate the participants' awakening time and bedtime. For example, Participant A's activity decreased early Sunday morning ((d) in Figure 5) and increased around 15:00 ((e) in Figure 5). In the interview after the experiment, Participant A confirmed that this activity pattern corresponded to his sleep pattern. Similarly, the

other participants' sleep patterns could also be estimated from their activity logs. These results demonstrate the usefulness of the users' stored behavioral log data.

On the other hand, it was not possible to fully estimate Participant G's weekend sleep pattern. Figure 5 suggests that Participant G's bedtime was 3:00 on Saturday, and her waking time was around 19:00. However, in the interview after the experiment, we learned that she was awake but not in the room with the sensor at certain times of the day, such as from 7:30 to 17:00 (interval (f) in Figure 5) on Saturday and from 8:00 to around 18:00 (interval (g) in Figure 5) on Sunday. The fact that Participant G spent less time in the room where the CPCS was installed may have contributed to her lower amount of activity.

These results demonstrate that the use of sensor logs and gameplay histories obtained by the CPCS allows us to estimate not only daily habits such as waking and sleeping times but also whether users go out for long periods of time.

5.2.3 *Using Sensor Logs to Capture Users' Lifestyle Habits*

This section focuses on the sensor logs obtained in the IDA, a component of the CPCS, while playing the memory game on the MCT, and before and after playing. Figure 6 shows the amount of activity recorded every one minute for about 20 min while Participant B was playing the memory game. The memory game has 13 levels of difficulty and records three degrees of fatigue: 0, Bad; 1, Normal; 2, Good. Figure 7 shows the transitions in difficulty level and degrees of fatigue during the time that Participant B played the memory game. Note that all the intervals in Figure 7 correspond to interval (h) in Figure 6. The red circles on the red line plot in Figure 7 indicate the times at which the memory game was started. In Participant B's interview after the experimental period, she noted that she played the memory game at her desk (near the location where the CPCS was installed). During and after the gameplay (interval (h) and the second half of (h) in Figure 6), the amount of activity increased compared to the period of time before play, while after 22:12 (interval (i) in Figure 7), the participant temporarily stopped playing the memory game. Figure 6 confirms that the amount of activity remains high. We can speculate that the cause of this pattern was that Participant B was feeling fatigued from the memory game and may have stopped playing the game to do some other task or take a break.

Figure 8 depicts the amount of activity recorded every one min for about 20 min while Participant E was playing the memory game. In contrast to the results obtained for Participant B, Participant E's activity tended to decrease during the game-playing time ((j) in Figure 8) compared to his activity before and after playing. The interviews revealed that Participant E tended to play memory games on the bed. We inferred that this position was responsible for the low activity levels recorded by the system, since the location of the sensor would have made it difficult to detect someone on a bed. The interviews of the other participants indicated that the sensor logs adequately expressed their behaviors. It was clear from these results that the data obtained from the IDA and the MCT can be used to extrapolate detailed information about the user's circumstances.

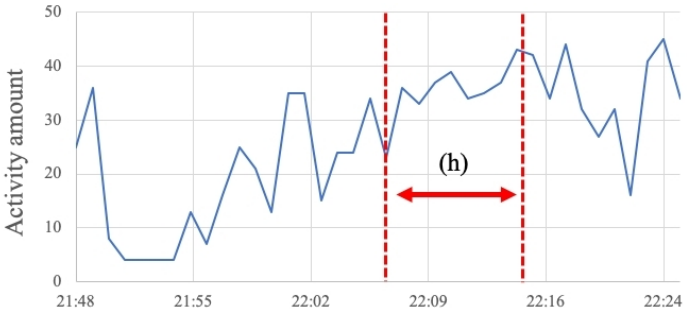


Figure 6: Activity before, during, and after Participant B's game playing.

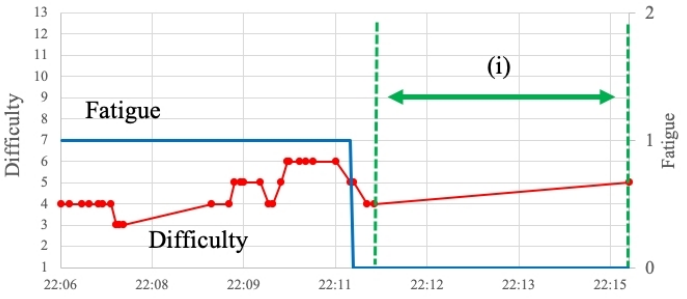


Figure 7: Changes in difficulty and fatigue during the memory game (Participant B).

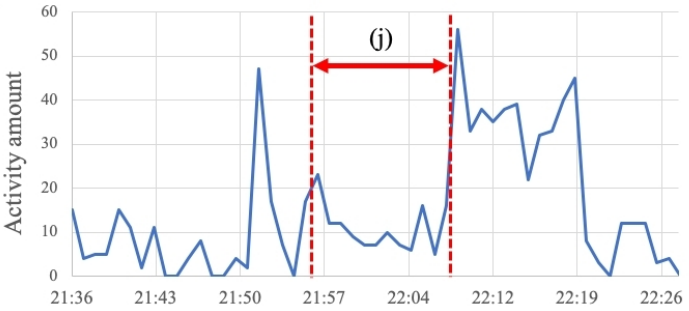


Figure 8: Activity patterns for Participant E.

5.2.4 *Questionnaire Investigation of the CPCS*

Next we will discuss the participants' impressions of the CPCS and the synergistic effects of combining the IDA and the MCT based on the results of subjective evaluations of the nine participants (college students) conducted after the experimental period. Among the nine participants, five stated that they felt there was a need to use the CPCS; however, there was only one participant who wanted to use both the IDA and the MCT as individual components of the CPCS. Thus, although in some cases the student participants expressed interest in the components of the CPCS, it can be said that, on the whole, they did not have a positive impression of the prototype version of the CPCS. One reason for this may be that students attend classes on a daily basis and have frequent opportunities to talk with the people around them. Further research needs to be conducted to investigate differences in the trends across age groups.

Table 5 shows the average number of times the memory game provided by the MCT was played per day by each participant with

Table 5: Differences in the number of times the memory game was played with and without speech from the IDA.

Participant	A	B	C	D	E	G	H	I	J	Average
Not talking	14.7	5.0	23.0	21.0	0.0	37.0	21.0	31.0	30.0	20.3
Talking	**20.7**	**10.0**	**26.7**	**31.0**	**15.7**	**41.0**	**29.0**	**40.3**	**41.3**	**28.4**
Difference (talking - not talking)	6.0	5.0	3.7	10.0	15.7	4.0	8.0	9.3	11.3	8.1

and without speech from the IDA. When the IDA spoke, the number of times each participant played was greater than that when the IDA did not speak (average: 8.1 times). We consider that the above result was caused by the synergy of the combination of the IDA and the MCT, because the dialogue between the IDA and the participant promoted their habitual game playing. The other questionnaire results can be seen in Appendix (C).

6 Conclusion

In the present study, we introduced the Speech Content Coordinating Function (SCCF) into the Intelligent Dialogue Agent (IDA), and developed a prototype version of the Comprehensive Preventive Care System (CPCS) consisting of the IDA and the Mechanism for Cognitive Training (MCT). The SCCF was developed as a text-based agent in the form of a chat-bot that employed reinforcement learning based on the policy gradient method. This system was tested in an experiment with student participants. The experimental results confirmed that the SCCF acquires appropriate policies based on the user's preferences and conditions. Furthermore, the results of experiments using the prototype version of the CPCS confirmed that taking advantage of data obtained from both the IDA and the MCT can lead to synergistic effects to monitor users in greater detail, better understand their circumstances, and increase the frequency of cognitive training.

In future research, we will further improve the convenience and familiarity of the SCCF by determining not only the content of speech but also the appropriate frequency and timing of speech based on the user's circumstances at any given time. At the same time, because this improvement may cause an exponential increase in the number of states and actions compared to the learning environment of the experiments conducted in the present study, we plan to improve the current learning algorithm or apply more powerful ones, such as deep reinforcement learning algorithms. The adaptability and familiarity of the improved SCCF will need to be evaluated through more practical experiments with older adults in a setting that more closely replicates real-life circumstances.

We also expect that further synergistic effects will occur by allowing the components of the prototype CPCS to communicate and share their data. It will also be necessary investigate the effects of the CPCS on older adults and their impressions of this system by fully examining the different trends in people of different age groups. We also plan to integrate a fall prevention system, which is currently being developed in parallel by our research group, into the CPCS; however, since this system involves physical exercise and a certain risk of injury, we will proceed with its integration in stages, starting with older adults who are at little risk of injury. Similarly, we will continue to work on the introduction of other preventive care systems and expect to see synergies similar to those obtained in the present study.

Acknowledgment The authors would like to thank the staff of the Fujisawa Care Center, and the members of the Hachioji Senior Citizens' Club and Cafe Kajiyashiki for their cooperation.

References

[1] D. Kitakoshi, S. Hirose, A. Yamashita, M. Suzuki, K. Suzuki, "Development of an Intelligent Dialogue Agent for Older Adults: Evaluation of Functions to Control Spontaneous Talk and Coordinate Speech Content," in 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 1–6, IEEE, 2019, doi:10.1109/TAAI48200.2019.8959870.

[2] S. Bureau, The Statistical Handbook of Japan 2019, Ministry of Internal Affairs and Communications, 2019.

[3] D. Kitakoshi, T. Okano, M. Suzuki, "An empirical study on evaluating basic characteristics and adaptability to users of a preventive care system with learning communication robots," Soft computing, **21**(2), 331–351, 2017, doi:10.1007/s00500-015-1631-7.

[4] T. Suzuki, H. Kim, H. Yoshida, T. Ishizaki, "Randomized controlled trial of exercise intervention for the prevention of falls in community-dwelling elderly Japanese women," Journal of bone and mineral metabolism, **22**(6), 602–611, 2004, doi:10.1007/s00774-004-0530-2.

[5] A. Sixsmith, N. Johnson, "A smart sensor to detect the falls of the elderly," IEEE Pervasive computing, **3**(2), 42–47, 2004, doi:10.1109/MPRV.2004.1316817.

[6] N. K. Suryadevara, S. C. Mukhopadhyay, "Wireless sensor network based home monitoring system for wellness determination of elderly," IEEE sensors journal, **12**(6), 1965–1972, 2012, doi:10.1109/JSEN.2011.2182341.

[7] W.-H. Wang, P.-C. Chung, Y.-L. Hsu, M.-C. Pai, C.-W. Lin, "Inertial-sensor-based balance analysis system for patients with Alzheimer's disease," in 2013 Conference on Technologies and Applications of Artificial Intelligence, 128–133, IEEE, 2013, doi:10.1109/TAAI.2013.36.

[8] R. Naqvi, D. Liberman, J. Rosenberg, J. Alston, S. Straus, "Preventing cognitive decline in healthy older adults," CMAJ, **185**(10), 881–885, 2013, doi:10.1503/cmaj.121448.

[9] Y. S. Delahoz, M. A. Labrador, "Survey on fall detection and fall prevention using wearable and external sensors," Sensors, **14**(10), 19806–19842, 2014, doi:10.3390/s141019806.

[10] T. Watanabe, K. Kamata, S. A. Hasan, S. Shibusawa, M. Kamada, T. Yonekura, M. Yamada, "Design of an antagonistic exercise support system using a depth image sensor," in Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, 162–169, 2016.

[11] A. Chelli, M. Patzold, "A Machine Learning Approach for Fall Detection and Daily Living Activity Recognition," IEEE Access, **7**, 38670–38687, 2019, doi:10.1109/ACCESS.2019.2906693.

[12] D. Yacchirema, J. DePuga, C. Palau, M. Esteve, "Fall detection system for elderly people using IoT and ensemble machine learning algorithm," Personal and Ubiquitous Computing, **23**(5-6), 801–817, 2019, doi:10.1007/s00779-018-01196-8.

[13] G. López, L. Quesada, L. A. Guerrero, "Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces," in International Conference on Applied Human Factors and Ergonomics, 241–250, Springer, 2017, doi:10.1007/978-3-319-60366-7_23.

[14] M. K. Wolters, F. Kelly, J. Kilgour, "Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia," Health Informatics Journal, **22**(4), 854–866, 2016, doi:10.1177/1460458215593329.

[15] S. Nakatani, S. Saiki, M. Nakamura, K. Yasuda, "Generating Personalized Virtual Agent in Speech Dialogue System for People with Dementia," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10917 LNCS, 326–337, 2018, doi:10.1007/978-3-319-91397-1_27.

[16] A. Russo, G. D'Onofrio, A. Gangemi, F. Giuliani, M. Mongiovi, F. Ricciardi, F. Greco, F. Cavallo, P. Dario, D. Sancarlo, V. Presutti, A. Greco, "Dialogue Systems and Conversational Agents for Patients with Dementia: The Human-Robot Interaction," Rejuvenation Research, **22**(2), 109–120, 2019, doi:10.1089/rej.2018.2075.

[17] F. Tang, I. Uchendu, F. Wang, H. H. Dodge, J. Zhou, "Scalable diagnostic screening of mild cognitive impairment using AI dialogue agent," Scientific Reports, **10**(1), 2020, doi:10.1038/s41598-020-61994-0.

[18] S. Yamada, D. Kitakoshi, A. Yamashita, K. Suzuki, M. Suzuki, "Development of an Intelligent Dialogue Agent with Smart Devices for Older Adults: A Preliminary Study," in 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI2018), 50–53, IEEE, 2018, doi: 10.1109/TAAI.2018.00020.

[19] D. Kitakoshi, R. Hanada, K. Iwata, M. Suzuki, "Cognitive training system for dementia prevention using memory game based on the concept of human-agent interaction," Journal of Advanced Computational Intelligence and Intelligent Informatics, **19**(6), 727–737, 2015, doi:10.20965/jaciii.2015.p0727.

[20] H. Eguchi, S. Yoshikawa, A. Yamashita, D. Kitakoshi, Y. Ho, Y. Hujimoto, T. Yamaguchi, M. Kanae, T. Sato, "Analysis of pyroelectric sensor data installed in thehomes of elderly individuals," in Proc. the First International Symposium on AI for ASEAN Development, 2018.

[21] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, MIT press Cambridge, 1998.

[22] R. J. Williams, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," Machine Learning, **8**(3-4), 229–256, 1992, doi:10.1007/BF00992696.

[23] S. M. Kakade, "A natural policy gradient," in Advances in neural information processing systems, 1531–1538, 2002.

[24] H. Kimura, K. Miyazaki, S. Kobayashi, "Reinforcement learning in POMDPs with function approximation," in International Conference on Machine Learning, volume 97, 152–160, 1997.

[25] C. J. Watkins, P. Dayan, "Q-learning," Machine learning, **8**(3-4), 279–292, 1992, doi:10.1007/BF00992698.

[26] G. A. Rummery, M. Niranjan, On-line Q-learning using connectionist systems, volume 37, University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[27] R. Nakano, "Efficient learning of behavioral rules," in From Animals to Animats 6: SAB2000 Proceedings Supplement of the Sixth International Conference on Simulation of Adaptive Behavior, p178-184. Paris, France, 2000.

Appendix

We administered three questionnaires concerning (A) older adults' impression of the smart speaker and the IDA, (B) the evaluation of the SCCF, and (C) the preliminary experiments with the CPCS, before, during, or after the experiments described in Section 4. Several major results were already given in Section 4, and other results related to the discussion in Section 4 are shown below. Tables A1 - A4, B1 - B3, and C1 - C8 list the questions and typical answers for each.

(A) A questionnaire administered to investigate older adults' impression of the smart speakers and the IDA

Affiliation of the participants

A preventive care circle for older adults and a senior citizens' club

Number of participants

28 (9 from a preventive care circle, and 19 from the senior citizens' club)

Gender and age

male: 12; female: 16; average age: 76.6 (male: 76.4; female: 77.4)

Table A1: Items in the questionnaire administered to older adults.

Item	Question
Awareness of smart speakers.	(1) Are you familiar with smart speakers? (2) What kind of speech content would you like the IDA to spontaneously talk about? (multiple choice allowed) (3) Would you want to use a smart speaker if one was in your home?
Impression of the spontaneous talking function	(4) If a smart speaker talked to you spontaneously, what would be your reaction? Could you select all items applicable to you? (5) If a smart speaker could talk to you spontaneously, what would you like to talk about?
Information about the environment	(6) Do you have a Wi-Fi environment at home?
Information about the participant	(7) Could you tell me your daily routine? (8) Could you tell me your hobbies?

Table A2: Answers to question (1) in Table A1 (Answerer: 9 older adults in preventive care circle).

Answer	Percentage of respondents (%)
I know about them.	0.0
I have heard the name.	11.1
I don't know about them.	77.8
No answer	11.1

Table A3: The number of contents the older adults selected for question (2) in Table A1.

Index	Number of responses
Maximum	6.0
Minimum	1.0
Average	3.4

Table A4: Answers to question (6) in Table A1.

Answer	Percentage of respondents (%)
I have Wi-Fi.	35.7
I don't have Wi-Fi.	46.4
I don't know.	7.1
No answer	10.7

(B) A questionnaire administered to evaluate the SCCF

Affiliation of the participants

National Institute of Technology, Tokyo College

Number of participants

10

Gender and age

male: 8; female: 2; average age: 18.9

Table B1: Items in the questionnaire administered after the experiment to evaluate the SCCF.

Item	Question
Effect of the learning	(1) Comparing the first and third of these 3 days, did you feel the speech contents from the chat-bot changed? (2) The IDA's setting to learn your preferences for its speech content was different between the first 3 days and the last 3 days. In which period do you think the IDA provided more interesting speech contents, the first 3 days or the last 3 days? (3) Could you tell me the reason for your answer to the above question?
Impression of the chat-bot's speech content	(4) Which of the following best describes your impression of the IDA's speech content on this particular day?
Participants' interest in the chat-bot's speech content	(5) Could you rank the topics provided by the chat-bot in terms of how much they interested you? (6) Could you rate each topic provided by the chat-bot during the experiment on a 5-point scale?
Participants' interest in the topics in daily life	(7) Could you rank the topics provided by the chat-bot in order according to how much they usually interest you? (8) Could you rate the topics provided by the chat-bot on a 5-point scale based on how much they interest you in your daily life?

Table B2: Older adults' impression of the chat-bot at each reward setting (answers to questions (1) and (2) in Table B1, 3-point scale, with 3 being best).

Item / Reward setting	R_A	R_B
Changes in speech content from chat-bot	2.3	1.9
Impression of the chat-bot's speech content	2.4	2.3

Table B3: Which reward settings R_A or R_B was preferred by older adults (answers to question (6) in Table B1).

Reward setting	Number of choosers
R_A	3
R_B	6
Neither	1

(C) A questionnaire administered regarding the preliminary experiments with the CPCS

Affiliation of the participants

National Institute of Technology, Tokyo College

Number of participants

10 (As for the questionnaires during and after the preliminary experiment, the number of participant equals 9.)

Gender and age

male: 8; female: 2; average age: 18.9

The parts of Tables C7 and C8 expressed as red-colored text are the questions and the answers discussed in Section 5.2.4.

Table C1: Items in the questionnaire administered before the preliminary experiment

Item	Question
Information about the environment	(1) Please place a check mark beside the following items if they were located at the system installation site.
Information about the participant	(2) Could you tell me your daily routine? (3) Could you tell me the approximate time when you go home? (4) Could you tell me your schedule after going home on mm/dd? (5) Could you tell me how many people you live with?
Awareness of smart speakers	(6) Do you know about smart speakers? (7) Do you have a smart speaker? (8) Would you want to use a smart speaker if you had one at home? (If you already have one, do you want to keep using it?)

Table C2: Equipment in the participant's home. Answers to question (1) in Table C1.

Equipment	The number of people with the equipment.
Display that can be connected with HDMI	9
HDMI cable	9
Wired/Wireless keyboard.	8
Wired mouse.	7

Table C3: Awareness of smart speakers. The number of respondents by choice for each question ((6), (7), (8) in Table C1).

Question	Answer	The number of respondents
(6)	I know about them.	5
	I only know the name.	5
	I don't know about them.	0
(7)	Yes.	2
	No.	8
(8)	I want to use one.	6
	I'm not sure.	2
	I don't want to use one.	0
	I have one and I want to keep using it.	2
	I have one, but I don't want to keep using it.	0

Table C4: Items in the questionnaire administered during the preliminary experiment

Item	Question
Frequency of listening to the IDA's speech	(1) Could you tell me the average number of times per day you listened to the speech from the IDA?
Impression of the spontaneous talking function	(2) Which of the following best describes your impression of the spontaneous talk from the IDA? (3) What is your impression of the IDA's spontaneous talk? (4) Could you tell me the reason you chose that item as the answer to the above question? (5) Could you tell me the contents that you want the IDA to talk about? (if any)
Willingness to keep using the CPCS	(6) Do you want to keep using the cognitive training mechanism and the IDA?
Impression of the CPCS	(7) Which of the following best describes your impression of the prototype of the CPCS (the cognitive training mechanism and the IDA)? (8) Could you tell me the reason for your answer?

Table C5: The number of times each older adult listened to the IDA's speech during the experimental period (answers to question (1) in Table C4).

Index	Value
Maximum	16.0
Minimum	1.0
Average	7.0

Table C6: The number of respondents by choice for each question ((2), (3), (6), (7) in Table C4)

Index	Answer	The number of respondents
(2)	I'd like the IDA to talk more frequently.	3
	The frequency of speech from the IDA is high.	2
	I'd like the IDA to increase its amount of speech.	3
	I'd like the IDA to decrease its amount of speech.	1
	The IDA provides useful information.	3
	The IDA provides uninteresting information.	4
(3)	Very good. It would be useful.	0
	A bit good.	3
	No opinion.	5
	Usually not good.	1
	Very bad. Unnecessary.	0
	Yes, both of them very much.	0
(6)	Yes, both of them a little.	1
	Yes, only the cognitive training mechanism very much.	2
	Yes, only the cognitive training mechanism a little.	1
	Yes, only the IDA very much.	0
	Yes, only the IDA a little.	4
	I don't want to use either of one of them very much.	1
	I don't want to use either one of them at all.	0
(7)	I thought it was interesting and a necessary part of the system.	4
	I thought it was interesting, but not really necessary for the system.	0
	I didn't feel the necessity of the IDA.	3
	I didn't feel the necessity of the cognitive training mechanism.	1
	I thought it was not interesting at present, but the system would be needed in the future.	1
	I thought it was not interesting and I do not feel that it is necessary.	0

Table C7: Items in the questionnaire conducted after the preliminary experiment

Item	Question
Impression of the training difficulty adjustment function	(1) The cognitive training mechanism has a function to coordinate training difficulty. How did you feel the difficulty of training (memory game) was coordinated by the function? (2) How would you like to the training difficulty to changed?
Required functions in the cognitive training mechanism	(3) Could you tell me your opinion about the functions you want us to add and/or your requests for improvement?
Willingness to keep using the CPCS	(4) Which of the following systems would you want to use on a regular basis?
Impression of the spontaneous talking function	(5) Could you tell me your impression of the IDA having the spontaneous talking function compared to that without the function? (6) Could you tell me the reason for your answer to the above question?
Required functions in the CPCS	(7) Could you tell me your opinion about the functions you want us to add and/or your requests for improvement in the CPCS?

Table C8: The number of respondents by choice for each question ((1), (2), (4), (5) in Table C7).

Index	Answer	The number of respondents
(1)	It was easy throughout the game.	0
	It was appropriate.	8
	It was difficult throughout the game.	0
	It was sometimes easy, sometimes difficult.	1
(2)	I want to be able to change the difficulty by myself.	3
	It is better that the cognitive training mechanism changes the difficulty.	3
	Both are OK.	0
	It depends on the situation.	3
(4)	The IDA and the cognitive training mechanism.	1
	The IDA.	3
	The cognitive training mechanism.	5
	I don't want to use either of them.	0
(5)	I thought it was interesting and felt that it was a necessary part of the system.	3
	I thought it was interesting, but didn't feel that it was necessary.	2
	I didn't feel the necessity of the IDA.	2
	I thought it was not interesting at present, but the function will be needed in the future.	2
	I thought it was not interesting and also not necessary.	0

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 10-16 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Synthesis and Characterization of Graphene Oxide Under Different Conditions, and a Preliminary Study on its Efficacy to Adsorb Cu^{2+}

Olayinka Oluwaseun Oluwasina, Surjyakanta Rana, Sreekantha Babu Jonnalagadda, Bice Susan Martincigh*

School of Chemistry and Physics, University of KwaZulu-Natal, Westville Campus, Private Bag X54001, Durban, 4000, South Africa

ARTICLE INFO

Article history:
Received: 22 August, 2020
Accepted: 30 October, 2020
Online: 10 January, 2021

Keywords:
Graphite
Sodium Nitrate
Characterization

ABSTRACT

Graphene oxide (GO) was prepared by the modified Hummer's method, but the mass ratio of graphite to sodium nitrate ($NaNO_3$) was varied from 2:1, 1:1, and 1:2. The primary reason for the variation was to determine the optimum conditions that would afford more oxygen functional groups to improve the material for application in adsorption. The final products, termed $GO_{2:1}$, $GO_{1:1}$ and $GO_{1:2}$, were analyzed by several instrumental techniques. The layered structure of the GO sheet was established by transmission electron microscopy, while powder X-ray diffraction showed that the $GO_{2:1}$ material was more crystalline than either $GO_{1:1}$ or $GO_{1:2}$. Raman spectroscopy revealed the presence of a greater defect density in $GO_{2:1}$. The presence of oxygen functional groups was verified by Fourier transform infrared spectroscopy, and these were quantified by the Boehm titration method. Overall, $GO_{2:1}$ had a larger oxygen content than either $GO_{1:1}$ or $GO_{1:2}$, and a larger specific surface area. A preliminary study on the adsorption properties of the samples revealed that $GO_{2:1}$ exhibited the highest percentage removal for Cu^{2+} in aqueous solution. Thus, the preparation of graphene oxide with a smaller amount of $NaNO_3$ yielded a material with a greater oxygen content, which showed suitable properties for the adsorption of contaminants in wastewaters.

1. Introduction

Graphene is a 2D nanomaterial that consists of sp^2 hybridized carbon atoms arranged in a hexagonal pattern [1]. This ultra-thin material is crystalline and has several remarkable properties. It has an outstandingly high surface area (2630 m^2 g^{-1} for a one-atom-thick flat sheet of graphene) [2], electrical and thermal conductivities, electron transport capabilities, and excellent tensile strength. These exceptional features have created considerable research interest and industrial implementation in areas such as supercapacitors [3], hydrogen production [4], sensors [5], adsorption [6], and support fillers [7] for nanocomposites [8].

Graphene oxide (GO) comprises of a single sheet of graphite oxide and is prepared by the action of strong oxidizers on graphite [1, 9, 10]. GO can easily be exfoliated and dispersed in different solvents due to its hydrophilicity [11, 12]. The hydrophilic nature of GO enables it to be evenly distributed on a substrate in the form of thin films, which is important in the production of electronic devices [13]. The presence of oxygenic functional groups, including carbonyl, carboxylic, epoxide and hydroxyl groups, the high surface area and hydrophilicity make GO flakes an excellent prospective sorbent material [14, 15]. These oxygen functional groups can complex with toxic metal cations through sharing of lone pairs of electrons. A large adsorption capacity is exhibited as a consequence of the large surface area of GO [16-18]. The hydrophilicity of oxidized graphite allows water to be absorbed into the layered structure, which further increases the gap between the layers to 1.15 nm [19], which allows for improved metal binding.

Graphene oxide is synthesised by using Hummer's method in which graphite is oxidized by reacting it with potassium permanganate and sodium nitrate in concentrated sulfuric acid [1]. Numerous reports describe the high adsorption capacity of GO for antibiotics [20, 21], heavy metals [15, 16], and dyes [21, 22]. Graphene oxide has been stated to possess an improved adsorption efficiency than other frequently used sorbent materials as a consequence of the number of O-containing moieties it possesses [14, 16]. Adsorption is the most cost-effective and resourceful procedure for the sequestration of pollutants from contaminated water [22-24]. This water remediation technique has been studied with a wide variety of adsorbent materials, including waste

*Corresponding Author: Bice Susan Martincigh, Email: martinci@ukzn.ac.za

biomass [25], peat moss [26], zeolites [27] and hydrogels [28], to name only a few.

Although many investigations on the preparation of graphene oxide have been published, there are no studies that discuss the preparation of GO that will afford more oxygen functional groups for the elimination of contaminants from wastewater [14, 16, 23]. Hence, the objectives of this work were to synthesize GO via a modified Hummers method; to establish the optimum synthesis conditions that would provide more oxygenic functional groups on GO, and to apply the most functionalised product for the removal of contaminants from wastewater. The products obtained were characterized by several methods which include transmission electron microscopy (TEM), scanning electron microscopy (SEM), powder X-ray diffraction (XRD), thermogravimetric analysis (TGA), Raman spectroscopy, Fourier transform infrared spectroscopy (FTIR), Brunauer-Emmett-Teller (BET) surface area analysis, Boehm titration, and elemental analysis.

2. Experimental

2.1. Materials

Analytical grade chemicals, including graphite powder (<20 µm, synthetic), sodium hydroxide (NaOH), potassium permanganate ($KMnO_4$), 98% sulphuric acid (H_2SO_4), 30% hydrogen peroxide (H_2O_2), 32% hydrochloric acid (HCl), copper metal and nitric acid (HNO_3), were utilized without further purification. They were obtained from Sigma-Aldrich (Pty) Ltd, South Africa. Milli-Q H_2O was from a Millipore Milli-Q Integral water purification system.

2.2. Synthesis of graphene oxide

Different variants of GO were synthesized according to the modified Hummer's method [1] by varying the mass ratio of graphite to sodium nitrate from 2:1, 1:1 and 1:2 respectively. The products obtained were named $GO_{2:1}$, $GO_{1:1}$, and $GO_{1:2}$ respectively. The preparation method was as previously described [29].

2.3. Characterization

The $GO_{2:1}$, $GO_{1:1}$ and $GO_{1:2}$, materials were analyzed to authenticate the homogeneity and structure of the products prepared. The morphology of the materials was investigated by scanning electron microscopy (SEM) (ZEISS Ultra PLUS) and structural information was obtained from transmission electron microscopy (JEOL, JEM 1010). X-ray diffractograms of the materials were measured with a Bruker Advance D8 spectrometer utilizing Cu Kα radiation at 40 mA and 45 kV. Raman spectra of the materials were recorded with a DeltaNu Advantage 532TM® Raman spectrometer with a Nd:YAG laser source and excitation at a wavelength of 532 nm. The thermal stability of the materials was determined in a nitrogen atmosphere with a PerkinElmer Simultaneous Thermal Analyser STA 6000 instrument. The samples were heated from ambient temperature to 1000 °C at a rate of 10 °C/min with a nitrogen flow rate of 50 mL/min. The textural properties of the samples were determined with a Micromeritics Tristar ASAP 3020 surface area and porosity analyser by using nitrogen gas as the adsorbent. The functional groups present were identified by FTIR with a PerkimElmer Spectrum RX1 spectrophotometer, using a pressed disc of each powder mixed with KBr. Elemental analysis of the materials was performed with a Thermo Scientific CHNS/O LECO CHNS-932 elemental analyser. The Boehm titration method was used for quantifying the surface groups on the carbon materials as follows. The number of acidic functional groups on the surfaces of the GO samples were measured with base solutions of increasing strength known as reaction bases such as NaOH, Na_2CO_3, and $NaHCO_3$. To determine the acidic sites of the materials, 0.2 g of the samples were weighed separately into 20 mL aliquots of solutions of 0.1 mol dm^{-3} NaOH, 0.05 mol dm^{-3} Na_2CO_3, and 0.1 mol dm^{-3} $NaHCO_3$ and then agitated in a shaking water bath for 24 h. The solutions were filtered, and 5 mL of the filtrates were back-titrated against a standardized 0.05 mol dm^{-3} HCl solution. It is suggested that sodium hydroxide (NaOH) neutralizes the lactonic, phenolic and carboxylic groups, while sodium carbonate (Na_2CO_3) neutralizes the lactonic and carboxylic groups, and sodium bicarbonate ($NaHCO_3$) neutralizes the carboxylic groups. Thus, by difference, the quantity of each type of functional group can be determined.

2.4. Determination of adsorption capacity

A batch adsorption technique, with Cu^{2+} as the adsorbate, was used to determine the adsorption capacity of the GO materials. For this, a standard Cu^{2+} stock solution was prepared by dissolving 1.0 g of pure Cu metal into 10 mL of concentrated nitric acid (HNO_3). The stock solution was made up to a final volume of 1000 mL with deionized water. This solution was subsequently diluted to prepare working solutions of known concentrations.

A Metrohm 827 pH-meter was used for measuring the pH, and a Mettler AE 200 balance was used for weighing of samples.

The dependence of the adsorption of $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ was determined by varying the adsorbent dose from 10 to 60 mg in contact with 20 mL of 20 mg dm^{-3} of Cu^{2+} solution at a pH of 5. The samples were agitated in a thermostated shaking water bath set at 25 °C for 5 h, and subsequently filtered through Whatman No. 1 filter paper. The filtrates were analysed for their Cu^{2+} content at a wavelength of 220.353 nm with a PerkinElmer Optima 5300 DV inductively coupled plasma-optical emission spectrometer (ICP-OES).

The adsorption efficiency, q_e (in mg g^{-1}), was estimated from (1):

$$q_{eq} = \frac{V}{m} \times \left(C_i - C_{eq}\right) \quad (1)$$

where V is the volume (in dm^3) of the adsorbate solution, m is the mass of adsorbent (in mg), and C_i and C_{eq} are the initial and equilibrium concentration of adsorbate (in mg dm^{-3}) respectively. The percentage sequestration of Cu^{2+} was derived from (2).

$$\%\, adsorbed = \left(\frac{C_i - C_{eq}}{C_i}\right) \times 100 \quad (2)$$

3. Results and discussion

In this work, different variants of GO (2:1, 1:1, 1:2) were synthesised via a modified Hummer's method. The GO samples were characterized by several techniques, and then tested for the adsorption capacity of a representative heavy metal cation contaminant, namely Cu^{2+}, in aqueous solution.

3.1. Morphology

The surface morphologies of the different GO samples were measured by SEM techniques. The SEM micrographs of $GO_{2:1}$, $GO_{1:1}$ and $GO_{1:2}$ are shown in Figure 1. In order to acquire these images, the GO samples were coated with a gold film fractured surface under liquid nitrogen. The surface morphology showed how the GO sheets are exfoliated for the different GO materials. The SEM micrographs show that the GO materials have a 2D sheet-like structure with a multiple lamellar layer structure that distinguishes the edges of the individual sheets. The GO sheets were thick, particularly at the edges, and this thickness arises from the O-functional groups introduced.

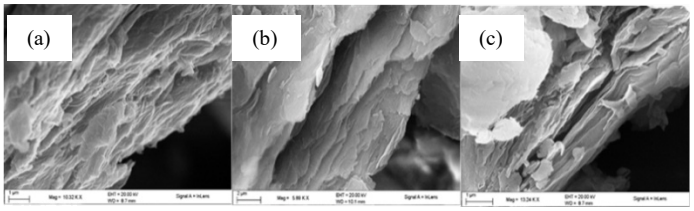


Figure 1: SEM images of GO samples: (a) $GO_{2:1}$, scale bar = 1 µm, (b) $GO_{1:1}$, scale bar = 1 µm, and (c) $GO_{1:2}$, scale bar = 1 µm

The TEM micrographs of the GO materials (Figure 2) revealed the anticipated layered structure of GO with a large sheet on top of the grids. The sheet resembles a transparent wrinkled, folded and silk-like veil due to the oxidation of graphite. This characteristic has been reported for the synthesis of GO by the modified Hummer's method regardless of the ratios of graphite to sodium nitrate [30]. The exfoliation of the graphite layer and introduction of defects on the GO layer is due to the introduction of oxygenic functional groups. From the TEM images, the level of transparency in the variants of GO might be as a result of the different ratios of graphite to sodium nitrate resulting in the intercalation of GO intermediates.

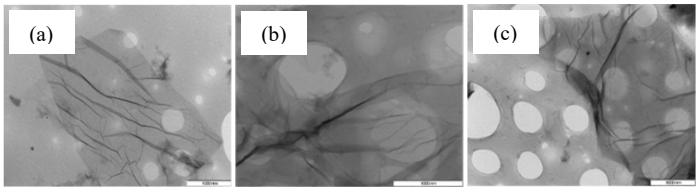


Figure 2: TEM images of GO samples: (a) $GO_{2:1}$, scale bar = 1000 nm (b) $GO_{1:1}$, scale bar = 1000 nm, and (c) $GO_{1:2}$, scale bar = 1000 nm

3.2. Crystallinity

The structural properties of the GO materials were examined by means of powder XRD. The diffractograms for $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ are shown in Figure 3. There is a sharp diffraction peak for graphite at $2\theta = 26.5°$ corresponding to a d-spacing of 0.34 nm of the graphite layers [31]. The characteristics of graphite at 42° disappeared after the oxidation process and a notable strong peak, for $GO_{2:1}$ at $2\theta = 10.49°$, $GO_{1:1}$ at $2\theta = 10.47°$, and $GO_{1:2}$ at $2\theta = 10.48°$, emerged. The increase in the interlayer spacing of GO to that of graphite arises from the introduction of O-containing functional groups, such as epoxide, hydroxyl, carboxylic and carbonyl groups, during the oxidative treatment [32]. The van der Waals interactions operative between the graphene sheets are weakened by the O-functional groups of GO and this increases the distance between the sheets [33-35]. In the diffractogram for $GO_{2:1}$ a strong peak at $2\theta = 43°$ is seen after oxidation indicating a random packing of the sheets [36]. The XRD results confirmed that the oxygenic functional groups were intercalated into the graphite interlayers.

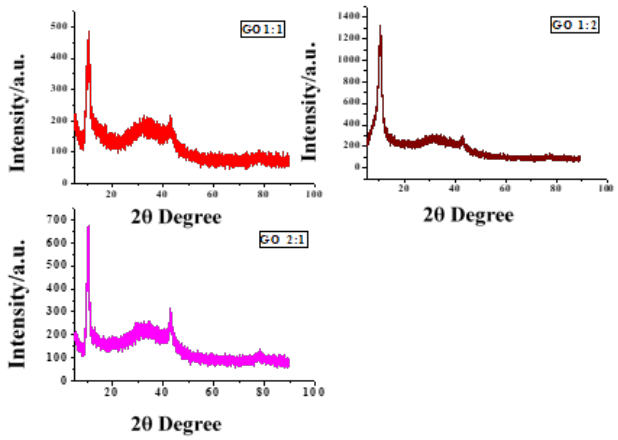


Figure 3: Powder X-ray diffractograms for GO samples

Selected area electron diffraction (SAED) allows for a diffraction pattern to be obtained with a transmission electron microscope [37]. It can be used to find the d-spacing of the crystal planes and also to obtain information about the crystallinity of the samples. The brighter the spots, the more crystalline is the sample. The reciprocal of the radius of the SAED ring is the d-spacing of that ring, and the d-spacings for $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ were determined as 0.2046, 0.1997 and 0.2131 nm respectively (Figure 4). From the SAED images, $GO_{2:1}$ was found to be the most crystalline of the three samples.

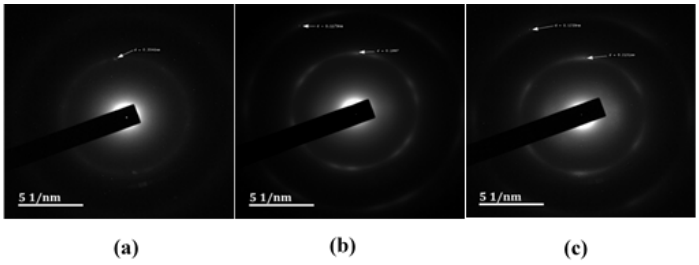


Figure 4: SAED patterns for (a) $GO_{1:1}$, (b) $GO_{1:2}$ and (c) $GO_{2:1}$

3.3. Graphitic and defective nature

Raman spectroscopy is an essential method for determining the defects, thermal conductivity, and strain in graphene and related materials [38]. The main aspects of the Raman spectra of graphitic C-based materials include the D and G bands, and their overtones. In the Raman spectra (Figure 5), the D bands of $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ have very strong peaks at 1353, 1350 and 1347 cm^{-1} respectively. The D-peaks are associated with the vibration of sp^3 carbon atoms of defects that are randomly arranged on the GO sheets. The D-peak intensity together with a large peak width indicates significant structural disorder in the GO samples. The wavenumbers at 1595, 1593 and 1596 cm^{-1}, respectively, represent the G bands, which are related to the vibration of sp^2 C-atoms in the graphitic 2D hexagonal lattice. This gives rise to in-plane

optical vibrations that relate to the E_{2g} phonons from the bond stretching of the sp^2 carbon atoms in a graphitic 2D hexagonal lattice. The ratio of the areas under the D- and G-bands (I_D/I_G) is used to measure the graphitic nature of the sample. For these samples, the order of the I_D/I_G ratio is $GO_{2:1}$ (1.02) > $GO_{1:2}$ (0.97) > $GO_{1:1}$ (0.93). This result shows that $GO_{2:1}$ shows a greater defective density.

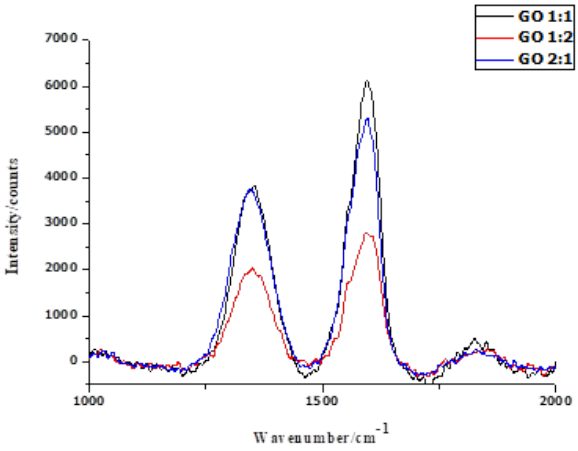


Figure 5: Raman spectra for GO samples

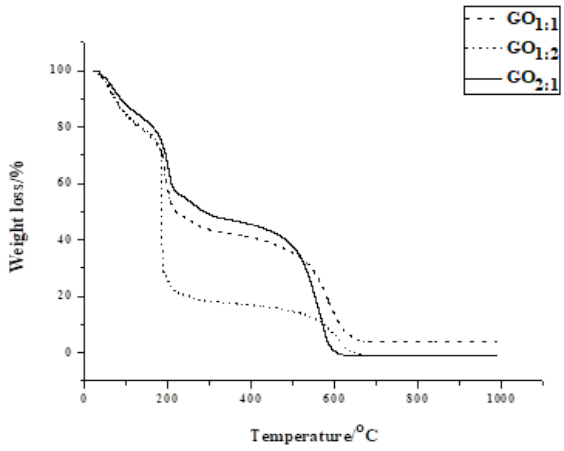


Figure 6: Thermograms of graphene oxide samples recorded in a N_2 atmosphere

3.4. Thermal stability

Thermal analysis was performed to assess the thermal stability of the $GO_{1:1}$, $GO_{1:2}$, and $GO_{2:1}$ sheets, since oxygen functional groups present in GO decompose on heating. Figures 6 and 7 depict the thermograms and derivative thermograms of the GO samples respectively. The first mass loss, that takes place at a temperature of 100 °C, is primarily due to the evaporation of adsorbed water that has interacted with the O-containing functional groups on GO [39, 40]. According to Dreyer *et al.* [41], the interaction of water with GO is by hydrogen bonding. The second mass loss arises from the pyrolysis of unstable O-containing functional groups such as carboxylic, hydroxyl and carbonyl groups to give CO_2, CO and H_2O [42, 43]. In $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ this loss occurred at a temperature of 202, 212 and 216 °C respectively. The final mass loss at 614, 617 and 587 °C for $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ respectively, shows that there is a reduction in the thermal stability of graphene oxide relative to graphite [44].

The derivative thermograms for the GO samples shown in Figure 7, revealed a small exothermic peak at 614, 617, and 587 °C. At 202, 212, and 216 °C, a larger exothermic peak is seen, and this is due to the loss of CO_2 and CO from the GO samples as a result of the decomposition of O-containing functional groups.

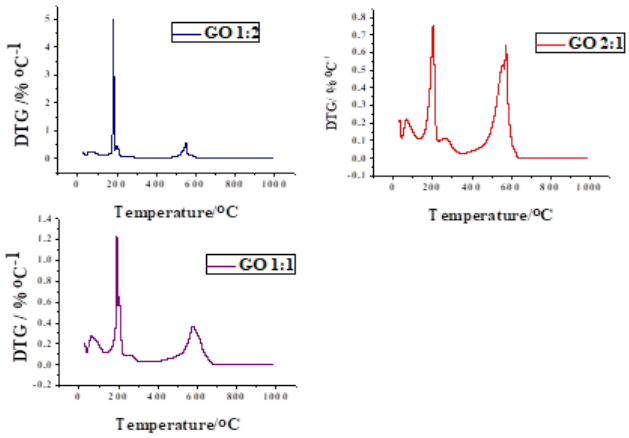


Figure 7: Derivative thermograms of graphene oxide samples

3.5. Textural properties

The BET method was used to determine the specific surface areas of the GO materials with N_2 as the adsorbed gas. In the first step, the materials were degassed at 120 K for 24 h. The results are shown in Table 1. The pores can be grouped into micro-, meso- or macro-pores, based on diameter. The results obtained for $GO_{2:1}$ (3.841 nm), $GO_{1:1}$ (7.350 nm) and $GO_{1:2}$ (19.861 nm) correspond to mesopores. The largest BET surface area was displayed by $GO_{2:1}$ (179.25 $m^2\ g^{-1}$), followed by $GO_{1:1}$ (82.68 $m^2\ g^{-1}$) and, lastly, by $GO_{1:2}$ (58.35 $m^2\ g^{-1}$).

Table 1: Physical properties of the $GO_{2:1}$, $GO_{1:1}$ and $GO_{1:2}$ materials

Sample	Surface area/$m^2\ g^{-1}$	Pore volume/$cm^3\ g^{-1}$	Pore diameter/nm
$GO_{2:1}$	179.25	0.051	3.841
$GO_{1:1}$	82.68	0.054	7.350
$GO_{1:2}$	58.35	0.122	19.861

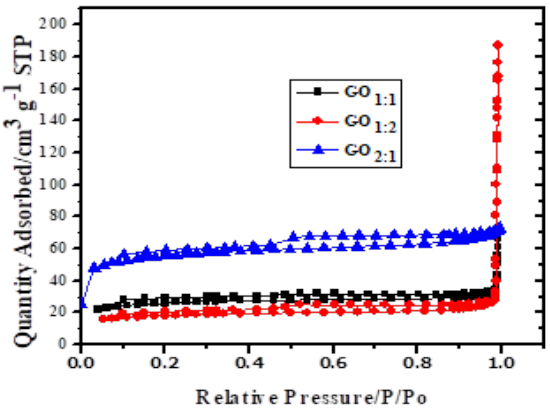


Figure 8: BET nitrogen adsorption-desorption isotherms

The N_2 adsorption-desorption isotherms for the GO samples are depicted in Figure 8. $GO_{1:1}$ and $GO_{1:2}$ exhibited type IV isotherms and type H3 hysteresis loops, signifying that the nanomaterials are mesoporous, as determined from the pore diameters. $GO_{2:1}$ also exhibited a type IV isotherm but has a H4 type of hysteresis loop. The type IV isotherm represents an adsorption isotherm with hysteresis related with capillary condensation in mesopores [45, 46]. The Barrett-Joyner-Halenda (BHJ) pore volumes of the samples are $GO_{1:1}$ (0.054 $cm^3\ g^{-1}$), $GO_{1:2}$ (0.122 $cm^3\ g^{-1}$) and $GO_{2:1}$ (0.051 $cm^3\ g^{-1}$) (Figure 9).

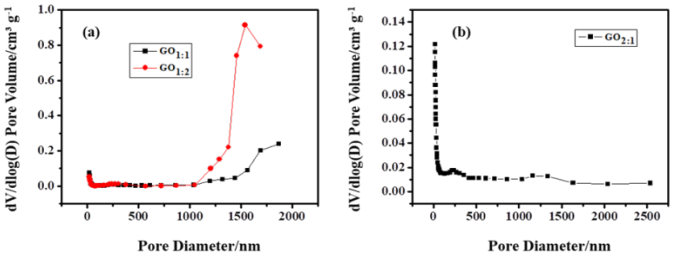


Figure 9: The BHJ pore volumes of (a) $GO_{1:1}$, $GO_{1:2}$ and (b) $GO_{2:1}$

3.6. Nature of oxygen functional groups

FTIR spectrophotometric characterisation was conducted to explore the nature of the functional groups on the GO products and to confirm the formation of GO by chemical treatment of graphite (Figure 10). From the spectra of the GO samples, it is clear that large amounts of O-containing functional groups are present on the surfaces of the samples. The broad peak at 3430 cm^{-1} can be attributed to the free and associated O-H stretch arising from the presence of adsorbed water molecules and structural hydroxyl groups (–COH and –COOH) of GO [47]. The peak at 1728 cm^{-1} is assigned to the C=O groups. The peaks at 1626 cm^{-1} can be assigned to C=C groups that are due to the structure of the graphitic layer retained previously and after the oxidation process [48]. The peaks at 1220 and 1092 cm^{-1} are related to the C-O groups. From the results obtained, it is apparent that $GO_{2:1}$ shows the strongest and sharpest peaks in the C-O region, but these peaks are absent in the spectrum of $GO_{1:1}$ at 1220 cm^{-1}. The presence of C-O and C=O groups after the oxidation process confirms the successful functionalization of the samples, and the conversion of graphite into GO. This is in agreement with previous reports [49, 50]. Additionally, peaks at 1728 cm^{-1} due to the C=O stretch (COOH of carboxyl or carbonyl groups), at 1626 cm^{-1} due to the C=C stretch, and at 1092 cm^{-1} due to the C-O groups, appear in all the GO samples. Hence, there is evidence that GO was successfully synthesized, and the functional groups present on the surfaces of the GO sheets can furnish abundant adsorption sites, to increase the adsorption capacity for both organic and inorganic pollutants.

3.7. Elemental analysis

Elemental analysis was undertaken to ascertain the C, H and O elemental composition of the samples. From the results, it is apparent that $GO_{2:1}$ has the largest oxygen content of the three products (see Table 2). The oxygen content of $GO_{2:1}$ was 40.01%. Also, the C/O ratio for $GO_{2:1}$ is the smallest, also indicating the greater oxygen content. The H present in the GO products mainly arises from the presence of O-containing functional groups. The S is a residual impurity from H_2SO_4 used in the synthesis of the GO samples.

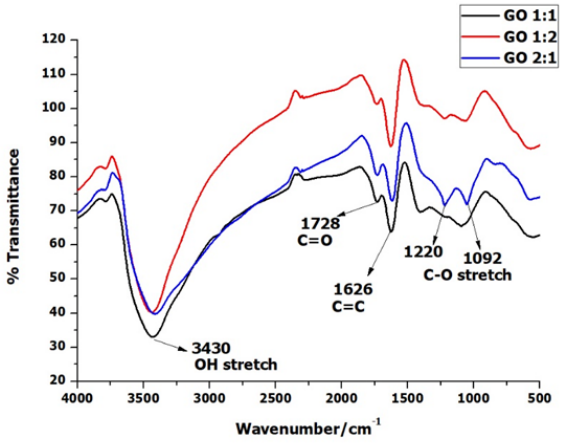


Figure 10: FTIR spectra of GO samples

Table 2: Elemental composition of GO samples

Material	C/%	H/%	S/%	O/%
$GO_{1:1}$	40.83	2.46	1.89	31.71
$GO_{1:2}$	42.35	2.59	1.97	32.02
$GO_{2:1}$	45.86	2.04	0.94	40.01

Table 3: Concentrations of acidic functional groups on the surfaces of $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ as determined by the Boehm titration method

Sample	Base solution	Concentration neutralized/ mmol g^{-1}	Functional group	Concentration/ mmol g^{-1}
$GO_{1:1}$	NaOH	2.42	Phenolic	0.43
	Na_2CO_3	1.99	Lactonic	0.40
	$NaHCO_3$	1.59	Carboxylic	1.59
$GO_{1:2}$	NaOH	3.15	Phenolic	1.26
	Na_2CO_3	1.89	Lactonic	0.40
	$NaHCO_3$	1.49	Carboxylic	1.49
$GO_{2:1}$	NaOH	2.42	Phenolic	0.43
	Na_2CO_3	1.99	Lactonic	0.40
	$NaHCO_3$	1.59	Carboxylic	1.59

3.8. Quantity of acidic oxygen functional groups

The surface properties of the different GO materials were examined by using the Boehm titration method. The acidic properties of the GO variants are determined by the carboxylic, lactonic and phenolic functional groups (Table 3). These functional groups vary in their acidities and can be differentiated by neutralization with different basic solutions, namely, $NaHCO_3$, Na_2CO_3 and NaOH. The concentration of strong acid groups on $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ was determined from the titre values of $NaHCO_3$, and that of strong and weak acidic functional groups was from that of Na_2CO_3 which gave 1.99, 1.89 and 1.99 mmol g^{-1} respectively. The concentration of weak acids was determined by the difference in the two titration values. The concentrations of the carboxylic acid groups were found to be 1.59, 1.49 and 1.59 mmol g^{-1} respectively. The overall concentration of the phenolic, and the strong and weak acidic functional groups was found from the

amount of NaOH neutralized. The total sum of the acidic functional groups on the $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ surface was found to be 2.42, 3.15 and 2.42 mmol g^{-1} respectively. The phenolic functional group concentrations were determined from the difference in the three titration methods. From the results obtained, $GO_{1:1}$ and $GO_{2:1}$ are similar, whereas $GO_{1:2}$ contains significantly more phenolic groups than the other two samples. Therefore, the acidic surface functional groups of the samples are suitable for the uptake of cationic pollutants from aqueous solution. For each sample, the order of the number of acidic groups is carboxylic > phenolic > lactonic.

3.9. Adsorption properties

The effect of the adsorbent dose on the adsorption of Cu^{2+} from aqueous solution was determined by varying the mass of the GO samples from 10 to 60 mg at a constant temperature of 25 °C for 5 h. The results, shown in Figure 11, indicate that as the sorbent dose increases, the percentage of Cu^{2+} ions adsorbed onto the GO samples increases. An increase in the adsorbent dose with a higher percentage of adsorption arises from the increase in surface area and the availability of more binding sites for adsorption. The sorption of Cu^{2+} onto $GO_{1:1}$ was initially low compared with the other materials; this is possibly due to its surface area and the availability of active sites on the adsorbent surface. The other two adsorbents showed similar behaviour. Overall, the removal efficiency of the three adsorbents was high, but $GO_{2:1}$ showed the highest adsorption with a percentage removal of 59.7%.

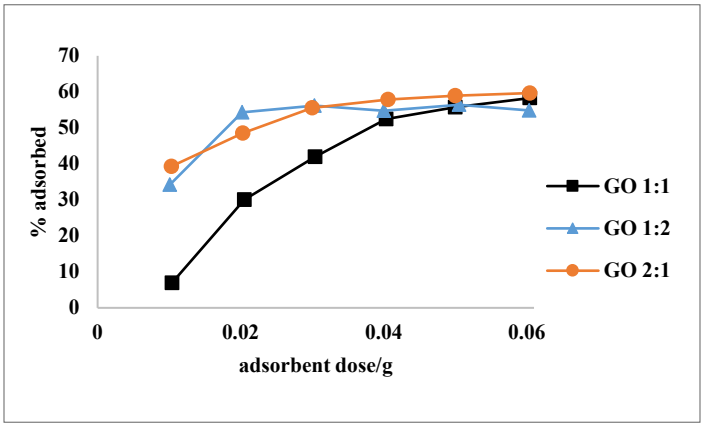


Figure 11: Effect of adsorbent dose on the adsorption of Cu^{2+} onto $GO_{1:1}$, $GO_{1:2}$ and $GO_{2:1}$ [conditions: 20 mL of 20 mg dm^{-3} Cu^{2+}, pH 5.0, agitation speed of 150 rpm, equilibration time 5 h, temperature 25 °C]

4. Conclusion

In this paper, GO was effectively prepared by making use of a modified Hummer's method. In the synthesis, mass ratios of 1:1, 1:2 and 2:1 of graphite powder and $NaNO_3$ were used. This was done in order to determine if it would alter the number of oxygenic functional groups on GO. The elemental composition of the $GO_{2:1}$ material, synthesized with a 2:1 mass ratio of graphite powder to $NaNO_3$, showed the greatest oxygen content. As indicated by SAED analysis, the $GO_{2:1}$ sample was also more crystalline than the other two products. It also exhibited the largest surface area. However, the Boehm titration showed that of the three samples, $GO_{2:1}$ did not contain the greatest number of acidic O-containing functional groups. Oxygen-containing moieties improve the hydrophilicity of GO so that it can form stable suspensions in aqueous media. The adsorption study revealed that the uptake of Cu^{2+} by $GO_{2:1}$ was greater than for the other two GO samples. Thus, the synthesis of graphene oxide with a smaller quantity of $NaNO_3$ offered a material with improved crystallinity, a larger surface area and a greater oxygen content, that showed enhanced adsorption properties. Therefore, this material was chosen for further use in the study of the uptake of both metal ions and dyes in aqueous solution.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This research was financially supported by the National Research Foundation of South Africa (NRF) and the University of KwaZulu-Natal Nanotechnology Platform. O.O.O. is grateful for the award of a bursary by NRF-TWAS.

References

[1] W.S Hummers Jr, R.E Offeman, Preparation of graphitic oxide, J. Am. Chem. Soc., **80**, 1339-1339, 1958. https//:doi.org/10.1021/ja01539a017

[2] K. Novoselov, A.K. Geim, S. Morozov, D. Jiang, M. Katsnelson, I. Grigorieva, S. Dubonos, A. Firsov, Two-dimensional gas of massless Dirac fermions in graphene, Nat., **438**, 197-200, 2005. https//:doi.org/10.1038/nature04233

[3] V. Strauss, K. Marsh, M. Kowal, M. El-Kady, R. Kaner, A simple route to porous graphene from carbon nanodots for supercapacitor applications, Adv. Mater., **30**, 1-10, 2018. https:doi.org/ 10.1002/adma.201704449

[4] P. Cheng, Z. Yang, H. Wang, W. Cheng, M. Chen, W. Shangguan, G. Ding, TiO_2–graphene nanocomposites for photocatalytic hydrogen production from splitting water, Int. J. of Hydrogen Energy, **37**, 2224-2230, 2012. https//:doi.org/10.1111/j.1750-3841.2012.02868.x

[5] A. Nag, A. Mitra, S. Mukhopadhyay, Graphene and its sensor-based applications: A review, Sens. and Actuators A: Physical, **270**, 177-194, 2018. https//:doi.org/10.1016/j.sna.2017.12.028

[6] Z. Li, F. Chen, L. Yuan, Y. Liu, Y. Zhao, Z. Chai, W. Shi, Uranium(VI) adsorption on graphene oxide nanosheets from aqueous solutions, Chem. Eng. J., **210**, 539-546, 2012. https//:doi.org/10.1016/j.cej.2012.09.030

[7] P. Bhawal, S. Ganguly, T. Chaki, N. Das, Synthesis and characterization of graphene oxide filled ethylene methyl acrylate hybrid nanocomposites, RSC Adv., **6**, 781-790, 2016. https//:doi.org/10.1039/C5RA24914G

[8] C. Soldano, A. Mahmood, E. Dujardin, Production, properties and potential of graphene, Carbon, **48**, 2127-2150, 2010. https//:doi.org/10.1016/j.carbon.2010.01.058

[9] V. Chandra, J. Park, Y. Chun, J.W. Lee, I.-C. Hwang, K.S. Kim, Water-dispersible magnetite-reduced graphene oxide composites for arsenic removal, ACS Nano, **4**, 3979-3986, 2010. https://doi.org/10.1021/nn1008897

[10] G. Shao, Y. Lu, F. Wu, C. Yang, F. Zheng, Q. Wu, Graphene oxide: the mechanisms of oxidation and exfoliation, J. of Mater. Sci., **47**, 4400-4409, 2012. https://10.1007/s10853-012-6294-5

[11] R. Ruoff, Graphene: Calling all chemists, Nat. Nanotechnol., **3**, 10-11, 2008. https//:doi.org/10.1038/nnano.2007.432

[12] H. Jiang, Chemical preparation of graphene-based nanomaterials and their applications in chemical and biological sensors, Small, **7**, 2413-2427, 2011. https://doi.org/10.1002/smll.201002352

[13] T.N. Ly, S. Park, Highly sensitive gas sensor using hierarchically self-assembled thin films of graphene oxide and gold nanoparticles, J. of Ind. Eng. Chem., **67**, 417-428, 2018. https//:doi.org/10.1016/jiec.2018.07.016

[14] E.T. Mombeshora, P.G. Ndungu, V.O. Nyamori, Effect of graphite/sodium nitrate ratio and reaction time on the physicochemical properties of graphene oxide, New Carbon Mater., **32**, 174-187, 2017. https://doi.org/10.1016/S1872-5805(17)60114-8

[15] M. Ahmed, A. Giwa, S. Hasan, Challenges and opportunities of graphene-based materials in current desalination and water purification technologies, Nanoscale Materials in Water Purification 735-758, 2019. https//:doi.org/10.1016/B978-0-12-813926-4.00033-1

[16] R. Sitko, E. Turek, B. Zawisza, E. Malicka, E. Talik, J. Heimann, A. Gagor, B. Feist, R. Wrzalik, Adsorption of divalent metal ions from aqueous solutions

using graphene oxide, Dalton Trans., **42**, 5682-5689, 2013. https//:doi.org/10.1039/c3dt33097d
[17] M.A. Atieh, O.Y. Bakather, B.S. Tawabini, A.A. Bukhari, M. Khaled, M. Alharthi, M. Fettouhi, F.A. Abuilaiwi, Removal of chromium(III) from water by using modified and nonmodified carbon nanotubes, J. Nanomater., **2010**, 232378, 2010. https://doi.org/10.1155/2010/232378
[18] M. Machida, T. Mochimaru, H. Tatsumoto, Lead(II) adsorption onto the graphene layer of carbonaceous materials in aqueous solution, Carbon, **44**, 2681-2688, 2006. https://doi.org/10.1016/j.carbon.2006.04.003
[19] A. Lerf, A. Buchsteiner, J. Pieper, S. Schöttl, I. Dekany, T. Szabo, H. Boehm, Hydration behavior and dynamics of water molecules in graphite oxide, J. Phys Chem. Solids, **67**, 1106-1110, 2006. https://doi.org/10.1016/j.jpcs.2006.01.031
[20] Y. Gao, Y. Li, L. Zhang, H. Huang, J. Hu, S.M. Shah, X. Su, Adsorption and removal of tetracycline antibiotics from aqueous solution by graphene oxide, J. of Colloid Interface Sci., **368**, 540-546, 2012. https://doi.org/10.1016/j.jcis.2011.11.015
[21] A.C. Sophia, E.C. Lima, N. Allaudeen, S. Rajan, Application of graphene-based materials for adsorption of pharmaceutical traces from water and wastewater- a review, Desalin. Water Treat., **57**, 27573-27586, 2016. https://doi.org/10.1080/19443994.2016.1172989
[22] H. Yan, X. Tao, Z. Yang, K. Li, H. Yang, A. Li, R. Cheng, Effects of the oxidation degree of graphene oxide on the adsorption of methylene blue, J. Hazard. Mater., **268**, 191-198, 2014. https://doi.org/10.1016/j.jhazmat.2014.01.015
[23] B. Kakavandi, R.R. Kalantary, M. Farzadkia, A.H. Mahvi, A. Esrafili, A. Azari, A.R. Yari, A.B. Javid, Enhanced chromium(VI) removal using activated carbon modified by zero valent iron and silver bimetallic nanoparticles, J. Enviro. Health Sci. and Eng., **12**, 1-15, 2014. https://doi.org/10.1186/s40201-014-0115-5
[24] P. Sharma, N. Hussain, D.J. Borah, M.R. Das, Kinetics and adsorption behavior of the methyl blue at the graphene oxide/reduced graphene oxide nanosheet–water interface: a comparative study, J. Chem. Eng. Data, **58**, 3477-3488, 2013. https://doi.org/10.1021/je400743r
[25] A.D. Schiller, A.C. Goncalves, A.D. Braccini, D. Schwantes, M.A. Campagnolo, E. Conradi, J. Zimmermann, Potential of agricultural and agroindustrial wastes as adsorbent materials of toxic heavy metals: a review, Desalination Water Treat., **187**, 203-218, 2020. https//:doi.org/10.5004/dwt.2020.25094
[26] T. Gosset, J.-L. Trancart, D.R. Thévenot, Batch metal removal by peat. Kinetics and thermodynamics, Water Res., **20**, 21-26, 1986. https://doi.org/10.1016/0043-1354(86)90209-5
[27] M.V. Mier, R.L. Callejas, R. Gehr, B.E.J. Cisneros, P.J. Alvarez, Heavy metal removal with Mexican clinoptilolite: multi-component ionic exchange, Water Res., **35**, 373-378, 2001. https://doi.org/10.1016/S0043-1354(00)00270-0
[28] C.B. Godiya, S.M. Sayed, Y. Xiao, X. Lu, Highly porous egg white/polyethyleneimine hydrogel for rapid removal of heavy metal ions and catalysis in wastewater, React. Funct. Polym., **149**, 104509, 2020, https://doi.org/10.1016/j.reactfunctpolym.2020.104509
[29] O.O. Ojo, S. B. Jonnalagadda and B. S. Martincigh, Synthesis of graphene oxide under differing conditions and its characterization, 2018 IEEE 8th International Conference on Nanomaterials: Application and Properties (NAP-2018), Zatoka, Ukraine, pp. 15-18
[30] S. Wang, P.J. Chia, L.L. Chua, L.H. Zhao, R.Q. Png, S. Sivaramakrishnan, M. Zhou, R.G.S. Goh, R.H. Friend, A.T.S. Wee, Band-like transport in surface-functionalized highly solution-processable graphene nanosheets, Adv. Mater., **20**, 3440-3446, 2008. https://doi.org/10.1002/adma.200800279
[31] G. Titelman, V. Gelman, S. Bron, R. Khalfin, Y. Cohen, H. Bianco-Peled, Characteristics and microstructure of aqueous colloidal dispersions of graphite oxide, Carbon, **43**, 641-649, 2005. https//:doi.org10.1016/j.carbon.2004.10.035
[32] M. Wojtoniszak, X. Chen, R.J. Kalenczuk, A. Wajda, J. Łapczuk, M. Kurzewski, M. Drozdzik, P.K. Chu, E. Borowiak-Palen, Synthesis, dispersion, and cytocompatibility of graphene oxide and reduced graphene oxide, Colloids and Surf., B: Biointerfaces, **89**, 79-85, 2012. https://doi.org/10.1016/j.colsurfb.2011.08.026
[33] J. Chen, Y. Li, L. Huang, C. Li, G. Shi, High-yield preparation of graphene oxide from small graphite flakes via an improved Hummers method with a simple purification process, Carbon, **81**, 826-834, 2015. https//:doi.org/10.1016/j.carbon.2014.10.033
[34] C. Wang, Z. Liu, S. Wang, Y. Zhang, Preparation and properties of octadecylamine modified graphene oxide/styrene-butadiene rubber composites through an improved melt compounding method, J. Appl. Polym. Sci., **133**, 1-9, 2016. https://doi.org/10.1002/app.42907
[35] W. Gao, Graphene oxide: reduction recipes, spectroscopy, and applications, Springer, 2015. https://doi.org/10.1007/978-3-319-15500-5
[36] Z. Li, W. Zhang, Y. Luo, J. Yang, J.G. Hou, How graphene is cut upon oxidation?, J. Am. Chem. Soc., **131**, 6320-6321, 2009. https://doi.org/10.1021/ja8094729
[37] J.-M. Zuo, M. Gao, J. Tao, B.Q. Li, R. Twesten, I. Petrov, Coherent nano-area electron diffraction, Microsc. Res. Tech., **64**, 347-355, 2004. https//:doi.org/10.1002/jemt.20096
[38] L. Malard, M. Pimenta, G. Dresselhaus, M. Dresselhaus, Raman spectroscopy in graphene, Phys. Rep., **473**, 51-87, 2009. doi.org/10.1016/j.physrep.2009.02.003
[39] X. Du, P. Guo, H. Song, X. Chen, Graphene nanosheets as electrode material for electric double-layer capacitors, Electrochim. Acta, **55**, 4812-4819, 2010. https://doi.org/10.1016/j.electacta.2010.03.047
[40] S.A. El-Khodary, G.M. El-Enany, M. El-Okr, M. Ibrahim, Preparation and characterization of microwave reduced graphite oxide for high-performance supercapacitors, Electrochim. Acta, **150**, 269-278, 2014. https://doi.org/10.1016/j.electacta.2014.10.134
[41] D.R. Dreyer, S. Park, C.W. Bielawski, R.S. Ruoff, The chemistry of graphene oxide, Chem. Soc. Rev., **39**, 228-240, 2010. https//:doi.org/ 10.1002/ange.201002160
[42] V.B. Mohan, R. Brown, K. Jayaraman, D. Bhattacharyya, Characterisation of reduced graphene oxide: Effects of reduction variables on electrical conductivity, Mater. Sci. Eng. B, **193**, 49-60, 2015. https://doi.org/10.1016/j.mseb.2014.11.002
[43] H.A. Becerril, J. Mao, Z. Liu, R.M. Stoltenberg, Z. Bao, Y. Chen, Evaluation of solution-processed reduced graphene oxide films as transparent conductors, ACS Nano, **2**, 463-470, 2008. https://doi.org/10.1021/nn700375n
[44] C. Zhu, S. Guo, Y. Fang, S. Dong, Reducing sugar: new functional molecules for the green synthesis of graphene nanosheets, ACS Nano, **4**, 2429-2437, 2010. https//:doi.org/10.1002/adma.201001068
[45] P. Ball, R. Evans, Temperature dependence of gas adsorption on a mesoporous solid: capillary criticality and hysteresis, Langmuir, **5**, 714-723, 1989. https//:doi.org/10.1021/jp111263p
[46] K. A. Cychosz, M. Thommes, Progress in the physisorption characterization of nanoporous gas storage materials, Eng., **4**, 559–566, 2018. https://doi.org/10.1016/j.eng.2018.06.001
[47] H.K. Koolivand, A. Sharif, M.R. Kashani, M. Karimi, M.K. Salooki, M.A. Semsarzadeh, Functionalized graphene oxide/polyimide nanocomposites as highly CO_2-selective membranes, J. of Polym. Res., **21**, 599, 2014. https://doi.org/ 10.1007/s10965-014-0599-9
[48] Z. Pei, L. Li, L. Sun, S. Zhang, X. Shan, S. Yang, B. Wen, Adsorption characteristics of 1,2,4-trichlorobenzene, 2,4,6-trichlorophenol, 2-naphthol and naphthalene on grapheme and graphene oxide, Carbon, **51**, 156 –163, 2013. https://doi.org/10.1016/j.carbon.2012.08.024
[49] J. Shen, M. Shi, B. Yan, H. Ma, N. Li, M. Ye, Ionic liquid-assisted one-step hydrothermal synthesis of TiO_2-reduced graphene oxide composites, Nano Res., **4**, 795-800, 2011. doi: 10.1007/s12274-011-0136-7
[50] K. Zhou, Y. Zhu, X. Yang, X. Jiang, C. Li, Preparation of graphene–TiO_2 composites with enhanced photocatalytic activity, New J. Chem., **35**, 353-359, 2011. https//:doi.org/10.1039/CONj00623H

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 11-15 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Case Study to Determine the Causes of Fire in Agriculture

Marianna Tomašková[*,1], Darina Matisková[2], Michaela Balážiková[1]

[1]*Department of Production Safety and Quality, Technical University of Kosice, Košice, 040 01, Slovakia*

[2]*Department of Industrial Engineering and Informatics, Faculty of Production Technologies in Prešov, Košice, 040 01 Slovakia*

ARTICLE INFO

Article history:
Received: 18 June, 2020
Accepted: 11 August, 2020
Online: 08 September, 2020

Keywords:
Fire analysis in Agriculture
Machine risks
Man - machine - environment system

ABSTRACT

The aim of this paper is to identify a critical link in the man - machine - environment system in the case of an adverse event, such as a hay baling fire, based on a comprehensive risk assessment method. The rate of spread of fires in agriculture depends on meteorological conditions, with large areas affected and potentially endangering the surrounding buildings, facilities. Access to fires is difficult and can extend to forests. Water is often lacking at the scene of a fire, which should be extinguished, and water sources are usually located over long distances. The paper addressed a specific example using a comprehensive method. The process of risk assessment in the work process was determined by the following steps: assessment of the overall risk of the work equipment, assessment of environmental impact, assessment of the person's ability to manage risk, calculation of the resulting risk value, comparison of calculated risk value and acceptability of risk value, proposal of measures. The result of the analysis was the finding that the primary cause of the fire is the environment, i.e. high ambient temperature. The critical element in hay baling work system is the work environment. The risk ratio was estimated at 5.78. The level of risk was low due to the rapid intervention of the human factor. Based on the results, the technical measures mentioned in the paper were proposed to the operator. The paper found that maintenance of the machine is important for protection against agricultural fires, where the human factor plays an important role in the man - machine - environment system.

1. Introduction

Agriculture is one of the most dangerous sectors in terms of accidents at work. There is an accident rate for employees in agriculture without fatalities 1.7 times higher than the average and accident rate with fatalities is three times higher than the average [1].

Fires pose a risk of destroying the environment and human lives [2]. The causes of fire in agricultural machinery are various. The literature [3] lists several possibilities in which a fire may occur: e.g. in the engine, in the bearings, brakes. Maintenance of these machines is an important fire protection.

Maintenance activities in agriculture are very diverse, including the following activities:

- maintenance and repair of machinery, equipment and vehicles,
- maintenance of farmyards and buildings,
- maintenance of silos, tanks, manure tanks and grain tanks,
- maintenance of electrical installations,
- maintenance of drainage and irrigation systems,
- maintenance of paved and unpaved roads.

The following hazards can be identified during these maintenance activities:

- mechanical hazards when working with maintenance machinery, for example crushing, connecting and high-pressure injection means,
- electrical risks,
- thermal risks,
- chemical risks associated with the use of hazardous substances during maintenance or maintenance of equipment containing hazardous substances,
- risk of explosion or fire or during maintenance,
- biological risks in the maintenance of contaminated equipment,
- ergonomic risks, e.g. incorrect construction of tools,
- work activities in enclosed spaces.

[*]Corresponding Author: Marianna Tomašková, marianna.tomaskova@tuke.sk

Wheeled machines, mostly agricultural, are responsible for up to 50% of fatal accidents worldwide. Work safety, work activities and these machines are very important.

For the detection of fires in agriculture, there are almost no studies aimed at identifying the causes in harvesters, respectively. tractors with accessories. Studies to date [4] in which the authors investigated more than 4,000 fires in combines and tractors. They gained access to machines to conduct a detailed study for 265 cases. The authors concluded that 74% of fires occurred in the engine area (e.g. surface heating, flue gas outlet and electrical components). They found that the material that began to burn was crop residues, followed by fuel and oil residues [5].

The tractor is used as a means of accessing hard-to-reach areas, driving and towing connected equipment. For carrying out various agricultural activities, such as harvesting, plowing, fertilizing. Although there are innovative machines and processing methods, there are still risks, such as machine fires, due to low investment in maintenance or machine replacement [6].

To reduce the occurrence of fire, it is possible to apply various systems, e.g. FOGMAKER TM system, which is highly innovative but not available for small businesses. The contribution of the presented paper is to evaluate the risk of hay baling by a complex method in which the bearing was damaged. The press is one of the most risky machines in agriculture. Fires and malfunctions are caused by bearing failure or overheated machine parts.

2. Material and Methods

2.1. Fire safety measures for machines from the Baler company

During harvesting season, fires on agricultural machinery or buildings occur more frequently than in the off-season.

The task of the balers is to continuously pick up the wilted or more often dry material (hay, straw) from the summarized rows, press it and tie it into bales [1,6].

In most cases, presses with variable chambers pose the greatest risk of fire. The fault is most often caused by tensioning and control rollers, which are responsible for the operation of packaging belts [2-7].

Bearings called seasonal bearings used in such machines must be replaced each year before the start of the summer season. Overheated bearings can come into contact with straw, hay and the right hot surface can initiate a negative phenomenon. Before starting work with the press in the field, it is recommended to switch on the press for 5-10 minutes and then check the bearings [7].

Due to the presence of dry straw or hay in the baling chamber, which has a flash point of approximately 300°C depending on the moisture content, as well as rubber bands on the rollers, which are constantly in contact with the hot surface and can heat up during normal operation, they create all the conditions for a fire. The operator of the machine is often able to watch the process only from a mirror.

With such an amount of dry material (straw, hay) it is difficult to extinguish the fire in 1-2 minutes so that the fire does not spread further. In such fires, the greatest chance of rescuing the tractor in the event of a quick uncoupling of the baler [8].

Another important parameter is the parallel guidance of the forming strips at the same distance. If these belts are not lowered in parallel, sooner or later their friction will lead to a fire. Gear bearings also deserve attention even if they are less exposed to direct contact with straw, but accidental dust and other contaminants can also cause a flammable substance to ignite when the bearings are overheated due to wear [9]. There is illustrated in Fig. 1 the baler machine KRONE Big Pack and the same kind of machine damaged after a fire accident is in Fig. 2.



Figure 1: KRONE Big Pack 1290XCHDP



Figure 2: Damaged Big Pack 750 after the fire accident [10]

2.2. Risk analysis of the selected equipment

A comprehensive workplace risk assessment method was used to assess the risks of the selected press [8-10].

The risk assessment at work was performed using a comprehensive method of risk assessment at the workplace [11]. Common practice in small and medium-sized enterprises requires time-saving methods, which, however, presuppose knowledge of the actual state of the monitored technology. The method developed by the international team of the Security Section (represented by France, Germany, Italy and the Czech Republic) can be satisfactory in this respect. The method is based on the knowledge that an injury occurs in most cases for several reasons. The work is performed in the system man (P) - machine (M) - environment (U) and the level of safety depends on all three elements - parameters. These elements are assessed individually by this method, by appropriately assigning a point method and defining an acceptable risk. The method is mainly suitable for immediate risk assessment in order to apply immediate, non-complex measures. The record is kept in the form of a questionnaire [12]. The values in the risk assessment were determined by direct consultation with the tractor driver.

Profession: tractor driver
Equipment under assessment: baler

Description of job: The damaged bearing got stuck and the action of frictional force led to the accumulation of heat, which passes into the chamber of the press and subsequently a machine fire occurred.

Description of the work procedure: After observing the fire, the tractor driver stopped the tractor and the press brake, switched off the drives and exits the tractor, then extinguishes the fire with the help of fire extinguishers**.** Table 1 shows the risk assessment of machinery, namely the baler.

Table 1: Risk assessment of the equipment

Risk element - M Equipment (baler) risk assessment		**Rating**	**Label**	**Range**
Identification of possible damages	Severe consequences (Burns)	S	7	1 - 10
Danger exposure	Frequently repeated exposure	Ex	**1.5**	1 - 2
Probability of a dangerous situation	Central	Wa	**1.5**	0.5-1.5
Possibility of prevention	United	Ve	**0.5**	0.5-1

$$M = S \times Ex \times Wa \times Ve \quad (1)$$
$$\boldsymbol{M = 5.25}$$

The resulting risk level of machinery from the range (0.25-30) is 5.25.
Table 2 shows the risk assessment of the working environment.

Table 2: Environmental impact

Environmental impact - U		**Result**	**Label**	**Range**
Work organization area	On one level	Ua	**0.6**	0.5-1
Work environment	Dust, heat	Ub	**0.6**	0.3-0.6
Other loads	A narrow space	Uc	**0.4**	0.2-0.4

$$U = Ua + Ub + Uc \quad (2)$$
$$\boldsymbol{U = 1.6}$$

The resulting risk level of the working environment from the range (1-2) is 1.6. Here it can be seen that the risk of the working environment is higher. Table 3 shows the risk assessment of the human factor.

The resulting risk factor for human factor from the range (15 - 0) is 15. Here it can be seen that the risk of human factor failure is low.

Table 3: Operator competence manage the risk

Ability of the operator to handle the risk-P		**Rating**	**Label**	**Range**
Person classification	An educated person with experience	Q	**10**	10 - 0
Psychological factors	Appropriate mental ability	φ	**3**	3 - 0
Work organization	The prescribed workflow is not always used	O	**2**	5 - 0

$$P = Q + \varphi + O \quad (3)$$
$$\boldsymbol{P = 15}$$

The resulting degree of risk of the human-machine-environment work system is determined by the following relation (4).

$$R = M \times U - P \times (M/30) \quad (4)$$
$$R = 5.25 \times 1.6 - 15 \times (5.25/30)$$
$$\boldsymbol{\underline{R = 5.78}}$$

The resulting level of risk when baling straw using a tractor with an attachment is low, but the method found that the cause of the fire was the working environment, i.e. high ambient temperature.

Based on the risk assessment, the following measures have been proposed to the operator. The first place of protection against agricultural fires is the maintenance of the machine.

2.3. Proposal of risk minimization measures

This phenomenon can be prevented by inserting a bearing temperature measuring device into a combine or press, thus reducing the risk of fire. Bearing-mounted temperature sensors transmit data to the device, where they are displayed. On the screen it is possible to monitor the bearing temperatures during operation can avoid the risk. It is possible to set the temperature limit, if a bearing with a red circle is signalled, then the temperature has exceeded the value that is set [13].

These bearing thermometers are also designed to detect the temperature of the bearing housing. The tip of the sensor is designed to make the best contact with the measured surface. The sensor should be placed directly in the hole. The clamping ring together with the spring ensures clamping on the surface so that information about the condition of the bearing is available. In the

event of overheating by the sensor signal, the machine can be stopped to prevent damage.

Parameters of the given sensor are:

- maximum ring diameter: 8 mm,
- measuring range: from -25°C to + 250°C,
- wire length: 2 m.

The innovative baler process already has this system built in, but older types are only cleaned mechanically. The hydraulically driven turbo fan, Fig. 3 and Fig. 4, system protects the knotters from pollution and reduces the risk of dust explosion [14].



Figure 3: The hydraulically driven turbo fan system protects the knotters from pollution, a strong air flow with speed 140 km/h. [14]



Figure 4: Two oil pumps supply the input gearbox and the TURBO FAN knot cleaning system with hydraulic oil. [14]

Fires arising from agricultural machinery occur mainly in the engine compartment. The consequence, especially on the damaged construction of the machine, often exceeds the value of the machine. Increased requirements to reduce noise and exhaust emissions are the main reasons for greater insulation and enclosure of the engine compartment, increased fuel pressure and thus an increase in engine compartment temperature.

The fire in the engine compartment spreads often and very quickly and intensively, so it cannot be extinguished with a fire extinguisher. Vehicles and equipment in agriculture, such as harvesters, tractors, loaders, etc. are high power machines that work at a constant load. An increasing number of more powerful electrical and hydraulic equipment pose an increased risk, and it is therefore possible to speak of these machines as potentially dangerous sources. [15-16]

If the engine fire start, the whole machine will burn during 20 minutes. Due to the nature of the work area, it can quickly fall victim to a fire, completely isolating the burning vehicle and its operator.

The FOGMAKER system ™ is an automated closed system that can be operated by direct human intervention and without power supply. On the principle of water mist, the fire in the engine compartment of the machine is extinguished. FOGMAKER ™ may consist of one, two or three fire extinguishing bottles. The construction allows full use of the content of the extinguishing mixture and thus emptying of the entire extinguishing cylinder. This is a positive advantage because the extinguishing roller can be placed vertically or horizontally in confined spaces. [17-18]

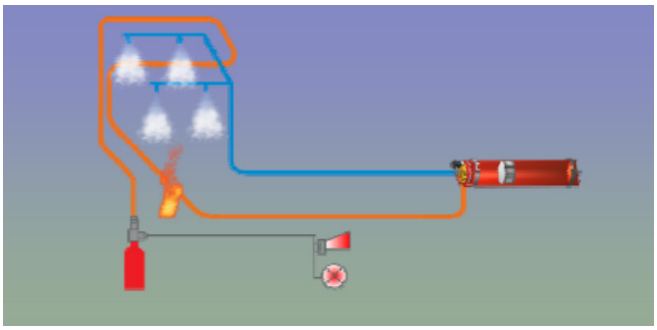


Figure 5: Hydro-pneumatic detection and activation [18]

Detection and activation of the system is performed hydro-pneumatically.

When a fire occurs, the detection hoses expand / burst with pressure and pressure drop, activating the valve on the extinguishing pressure cylinder. The pressure gauge, which is located on the detection cylinder, then alerts you with light and sound signals.

This power-independent system has an extraordinary cooling effect of - 500 ° C in less than 3 seconds. Water high-pressure mist with a pressure of 100 bar and special nozzles turn water into microdroplets with a diameter of 50 μm. The volume of 1 drop with a diameter of 1 mm is sufficient to create 8,000 drops of water mist. Water with a volume of 1 liter absorbs 540,000 calories during evaporation, which gives the extinguishing a completely unique effect. Cooling contributes to rapid extinguishing and thus reduces the risk of re-ignition [19-20].

3. Discussion

The result of the analysis was a finding that the cause of the fire is the environment, i.e. the high ambient temperature. The critical element in hay baling processing system is the working environment. Although the primary cause of the fire was a damaged bearing, the high ambient temperature and the presence of hay in the working environment caused an adverse event, i.e. fire. The risk level of the working environment was estimated at the value 1.6 from the interval < 1; 2 >. The resulting risk level 5.78 was low due to rapid human intervention. The acceptance limit in this method is set at the value 15. Based on the results, the technical measures were proposed to the operator. There was determined in this article that the maintenance of the machine is important for protection against the agricultural fires, where the human factor plays an important role in the man - machine - environment system. As a damaged bearing or electrical installation contributes significantly to a fire in agriculture, maintenance is absolutely necessary in preventing of fires.

4. Conclusion

It can be stated that a damaged bearing, as stated for presses, plays a major role in the fires of agricultural machinery. For tractors, trucks, the main initiator is an electrical short or faults in the electrical installation. In the case of machines intended for harvesting such as a combine harvester or forage harvester, a damaged bearing or electrical installation contributes significantly to a fire. After the risk analysis with a comprehensive method, satisfactory results were obtained, but this does not mean that there is no danger. The proposed measures to improve the situation are feasible, but nowadays smaller companies cannot afford it. Companies with better financial possibilities hire a special service for machines [21-22].

After the service period, maintenance is performed in-house. Today's companies try to have as little as possible, there are 1 maintenance person on about 10 machines [23-24]. Under the pressure of time, maintenance workers carry out repairs to such an extent that the machine is functional-mobile, but safety is significantly neglected. Today's generation of tractor drivers and engineers also neglects the basic maintenance of machines such as bearing lubrication or engine oil change. Some companies motivate their employees by being paid for maintenance. Machines and systems can be improved from a safety point of view, but the human factor plays the biggest role. Fighting fires in agriculture are liquidated according to the methodological sheets of the Fire and Rescue Service.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

APVV 15-0351 Development and application of risk management models in the conditions of technological systems in accordance with the strategy Industry 4.0 and VEGA1/0121/18 Development of methods for implementation and verification of a comprehensive security solution in Smart Factory as part of the Industry 4.0 Strategy.

References

[1] C. Y. Chen, Y.C. Ling, J. T. Wang, H.Y. Chen, "SIMS depth profiling analysis of electrical arc residues in fire investigation", Applied Surface Science, New York, 2003, https://doi.org/10.1016/S0169-4332(02)00817-6

[2] M. P. Filippakou, C.G. Karagiannopoulos, D.P. Agoris, P.D. Bourkas, "Electrical contact overheating under short-circuit currents" 2001 Electric power systems research. https://doi.org/10.1016/S0378- 7796(01)00081-5

[3] M. Balara, D. Dupláková, D. Matisková, "Application of a signal averaging device in robotics" In: Measurement., 115, 125-132, 2018, https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=1&SID=C2GBN... https://doi.org/10.1016/j.measurement.2017.10.037

[4] D. Matisková, Ľ. Ambiško, "Optimization of cutting conditions and improvement of production in economic terms", TEM Journal., **6** (3), 584-590. 2017. ISSN 2217-8309

[5] V. Shah, D. Twidwell, C.L. Wonkka, M.T. Sindelar, J.R. Weir. "First Approximations of Prescribed Fire Risks Relative to Other Management Techniques Used on Private Lands". PLoS ONE 2015

[6] P. J. Val-Aguasca, M. Videgain-Marco, P. Martín-Ramos, M. Vidal-Cortés, A. Boné-Garasa, F. Javier García-Ramos. "Fire Risks Associated with Combine Harvesters: Analysis of Machinery Critical Points", Agronomy 2019, **9**, 877, https://doi.org/10.3390/agronomy9120877

[7] I. Hadade, T. M. Jones, F. Wang, L. di Mare, "Software Prefetching for Unstructured Mesh Applications" in 2018 IEEE/ACM 8th Workshop on Irregular Applications: Architectures and Algorithms (IA3), Dallas, TX, USA, 2018. https://doi.org/10.1109/IA3.2018.00009

[8] J. Shutske, W.E. Field. "An Integrated Loss Control Strategy for Grain Combine Fires". In 1988 International Winter Meeting of the American Society of Agricultural Engineers; American Society of Agricultural Engineers: Chicago, IL, USA, 2014; p. 17.

[9] H. Lihua, "Analysis of Fuel Cell Generation System Application," Ph.D Thesis, Chongqing University, 2005.

[10] Agroinform.hu, Erre vigyázzon bálázáskor, hogy ne gyulladjon ki a gépe, [cit.2018-4-1], https://www.agroinform.hu/gepeszet/erre-vigyazzunk-balazaskor-23427

[11] Agroforum.hu, Ezért gyulladnak ki a bálázók és a kombájnok – Tűzvédelmi tanácsok az aratás kezdetére, [cit. 2018-4-1], https://agroforum.hu/agrarhirek/gepinfo/ezert-gyulladnak-ki-a-balazok-es-a-kombajnok-tuzvedelmi-tanacsok-az-aratas-kezdetere/

[12] Magro.hu, Sok bálázó ég ki munka közben – erre figyeljen Ön is!, [cit. 2018-4-1],https://www.magro.hu/agrarhirek/szines/sok-balazo-eg-ki-munka-kozben-erre-figyeljen/

[13] Agrofarmelectro.com, Csapágy hőfok figyelő, [cit. 2018-4-1], https://www.agroinform.hu/gepeszet/automata-tuzvedelem-a-mezogazdasagban-12305,

[14] Tiszacsege.hu, Aratás előtti gépszemle, [cit.2018-4-1], http://tiszacsege.hu/wp-ontent/uploads/2014/07/betakaritas_tuzvedelem.pdf

[15] Agroinform.hu, Automata tűzvédelem a mezőgazdaságban, [cit. 2018-4-2],:https://www.agroinform.hu/gepeszet/elozzuk-meg-a-kombajntuzet-33150-001

[16] Fogmaker.com, R107, [cit. 2018-4-2], http://fogmaker.com/r107/

[17] Jandjpsv.eu, Meghódította a világot 6 kontinensen több mint 30 országban,[cit.2018-4-2],http://jandjpsv.eu/wp-content/uploads/sajto/Florian-exPress-2013-oktober.pdf

[18] Enatruck.cz, Fogmaker - unikátní hasicí systém, [cit. 2018-4-2], https://www.enatruck.cz/enatruck-cz/eshop/14-1-Signalizacni-a-hasici-systemy/-4-/5/147-Fogmaker-unikatni-hasici-system

[19] Internal documents of the company, consultation with the security technician and mechanizer of the company, 2015.

[20] Maintenance in Agriculture - A Safety and Health Guide, European Agency for Safety and Health at Work, ISBN 978-92-9191-667-2 https://osha.europa.eu/en/publications/reports/maintenance-in-agriculture-a-safety-and-health-guide, 2011.

[21] G.R. Quick. "An Investigation into Combine Harvester Fires"; Grains Research and Development Corporation: Canberra, Australia, 2010.

[22] K. Neubauer, Stroje pro rostlinnou výrobu, (Machines for Agricultural Production), Státní zemědělské nakladatelství, Praha, ISBN 80-209-0075-6, 1989.

[23] J. Sinay, M. Oravec, H. Pačaiová. "Risk assessment methods". Košice: TU, 2001.

[24] T. Henschel, Erfolgreiches Risikomanagement im Mittelstand. Strategien zur Unternehmenssicherung, Erich Schmidt Verlag, Berlin, ISBN 978 3 503 11648 5, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 2, 19-28 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Automatic Comprehension and Summarisation of Legal Contracts

Sibusiso Kubeka*, Abejide Ade-Ibijola

Formal Structures, Algorithms and Industrial Applications Research Cluster, Department of Applied Information Systems, University of Johannesburg, Bunting Road Campus, Johannesburg, 2006, South Africa

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 09 September, 2020
Online: 10 March, 2021

Keywords:
contract
clauses
automatically comprehend and summarise
context-free grammars

ABSTRACT

Contracts may range from a simple agreement between a tenant and a landlord or a gym contract, or it could be as important as an employment or marital contract. No matter the level of importance, individuals are legally obligated to obey and carry out all clauses in the contract. In this paper, we have identified that the majority of people seldom read through the entire contracts for several reasons such as the size of the contracts i.e. bulky contracts or the inability to fully comprehend a contract. As a solution to the identified problem, this paper presented a software tool that automatically comprehends and summarises legal contracts. We designed context-free grammar (CFG) rules for the recognition of critical clauses found in contracts. These CFG rules were implemented in the software tool. An evaluation of this tool showed that it was able to identify critical clauses in contracts to an accuracy of 79.2%.

1 Introduction

In this section, we present an overview and contributions of the paper. The history of contract laws dates back to prehistoric civilizations which was profoundly influenced by ancient Roman and Greek thoughts [1]–[2]. At the time contract laws were established around a form of action identified as the action of "assumpsit", which was put into effect in the early 16th century as a solution for the breaches in informal contracts made by word of mouth [3, 4]. As the years have gone by contract laws have evolved and there have been several developments with regards to the type of contracts [5, 6].

With all these new developments of contract laws comes several challenges and unwanted problems, such as the recent discovery that majority of contract receivers fail to thoroughly comprehend bulky contracts which has led to a large number of law cases related to breaches in modern day contracts by contract receivers [7, 8]. Contract receivers are easily exploited by organisations given that they do not have a full understanding of the clauses identified in the contract [9, 10].

In this paper we discussed in great detail a feasible solution to the problems at hand. We mainly focused on developing a web application that is capable of comprehending any legal contract, extracting semantic information as well as generating accurate summaries that includes all the important clauses stated in a contract. Furthermore, we incorporated an Artificial intelligence concept of natural language processing (NLP) into our proposed solution in order to automatically comprehend and extract critical sections as well as to find patterns from within a contract [11]–[12].

Natural language processing is a phenomenon that has been in existence for close to 50 years and focuses on techniques such as automatic comprehension, summarisation and information extraction. NLP is by far more efficient and will help save time as compared to processing the bulky contracts manually [13].

In order to develop the web application we needed to start with the cleaning out of unnecessary information and white spaces by performing lexical analysis on the contracts, we then outlined a context-free grammar to identify syntactic patterns in the contracts, and finally we had semantic rules that were used to make conclusions on the basis of the identified patterns [14]–[15].

Figure 1 is a graphical representation of the logical flow, that we used to automatically comprehend and summarise a legal contract document. The diagram outlines the different phases a contract will undergo, these phase include pre-processing, document slicing and finally compiling.

The contributions of the paper are as follows. We have:

1. conducted extensive literature reviews on contracts,

*Corresponding Author: Sibusiso Kubeka, University of Johannesburg, Bunting Road Campus, Johannesburg, South Africa, skubeka32@gmail.com

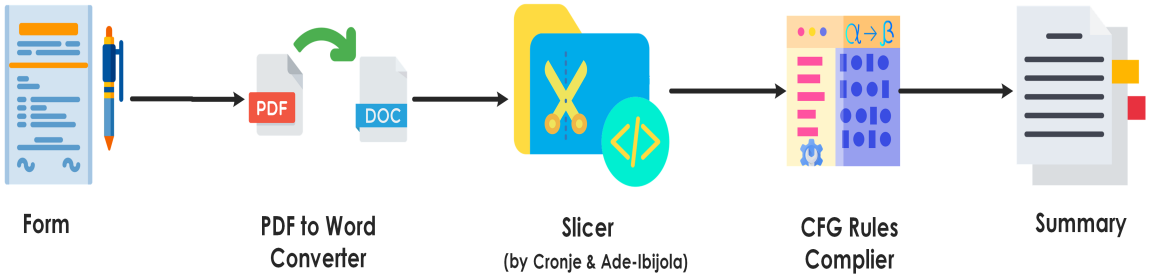


Figure 1: Flow diagram of the comprehension and summarisation process.

2. performed lexical and document slicing of contract documents,
3. identified any hidden implications as well as critical clauses that one should look out for in a contract,
4. formulated context-free grammar rules to help with the identification of patterns within contracts, and
5. implemented the CFG rules into the web application to automatically comprehend and summarise contract documents.

2 Background

In this section, we provided a detailed literature review on contracts, the structure of contracts, different types of contracts that exist, various critical clause found in contracts, information extraction, automatic text comprehension and summarisation, and related work.

2.1 Contracts

A contract can be defined as a promise enforceable by law, it is a legal written agreement that requires the mutual approbation of two or more parties [16, 17]. There are only two roles identified in contracts namely: the person making an offer and the person accepting the proposed offer. If for any reason one of the involved parties fails to adhere to the condition stated in the contract, the other party is entitled to legal reparation [18].

2.2 Structure of Contracts

In this section, we outlined the typical structure of a contract. Contract structures will differ depending on the type of contract.

There are many different types of contracts available and each contract serves a specific purpose [19, 20]. Types of contracts include: General business contracts (Partnership agreements, indemnity agreements, loan agreement and property and equipment lease), Sales-related contracts (Bill of sales, purchase order and security agreement), and lastly Employment contracts (General employment contract, Non-compete agreement and independent contractor agreement).

The structure of a contract typically includes the following categories. Parties to the contract: the details of all involved parties. Recitals: factual information as well as the purpose of the contract. Definitions: explanations of keywords. Conditions precedent: the state of affairs required. Operative provisions: the rights and obligations of the parties involved. Boilerplate (or Miscellaneous) clauses: important clauses found in a contract. Schedules: the appendix to a formal document. Signatures: distinctive pattern used for identification of parties. The main objective of identifying the structure of a contract is to gain a better understanding in order to be able to recognize the syntax or patterns and create semantic rules to make inferences on identified patterns [21].

2.3 Critical Clauses in Contracts

Within a contract, a clause is defined as a specific section or prerequisite that focuses on an aspect of an agreement. Generally, clauses are used to provide details on aspects such as party duties and roles, rights, regulations, as well as benefits [11, 22, 23]. In this section, we provide a list of critical clauses and important information typically found in legal contract documents. The objective is to identify which information is relevant and should be extracted from an original contract, to create an informative and accurate summary.

Recitals clause

This is a contract clause commonly referred to as the "whereas" clause, it is usually associated with the purpose of a contract. This clause provides the contract receiver with an idea of what the contract is about and why they should sign the contract.

Contract Parties

this section in a contract usually contains information about all parties involved in the agreement. It typically includes the company details, contract receiver information, and any other third party.

Dates

We must be able to extract dates such as "11 July 2020" or "11-07-2020", as they could be representing important information such as the Commencement or end date of a contract, or it could be a representation of important milestones or conditions that need to be accomplished i.e."lease period","termination dates", etc.

Period/Duration

This could indicate for instance the duration of the contract or number of hours an individual is required to work in a week. The extraction of keywords such as "10 Hours","15 days", or "three years" is critical and should, therefore, be

included in the summary.Duration and period in contracts are a representation of "working hours", "inspection period", and "eviction notice".

Money/ currency

It is important to extract money and different currency symbols such as "R10", or ZAR, USD, JYP, etc as these units could represent information such as fines or penalties, salaries, or rent amount. Within a contract money variables are a representation of "Cancellation fees", "Compensation and benefits",

Definitions

The summary must include all definition statements found in a contract, these statements normally begin with words such as "is defined as", "refers to", "shall mean", and "means that".

Conditions

Conditional sentences consist of condition clauses such as "if/ if not"," when/ when not"," where"," as soon as", "upon the occurrence", and many other clauses or conditions found in a contract [11].

Amounts

It is important to include non-financial amounts when extracting information for summaries. This could include quantities such as "Three hundred thousand".

2.4 Information Extraction

Information Extraction (IE) can be defined as the automatic extraction of structured information from unstructured source documents. It takes natural language inputs and can convert them into structured texts using a set of indicated criteria. Information extraction is normally associated with the following sub-tasks: Name Entity Recognition and Linking, Conference Resolution, and Relation Extraction which form part of the segments of NLP tasks like Text Summarisation, Machine Translation, etc. IE technologies are mainly tasked with discovering valuable and relevant information, to construct a significant representation of the original document [24, 25].

There are two types of approaches for automated information extraction using NLP, namely the rule-based approach, and the machine learning-based approach [24, 25]. The machine learning-based approach makes use of machine learning algorithms for text processing based on text features of a certain text [26]. On the other hand, the Rule-based approach employs manually coded rules for text processing. The rule-based approach creates reiterative and refined rules to expand the precision of text processing. The Rule-based approach generally requires more human effort and makes use of several NLP techniques such as tokenization, parts of speech (POS) tagging, phrase structure analysis, and sentence slicing [24]. In this paper, we will make use of context-free grammar (CFG) for syntactic analysis to reduce the number of patterns required in information extraction.

2.5 Automatic Text Comprehension and Summarisation

Automatic text comprehension is a phenomenon that has been in existence since the early 70s and has since become popular in today's society [27, 28]. Text comprehension is a far more complex task when compared to other computational tasks such as reading or writing (inputting). It is the ability to gain a thorough understanding of a given statement. Understanding a text automatically deduces that one has the ability to form an interconnected and cohesive interpretation of the information found in that particular text. There are several Natural language Processing (NLP) applications that require comprehension of source documents and it is important to note that the majority of those applications comprise of semantic processing of a natural language [28, 29].

Automatic text summarisation is a process related to the construction of a brief and precise version of an original text with the use of a computational software program. The aim is to create summaries that are considerably shorter and contain relevant and important information found within the subsequent document i.e. a contract [30, 31]. Automatic text summarisation can be achieved using two approaches:

Extraction

The extraction-based approach to automatic summarisation focuses on extracting important information based on keywords or key-phrases [32]. The information extracted is not modified nor represented in a new way or form.

Abstraction

The abstraction-based approach, on the other hand, is aimed at using techniques such as paraphrasing in order to provide summaries that are more structured and represent information in a new way, this usually includes information that was not present in the original document. In short, abstractive summaries are semantic representations of the source document.

In this paper, we looked into creating summaries for contract documents using the extraction approach. There is an all-encompassing range of practices that have been recommended for constructing summaries in automatic text summarisation. And it is said that most efficacious practices are based on either the positioning of words/ sentences in a text or on text retrieval. Text retrieval techniques are statistically more fruitful when implemented on source documents for text summarisation.

3 Related Work

From the research that we have conducted it is safe to conclude that very few attempts have been made to automate the comprehension and summarisation of legal documents. There are numerous articles and literature reviews have been published on automatic comprehension and summarisation of source documents using CFGs. Several papers have also been written on the use of natural language processing and machine learning on contracts and legal texts.

FINCHAN

In [29], the author presented an innovative technology that

used a Context-Free Grammar to create a subset of every statement in a given Instant message. This article focused on creating a software application for automatic comprehension and summarisation of financial chats.

LexNLP
In [11], the author proposed a new technology that uses natural language processing as well as machine learning to extract information in legal and regulatory documents. LexNLP is a software designed to enable end-users to turn an unstructured contract into a structured data object. The main aim of LexNLP is to make tasks such as regulation analysis and migration of legal documents a lot easier.

Source Code
Automatic Summarisation of Source Codes in [15], the author proposed a solution that enables software developers to read shorter and accurate source codes to save development time. The solution involves using techniques from automatic text summarisation. The paper states that a combination of summarisation techniques is more effective when summarising source codes.

Semantics in Law
Automatic Extraction of Semantics in Law Documents in [33] present a tool called SALEM, that uses Natural language Processing practices to execute two main tasks, namely classification of law paragraphs in relation to their regulatory substance, and extraction of relevant information in the form of text fragments. The article outlines the architecture of the system as well as a report based on a case study on Italian laws.

Rule Extraction
NLP Approaches for Rule Extraction from Legal Documents by Dragoni presented a software tool that intermixes several Natural Language Processing techniques in order to extract rules found in contracts or legal documents. The paper proposed the combination of linguistic information together with syntax-based extraction of rules from a contract. Furthermore, these techniques will be combined with the logic-based extraction of dependencies.

Summarisation of Legal Texts
In [34], the author proposed a method that automatically summarises legal texts using graph-based summarisation algorithms. Each legal document has its own set of connected graphs, the connected graphs show a cluster which shared the same clause or topic. This method does not require any self-developed linguistic features.

LetSum
An automatic Legal Text Summarising system in [35], the author presented a system that can establish the thematic structure of a legal document and group it into four sections, namely Introduction, context, legal text analysis and lastly the conclusion. From there, the system is able to recognize the relevant text for each section.

3.1 The Gap

People tend to accept or sign contracts without reading through them thoroughly which could result in; Exploitation of contract receivers, the legal cost acquired by contract receivers, etc. Most text summarisation tools are not designated for contract comprehension and summarisation. There has not been any technology developed to provide contract receivers with accurate textual summarisation of critical clauses found in contracts.

In this paper, we were able to identify a gap in that there are no automatic tools that have been developed to help with comprehending and summarising legal contracts. This tool will help the people that struggle with reading bulky contracts by providing them with relevant summaries of contracts.

3.2 Reason for using Context-free Grammars

In this section, we outline the two main reasons for using CFG rules to automatically comprehend contract documents.

Usage of a Document Slicer When developing the tool, we identified that using a document slicer to slice through a contract to recognize relevant sections would increase the accuracy of the findings of the applied CFG rules. For our web application, we made use of an automatic slicer API [36, 37].

Distribution of Grammar Rules The use of CFG rules ensures effortless distribution of grammar rules amongst other researchers and scientists that aim to comprehend or synthesize similar documents. Within the science field, it is critical to have work that can effortlessly be distributed, reusable, or modified by individuals occupied with similar content.

Use of FLAT Formal Languages and Automata Theory (FLAT) refers to a popular feature of Theoretical Computer Science. FLAT is known for having a lot of mathematical theories with very little real-life implementation [38]. The application of FLAT helps with identifying new applications within areas such as computer science education.

3.3 Definition of Terms

Definition 1 (Context-Free Grammar) *Context-free grammar (CFG) is a set of recursive rules utilized to create patterns of a given text. A CFG is formally defined as a 4 tuple, G = (V, Σ, P, S) where:*

1. *V represents the variables or nonterminal characters (the nonterminal characters indicates the types of clauses found in a sentence)*
2. *Σ represents a finite set of terminal characters (these are the alphabets of a given language).*
3. *P represents the production rules.*
4. *S is the start symbol. The relation S ϵ N must always hold. [39].*

Definition 2 (Lexical analysis) *Lexical analysis also referred to as tokenization or lexing is a computer science process that involves converting a sequence of characters into tokens. A token is usually divided into two sections namely the token name and an attribute value [40, 41]. A lexer is a program that performs lexical analysis and it is often used to find specific patterns such as white spaces, new lines, and digits or letters in any text or document.*

4 Context-free Grammar for Comprehension of Contracts

4.1 Basis

Here we specify the building block for the rules that is presented later in this section. We have three major building blocks namely n, spc, and lastly rand_text. n represents any numeric value, spc represents tabs and spaces found within a text, rand_text represents any random combination of letters that make up words.

4.2 CFG rules

The following are CFG rules for generating the subset of all critical clauses and important information found in contracts. Figure 2 is the syntax tree diagram outlining the structure of contracts. Contracts are grouped into three categories as discussed earlier, namely General business contracts <gb>, Sales-related contract <sr>, and Employment contracts <emp>. A contract <contr> includes components such parties to the contract <pc> and critical clauses <cc>. The production of these components is as follows:

$$\text{<contract>} \longrightarrow \text{<gb>} \mid \text{<sr>} \mid \text{<emp>} \tag{1}$$

$$\text{<gb>} \longrightarrow \text{<pc><recital><CC\_comp>} (\text{<CC\_gb\_opt>} \mid \lambda) \tag{2}$$

$$\text{<sr>} \longrightarrow \text{<pc><recital><CC\_comp>} (\text{<CC\_sr\_opt>} \mid \lambda) \tag{3}$$

$$\text{<emp>} \longrightarrow \text{<pc><recital><CC\_comp>} (\text{<CC\_emp\_opt>} \mid \lambda) \tag{4}$$

$$\text{<CC\_comp>} \longrightarrow \text{<term\_date><comm\_date>} \mid \tag{5}$$

$$\longrightarrow \text{<comm\_date><term\_date>} \tag{6}$$

All types of contracts contain a section that represents the parties to the contract, hence production Rules 2, 3, and 4 consist of <pc>. A parties to the contract section is made up of details about the company <com> and one or more contract receiver <cr>.

$$\text{<pc>} \longrightarrow \text{<com>}(\text{<cr>})^{+} \tag{7}$$

$$\text{<cr>} \longrightarrow \text{<first\_name><spc><last\_name><cell\_no.>}(\text{<email>} \mid \lambda) \tag{8}$$

$$\text{<com>} \longrightarrow \text{<com\_name><com\_info>} \tag{9}$$

$$\text{<com\_info>} \longrightarrow x \in (\text{<address>}, \text{<tel\_no>}, \text{<fax>})^3 \ni: {}^3P_3 \tag{10}$$

$$\text{<address>} \longrightarrow \text{<area\_no.><spc><street\_name>} \tag{11}$$

$$\longrightarrow \text{<city><spc><postal\_code>} \tag{12}$$

$$\text{<tel\_no>} \longrightarrow \text{<tel\_no.><n>} \tag{13}$$

Here ${}^3P_3$ is a k-permutation of x (i.e. address, tel_no, and fax notices without repetitions).

Recital clauses are found in all contract types. We observed that the word "WHEREAS" is often associated with the recital in a contract, and if a line in the contract start with "WHEREAS" (usually written in capital letters) , it often means that the purpose of the contract <poc> is to follow.

$$\text{<recital>} \longrightarrow (\text{<rand\_text>} \mid \lambda)\text{<WHEREAS><spc><poc>} \tag{14}$$

$$\text{<poc>} \longrightarrow (\text{<rand\_text>} \mid \lambda) \tag{15}$$

The symbols <CC_gb_opt> from production Rule 2 is composed of four optional critical clauses: deposit<deposit>, lease period <lease_period>, rent <rent>, and eviction notices <evic_not>.

$$\text{<CC\_gb\_opt>} \longrightarrow x \in (\text{<deposit>}, \text{<lease\_period>}, \text{<rent>}, \text{<evic\_not>})^4 \ni: {}^4P_4 \tag{16}$$

Here ${}^4P_4$ is a k-permutation of x (i.e. deposit, lease period, rent, and eviction notices without repetitions).

The symbols <CC_sr_opt> from production Rule 3 is composed of: description of goals <goal_desc>, tax/rate <tax>, delivery instructions <del_inst>, and inspection period <insp_period>.

$$\text{<CC\_sr\_opt>} \longrightarrow x \in (\text{<goal\_desc>}, \text{<tax>}, \text{<del\_inst>}, \text{<insp\_period>})^4 \ni: {}^4P_4 \tag{17}$$

Here ${}^4P_4$ is a k-permutation of x (i.e. description of goals, tax, delivery instructions, and inspection periods without repetitions).

The symbols <CC_emp_opt> from production Rule 4 is composed of: compensation <compen>, benefits <ben>,working hours <hours>, roles and responsibilities <roles>.

$$\text{<CC\_emp\_opt>} \longrightarrow x \in (\text{<compen>}, \text{<ben>}, \text{<hours>}, \text{<roles>})^4 \ni: {}^4P_4 \tag{18}$$

$$\text{<compen>} \longrightarrow \text{<salary>} \mid \text{<wages>} \mid \text{<pay>} \mid \text{<remuneration>} \mid \text{<earnings>} \tag{19}$$

$$\text{<roles>} \longrightarrow \text{<responsibility>} \mid \text{<duties>} \tag{20}$$

$$\text{<leave>} \longrightarrow \text{<sick\_leave>} \mid \text{<annual\_leave>} \mid \text{<maternity\_leave>} \tag{21}$$

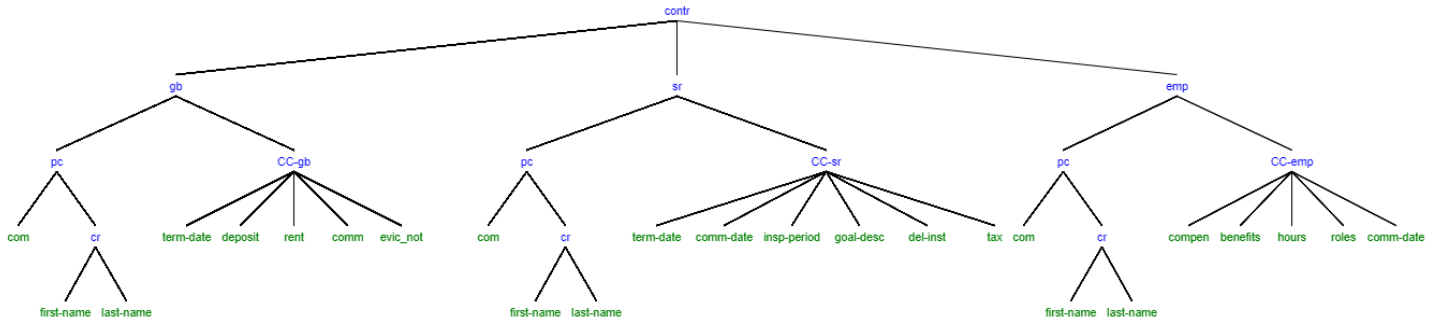


Figure 2: Syntax tree outlining the structure of contracts.

Here ${}^4P_4$ is a k-permutation of x (i.e. compensation, benefits, working hours, and roles and responsibilities without repetitions).

We also observed that the words "shall mean", "is defined as"," means that" is often associated with definitions found in a contract, and if a line in the contract contains these words it often means a definition is to follow.

$$\begin{aligned} \texttt{<def>} &\longrightarrow (\texttt{<rand\_text>} \mid \lambda)\ \texttt{(shall mean} \mid \\ &\quad \texttt{is defined as} \mid \texttt{means that)} \\ &\quad \texttt{<spc><definition>} \qquad (22) \\ \texttt{<definition>} &\longrightarrow (\texttt{<rand\_text>} \mid \lambda) \qquad (23) \end{aligned}$$

Money or Currency:

$$\begin{aligned} \texttt{<cur>} &\longrightarrow \texttt{ZAR} \mid \texttt{INR} \mid \texttt{ZMW} \mid \texttt{JPY} \mid \qquad (24) \\ &\longrightarrow \texttt{MUR} \mid \texttt{KES} \mid \texttt{UGX} \mid \texttt{USD} \mid \texttt{EUR} \mid \qquad (25) \\ &\longrightarrow \texttt{CNY/RMB} \mid \texttt{GBP} \qquad (26) \end{aligned}$$

Date and Duration:

$$\begin{aligned} \texttt{<date>} &\longrightarrow \texttt{<day><spc><mon><spc><year>} \qquad (27) \\ \texttt{<day>} &\longrightarrow \texttt{<Mon>} \mid \texttt{<Tue>} \mid \ldots \mid \texttt{<Sun>} \qquad (28) \\ \texttt{<mon>} &\longrightarrow \texttt{<Jan>} \mid \texttt{<Feb>} \mid \ldots \mid \texttt{<Dec>} \qquad (29) \\ \texttt{<year>} &\longrightarrow \texttt{<n><n><n><n>} \qquad (30) \\ \texttt{<dur>} &\longrightarrow \texttt{<year>} \mid \texttt{<mon>} \mid \texttt{<days>} \qquad (31) \end{aligned}$$

Acts and laws: This section includes the important acts and laws that are found in a contract document.

$$\begin{aligned} \texttt{<Act>} &\longrightarrow \texttt{<Act>(<n>}^{+}\texttt{)<of>} \qquad (32) \\ &\longrightarrow \texttt{<year>(<rand\_text>} \mid \lambda) \qquad (33) \\ \texttt{<Act>} &\longrightarrow (\texttt{<rand\_text>} \mid \lambda) \qquad (34) \end{aligned}$$

5 Implementation and Results

In this section, we outline the implementation details as well as the results of the proposed tool for automatic comprehension and summarisation of legal contract documents.

5.1 Implementation Details

The Context-Free Grammar rules defined in Section 4 of this paper were executed in a web application using Microsoft's Visual Studio 2019 as the development environment. We used the .Net framework with Visual Basic as our programming language. Furthermore, we made use of a document slicer API by Cronje and Ade-Ibijola [36].

5.2 Web Application

In section, we present a brief explanation on the sections found in our web application, we will also provide screenshots of the relevant parts of the web application. The web application is made up of three major sections, namely, the home page section , contract summarisation section and the summary section. Figure 3-5 below are the respective screenshots of each of mentioned sections. Table 1 is an example of a summary produced by the tool.

6 Evaluation of Tool

In this section, we presented an in-depth evaluation of our tool based on a performance analysis we had conducted on the tool. We also present the user perception of the tool based on the results gather from a survey we had conducted.

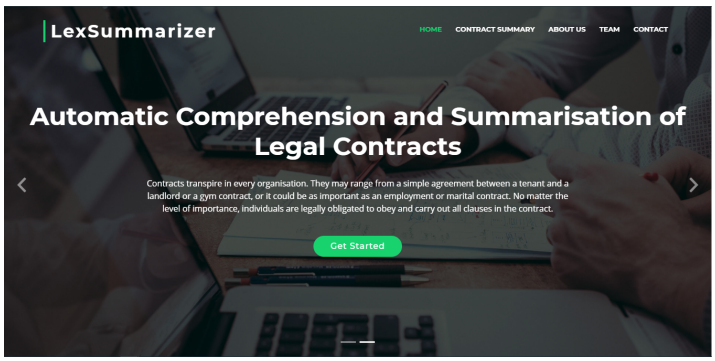


Figure 3: Homepage of the web Application.

Table 1: Example of a summary produced by the tool.

Contract section	Clauses from the tested contract	Information extracted from the contract	Production rule(s)
Acts and Laws	1	The terms and conditions set out herein will constitute the employee's contract with the company. Where a basic condition of employment is not specifically mentioned, the relevant legislation will be applicable (e.g. the Basic Conditions of Employment Act, Act 75 of 1997, the Labour Relations Act, Act 66 of 195, etc.).	32
	11	The payment of maternity leave will be determined in terms of the provisions of the Unemployment Insurance Act, Act 30 of 1966 read together with the provisions of the Basic Conditions of Employment Act, Act 75 of 1997.	34
	12	in the event of death of the employee's spouse or life partner, parent or adoptive parent, grandparent, child, adopted child, grandchild or sibling.	32-34
	17	The company subscribes to the principle of freedom of association as stipulated in the Constitution of the Republic of South Africa, Act 108 of 1996, and thus the employee may join any organisation/trade union of his/her own choice as regulated in the Labour Relations Act, Act 66 of 1995.	32-34
Leave	8	Leave must be applied for in writing in the form and manner prescribed by the company from time to time, and may only be taken after approval by the company or its delegated authority.	21
	9	An application for sick leave must be supported by a certificate from a registered medical practitioner.	21
Remuneration	4	The employee's total monthly remuneration will be R , payable in arrears on.	19
	5	Benefits	18
	6	The remuneration method in 7.2.1 above will be the normal method of remuneration. Should the employer need to use one of the other options due to circumstances, he/she will inform the employee accordingly, preferably in writing, before the commencement of overtime.	19
	13	The employer may not deduct any monies from the employee's salary unless the employee has agreed thereto in this contract or in writing on each occasion.	19
Working Hours	7	The employer may also by agreement grant two (2) paid working days off in lieu of payment.	17
Termination	3	The employer will be entitled to terminate this contract in terms of the Disciplinary Procedures referred to in paragraph 20 hereof.	5-6
Dates and Duration	16	The appointment under this contract is a full time appointment and the employee shall devote his/her full commitment, energy and attention to the employer's business.	27-31
	10	on a date from which a medical practitioner or a midwife certifies that it is necessary for.	27-31
False positive error	2	During the period of employment within the company the employee will report to (the) and obey instructions given by him/her and any other person duly.	N/A
	15	This limitation of trade is restricted to the nature of the employer's business, products and services.	N/A
	14	The employee will not be liable to the employer for information divulged in terms of legislation or a court order compelling him/her to do so.	N/A

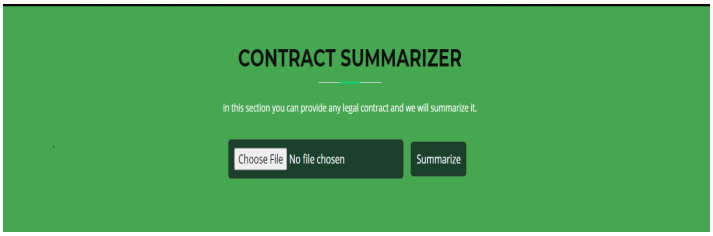


Figure 4: Contract summarisation section of the web Application.

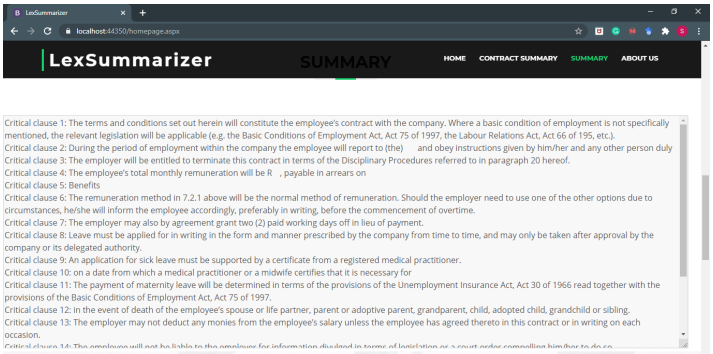


Figure 5: Summary produced by our CFG rules.

6.1 Accuracy of Tool

The analysis was founded by one perspective, which was, the accuracy in identifying critical clauses in a contract. To measure the accuracy of our tool, we tested it on five different contract documents in which all critical clauses were manually identified before the commencement of the test.

From the first contract were able to identify a total of 22 clauses and our tool was able to identify 17 clauses in the same contract document giving us an accuracy of 77.3%. In the second contract, we manually recognised a total of 14 clauses, and the tool identified 11 critical clauses, which gave us an accuracy percentage of 78.6%. Contract three consisted of five critical clauses and the tool recognized four clauses giving us an accuracy of 80%. The fourth contract we tested consisted of 24 manually recognised clauses and the tool was able to recognise 21 clauses from the same contract. The last contract consisted of seven critical clauses and the tool identified four of those clauses which gave an accuracy of 57.1%.

In total, we were able to manually identify 72 critical clauses from the five tested contracts and our tool was able to identify 57 of these clauses. From these results, we were able to calculate an aggregate accuracy percentage of 79.2%. Given that there are different types of contracts that exist, we can observe the vast difference in the tool's accuracy with every test we conducted.

When conducting our tests on the five different contract documents, we recognised that in some cases the tool presented a few

statements that did not fall under the critical clause category in the summary, which indicated that our tool produced false-positive errors. To calculate the rate of the false-positive errors we use the formula $\frac{FP}{FP+TN}$, where FP represented the number of false positives and TN was the number of true negatives. The total number of false-positive errors in each of the tested contracts is as follows, contract one had three, contract two had no false-positive errors, contract three consisted of two, contract four had a total of four false-positive errors and lastly contract five consists of two false-positive errors. After calculations, we identified that the tool produced an aggregate false positive rate of 0.13 (13%).

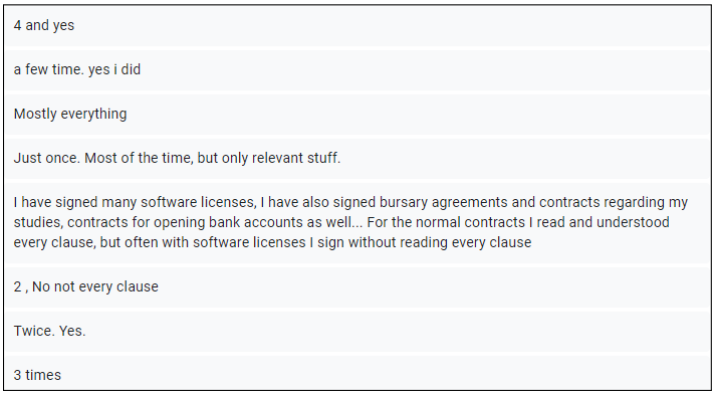


Figure 6: Number for time participants have received a contract.

6.2 *User Perception of Tool*

To make our data collection and analysis process less complicated we divided our online survey into two sections, the first section contains questions based on general information and the user's general knowledge and experience on contract documents. The second section included questions that were based on the user's experience of our automatic comprehension and summarisation tool.

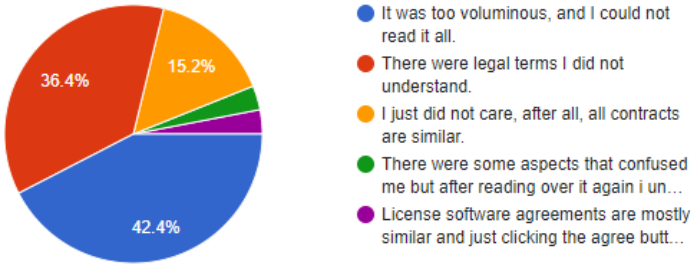


Figure 7: The challenges individual have when reading a contract.

From the raw data, we were able to recognize that of the 39 individuals that have read a contact, at least 26 respondents (66.7%) have been exposed to at least 2 contracts in their lives. We were also able to determine that 29 individuals out of the 39 respondents did not understand every clause in the contract documents they have received and read. This gave us a total percentage of 74.3 respondents that struggled to fully comprehend the clauses in a contract document. We were then able to conclude that the introduction of a tool that can automatically comprehend and summarize legal contract documents could provide great assistance to the majority of individuals who receive contracts but struggle to fully comprehend the terms and clauses found in these contract documents.

Figure 7 outlines that voluminous contracts and the inability to understand legal terms as the main problems individuals generally face when reading a contract. Most of the survey participants agree that tool is useful and relevant.

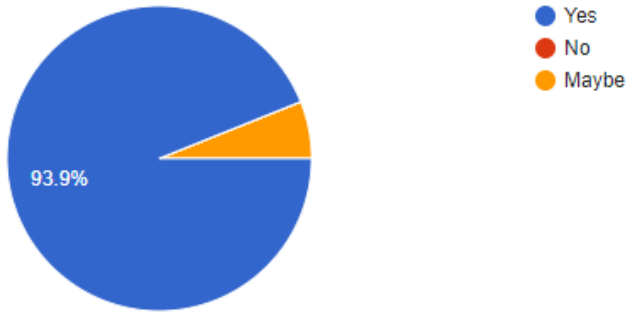


Figure 8: Usefulness of the tool.

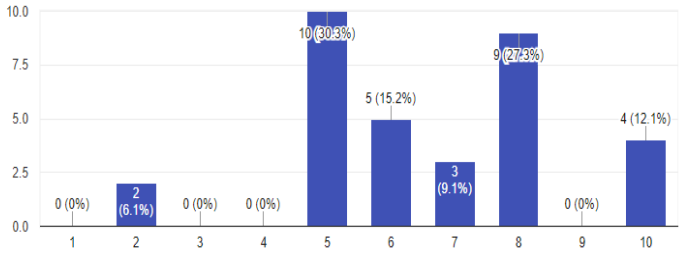


Figure 9: Accuracy of to from participants that tested the tool.

7 Conclusion and Future Work

In this paper we presented a newly formulated approach to automatically comprehend and summarise legal contract documents using formal grammar rules. We started by defining the structure of contracts to gain an in-depth understanding of the different sections found in a contract and their level of importance. We were than able to identify critical clauses that one would normally find in a contract document. From these critical clauses we were able to design CFG rules to outline the syntax of a legal contract document.

Our concept was implemented in a web applications using the .Net framework. When implemented, the web application was able to take any contract document, pre-process the document using the following techniques: PDF-to-Word conversion and lexical analysis. From there the contract is sliced into important sections using a document slicer API. Lastly, the contract is comprehended using our grammar rules and then an accurate summary can be provided.

We were able to perform an in-depth evaluation of our tool based on a performance analysis to measure the accuracy of the tool. The tool was tested on five different contracts and from these tests, we calculated an aggregate accuracy percentage of 79.2%. Furthermore, we identified that the tool produced an aggregate false positive rate

of 0.13 (13%). There was a respectable difference in each of the tests we conducted and we believe these differences were because of the various types of contracts that exist and the way they are structure.

In the near future, we will improve on the comprehension aspect of the presented CFG by designing random forbidden and permitting context grammars.

References

[1] F. Kessler, "Contracts of Adhesion–Some Thoughts about Freedom of Contract," Columbia Law Review, **7**(2), 1–18, 1943, doi:10.2307/1117230.

[2] J. W. Chisholm, "The 1990 contract: Its history and its content," British Medical Journal, 1990, doi:10.1136/bmj.300.6728.853.

[3] A. W. B. Simpson, "The Horwitz Thesis and the History of Contracts," The University of Chicago Law Review, **35**, 115, 1979, doi:10.2307/1599448.

[4] M. J. Horwitz, "The Historical Foundations of Modern Contract Law," Harvard Law Review, 1974, doi:10.2307/1340045.

[5] R. M. Parizi, Amritraj, A. Dehghantanha, "Smart contract programming languages on blockchains: An empirical evaluation of usability and security," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, doi:10.1007/978-3-319-94478-4_6.

[6] P. Steyaert, C. Lucas, K. Mens, T. D'Hondt, "Reuse contracts: Managing the evolution of reusable assets," in Proceedings of the 11th ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, OOPSLA 1996, 1996, doi:10.1145/236337.236363.

[7] L. Budworth, A. Prestwich, B. Sykes-Muskett, K. Khatun, J. Ireland, F. Clancy, M. Conner, "A feasibility study to assess the individual and combined effects of financial incentives and monetary contingency contracts on physical activity," Psychology of Sport and Exercise, 2019, doi:10.1016/j.psychsport.2019.04.021.

[8] W. Zou, D. Lo, P. S. Kochhar, X.-B. D. Le, X. Xia, Y. Feng, Z. Chen, B. Xu, "Smart Contract Development: Challenges and Opportunities," IEEE Transactions on Software Engineering, 2019, doi:10.1109/tse.2019.2942301.

[9] T. Wilhelmsson, C. Willett, "Unfair terms and standard form contracts," in Handbook of Research on International Consumer Law, Second Edition, 2018, doi:10.4337/9781785368219.00012.

[10] F. Gómez Pomar, J. Ganuza, M. Artigot, "Standard contract terms and legal controls: a view from the law and economics theory of contract (forthcoming)," 2018.

[11] M. J. Bommarito, D. M. Katz, E. M. Detterman, "LexNLP: Natural language processing and information extraction for legal and regulatory texts," 2018, doi:10.2139/ssrn.3192101.

[12] N. E. Okwunma, Automatic Comprehension of Customer Queries for Feedback Generation, Ph.D. thesis, 2019.

[13] Y. Kim, J. Lee, E. B. Lee, J. H. Lee, "Application of Natural Language Processing (NLP) and Text-Mining of Big-Data to Engineering-Procurement-Construction (EPC) Bid and Contract Documents," in Proceedings - 2020 6th Conference on Data Science and Machine Learning Applications, CDMA 2020, 2020, doi:10.1109/CDMA47397.2020.00027.

[14] K. Merchant, Y. Pande, "NLP Based Latent Semantic Analysis for Legal Text Summarization," in 2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, 2018, doi:10.1109/ICACCI.2018.8554831.

[15] S. Haiduc, J. Aponte, L. Moreno, A. Marcus, "On the use of automated text summarization techniques for summarizing source code," in 2010 17th Working Conference on Reverse Engineering, 35–44, IEEE, 2010.

[16] S. R. M. Rahim, Z. Z. Mohamad, J. A. Bakar, F. H. Mohsin, N. Isa, "Artificial intelligence, smart contract and islamic finance," Asian Soc Sci, **14**(2), 145, 2018.

[17] C. Dirican, "The impacts of robotics, artificial intelligence on business and economics," Procedia-Social and Behavioral Sciences, **195**, 564–573, 2015.

[18] J. Levin, "Relational incentive contracts," American Economic Review, **93**(3), 835–857, 2003.

[19] M. Suprapto, H. L. Bakker, H. G. Mooi, M. J. Hertogh, "How do contract types and incentives matter to project performance?" International Journal of Project Management, 2016, doi:10.1016/j.ijproman.2015.08.003.

[20] H. Dagan, "Types of Contracts and Law's Autonomy-Enhancing Role," Tourism Tribune, 104–113, 2018.

[21] S. Lyu, et al., "The legal structure of the right of unilateral change in travel contracts," Tourism Tribune, **33**(5), 104–113, 2018, doi:10.1109/20183167907.2018.27.

[22] U. Mattei, "The Comparative Law and Economics of Penalty Clauses in Contracts," The American Journal of Comparative Law, 1995, doi:10.2307/840646.

[23] T. Eisenberg, G. P. Miller, E. L. Sherwin, "Arbitration's Summer Soldiers: An Empirical Study of Arbitration Clauses in Consumer and Nonconsumer Contracts," SSRN Electronic Journal, 2011, doi:10.2139/ssrn.1076968.

[24] J. Zhang, N. M. El-Gohary, "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking," Journal of Computing in Civil Engineering, 2016, doi:10.1061/(asce)cp.1943-5487.0000346.

[25] J. Zhang, N. El-Gohary, "Extraction of construction regulatory requirements from textual documents using natural language processing techniques," in Congress on Computing in Civil Engineering, Proceedings, 2012, doi:10.1061/9780784412343.0057.

[26] P. J. Tierney, "A qualitative analysis framework using natural language processing and graph theory," International Review of Research in Open and Distributed Learning, **13**(5), 173–189, 2012, doi:10.1016/j.irrodl.2012.08.065.

[27] W. Bakari, P. Bellot, M. Neji, "Towards an automatic text comprehension for the arabic questionanswering: Semantic and logical representation of texts," 2018.

[28] F. Bravo-Marquez, G. L'Huillier, P. Moya, S. A. Ríos, J. D. Velásquez, "An automatic text comprehension classifier based on mental models and latent semantic features," in ACM International Conference Proceeding Series, 2011, doi:10.1145/2024288.2024317.

[29] A. Ade-Ibijola, "FINCHAN: A grammar-based tool for automatic comprehension of financial instant messages," in ACM International Conference Proceeding Series, 2016, doi:10.1145/2987491.2987518.

[30] P. W. McBurney, C. McMillan, "Automatic source code summarization of context for java methods," IEEE Transactions on Software Engineering, **42**(2), 103–119, 2015.

[31] A. T. Ying, M. P. Robillard, "Code fragment summarization," in Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, 655–658, 2013.

[32] J. Fowkes, P. Chanthirasegaran, R. Ranca, M. Allamanis, M. Lapata, C. Sutton, "Autofolding for source code summarization," IEEE Transactions on Software Engineering, **43**(12), 1095–1109, 2017.

[33] C. Soria, R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli, "Automatic extraction of semantics in law documents," 253–266, European Press Academic Publishing, 2007.

[34] M. Y. Kim, Y. Xu, R. Goebel, "Summarization of legal texts with high cohesion and automatic compression rate," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, doi:10.1007/978-3-642-39931-2_14.

[35] A. Farzindar, G. Lapalme, "Letsum, an automatic legal text summarizing system," Legal knowledge and information systems, JURIX, 11–18, 2004.

[36] M. Cronje, A. Ade-Ibijola, "Automatic slicer API," personal communication.

[37] M. Cronje, A. Ade-Ibijola, "Automatic slicing and comprehension of CVs," in 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), 99–103, IEEE, 2018.

[38] J. Shallit, A second course in formal languages and automata theory, volume 179, Cambridge University Press Cambridge, 2009, doi:10.1017/CBO9780511808876.

[39] A. Ade-Ibijola, "Synthesis of integration problems and solutions," in 2019 6th International Conference on Soft Computing and Machine Intelligence, ISCMI 2019, 185–190, IEEE, 2019, doi:10.1109/ISCMI47871.2019.9004291.

[40] D. Watson, "Lexical Analysis," in A Practical Approach to Compiler Construction, 37–73, Springer, 2017.

[41] J. Vakhrusheva, S. Khan, R. Chang, M. Hansen, L. Ayanruoh, J. J. Gross, D. Kimhy, "Lexical analysis of emotional responses to "real-world" experiences in individuals with schizophrenia," Schizophrenia Research, 2020, doi:10.1016/j.schres.2019.11.045.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 22-29 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Supervised Learning Techniques for Stress Detection in Car Drivers

Pamela Zontone[1], Antonio Affanni[1], Riccardo Bernardini[1], Leonida Del Linz[2], Alessandro Piras[1], Roberto Rinaldo[*1]

[1]*DPIA, Polytechnic Department of Engineering and Architecture, University of Udine, Udine, 33100, Italy*

[2]*VI-grade Srl, Via G. Galilei 42, Tavagnacco, Udine, 33010, Italy*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 21 October, 2020
Online: 08 November, 2020

Keywords:
Stress Detection in Car Drivers
Electrodermal Activity
Heart Rate Variability
Motion Artifact
Supervised Learning Algorithm
Machine Learning
Deep Learning

ABSTRACT

In this paper we propose the application of supervised learning techniques to recognize stress situations in drivers by analyzing their Skin Potential Response (SPR) and the Electrocardiogram (ECG). A sensing device is used to acquire the SPR from both hands of the drivers, and the ECG from their chest. We also consider a motion artifact removal algorithm that allows the generation of a single cleaned SPR signal, starting from the two SPR signals, which could be characterized by artifacts due to vibrations or movements of the hands on the wheel. From both the cleaned SPR and the ECG signals we compute some statistical features that are used as input to six Machine Learning Algorithms for stress or non-stress episodes classification. The SPR and ECG signals are also used as input to Deep Learning Algorithms, thus allowing us to compare the performance of the different classifiers. The experiments have been carried out in a firm specialized in developing professional car driving simulators. In particular, a dynamic driving simulator has been used, with subjects driving along a straight road affected by some unanticipated stress-evoking events, located at different positions. We obtain an accuracy of 88.13% in stress recognition using a Long Short-Term Memory (LSTM) network.

1 Introduction

Lately[1], the assessment of stress in car drivers is getting more attention, because a highly stressed driver could be less focused, more likely to drive in a risky way [1], and/or to get involved into accidents [2]. The emotional state, corresponding to a condition of tension and pressure in an individual, causes reactions of the sympathetic nervous system. These are defined as stress in literature [3].

Different approaches have been proposed to detect stress in individuals, using physical or physiological parameters [4]. Machine Learning (ML) algorithms are often employed [5]– [8]. In [9], a Support Vector Machine (SVM) classifier is used, with Skin Conductance Response (SCR), Electroencephalogram (EEG), and Heart Rate (HR) as inputs. In [10], various signals are utilized, such as Blood Volume Pulse (BVP), Skin Conductance (SC), Skin Temperature (ST), Electromyography (EMG), and Respiration (RESP). Deep Learning (DL) algorithms have also been successfully applied for stress recognition. Physiological signals are used, such as Electrocardiogram (ECG) [11] and Electrodermal Activity (EDA) [12], or facial expressions [13], or a collection of physiological and motion data, including ECG, EDA, EMG, RESP, Temperature (TEMP), and Acceleration (ACC) [14]. The system described in [15] uses the time differences between consecutive R-waves of QRS complexes, also called RR intervals, for mental stress identification in firefighters. In several works, a comparison between ML and DL methods for stress recognition has been carried out, also showing that Convolutional Neural Networks (CNNs) can outperform ML conventional methods [16, 17].

There are various stress analysis approaches when we consider car driving scenarios as well. Many use physical characteristics, e.g., facial or eye images [18, 19]. In others, physiological characteristics are considered [20–22]. Among these, EDA can be used jointly with different signals [23, 24]. Authors in [25] study the relation between the vehicle's acceleration and the driver's heart rate in elders. In [26], EEG patterns, belonging to different driving conditions, are classified using an SVM, a Neural Network (NN), and a Random Forest (RF). Some DL algorithms have also been successfully applied for stress recognition in the driving context as well, and are particularly relevant to this work. Authors in [27] present an end-to-end architecture, composed of convolutional layers, that uses ECG signals for stress detection. Their experiments are designed for identification of mental fatigue and for discrimina-

[*]Corresponding Author: Roberto Rinaldo, roberto.rinaldo@uniud.it
[1]This paper is an extension of the work originally presented in [43].

tion between driving and resting conditions. A multimodal system based on CNNs and Long Short-Term Memory (LSTM) networks, using ECG, vehicle data, and contextual data has been proposed in [28]. In [29], CNNs are considered to detect a driver's braking intention, in different driving conditions, using EEG data.

In this work we present a system that allows the recognition of the emotional and stress conditions of subjects while driving. The main features of our system, which differentiate our work from others in the literature, are the relative simplicity of the sensors used, the possibility of being used nearly in real time on 15 s signal blocks, and the ability to discriminate stressful events of short duration during driving. The experiments are organized in a company that designs professional driving simulators. We use a simulator that reproduces a road track in a 3D environment. Along the track we place different tasks to go through, which are expected to cause a reaction in the subjects. Two different signals are considered, the Skin Potential Response (SPR) and the ECG. We have proven that these signals are suitable for the analysis of the emotional condition of a driver and that can be acquired with an acceptably intrusive setup, with no intervention on the vehicle [30].

We analyze the SPR signal instead of the widely utilized SCR, because it can be logged in a more straightforward way, with no electric current in the skin. In addition, the electrode impedance and skin impedance changes have a lower impact on this signal [31]. In detail, our objective is to compare the performance of several ML algorithms in recognizing the stress responses of the drivers that undergo various stress situations. To do this, some features are extracted from both the SPR and ECG signals. In addition, we consider deep learning techniques for driver stress recognition. These require the raw recorded signals as input, contrary to what happens in traditional machine learning techniques, where a feature extraction process is needed. Specifically, we use a CNN and an LSTM network. Compared to other similar studies cited above [24, 27], we apply machine and deep learning techniques for detecting stress during short tasks in a controlled and simulated environment.

The paper is structured as follows. In the next section, a description of the fundamental blocks of the proposed system is provided. We introduce the sensing device utilized for both SPR and ECG recordings, the Motion Artifact (MA) removal algorithm developed to reduce motion artifacts, and the learning algorithms implemented for classification. Section 3 reviews the experimental setup. Section 4 presents the results of our comparative study, using different classification algorithms, on drivers' stress recognition. Finally, in Section 5 we draw some conclusions.

2 Description of the system

The block diagram of the developed system is shown in Figure 1. The inputs are the two SPR signals, recorded from the hands of a test subject ($s_1(t)$ and $s_2(t)$ in the figure), and the ECG signal recorded from the chest of the subject. The signals are acquired synchronously through a sensing device that uses a textile garment with electrodes, in addition to other electrodes placed under the gloves on both hands.

The two SPR signals can exhibit some artifacts that can be generated when the subject is driving, e.g., due to the hands movements or vibrations. To overcome this problem, we designed a MA removal algorithm that, taking into account the local energy of the two SPR signals, picks the smoother one, with the aim to remove spurious peaks and asymmetric signal spikes, possibly related to MA. As a matter of fact, the SPR component due to the sympathetic nervous system activity in the two hands is highly correlated. In the end, a learning classification algorithm is used to recognize the presence of stress in certain time intervals. In detail, every single time interval is classified into two possibile classes, which are the stress class (or "1") and the non-stress class (or "0"). In case of ML algorithms, this is carried out by selecting some statistical features from the cleaned SPR signal ($s(t)$ in Figure 1) and from the ECG signal. These features have been picked out, among others, since they well represent the stress signal peculiarities. In case of DL algorithms, instead, the raw SPR and ECG recordings are directly processed. In the next subsections, we will describe in detail each one of these fundamentals blocks.

2.1 The sensing device

The sensor used for data acquisition is the VI-BioTelemetry system (https://www.vi-grade.com/en/solutions/vi-biotelemetry/). This commercial sensor, shown in Figure 2, has been designed in our group and engineered by the VI-grade company. Its detailed description can be found in several recent scientific papers (e.g., [31, 32]).

Summarizing its features and specifications, the sensor is battery operated and transmits data to a laptop via WiFi connection. It acquires two SPR channels and three ECG channels. The SPR signals are acquired from the hands of the tested subject (through Ag/AgCl wet electrodes positioned on the palm and the back of each hand) and the ECG data are acquired using a commercial vest having textile electrodes in contact with the chest of the subject (using an adhesive conductive gel). The analog front end conditions the SPR and ECG signals with proper bandwidth and gain. Analog signals are converted using a 12-bit A/D and an on board DSP which acquires the signals at 1 kSa/s rate. The A/D converter is characterized by an Integral Nonlinearity (INL) lower than 1 LSB (Least Significant Bit) and an Effective number of bits (ENOB) equal to 11.1. The DSP sends the acquired signals to the WiFi module which is responsible to send all the time-aligned data to a server in the local area network.

Regarding the ECG channels, they are basically designed as band-pass differential amplifiers with maximum input range ± 5 mV (i.e., gain 370) and bandwidth in the [0.03, 160] Hz range. The input impedance of each channel is 100 MΩ, in order to reduce the load error to less than 1%, since the skin impedance is lower than 1 MΩ. The resolution of the ECG signals acquired on the chest skin results to be in the order of 2.5 μV and the accuracy of the ECG readout has been characterized in the past [32] to be as low as 0.1%. As for the ECG signal, we first detect the R-peak locations using the Pan-Tompkins algorithm [33].

The instantaneous RR signal is obtained by converting the RR interval time series with cubic spline interpolation at 100 Sa/s. Similarly to what is done in [34], where RR values and DL are used for Congestive Heart Failure (CHF) detection, interpolation is needed to synchronize signal segments, in our case with the SPR signal blocks. Note that this can be important to preserve correct time and

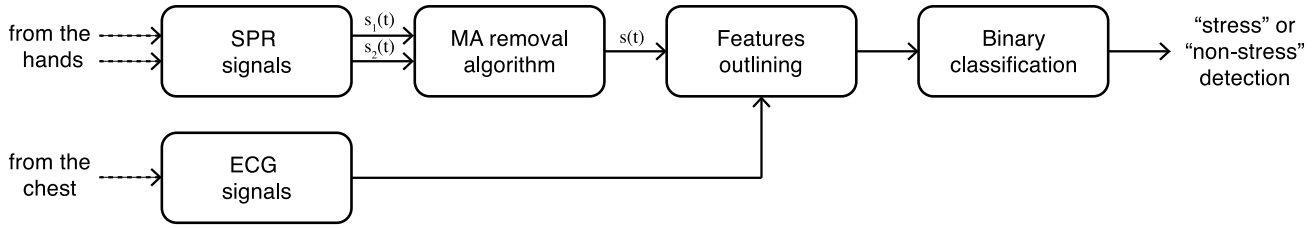


Figure 1: Block diagram of the presented system.

frequency information (see [35] and references therein), which the DL classifiers can possibly extrapolate from the data.

Regarding the SPR channels, the sensors acquire the electrodermal signals on both hands as differential voltages between the palm and the back of each hand. The SPR signals are supposed to be in the ±10 mV range. They are amplified with a gain equal to 160. The acquisition bandwidth is in the [0.08, 40] Hz range; the input impedance of each channel is 100 MΩ. The resolution of the SPR data acquired on the hands results in the order of 5 μV and the accuracy of the SPR readout has been characterized in the past [31] to be in the order of 0.2%. The SPR signals are subsampled at 100 Sa/s before further processing.

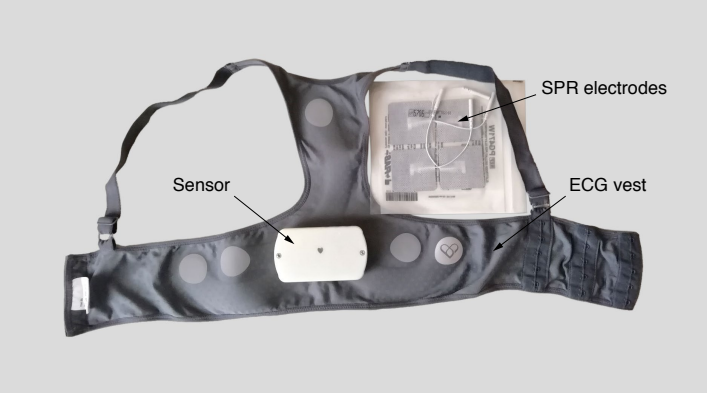


Figure 2: The VI-BioTelemetry sensor used in the experiments. The system is composed by an acquisition box, a vest with textile electrodes for ECG acquisition on the chest and Ag/AgCl electrodes for SPR acquisition on the hands.

2.2 Motion artifact removal block

As introduced above, some artifacts due to the movements and actions of the subject's hands while driving can become visible on the SPR signals, and could alter their shape. To take this issue into account, we propose an MA algorithm (see also [36, 37]) that is built on two assumptions: the first being that motion artifact enhances the local energy of the signal. The second being that the motion artifacts rarely appear simultaneously in the SPR signal of both hands. We observed this behaviour in our testing most of the time, and the proposed solution appears to be a good compromise to lessen the MA. Considering the RMS values σ_1 and σ_2 of the SPR input signals s_1 and s_2 on a $L = 100$ samples wide moving window, corresponding to 1 s at a sample rate of 100 Sa/s, we establish a smoothing threshold function $f(x)$ in order to obtain $f(x) \approx 0$ when $0 \leq x \ll 1$ and $f(x) \approx 1$ when $1 \ll x \leq 2$, namely,

$$f(x) = \frac{1}{1 + e^{-2(x-1)}} \tag{1}$$

This function is employed in a correction term β, that enables us to reduce motion artifacts. At the nth sample, this correction term is calculated as:

$$\beta(n) = \begin{cases} f\left(\frac{\sigma_1(n)}{\sigma_2(n)}\right) & \text{if} \quad \sigma_2(n) \neq 0 \\ 1 & \text{if} \quad \sigma_2(n) = 0 \end{cases} \tag{2}$$

We can deduce that $\beta \to 0$ when $\sigma_1 < \sigma_2$ and $\beta \to 1$ when $\sigma_1 > \sigma_2$. The correction term β, as in (2), is then used to determine the clean single output $s(n)$ of the procedure as

$$s(n) = \beta(n) \cdot s_2(n) + \left[1 - \beta(n)\right] \cdot s_1(n) \tag{3}$$

In the end, the output in (3) imitates the signal characterized by the lower energy in the 1 s examined interval.

2.3 Classification algorithms

In the proposed system, we utilize a classification technique to detect the presence (or absence) of a driver's stress condition in a given time interval during a car driving simulation. For this classification task, we utilize and compare six different supervised machine learning algorithms: an SVM classifier, a Naïve Bayes classifier, a k-Nearest Neighbors (k-NN) classifier, a Decision Tree (DT), an RF classifier, and an Artificial Neural Network (ANN). All these algorithms have been widely used in the literature for classification purposes. We also use two deep learning algorithms, in particular a 1D CNN and an LSTM network.

Our goal is to classify all of the time intervals in two possible classes: the stress class (or "1") and the non-stress class (or "0"). For this, we use the processed and cleaned SPR signal, and the ECG signal, with the latter one processed to remove the ectopic beats similarly to what is proposed in [35]. In addition, a further normalization preprocessing step is applied to both SPR and RR signals, because they can differ considerably among the individuals. This step consists of the standardization of each signal, based on the mean and variance of the preceding 5 minutes of the current signal value. We decided to process a 15 s long interval at a time, and we take a new interval every 5 s. In this way each interval covers the preceding one by 10 s. We do this to keep a reasonably low processing delay. The overall number of features extracted from these time intervals is nine: five from the SPR signal and four from the ECG signal. The five statistical features considered for the SPR signal

are: the interval variance, the energy, the mean absolute value, the mean absolute derivative and the max absolute derivative. The four statistical features from the ECG signal are: the mean value of RR intervals, the standard deviation of RR intervals (SDNN), the root mean square of subsequent RR interval differences (RMSSD) and the mean value of the Heart Rate [38]. We gather the features of the two signals together to construct the feature vectors. Each feature is then normalized in the range [0, 1]. All of the ML classifiers have been set up using Matlab (2017.a), except for the ANN that has been set up using Python. Hyperparameters are optimized automatically by the Matlab routine functions used for classification. In addition, for the SVM classifier, a Radial Basis Function (RBF) kernel has been used. The Bayesian optimization has also been used during the training procedure. For the ANN, we implement a four-layer perceptron trained with the back-propagation algorithm. We utilize Keras and the optimization package called "hyperopt" [39], allowing the use of the Bayesian optimization in this case as well. The optimization process also allows us to select the number of nodes in the two hidden layers (ranging from 16 to 128). The drop-out percentage has also been computed by the optimization package. Regarding the cross validation methodology, the 10-fold cross validation is used for all of the ML algorithms. For the ANN classifier we set the batch-size equal to 128.

The DL algorithms are also set up using Python. In detail, the architecture of the CNN used for classification has been inspired by the one proposed in [28], and its parameters are briefly reported in Table 1. The CNN contains two identical sequences of four operations: a convolutional layer, an Exponential Linear Unit (ELU), a batch normalisation layer and a max pooling layer. These are followed by a fully-connected layer (of 128 nodes), and the output layer where the softmax activation function is used. As anticipated, the inputs of the CNN are no longer the feature vectors, but the unprocessed samples of both the SPR and RR signals recorded from the subject at the same moment, in time intervals 15 s long. For each time interval, the values of the SPR and the RR signals must be arranged side by side. Since each 15 s long time interval is made of N = 1500 samples, with a sample rate of 100 Sa/s for both the SPR and RR signals, we obtain an input matrix with a size of 1500×2, to be sent as input to the network. We proceed picking a new interval every 5 seconds and applying the same procedure. The labels corresponding to each interval are the same already derived for the ML algorithms. Other details about how the CNN model is defined are the following: we use the Adam optimizer, the learning rate has been set equal to 0.001, the categorical cross-entropy has been chosen as loss function, the number of epochs has been set equal to 20 and the batch-size to 64 (https://keras.io).

Another DL algorithm, i.e., an LSTM network, has been used in the experiments. This is a particular type of Recurrent Neural Network (RNN) that overcomes the issue of the short-term memory of the RNNs (also known as the vanishing gradient problem). The LSTM internal gates keep the relevant information in a long sequence of data, throwing away the non-relevant data, and this can help us recognizing stress with a long dynamic temporal behavior. The architecture of the implemented LSTM network is reported in Table 2. The same settings introduced before for the CNN (the learning rate, the number of epochs, etc.) have been used for the LSTM as well, with batch-size 128. Regarding this network, we also consider a variant of it which uses the hyperopt optimization package (it will be denoted as LSTM2 in Section 4). In this case the number of nodes (ranging from 16 to 128) and the dropout value (varying between 0 and 1), for each layer, have been computed by the optimization package.

Table 1: Architecture of the implemented CNN

CNN	
Convolutional layer	Filters = 20, Kernel size = 10, Stride = 2
Exponential Linear Units (ELU)	Alpha = 0.1
Batch norm. + dropout (0.15)	
Max pooling	Pool size = 2, Stride = 2
Convolutional layer	Filters = 20, Kernel size = 10, Stride = 2
Exponential Linear Units (ELU)	Alpha = 0.1
Batch norm. + dropout (0.15)	
Max pooling	Pool size = 2, Stride = 2
Flatten	
Dense + dropout (0.5)	Nodes = 128, Activation = ReLU
Dense	Nodes = 2, Activation = Softmax

We employ the "leave-one-person-out" technique for all of the learning algorithms, such that the training of each classifier is performed using the data of all individuals, except one, on which the classifiers will be tested. This process is carried out for each individual, and in the end the overall system performance is determined by computing the average of the results originated from all of the individuals.

Table 2: Architecture of the implemented LSTM network

LSTM	
LSTM layer	Nodes = 128, Activation = ReLU
Dropout	0.4
LSTM layer	Nodes = 128, Activation = ReLU
Dropout	0.4
Dense	Nodes = 2, Activation = Softmax

3 Experimental setup

The experiments, as introduced above, are carried out in a firm that develops professional driving simulators, both in hardware and software. For our tests, we use a dynamic driving simulator that allows the reproduction of the same movements of a real car. In fact, it provides motion feedback to the drivers through a nine degrees-of-freedom moving platform. We enrolled 16 healthy individuals, 22-47 years old, some from the University of Udine and some from the University of Padua. The subjects were asked to drive on a 28 km long straight highway track, with four stress-inducing obstacles demanding some concentration to get over. These four obstacles are shown in Figure 3. From left to right and from top to bottom, they are: Sponsor block (from right to left), Tire labyrinth, Double lane change (right to left) and Sponsor block (from left to right).

All subjects prepare themselves before entering the simulator. They wear the sensing devices both on chest and hands. They gave consent having their physiological signals recorded, and the test were performed according to the principles of the Declaration of Helsinki.

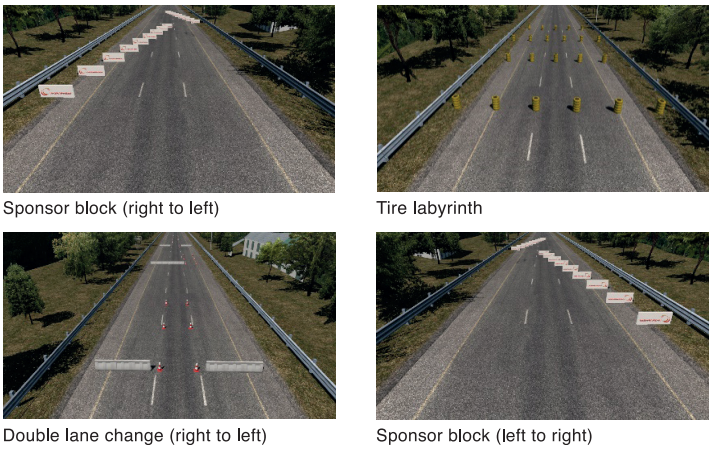


Figure 3: The four stress-inducing events located along the road.

Figure 4 displays the cleaned SPR signal, after the MA removal block, and the instantaneous HR signal of a subject (derived from the RR signal) during the simulation. The start of an event is signaled with a vertical line with a circle marker, and the end of an event with a vertical line with a square marker. We also assume an event starts when the obstacle enters the field of view of the subject (and this happens approaching a distance of 800 meters from the obstacle), and ends 30 seconds after surpassing the obstacle. Passing each obstacle takes approximately 65 seconds. The peculiar stress peaks shape in the SPR signal during the stress event can be seen evidently in the figure, just as the distinctive rise in Heart Rate.

4 Experimental Results

The classifiers, which allow the recognition of stress in each 15 s long interval, are built using the cleaned SPR signal and the ECG signal of each subject. For the traditional ML algorithms the features of the signals are extracted from each interval. For the DL algorithms there are no features to be extracted, and the raw signals themselves are considered for each interval. In all cases, the intervals overlap each other by 10 s as already specified.

We assume that all the intervals happening in a stretch of the road where obstacles are present, to have the value "1", with stress, and that all the intervals happening in a stretch of the road without obstacles to have the value "0", with no stress. As already mentioned, we employ the "leave-one-person-out" method in the classification process. So, for the training, we examine 15 subjects at a time out of the 16 in total. In this way, we obtain a total number of intervals ranging from 1176 to 1209, for both stress and non-stress classes. The total number of intervals for training is not fixed because each individual overcomes the obstacles with slightly different timing and speed. During the test we obtain approximately 187 intervals to analyze for each individual (this is an average value taking all of the subjects into consideration), where the majority of them, about two thirds, belong to the non-stress class.

According to the assumed ground-truth, the True Positives (TP), the False Negatives (FN), the True Negatives (TN) and the False Positives (FP) can be computed, as well as the performance indicators like Accuracy, Sensitivity and Specificity:

$$\text{Accuracy } (\%) = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \cdot 100 \tag{4}$$

$$\text{Sensitivity } (\%) = \frac{\text{TP}}{\text{TP+FN}} \cdot 100 \tag{5}$$

$$\text{Specificity } (\%) = \frac{\text{TN}}{\text{FP+TN}} \cdot 100 \tag{6}$$

Due to the fact that our test data are unbalanced (in fact we have less stress intervals than non-stress intervals), two additional indicators are also important, the Balanced Accuracy (BA) and the Geometric Mean (GM), defined as

$$\text{BA } (\%) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{FP+TN}}\right) \cdot 100 \tag{7}$$

$$\text{GM } (\%) = \sqrt{\frac{\text{TP}}{\text{TP+FN}} \cdot \frac{\text{TN}}{\text{FP+TN}}} \cdot 100 \tag{8}$$

Table 3 presents the overall performance of our system (MEAN ± STD), calculated by averaging the results of the 16 subjects, for all of the considered ML algorithms. The accuracy and balanced accuracy values among the classifiers are close, even if the ANN classifier is slightly better than the other ones in terms of accuracy, and the RF in terms of balanced accuracy. Moreover, the sensitivity of the RF is higher than the others, as it is the specificity of the Naïve Bayes classifier. The GM values are similar for all of the ML algorithms, with the RF classifier still performing marginally better than the others.

As introduced before, as a first attempt to use a deep learning algorithm to recognize stress and non-stress episodes in our scenario, we set up a 1D CNN and an LSTM. We compute the overall performance of the CNN and the LSTM using the test set, and we obtain the values reported in Table 3. So, comparing the results of the DL algorithms with the ones obtained applying the aforementioned ML algorithms, we can notice that the CNN architecture performs similarly to the other ML architectures. In particular, the accuracy of the ANN and the specificity of the Naïve Bayes classifier exceed the values corresponding to the CNN classifier, whereas the CNN sensitivity is better than that obtained with the other ML classifiers. However, when looking at the LSTM, we can see that it works well in detecting the stress episodes, outperforming all the other algorithms in terms of accuracy, sensitivity, balance accuracy and geometric mean. The specificity of the LSTM, instead, is lower than the Naïve Bayes one. The variant LSTM2 performs even better, and all of the performance indicators are higher than the other DL and ML algorithms.

The overall performance of the system looks favourable. It should be noted that some of our hypotheses are critical and can affect the performance indicators. The first hypothesis we make is that all the time intervals happening during the overcoming of an obstacle by a driver should originate stress, even if it is possible that the driver is not constantly stressed during the manoeuvre. The second one is that a driver is not stressed when not overcoming an obstacle. This might not be true at all times, since a driver can be in a condition of stress during those time intervals for causes unknown to us.

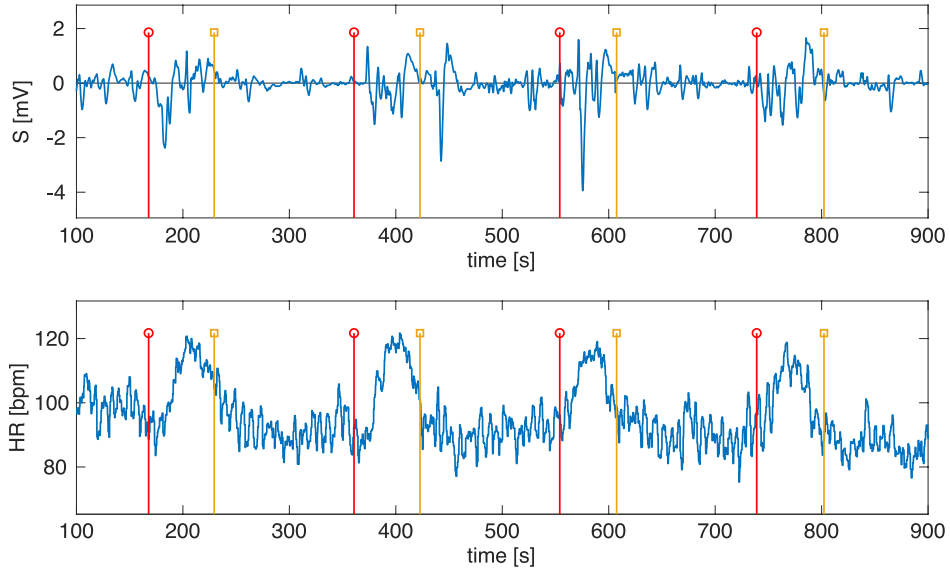


Figure 4: SPR and HR signals (at the top and the bottom, respectively) for a subject during the driving simulation. Each event onset is evidenced with a red vertical line with a circle marker, and each event offset with a yellow vertical line with a square marker.

Table 3: Performance of all the supervised learning algorithms

	Accuracy (%)	Sensitivity (%)	Specificity (%)	BA (%)	GM (%)
SVM	72.86 ± 8.77	86.69 ± 7.79	66.46 ± 13.02	76.57 ± 6.91	75.40 ± 7.85
Naïve Bayes	74.04 ± 6.53	82.83 ± 8.04	70.27 ± 10.85	76.55 ± 4.65	75.81 ±5.01
k-NN	72.07 ± 6.01	86.60 ± 8.30	65.99 ± 8.77	76.30 ± 4.96	75.26 ± 5.23
DT	70.76 ± 8.88	85.16 ± 6.61	64.80 ± 13.13	74.98 ± 5.98	73.69 ± 7.20
RF	74.25 ± 6.17	87.05 ± 7.60	68.83 ± 8.83	77.94 ± 5.03	77.11 ± 5.35
ANN	74.45 ± 8.65	86.19 ± 8.13	69.00 ± 13.38	77.59 ± 6.50	76.56 ± 7.30
CNN	73.97 ± 6.23	87.99 ± 6.40	67.76 ± 8.30	77.87 ± 4.82	76.99 ± 5.27
LSTM	82.56 ± 3.04	88.61 ± 7.55	69.79 ± 17.61	79.20 ± 5.96	77.46 ± 9.23
LSTM2	88.02 ± 0.97	93.60 ± 1.92	76.26 ± 4.22	84.93 ± 1.55	84.43 ± 1.78

Another possible critical aspect of our hypotheses is that part of the signal responses may be due to the increased physical effort during the events. However, in some previous works, e.g., [40], we observed that the intensities of the signals during designated stress events were significant, and could not be confused with the levels recorded during non-stress intervals, even if a comparable physical activity was required for driving. In [41], we also evaluated a car driver's stress condition in a simulated autonomous driving scenario. In that experiment, we submitted questionnaires to the drivers and evaluated other parameters such as the User Experience (UX) measure and the NASA Task Load Index. Questionnaires supported our findings derived by analyzing the SPR and ECG signals (see also [42]).

To account for these considerations, and enhance the stress recognition ability, we apply a re-label step to the output of the classifiers (see Figure 5). To do so, we assume stress is happening if a minimum of four consecutive intervals are marked as positive (thus covering a time span of 30 s). In the end, the blocks of four or more consecutive positive intervals are kept, whereas blocks of up to three consecutive positive intervals are instead re-labelled as negative.

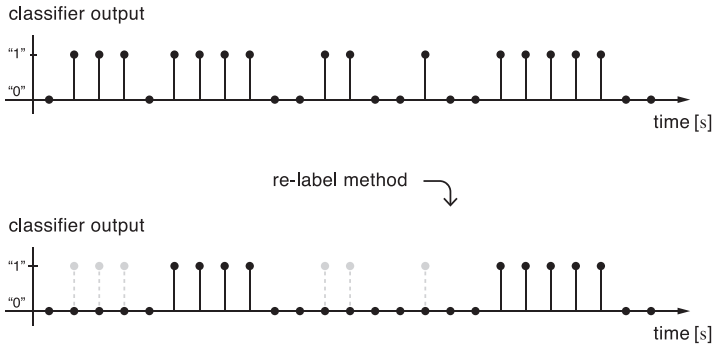


Figure 5: The re-label method applied to the output of each classifier.

In Table 4 the performance indicators applying the re-label method are presented. The values are slightly higher than the ones in Table 3. On the other hand, sensitivity is a little smaller for all of the classifiers, since we noticed that a number of stress intervals, which were correctly detected before, are now discarded by the re-label step. The performance indicators of the LSTM2 are clearly higher, when compared to the other algorithms, in this scenario as well.

Table 4: Performance of all the supervised learning techniques when the re-label method is applied to the output of each classifier

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Balanced Accuracy (%)	Geometric Mean (%)
SVM	77.61 ± 8.14	85.11 ± 8.82	74.05 ± 11.67	79.58 ± 6.89	79.00 ± 7.30
Naïve Bayes	77.39 ± 6.69	82.83 ± 8.04	70.27 ± 10.85	76.55 ± 4.65	75.81 ±5.01
k-NN	75.86 ± 5.84	83.85 ± 11.04	72.61 ± 8.95	78.23 ± 5.47	77.56 ±5.63
DT	76.17 ± 7.87	82.42 ± 8.09	73.69 ± 11.41	78.06 ± 5.85	77.50 ±6.16
RF	77.15 ± 6.39	86.16 ± 7.42	73.32 ± 8.89	79.74 ± 5.35	79.22 ±5.59
ANN	78.17 ± 9.13	84.73 ± 9.72	75.16 ± 14.23	79.94 ± 6.88	79.15 ± 7.50
CNN	79.47 ± 6.45	85.31 ± 8.33	77.09 ± 8.48	81.20 ± 5.79	80.87 ± 5.90
LSTM	83.21 ± 3.27	88.23 ± 7.78	72.62 ± 17.39	80.42 ± 5.94	78.91 ± 9.00
LSTM2	88.13 ± 0.87	93.30 ± 1.92	77.20 ± 4.05	85.25 ± 1.42	84.83 ± 1.60

5 Conclusions

We presented a scheme to recognize individuals' stress while driving in a car driving simulator. The proposed system uses relatively simple sensors, with an acceptably intrusive setup. Due to the short signal block analysis, it can be used nearly in real time and can discriminate stressful events of short duration. In our system, we record two SPR signals for each individual, one for each hand, and the ECG signal. We then apply an MA removal algorithm to remove the artifacts that could appear in the SPR signals during the drive, so that the output is a single cleaned SPR signal. Some features are extracted from both the cleaned SPR signal and the ECG, and sent as input to an ML algorithm. The raw signals are instead directly used as input to DL algorithms. In particular, we evaluate the performance of an SVM classifier, a Naïve Bayes classifier, a *k*-NN classifier, a DT, an RF classifier and an ANN. The performance of all the ML classifiers are similar one to each other, however, the ANN classifier performs better than the others in terms of accuracy. The RF classifier outperforms the others in terms of sensitivity, balanced accuracy and geometric mean, while the Naïve Bayes classifier in terms of specificity. We also report the results obtained with a 1D CNN and an LSTM. The CNN outperforms the other ML classifiers in terms of sensitivity. The LSTM, when some parameters are selected by the optimization procedure, outperforms all the other learning algorithms. A re-labelling procedure is also considered, and the performance indicators of all of the algorithms are compared again in this case. We obtain the highest accuracy, equal to 88.13%, using the optimized LSTM.

Some critical aspects of the proposed experiment are due to the assumption that stress is present only during the arranged events, and that part of the signal responses may be due to the increased physical effort while driving through the events. We have observed that these issues still allow us to interpret the results with a good level of confidence. We conclude that the overall performance of our system is encouraging, showing that by using the SPR and ECG signals we can identify the stress in drivers, deriving from some demanding driving activities, with acceptable accuracy.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We are grateful to VI-grade S.r.l., Tavagnacco (UD), Italy, and in particular to Fabio Formaggia, Diego Minen, Michela Minen, and Carlo Savorgnan, for their assistance in data acquisition with their professional driving simulator.

References

[1] M. Navon, O. Taubman-Ben-Ari, "Driven by emotions: The association between emotion regulation, forgivingness, and driving styles," Transportation Research Part F: Traffic Psychology and Behaviour, **65**, 1–9, 2019, doi:10.1016/j.trf.2019.07.005.

[2] C. D. Katsis, N. Katertsidis, G. Ganiatsas, D. I. Fotiadis, "Toward emotion recognition in car-racing drivers: A biosignal processing approach," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, **38**(3), 502–512, 2008, doi:10.1109/TSMCA.2008.918624.

[3] W. Boucsein, Electrodermal Activity, Springer, 2012.

[4] S. Greene, H. Thapliyal, A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," IEEE Consumer Electronics Magazine, **5**(4), 44–56, 2016, doi:10.1109/MCE.2016.2590178.

[5] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, J.-H. Chen, "EEG-based emotion recognition in music listening," IEEE Transactions on Biomedical Engineering, **57**(7), 1798–1806, 2010, doi:10.1109/TBME.2010.2048568.

[6] J. Zhang, W. Wen, F. Huang, G. Liu, "Recognition of Real-Scene Stress in Examination with Heart Rate Features," in 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2017, doi:10.1109/IHMSC.2017.13.

[7] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, L. Albera, "Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources," IEEE Transactions on Affective Computing, **11**(2), 244–257, 2020, doi:10.1109/TAFFC.2017.2768030.

[8] Z. Qin, M. Li, L. Huang, Y. Zhao, "Stress level evaluation using BP neural network based on time-frequency analysis of HRV," in 2017 IEEE International Conference on Mechatronics and Automation (ICMA), Takamatsu, 2017, doi:10.1109/ICMA.2017.8016090.

[9] M. Moghimi, R. Stone, P. Rotshtein, "Affective recognition in dynamic and interactive virtual environments," IEEE Transactions on Affective Computing, **11**(1), 45–62, 2020, doi:10.1109/TAFFC.2017.2764896.

[10] C. Maaoui, A. Pruski, F. Abdat, "Emotion Recognition for Human-Machine Communication," in 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, 2008, doi:10.1109/IROS.2008.4650870.

[11] G. Giannakakis, E. Trivizakis, M. Tsiknakis, K. Marias, "A novel multi-kernel 1D convolutional neural network for stress recognition from ECG," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, United Kingdom, 2019, doi:10.1109/ACIIW.2019.8925020.

[12] K. T. Chui, M. D. Lytras, and R. W. Liu, "A generic design of driver drowsiness and stress recognition using MOGA optimized deep MKL-SVM," Sensors, **20**(5):1474. 2020, doi:10.3390/s20051474.

[13] L. Lombardi, F. Marcolin, Psychological Stress Detection by 2D and 3D Facial Image Processing, Progresses in Artificial Intelligence and Neural Systems, Springer, 2020.

[14] M. T. Uddin, S. Canavan, "Synthesizing Physiological and Motion Data for Stress and Meditation Detection," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, United Kingdom, 2019, doi:10.1109/ACIIW.2019.8925245.

[15] A. Oskooei, S. Mai Chau, J. Weiss, A. Sridhar, M. Rodríguez Martínez, B. Michel, "DeStress: deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (HRV) data," arXiv:1911.13213v1, 2019.

[16] B. Hwang, J. You, T. Vaessen, I. Myin-Germeys, C. Park, B.-T. Zhang, "Deep ECGNet: An optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals," Telemedicine and e-Health, **24**(10), 753–772, doi:10.1089/tmj.2017.0250.

[17] F. Al Machot, A. Elmachot, M. Ali, E. Al Machot, K. Kyamakya, "A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors," Sensors, **19**(7):1659, 2019, doi:10.3390/s19071659.

[18] T. Kundu, C. Saravanan, "Advancements and recent trends in emotion recognition using facial image analysis and machine learning models," in 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, 2017, doi:10.1109/ICEECCOT.2017.8284512.

[19] Y. Ishii, T. Ogitsu, H. Takemura, H. Mizoguchi, "Real-time eyelid open/closed state recognition based on HLAC towards driver drowsiness detection," in 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO), Bali, 2014, doi:10.1109/ROBIO.2014.7090707.

[20] D. S. Lee, T. W. Chung, B. G. Lee, "Stress events detection of driver by wearable glove system," IEEE Sensors Journal, **17**(1), 194–204, 2017, doi:10.1109/JSEN.2016.2625323.

[21] Y. J. Yu, Z. Yang, B.-S. Oh, Y. K. Yeo, Q. Liu, G.-B. Huang, Z. Lin, "Investigation on driver stress utilizing ECG signals with on-board navigation systems in use," in 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), Phuket, 2016, doi:10.1109/ICARCV.2016.7838780.

[22] D. Tomoi, W. Wen, H. Yamakawa, S. Hamasaki, K. Takakusaki, Q. An, Y. Tamura, A. Yamashita, H. Asama, "Estimation of stress during car race with factor analysis," in 2015 International Symposium on Micro-NanoMechatronics and Human Science (MHS), Nagoya, 2015, doi:10.1109/MHS.2015.7438258.

[23] G. Rigas, Y. Goletsis, D. I. Fotiadis, "Real-time driver's stress event detection," IEEE Transactions on Intelligent Transportation Systems, **13**(1), 221–234, 2012, doi:10.1109/TITS.2011.2168215.

[24] J. A. Healey, R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," IEEE Transactions on Intelligent Transportation Systems, **6**(2), 156–166, 2005, doi:10.1109/TITS.2005.848368.

[25] D.-W. Koh, S.-G. Lee, "An evaluation method of safe driving for senior adults using ECG signals," Sensors, **19**(12):2828, 2019, doi:10.3390/s19122828.

[26] Z. Halim, M. Rehan, "On identification of driving-induced stress using electroencephalogram signals: A framework based on wearable safety-critical scheme and machine learning," Information Fusion, **53**, 66–79, 2020, doi:10.1016/j.inffus.2019.06.006.

[27] H.-M. Cho, H. Park, S.-Y. Dong, I. Youn, "Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network," Sensors, **19**(20):4408, 2019, doi:10.3390/s19204408.

[28] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, V. Chandran, "Automatic driver stress level classification using multimodal deep learning," Expert Systems with Applications, **138**:112793, 2019, doi:10.1016/j.eswa.2019.07.010.

[29] L. G. Hernández, O. M. Mozos, J. M. Ferrández and J. M. Antelis, "EEG-based detection of braking intention under different car driving conditions," Frontiers in Neuroinformatics, **12**:29, 2018, doi:10.3389/fninf.2018.00029.

[30] P. Zontone, A. Affanni, R. Bernardini, A. Piras, R. Rinaldo, F. Formaggia, D. Minen, M. Minen, C. Savorgnan, "Car driver's sympathetic reaction detection through electrodermal activity and electrocardiogram," IEEE Transactions on Biomedical Engineering, 1–12, 2020, doi:10.1109/tbme.2020.2987168.

[31] A. Affanni, G. Chiorboli, "Design and characterization of a real-time, wearable, endosomatic electrodermal system," Measurement, **75**, 111–121, 2015, doi:10.1016/j.measurement.2015.07.047.

[32] A. Affanni, "Wearable instrument to measure simultaneously cardiac and electrodermal activities," in 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Benevento, Italy, 2016, doi:10.1109/MeMeA.2016.7533749.

[33] J. Pan, W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, **BME-32**(3), 230–236, 1985, doi:10.1109/TBME.1985.325532.

[34] W. Chen, G. Liu, S. Su, Q. Jiang, and H. Nguyen, "A CHF detection method based on deep learning with RR intervals," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, 2017, doi:10.1109/EMBC.2017.8037578.

[35] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, P. A. Karjalainen, "Kubios HRV – Heart rate variability analysis software," Computer Methods and Programs in Biomedicine, **113**(1), 210–220, 2014, doi:10.1016/j.cmpb.2013.07.024.

[36] A. Affanni, A. Piras, R. Rinaldo, P. Zontone, "Dual channel Electrodermal activity sensor for motion artifact removal in car drivers' stress detection," in 2019 IEEE Sensors Applications Symposium (SAS), Sophia Antipolis, France, 2019, doi:10.1109/SAS.2019.8706023.

[37] P. Zontone, A. Affanni, R. Bernardini, A. Piras, R. Rinaldo, "Low-complexity classification algorithm to identify drivers' stress using electrodermal activity (EDA) measurements," Biomedical Engineering and Computational Intelligence, **20**, 25–33, 2020, doi:10.1007/978-3-030-21726-6_3.

[38] M. Malik, "Heart rate variability. Standards of measurement, physiological interpretation, and clinical use," European Heart Journal, **17**(3), 354–381, 1996, doi:10.1161/01.CIR.93.5.1043.

[39] J. Bergstra, D. Yamins, D. D. Cox, "Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms," in 2013 12th Python in Science Conference (SCIPY), Austin, Texas, 2013, doi:10.1088/1749-4699/8/1/014008.

[40] A. Affanni, R. Bernardini, A. Piras, R. Rinaldo, P. Zontone, "Driver's stress detection using skin potential response signals," Measurement, **122**, 264–274, 2018, doi:10.1016/j.measurement.2018.03.040.

[41] P. Zontone, A. Affanni, R. Bernardini, L. Del Linz, A. Piras, R. Rinaldo, "Stress evaluation in simulated autonomous and manual driving through the analysis of skin potential response and electrocardiogram signals," Sensors, **20**(9):2494, 2020, doi:10.3390/s20092494.

[42] A. Affanni, P. Zontone, R. Fenici, D. Brisinda, D. Bacchin, L. Gamberini, P. Pluchino, M. Bruschetta, C. Savorgnan, F. Formaggia et al., "Assisted / autonomous vs. human driving assessment on the DiM driving simulator using objective / subjective characterization," in 2019 10th International Munich Chassis, Munich, Germany, 2019, Pfeffer P. (eds), Springer Vieweg, Wiesbaden, 307–321, 2020, doi:10.1007/978-3-658-26435-2_23.

[43] P. Zontone, A. Affanni, R. Bernardini, A. Piras, R. Rinaldo, "Stress Detection Through Electrodermal Activity (EDA) and Electrocardiogram (ECG) Analysis in Car Drivers," in 2019 European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2019, doi:10.23919/EUSIPCO.2019.8902631.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 30-35 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Analysis of Green Building Effect on Micro grid Based on Potential Energy Savings and BIM

Ihsan Mizher Baht[*,1], Petre Marian Nicolae[1], Ileana Diana Nicolae[2], Nameer Baht[3]

[1]*Department of Electrical engineering, Craiova University, Craiova, 200396, Romania*

[2]*Department of Computer Science and IT, Craiova University, Craiova, 200396, Romania*

[3]*Department of Computer Science, Universities west of Timisoara, Timisoara, 300088, Romania*

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 08 October, 2020
Online: 08 November, 2020

Keywords:
Energy Management Systems
Micro Grids
Green Buildings
Optimization Software

ABSTRACT

In this paper we are demonstrating the importance of building and achieving a renewable electrical power system that works in synchronization with the grid. The average of consumption, generation and the negative environmental effect of generating power were analyzed. This paper is divided into three parts, one was analyzed with Green Building Studio software and the other with Building Information software. Using BIM software, the levels of energy consumption were analyzed for a building in Baghdad – Iraq and in Craiova _ Romania, alongside pollution levels produced, and the possibility of energy generation utilizing PV cells and wind, the output power that can be achieved in respect to temperature and irradiance levels. Using Green Building software, energy consumption system was analyzed in the level of green building.

1. Introduction

The concept of power and energy management refers to the use of highly effective and reliable blueprints of managing the available power and energy in such a way that the outcomes are optimized while also preventing wastage. Despite the fact that there are many factors and issues that cause power and energy loss, the absence of an effective management blueprint is one of the main contributors [1]. When an effective blueprint is absence, it then follows that the framework of power loss is heighted. Following is an evaluation of the distinctive factors and issues affecting the management of power and energy. It is vitally important to accentuate the fact that there are numerous issues and factors that impact the management of power and energy. While some of the factors are preventable in the short-term, others can only be addressed in the long-term. Unless these factors are identified in the first place, the blueprint for management of power and energy cannot be enhanced. Over the years, many problems have characterized power and energy systems in different parts of the world because such systems were not aligned towards an effective management blueprint [2]. There is thus the need to identify the distinctive issues and factors that play a direct role in the framework of power and energy management.

[*]Corresponding Author: Ihsan Baht, Contact No. +9647705808038 Email: imb121378@yahoo.com.

2. Factors Effecting Power and Energy Management

As far as the factors affecting the management of power and energy are concerned, one of the most important issues of emphasis pertains to location. Depending on the location, a given blueprint for management of power and energy can either be effective or ineffective [3]. In line with this attribute, there is always the need to ensure that the process of management of power and energy is strongly aligned towards the specific needs and priorities of a given location. Such an approach is essential towards the prevention of potential problems such as power and energy loss. The factors affecting the framework of management of power and energy also encompass the source of power. Different sources of power are characterized by distinctive attributes and issues. It then follows that the specific source must be determined in order to develop a framework of management of power and energy that is effective in the short-term as well as in the long-term [4]. A notable way in which the source of power impacts the framework of management of power and energy is that it determines whether a given management strategy can help in the reduction of energy wastage or not. This goes a long way towards determining the quality of outcomes in the long-term as well as in the short-term.

Competence of personnel is a major issue that determines the approach used within the context of management of power and

energy. For instance, the competence of staff members in a power grid determines the manner in which they approach decision making processes concerning power management [5]. Additionally, it is also notable that the competence of personnel goes a long way towards determining whether they understand the technical elements of management of power and energy or not.

As far as the factors influencing the blueprint of management of power and energy are concerned, another notable aspect of emphasis pertains to the availability of resources. In some instances, the scarcity of resources can compromise the extent to which the management of power and energy is aligned towards the intended goals. This is because there are extensive resource requirements at any stage of managing energy and power. It is also vitally important to highlight the fact that the availability of resources impacts the manner in which planning is actualized into the blueprint for management of power and energy [6]. In view of these stipulations, it is hence evident that the availability of resources is always a notable aspect of emphasis that determines the manner in which the blueprint for management of power and energy is actualized. The engineering vision of analyzing and constructing green buildings is a foundation for energy management such as to guarantee the achieving of great economic benefits [7, 8].

3. Building Performance Analysis (BPA) and Optimization Energy (OE) in Green Building Studio (GBS)

The Green building studio (GBS) which powers the energy analysis through (cloud-based service), the GBS with Insight 360 Revit Autodesk is able to test daily and weekly then make review, comparing with the previous results, there are two types of testing in green building studio (GBS):

Service side: energy results from GBS compare against the baseline. Analytics: written the energy results as simulation and compared with the baseline. The green building studio analyze the performance of the building as total annual energy consumption (electrical, fuel, CO2), total annual cost and potential energy savings.

The GBS analyzes automatically 37 separate energy simulations that can run in the cloud through the weather data station (in Baghdad use for analyzing the weather station GBS_06M12_12_002295).

The performance factors of analysis are:

- Light efficiency (from 0.7w/sf-1.9w/sf)
- Plug load efficiency (from 0.6w/sf-2.6w/sf)
- Building orientation (0 degree, 135 degree….etc.)
- Operations schedule (24/7, 12/7, 12/5 ….etc.).
- Daylighting occupancy and non-control.
- Infiltration load.
- HVAC.
- Other (civil and mechanical parameters).

3.1. Green Building Analysis Based On Bulding Orientation

The first effective factor before creating the green building, the Building orientations, selects the optimum degree to the building construction to provide the best energy use intensity, total annual cost and total annual energy. Table 1 represents the orientation of a green building in Baghdad city.

Table 1: The Total Annual Energy Cost (Building Orientations)

Degree	EUI MJ/M^2 /YE.	Total annual cost $		
		Electrical	Fuel	Energy
0	1030	7233	120	7354
135	1054	7201	130	7332
180	1026	6752	136	6889
225	985	6401	134	6535
270	1024	6976	127	7104
315	1016	7057	121	7178
45	1012	6982	123	7106
90	1060	7346	127	7473

From the Table 1 at 225 degree building orientation, the energy use intensity (EUI) 985 MJ/m^2 /year, total annual cost electric 6401 and the energy cost for one year 6535. The solar irradiance, temperature, flow air, infiltrations etc. at 225 degree is the optimum point for construction of green building to save the cost of energy and fuel.

Table 2 represents the annual energy consumptions at a green building with the different degree orientation at the same location. At 225 degree the total annual energy consumption in kWh 46.590(minimum consumptions), and at 90-degree orientation 53.465(maximum consumptions)

Table 2: The Total Annual Energy Consumptions (Building Orientations)

Degree	Total annual energy	
	Electrical kWh	Fuel MJ
0	52.649	87.929
135	52.416	95.127
180	49.146	99.485
225	46.590	97.616
270	50.777	93.077
315	51.362	88.650
45	50.822	89.573
90	53.465	92.914

From the results (Table 1, Table 2) the minimum energy consumption, minimum EUI, and minimum energy fuel cost is at the 225 degree. Green Building Studio (GBS), Insight 360 is able to analyze the weather effect on the building to provide the best benefit from the environment and the effects of climate, wind …etc.

The parameters analysis extensions a scope of possibilities to find very active building as a green building, for example the climate can determine the type and capacity of the lighting installation depending on the degree of the building orientations.

Figure 1 and Figure 2 Represent the total annual energy consumption MJ and total annual electrical consumption. The Y-axis total energy (MJ), X-axis time (year) for example in Jan the red bar shows the space heat ,yellow bar area light, blue bar space cooling.

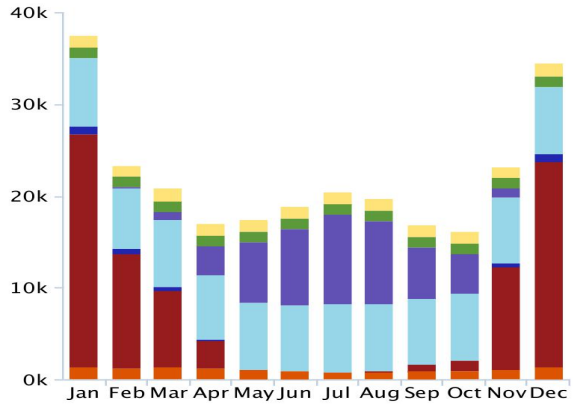


Figure 1: The Total Annual Energy Consumption (Mj)

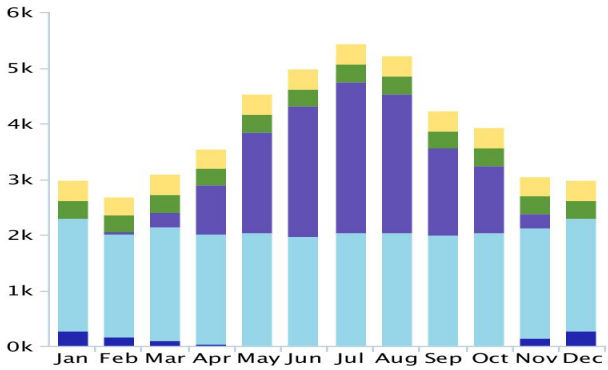


Figure 2: The Total Annual Electrical Consumption (Kwh)

3.2. Green Building Analysis Based On Daylighting

The second factor effect on the energy consumption in green building is daylighting analysis, the daylighting analysis is divided into four parts:

- Daylighting and occupancy control daylighting & occupancy control.
- Daylighting and occupancy control daylighting control.
- Daylighting and occupancy control none.
- Daylighting and occupancy control & occupancy control.

Table 3 presents the amount of total annual cost and total energy consumption for the four parts of daylights control regarding to GBS weather station (cloud-based service).

Table 3: The Amount of Total Annual Cost and Total Energy Consumption

		Total Annual Cost $			Total Annual Consumptions Energy	
Day Light	Eui Mj/M²/Y	Ele.	Fue.	Ener.	Elec. Kwh	Fuel Mj
1	1016	7.020	123	7.143	51.04	89.58
2	1021	7.105	122	7.227	51.710	88.89
3	1030	7.233	120	7.354	52.649	87.92
4	1023	7128	121	7250	51.883	88.74

From the Table 3 the green building studio analysis improves the maximum energy potential savings in Daylighting and occupancy control daylighting & occupancy control.

Figure. 3 and Figure. 4 represent the total energy (MJ), electrical consumption kwh for Daylighting and occupancy control daylighting & occupancy control for the one-year period.

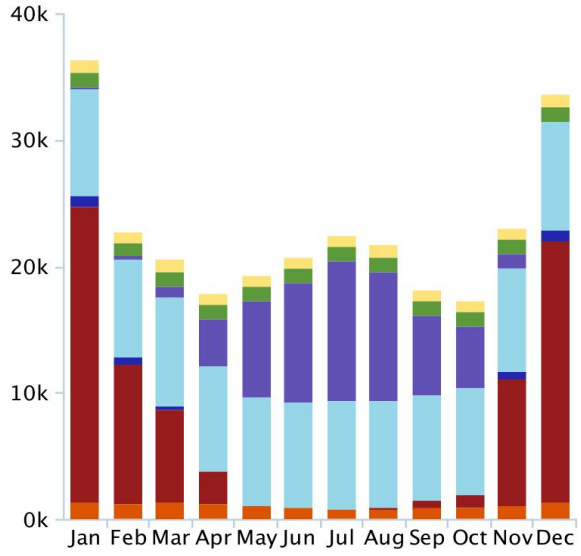


Figure. 3: The Total Energy (Mj)

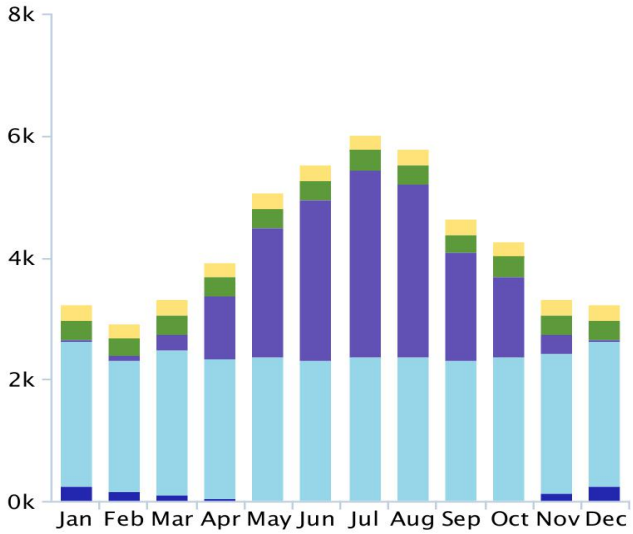


Figure 4: Electrical Consumption (Kwh)

3.3. Green Building Analysis Based On potantial energy (cost)savings

For the design of a green building and an existing building analysis, the reduce of the energy used is one of the main targets of the GBS, potential energy savings (PES) analyze all buildings energy performance (more than 37 separate energy simulations).

For the photovoltaics analysis (in Baghdad, weather station GBS_06M12_12_002295) at 225-degree orientation building, installed on the roof of the building, the panel type single crystalline -13.8 efficiency, installed panel cost 8$ per watt, applied electric cost 0.14$, the maximum payback period 50 years, installed panel cost 162.266$, installed panel area 147m^2, the annual energy production 36 kWh.

Table 4 presents the potential cost saving and annual energy with payback per surface per year, the annual energy produced depends on two factors, the panel area and solar exposure %, for example at the annual energy produce of 984(kWh), the panel area 4 m^2, solar exposure 69.9%.

Table 4: Potential Cost Saving and Annual Energy with Payback Per Surface Per Year

	Potential cost savings $		
Annual energy kWh	Per year/m^2	Per year	payback per surface per years
984	36.04	133	23.9
2914	36.04	408	23.9
364	36.04	51	23.9
447	36.04	63	23.9
16.703	32.23	2.338	26.1
3.251	32.14	455	26.2
3.207	31.80	499	26.3
3.563	31.8	499	26.3
2.767	31.80	387	26.3

3.4. Implications of Power and energy management

In view of the various factors and issues affecting the process of management of power and energy, there is also the need to identify the distinctive implications of management of power and energy [9]. These implications re essential in that they influence the planning process especially when it comes to decision making as well as allocation of resources [10]. Additionally, clarity of the implications helps in determining the best possible strategy to apply within the confines of management of power and energy.

The specific implications of management of power and energy also encompass sustainability. This refers to the use of energy resources in such a way that the needs of the future generations are considered even while scarce resources are being utilized by current generations [11]. Despite the fact that there are many factors and issues that contribute towards the sustainable use of power, the framework of management is always important. Additionally, the fact that the process enhances the extent to which energy and power is is optimized is also indicative of the manner in which the blueprint for management of power and energy enhances suitability.

3.5. Evaluation of CO2 emissions when only fuel is used for power generation

One of the most important criteria for creating a system that depends on improving the environmental levels is the effect of CO2 gas that is a by-product of electrical energy production.

Figure 5. Depicts the effect of energy production. Fuel consumption in Baghdad – Iraq is up to 6 tons [12] [13].

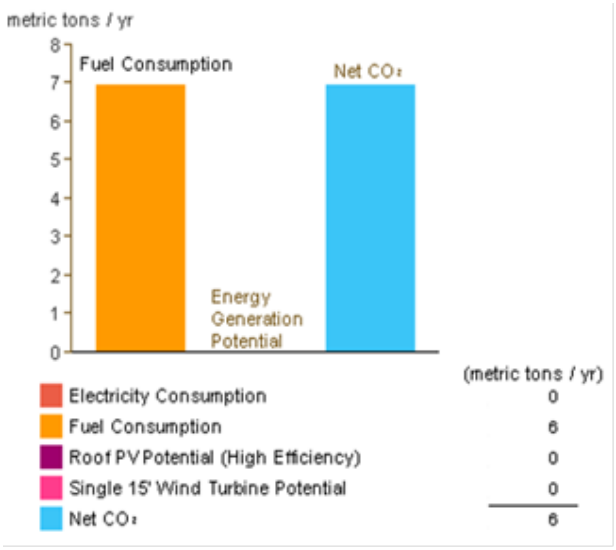


Figure 5: Energy and CO2 Consumptions (Baghdad)

Figure 6 depicts that fuel consumption is 4 tons in the city of Craiova – Romania. The generation system above was analyzed using BIM software for the same area, utilization and population. But the average effect in Craiova is less than Baghdad [12] [13].

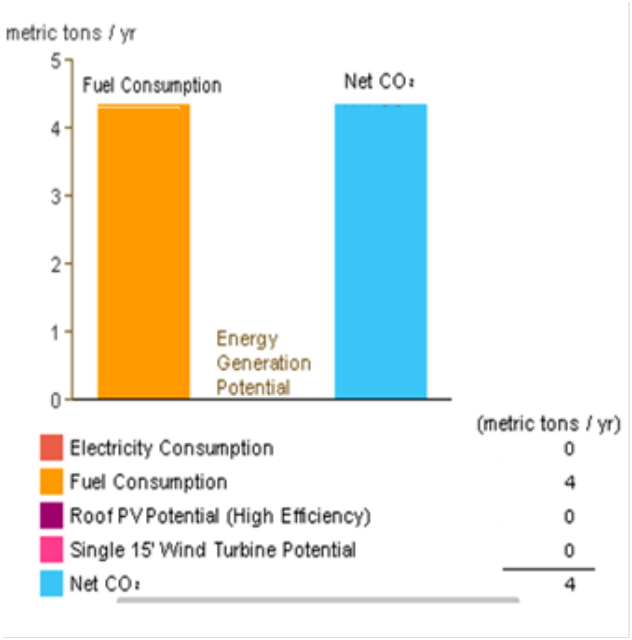


Figure 6: Energy and CO2 Consumptions (Craiova).

The table 5 below states the expected average electrical power generation if renewable energy (PV) is utilized, where these data were analyzed and average temperature and irradiance

were calculated for both cities of Baghdad and Craiova for several types of solar panels and also for wind turbine. The average production in Baghdad is higher than that in Craiova because of the nature of weather and intervals of sun energy availability and higher levels of irradiance [12] [13].

Table 5: Estimations energy produce by Pv cells And Wind

Energy Estimation	**BGW**(kwh/yer)	**CRA.** (kwh/yer)
PV cells (low efficiency)	12,377	9,009
PV cells (medium efficiency)	24,755	18,017
PV cells (high efficiency)	36,132	27,026
wind turbine Single 15’	1,956	691

4. Analysis by using BIM software

BIM software functions through analyzing of buildings and stating the effects of civil and mechanical works on the electrical power consumption. BIM software includes several criteria, in total 26 factors which improve grid performance and decrease consumption. Figure 7 illustrates the average annual energy consumption of the buildings taking into consideration building utilization, for example as a school, hospital or residence.

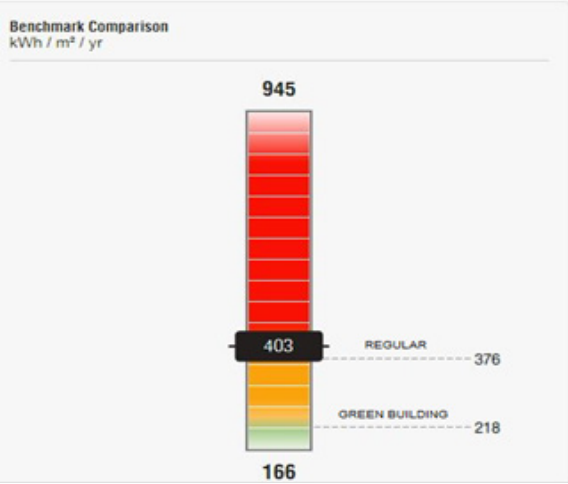


Figure 7: Energy Consumption

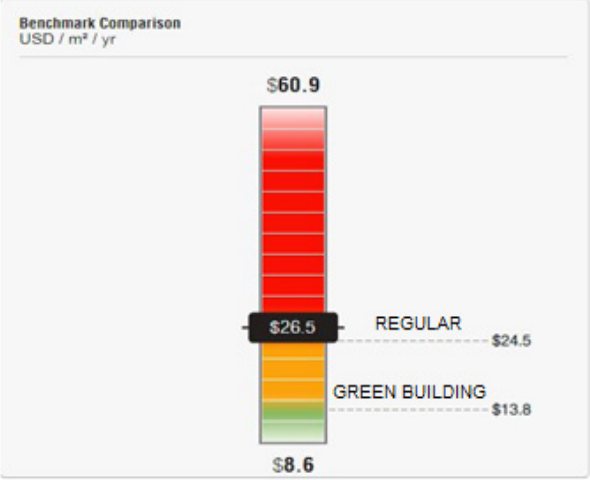


Figure 8: Costs Analysis

We can observe that average energy consumption reaches 376 kW, while when using BIM software analysis system consumption may be 218 kW/m^3 at the same efficiency. Estimated cost changes also from $24.5 to $13.6 (Figure. 8).

Figure 9 represents a comparison in analysis of a building in Baghdad where average consumption is highest through normal analysis, but when Green Building software is utilized as represented in yellow color which reaches a consumption of 310kwh/m^3, and green color when analyzed using BIM software with an annual consumption of up to 120kwh/m^3 [12], [13].

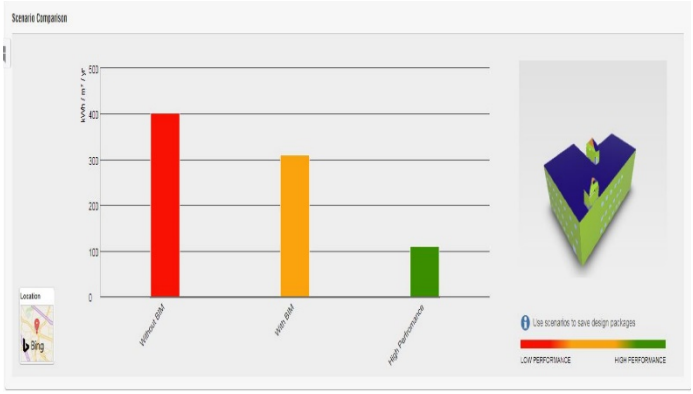


Figure 9: Analysis Performance

5. Conclusions

In this paper, the work was divided into three parts, first part analyzes a building and the effect of weather using Insight software. In the second part an analysis is done via Green Building, where a total cost was calculated to build a solar energy generation system to reduce electrical power consumption, and utilization of building angle which serves to better usage of weather. Third part is an analysis of the systems above. And depicting the economic value that BIM software provides to building efficiency, decrement of consumption, and decrement of costs

References

[1] A. Gulagi , Dmitrii Bogdanov & Christian Breyer, "A Cost Optimized Fully Sustainable Power System for Southeast Asia and the Pacific Rim," Energies, MDPI, Open Access Journal, **10**(5), 1-25, April 2017, doi:10.3390/en10050583. 2017. 10.583 .

[2] T. Dragicevic, J. Guerrero, J.Vasquez, and D. Skrlec, “Supervisory control of an adaptive-droop regulated dc microgrid with battery management capability,” IEEE Trans. on Power Electron., **29**(2), 695–706, 2014. doi: 10.1109/TPEL.2013.2257857.

[3] H. Yousefi, M. Ghodusinejad, " Feasibility Study of a Hybrid Energy System for Emergency Off-grid Operating Conditions. Majlesi Journal Of Electrical Engineering", **11**(3), 7-14 ,2017.

[4] A. Wentland,"Imagining and enacting the future of the German energy transition: electric vehicles as grid infrastructure. Innovation" The European Journal Of Social Sciences, **29**(3), 285-302, 2016, doi: 10.1080/13511610.2016.1159946 .

[5] E. Ganesan., "Modeling, control, and power management for a grid-integrated photo voltaic, fuel cell, and wind hybrid system", Turkish Journal Of Electrical Engineering & Computer Sciences, **24**(6), 4804-4823, 2016 , doi: 10.3906/elk-1404-409.2016.

[6] N. Çetinkaya, "Improving of renewable energy support policy and a performance analysis of a grid-connected 1 mwp pv power plant in Konya Selcuk University" , Journal Of Engineering, Science & Technology / Selçuk Üniversitesi Mühendislik, Bilim Ve Teknoloji Dergisi, **5**(3), 251-261, 2017,doi:10.1273/tused.10159.2016

[7] T. Dragicevic, J. Guerrero, J.Vasquez, and D. Skrlec, “Supervisory control of an adaptive-droop regulated dc microgrid with battery management capability,” IEEE Trans. on Power Electron., **29(**2), 695–706 , 2, doi: 10.1109/TPEL.2013.2257857,2014.

[8] F. Zhang, T. Zhang, R. Wang, "Optimizing Hierarchical Power Distribution of Multiple Local Energy Network Systems in GridConnected Mode”, Polish Journal of EnvironmentalStudies, **26**(5), 1981-2000, doi:org/10.15244/pjoes/69440.2017.

[9] A. Al Ameri , Aouchenni Ounissa , Cristian Nichita & Aouzellag Djamal, "Power Loss Analysis for Wind Power Grid Integration Based on Weibull Distribution," Energies, MDPI, Open Access Journal, **10**(4), 1-16, April, 2017, doi:org/10.3390/en10040463.2017.

[10] A. Lösch, & C. Schneider,"Transforming power/knowledge apparatuses: the smart grid in the German energy transition. Innovation". The European Journal Of Social Sciences, **29**(3), 262-284, 2016, doi:org/10.1080/13511610.2016.1154783.2016.

[11] M. Serraji, El Amine D. O., & Boumhidi J., "Multi swarm optimization based adaptive fuzzy multi agent system for microgrid multi-objective energy management",International Journal of Knowledge-based and Intelligent Engineering Systems, **20**(4), 229-243, 2016, doi: 10.3233/KES-160350.2016.

[12] I. M. Baht, P.M. Nicolae, M. Ş. Nicolae, and Ileana Diana Nicolae, Analysis of Load Demand and Economic Assessment Based on Weather Forecasting in Green Buildings, Annals of the University of Craiova, Electrical Engineering series, **43**(1), 147-152, 2019 .

[13] I.M. Baht, P.M. Nicolae and M. Ş. Nicolae, "Impact of Weather Forecasts and Green Building on Micro Grid Energy Management System", Proceedings of SIELMEN 2019, Chişinău, Oct , POCU380/6/13/123990. 2019.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 36-44 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Effective Segmented Face Recognition (SFR) for IoT

Fei Gao, Jiangjiang Liu*

Department of Computer Science, Lamar University, Beaumont,77710, USA

ARTICLE INFO

Article history:
Received: 01 September, 2020
Accepted: 07 October, 2020
Online: 08 November, 2020

Keywords:
Fully Convolutional Networks (FCN)
Convolution Neural Network (CNN)
Classification
Face Recognition

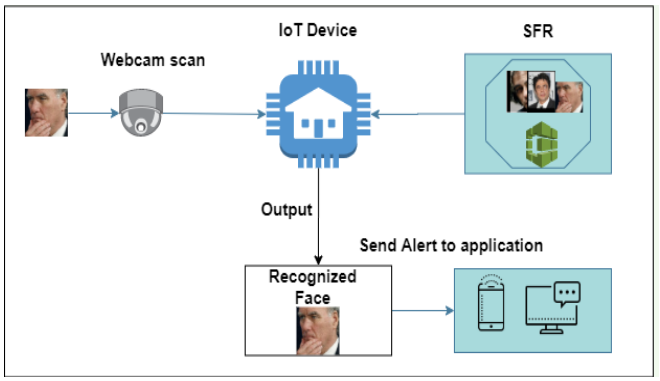
ABSTRACT

Face recognition technology becoming pervasive in the fields of computer vision, image processing, and pattern recognition. However, face recognition accuracy rates will decrease if training is done on disguised images with covered objects on a face area. This paper aims to propose a state-of-the-art face recognition methodology which could be applied in Internet of Things (IoT) devices as an input source; then face segmentation and training process will be executed in the cloud via internet; the recognition result will be sent to the connected applications which determines a safety check for personal or public security. This paper focuses on implementation of face segmentation and training process for IoT. Face extraction from the background and disguised part is applied by Fully Convolutional Networks (FCN), and then deep convolution neural network is employed for face training and testing process. This algorithm has been experimented on a challengeable face dataset. The proposed face recognition system is applied to IoT services which have multiple applications, such as, personal home security and public library space management. Compared to recognition without face segmentation, the results of proposed methodology indicate a better accuracy regarding recognition rate.

1. Introduction

IoT is the combination of hardware environment, software services and Internet. It is composed of sensors, servers, hardware equipment, software equipment, and Internet connections [1]. With the fast improvement of sensors and devices, IoT could be applied to multiple fields such as wearables, smart home, and city. Face recognition is one of the popular personal security topics for IoT system, which could be widely used in personal home and public places security fields. Also, it has been used as an identification task in different areas, especially in corresponding with computer-based security and safety systems in homes, criminal identification, and smart phone devices' face identification. For instance, cameras were used for image capturing and IoT processing devices, and Raspberry Pis, were used for comparing captured images with server database to provide directions over the GSM module and alert mobile phones [2]. Similar IoT systems with face recognition components were also proposed for security purposes in various applications, such as libraries and banks [3,4]. In this paper, we propose an effective segmented face recognition for IoT (SFR-IoT) shown in Figure 1.

Face recognition means recognize a specific person with a 2-D or 3-D face image or video base on available face data sets. If a camera is utilized to capture a suspect's face image, that image can then be used to identify the suspect by using biometrics. The application of biometrics technology such as face recognition could reduce crime rates. Instead, the lack of effective skills in capturing biometric data from face images will potentially increase the opportunity for criminal activity [1].



*Corresponding Author: Jiangjiang Liu, Email: jliu@lamar.edu

Figure 1: Segmented Face Recognition Architecture (SFR-IoT)

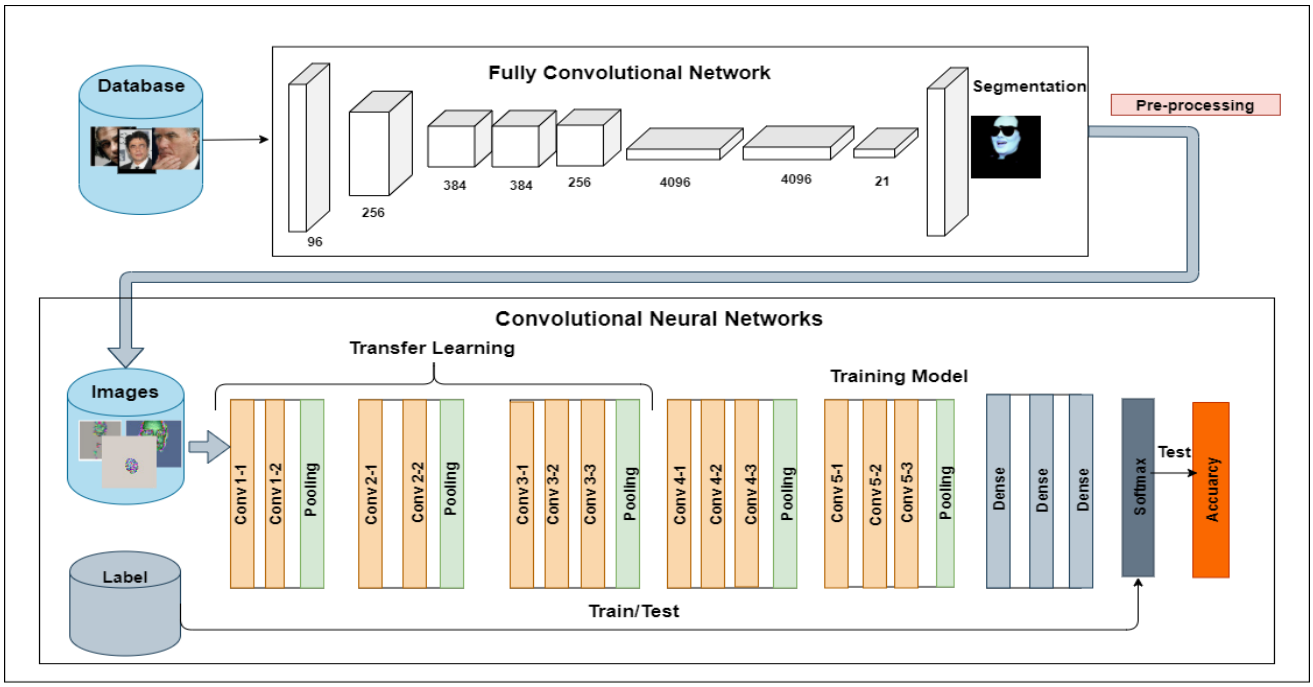


Figure 2: FCN-8Ss-VGG-16 System Architecture

Recently, a few researchers have published work related to crime detection. One study introduced a framework that first detects facial key points and then uses them to perform face recognition [2]. They mentioned that a larger number of images and camouflage images available in a data set can improve the training of the learning network and avoid the need to perform transfer learning. Results show that their framework outperforms the most advanced methods in critical point detection and facial camouflage classification.

For past investigation techniques, there are a few popular algorithms in regards to face cognition which include the geometric features method: principal component analysis (PCA) [3], linear discriminant analysis (LDA) [4], hidden Markov strategy (HMM) [5] and regular LBP features [6]. However, there are evident disadvantages in terms of utilizing these techniques. For instance, the accuracy rate of PCA will diminish significantly with lightness and a state of image change. Also, the strategy for inadequate data requires strict arrangement of information pictures, which is not applicable for basic application.

Due to the important factors of the accessibility of public data sets, as well as the improvement of Graphics Processing Unit (GPU) computation, neural networks has achieved a large renaissance, resulting in a significant increase in accuracy rates [7]. This leads to a major improvement in image recognition and finally in face recognition. For instance, Convolutional Neural Network (CNN) [8,9] is a popular model in neural networks. It is a state-of-the-art technique that has replaced the traditional algorithms on face recognition and consequently taken the network by storm, fundamentally enhancing the cutting edge in numerous applications. Several studies have applied CNN for face recognition purposes, such as Deep Face [10], DeepID3 [11] and Face Net [12]. They have achieved around 97%, 99.6%, and 99.5% success rates, respectively.

Recent studies that combined face segmentation and face recognition technology have provided an important concept for our research. For example, one study extracted face-like regions by using the extracted color information of an image in HSV color space, and the RHT algorithm is then applied to find face region [13]. Samantha proposed a technology of segmentation skin color through YcbCr as well, then applied artificial neural network classified face and non-face classes [14]. However, these researches did not integrate deep learning study for face segmentation and recognition, so this paper would like to propose a new methodology that improve the deficiency.

Even though current automated face recognition systems can recognize individual faces in controlled environments with a 99% accuracy, the accuracy drops to below 60% when images are in unconstrained environments [15].

This is caused by rich facial expressions and changing gestures; face movements with lights; angle; and distance [2].

In addition, disguised faces with glasses, scarves, and accessories will become challenging elements in face recognition tasks. Hence, in order to explore a more effective method to solve this problem, this study proposes a new system called FCN-8s-VGG-16, which is used for the segmentation technology face recognition before classification. The system contains two procedures, as shown in Figure 2. The process of this research is to build a training set of segmentation of the face area with FCN as an input of Convolution Neural Network (CNN), and then train the CNN network from VGG-Face with transfer learning. Lastly, the classification function SoftMax is applied for face probability distribution. The first step is to segment the original image for extracting face parts (eyes, nose, mouth, and skin) from the background, which applies a pre-trained FCN-8s model [16]. Secondly, the output images are fed into the network applying transfer learning [10] based on VGG-16 [8] with our fine-tuned

technology on face recognition task. Finally, the Softmax function calculates the probability distribution of the test image and generates the accuracy rate while comparing it to the label box.

2. Related Work

2.1. General Segmentation and Face Segmentation

Segmentation refers to the process of separating an image or frame into groups of pixels which are identical with respect to some standard. A few papers have applied segmentation for their research topic.

In [17], the author proposed a food calorie estimation model, which was applied on a smartphone to predicate calories of the food in images taken by a phone camera. Firstly, they estimated approximate positions of dishes by edge detection. Secondly, k-means based clustering was applied on the color pixels, thereby extracting a bounding box of a food area. Next, they used GrabCut [18] to obtain an accurate food area within the estimated bounding box. Then, they employed CNN as a calorie estimation task. They prepared 120-calorie annotated food photos as the data for the experiments. Finally, 60 test images with real food calories are estimated with the relative average error of 21.3%, regarding food calorie prediction achieved, and it is higher than the Japanese government definition.

In [19], the author applied different segmentation techniques in the aim of reorganization of the left ventricle in a VEF image. He compared two segmentation approaches: region-based segmentation and edge-based segmentation. The first detects homogeneities and the common features by simultaneously applying the chan_vese model and the thresholding techniques. In the second approach, he applied a certain algorithm such as Sobel, Canny, and Perwitt to obtain the interim between two related regions. Results showed that region-based segmentation using the chan_vese model in conjunction with thresholding gives a better performance.

Since Convolution Neural Network (CNN) has been widely researched by the scholarly field, various models and systems with different purposes are dramatically increased. Fully Convolution Network (FCN) [20] is one of most popular image segmentation technology that is widely used in studies. FCN extracts pixels from CNN layers of mixed scales, enlarges them to the size of the original image, and then applies a convolution layer to classify all this information.

There are research works pertaining to FCN for segmentation purposes that have achieved excellent results. One study presented a One-Shot Video Object Segmentation (OSVOS) based on the FCN network architecture, with transfer learned on ImageNet then fine-tuned on one training sample [21]. The performance was tested on DAVIS database, which consists of 50 full-HD video sequences and YouTube objects. The result shows that OSVOS is fast and improves the state-of-the-art technology by a significant margin (79.8% vs 68.0%). Based on these outstanding results, this paper chooses FCN as our face extraction segmentation tool.

In [22], the author proposed a method to achieve the segmentation of skin, hair, and background by applying FCN-8s and fully-connected CRF. Next, matting algorithm was facilitated in their experiment in order to receive clear hair and skin alpha masks. Finally, they demonstrated that state-of-the-art performance on LFW Parts dataset [23] holds more accurately than other approaches. The defect of the FCN algorithm influences our decision to apply Transfer Learning in our experiment, which we will talk about it in Chapter 3.

2.2. Face Recognition

Another study has shown that choosing an appropriate model in CNN architecture is vital for face recognition. This study evaluated two popular CNN architectures, Alex-Net [9] and VGG-16 [8], on face recognition. They accomplished their task by transferring learning idea to the networks trained for various classifying purposes. One study [24] used Alex-Net model to train the CASIA-WebFace [25] database, and the research shows that VGG performs better than Alex-Net. This paper inspired us to employ VGG-16 [8] as a face recognition model for our own experiment.

In [26], the author conducted face recognition by applying a CNN for feature extraction. First, they randomly selected patches from the STL-10 [27] database and used them to train a linear decoder and obtain a 400 × 192 learned weights matrix. The 64 × 64 RGB images are used to train the identifier through a convolutional layer, and then training features are extracted using these learned weights. Results show the proposed method reached a high rate of excellent sorting, ranging from about 80% to 100%.

There is another study [28] that has proposed a modified Convolutional Neural Network (CNN) architecture with the addition of two batch normalization operations on two of the layers. CNN architecture was utilized to separate particular face features, and Softmax classifier was utilized to recognize faces in the completely associated layer of CNN. The training and test process were tested on a Georgia Tech face database. After pre-processing, the researchers changed the sizes of input to 16×16×1, 16×16×3, 32×32×1, 32×32×3, 64×64×1, and 64×64×3 in order to receive the best result. Finally, 64x64x3 surpasses other sized images regarding the lowest Top-1 error rate. Our paper applied a similar strategy with this research, through altering the size of input in order to obtain the best result in our experiment.

However, to our knowledge, segmentation technology is not used in the face recognition field so far, except in our initial work [29]. This paper is an extension of work originally presented in the 2019 18th IEEE International Conference on Machine Learning and Applications. The accuracy rate will increase due to the interference decrease from unrelated face pixels, such as glasses. In order to decrease the useless effect pixels of disguised face parts, this study would like to propose an FCN-8s-VGG-16 system with segmented technology before classification. It only keeps face sections, and the other parts will be taken off. For the purpose of examination both before and after the change, the face section is sent to the classification model and accuracy will be tested for comparison.

3. The Proposed Approach

3.1. Methodology (CNN)

3.1.1. VGG-16

The core idea of convolutional networks is classifying target data samples to the different distances between classes. The

network structure includes three convolutions (conv1, conv2, conv3), two pooling layers (Pool1, Pool2), and a fully connected layer [8].

3.1.2. Convolution Layer

The input data is trained by employing a set of trainable neurons. The output of each of the feature maps corresponds to an image filter of the same size as the input of the convolution layer. The primary function of the convolution layer is to extract features from an image. Every convolution layer is trained on a feature map of the past layer, sequentially, and then it usually adds a bias parameter in order to increase accuracy by activating the function to translate results from linear to non-linear function. After that, the feature maps are fed into the next convolutional layer as input data.

3.1.3. Pooling Layer

Pooling layer decreases the dimensionality of each activation map but keeps the most vital data information. The input images are separated into a set of non-covering square shapes. Each field is downsampled by a non-linear function. For example, average or maximum is used the most. This layer accomplishes a better speculation and robust result for system.

3.1.4. ReLU Layer

Rectified linear units (ReLU) is a non-linear operation. It is an important function because it ends up with 0 if the input is less than 0. However, if the input is greater than 0, the output will be an original number. Research studies show that ReLU results can implement faster training for a huge network.

3.1.5. Fully Connected Layer

The output from convolution layer, pooling layer, and ReLU layer is high- level features of input data. The purpose of applying the Fully Connected Layer (FCL) is to sum these features for classifying the input data into different classes based on the probability of each class of each feature map. Next, FCL supports the features to a classifier, which is Softmax function. This function will conclude the probabilities of every target class over all possible target instances. Afterwards, the calculated probabilities will decide the target class for the given inputs.

3.1.6. Why VGG

There are a few CNN architectures, such as LeNet, AlexNet, Reset, and so on, which have been widely used since CNN was invented. Recently, the most state-of-the-art model [10] outperforms in the localization and recognition tracks, respectively, the ImageNet Challenge 2014 among them.

This VGG model is described by its simplicity, applying only 3×3 filters stacked over each other in increasing depth. Reducing parameter numbers and model size is employed by max pooling, two fully-connected layers, each with 4,096 nodes followed by a Softmax classifier.

3.1.7. Transfer Learning

Transfer learning is an area of artificial intelligence, which centers around the capacity for a machine learning calculation to enhance learning limits on a target data set through past exposure to an alternate way. The measured quality of a CNN implies that we can easily apply the weights from a pre-trained model and only re-prepare most elevated layers. In particular, we re-prepare all linear layers in the model and replace parameters of the highest layers in VGG-16 [8]. To get rid of the lack of geometric invariance of these approaches, fine-tuning [30] with an external data set can be utilized. The primary distinction between picture arrangement and picture recovery is the measure of information and its fluctuation. In classification, it is important to use huge data sets with high variability for different categories. However, in image recognition, the geometric invariance of an image is less essential for a training model. As the purpose of image recognition is to recognize the instance of example, less variability data will be needed.

In this paper, to use CNN functionally, we fine-tune the pre-trained CNN [10] a face data set for image recognition. Fine-tuning usually focuses on the higher layers while fixing the lower layers of a CNN. We use FCN-8s-VGG [16] architecture, which has been pre-fine-tuned for segmentation on PASCAL [31]. As the CNN model [16] is pre-trained on face dataset and performs an excellent result, we don't consider it for Fine-tuning while we are doing the same object segmentation purpose.

Table 1: The Architecture of VGG-Face Refined

Layer	Layer Type	Filters	Input	Training
1	Conv.	64	3x3x3	False
2	Conv.	64	3x3x64	False
3	Max-Pool	-	2x2	-
4	Conv.	128	3x3x64	False
5	Conv.	128	3x3x128	False
6	Max-Pool	-	2x2	-
7	Conv.	256	3x3x128	False
8	Conv.	256	3x3x256	False
9	Conv.	256	3x3x256	False
10	Max-Pool	-	2x2	-
11	Conv.	512	3x3x256	True
12	Conv.	512	3x3x512	True
13	Conv.	512	3x3x512	True
14	Max-Pool	-	-	-
15	Conv.	512	3x3x512	True
16	Conv.	512	3x3x512	True
17	Conv.	512	3x3x512	True
18	Max-Pool	-	2x2	-
19	FC	-	4096	True
20	FC	-	4096	True
21	FC	-	85	True
22	Softmax	-	-	-

3.1.8. VGG-FACE

VGG-Face [10] is a model that was pre-trained based on the VGG-16 structure and set up on a face data set acquired by the Visual Geometry Group that is comprised of more than 2.5 million pictures and 2,622 different labels. Detailed information in terms of VGG-16 model was shown in Table 1, including layer types, number of filters, and input image size.

The architecture of VGG-face contains a total of 22 layers: 13 convolutional layers for image and filter calculation, 5 max pooling layers for keeping feature information, 3 fully connected layers for summing up all previous features, and later a Softmax layer for classification [10]. The normal image resolution of the input layer is 224x224. Table 1 shows the detailed architecture of the VGG-face, including filter numbers and input size at each layer applied in our research. In this experiment, we have fixed the previous 10 layers and changed 21 fully connected layers from 1,000 to 85, since we have 85 classes in the experiment, which means 85 labels will be generated. Additionally, we fixed the parameters of the lower convolutional layers from 1-9 but trained the higher convolutional layers from 11-17 and the FCN layers from 19-21. Applied changes are highlighted in the table.

3.2. Methodology (FCN)

FCN [16] uses the identical convolutional network based on VGG-16 layers [8] and converts the fully connected layer to a convolutional layer. In the traditional CNN structure, the first five layers are convolutional layers, and the sixth and seventh layers are a one-dimensional vector with a length of 4096. The eighth layer is a one-dimensional vector with a length of 1,000, corresponding to 1,000 different probabilities of the category. Instead, FCN represents these three layers as a convolutional layer whose size (number of channels, width, height) is (4096, 1, 1), (4096, 1, 1), (1000, 1, 1), respectively. There does not seem to be any difference in numbers, but convolution is not the same concept, and calculation process as a fully-connected layer. The convolutional layer uses the weights and biases that CNN has pre-trained, but the difference is that the weights and offsets are their own scope and belong to their own convolution kernel. Therefore, all layers in the FCN network are convolutional layers and are called fully convolutional networks.

3.2.1. Upsampling

Upsampling is also called deconvolution at some points, as both of their operations are multiplication and addition. The upsampling of the accompanying factor f is a convolution operation on a fractional input with a step size of 1/f.

Backward convolution is called deconvolution. Meanwhile, the forward and backward propagation of upsampling can only be achieved by reversing the forward and backward propagation of the convolution. Consequently, it performs well no matter the optimization or backward propagation. Deconvolution can enlarge input size by learning parameters while training [20].

3.2.2. Skip Layers

Now that we have a 1/32 size, a 1/16 size, and a 1/8 size feature map, after upsampling the 1/32 size heatmap, the image restored by this operation is a convolutional kernel of conv5. The features are limited to the accuracy, so they cannot restore the features in the image well; consequently, forward propagation will be applied. Detailed information could be complemented by upsampling the conv4. Afterwards, conv3 does the exact same calculation as conv4. Finally, the process completes the restoration of the entire image. According to different strides with 32, 16, and 8, FCN splits into three kinds of results: FAC-32s, FCN-16, and FCN-8s [20]. Obviously, FCN-8s combined previous FCN-32s and FCN-16 results, which gives FCN-8s more accuracy. We decided to choose FCN-8s in our experiment.

3.2.3. Why choosing FCN

The reason why this research employs FCN are below:

- Traditional segmentation method segmented the image into similar sections, including coherent area-independent, low-level clues based on categories, such as pixel color or proximity. Semantic segmentation, on the other hand, assigned each pixel of the image to a semantic tag. This usually means classifying each pixel (for example: the pixel 1 belongs to the glass, the pixel 2 belongs to the hair). Pixel-level classification seems more efficient than cutting out each patch [32]. The FCN-8s model brings state-of-the-art performance on PASCAL VOC in 2011 and 2012 [20].
- No fully connected layer is used in this kind of architecture, which reduces the number of parameters and the computation time. In addition, the network works regardless of the original image size. It could be trained from small and big images, and it does not require any particular number of units at any stage. Additionally, all connections are local [20]. FCN brought great breakthroughs on semantic segmentation tasks when implemented on image segmentation [20].

It is tough to obtain millions of diverse images with ground truth segmentation labels, which could be difficult and mostly performed manually. However, it is convenient to use a pre-trained model for research purposes. Our method applies the FCN segment to the visible parts of faces from their specific circumstance and impediments, which is a fine-tuned system for face segmentation on PASCAL [20] on the data set of IARPA Janus CS2 [33] with 9,818 segmented faces achieved.

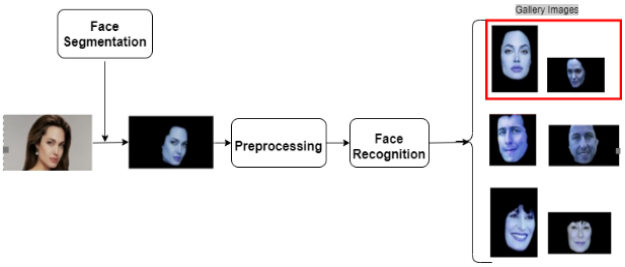


Figure 3: FCN-8s-VGG-16 System 2

This research has achieved an excellent segmentation performance that could be obtained with a standard FCN trained on plenty of rich and varied examples. FCN-8s combined previous FCN-32s and FCN-16s results, which gives FCN-8s more accuracy. We decided to choose FCN-8s in our experiment. Figure 3 shows the process of our experiment.

3.3. Methodology

3.3.1. Data selection

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [34] was dominant in giving this information to the general picture classification task. Moreover, analysts have made data sets accessible for object recognition [35]. Labeled Faces in the Wild (LFW) [36], as the most popular benchmark data set has dominated in the field of face recognition for many years. It is a

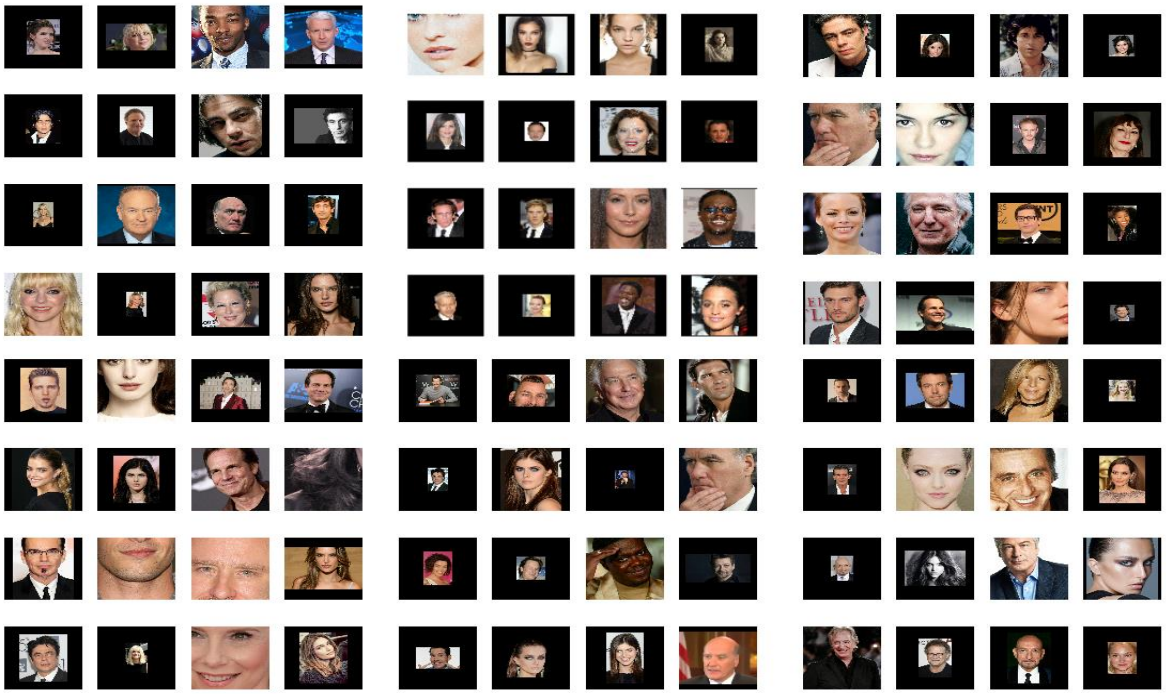


Figure 4: Gallery Images

huge scale open data set and the benchmark database of face photos intended for concentrating the issue of unconstrained face recognition, which has been widely applied in numerous research studies [10,12]. The data set contains more than 13,000 images of face pictures gathered from online.

Each face has been labeled according to the name of the individual envisioned. However, due to the shortage of obvious challenging face images, such as block face, tilt the head, wrinkle the eyebrows, and so on, it is not an ideal image set in terms of the purpose of the paper, which is to compare face recognition performance with and without segmentation. Furthermore, FRGC [37] MS-celeb-1M [38] and MOBIO [39] are other benchmark face data sets used to identify face images though various experiments with the same problem as LFW.

3.3.2. Applied Data Set

In this research, we decided to employ the Celebrity-Face-Recognition-Dataset, which consists of 1,100 famous celebrities with 8,000 images [40] in each. In this dataset, the total size of the data set is 172 GB with 800*800*3 pixels in each. Due to the limitations of our hard drive and memory, this research randomly and manually selected 85 individuals with 100 images for each as classes for the face recognition task.

Since accuracy difference along with and without segmentation face section from the background is in a significant index for the research, the images with obvious covers such as hats, accessories, or different poses and hairstyles, are our preferred selection among the 8,500 figures from the Celebrity-Face-Recognition dataset. Figure 4 shows partial data processed in our experiment.

3.3.3. Data Augmentation

Image recognition technology focuses on the work of few data sets and small variability in object images. This results in very few images that can train a particular CNN model, even with fine-tuning. One way to solve this problem is to augment the data, by randomly applying transformations, color perturbations, and other random transformations. Randomly rotating and flipping the images can make up the deficiency of diversity pixels of the model. In the experiment, the following values are applied:

- Brightness: Images are randomly changing brightness [-0.1, 0.1) [41].
- Flipping: Images are horizontally flipped from left to right.
- Scaling: The pixel values are scaled to the range [0, 1].
- Contrast: Images are randomly adding contrast to the range [0.2, 1.8].

The result shows below as Figure 5.

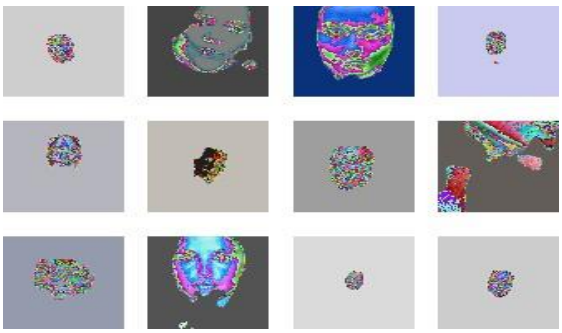


Figure 5: Pre-processing Result with FCN-8s Segmentation

3.3.4. Cross Validation

We found that in using the same data set, both training and model error estimation, the error estimation is mostly inaccurate, which is called the optimism of the model error estimation. To overcome this problem, cross-validation was proposed. According to Kohavi [42], cross-validation is a technology used to access predictive systems by dividing the original data set into a training set for training the parameters, with a test set for assessing the trained model. There are three more popular methods: leave-one-out, leave-P-out, and K-fold. Compared to the previous two technologies, K-fold only needs to calculate k times, which dramatically decreases code complexity. In K-fold cross-validation, the input data is split into k same size subsets, which is called folds as well. K-1 subset data will be trained into a system afterward, assessing the model on the subset that was not utilized for preparing. This procedure is repeated k times (the folds) until each subset has been used for evaluation purposes (and prohibited from training) each time finished. The results from k times calculation of K-folds can then be averaged (or joined) to receive a final estimation. The advantage of this method is that all data information is used for both training and evaluation, and each sub-dataset is used for validation exactly once. Based on the advantage of cross-validation, this paper has applied the 10-folds cross-validation on a data set.

3.3.5. Hardware Utility

Our implementation is based on Tensorflow Framework. AWS as a service platform that support the measurement of runtimes and performance. We implemented the experiment on the 16 GB memory of Nvidia Tesla 100V GPU.

4. Experimental Results

During this examination process, the implementation of the proposed method has been evaluated based on the percentage recognition rate of the identities on the testing data set. Initially, we set the learning rate as 0.001 for the CNN system, and a total of 10,000 steps are required for every k times calculation. Then for each step, 10 images (batch size) and 20 images (batch size) are sent for training through the network individually. At the same time, a 50% dropout was also been used in the experiment. Rectified Linear Unit (ReLU) as activation functions and Cross Entropy Loss function are applied to guide CNN training. The Adam algorithm calculates the gradient of entire data sets and updates values in the opposite direction to the gradients until a local minimum is found. Top-1 accuracy is applied to measure the performance of processes. This entire test experiment will be repeated ten times until the average of performance is obtained. The results of our experiment are used to compare with the performance of others. To check and analyze the implementation of the proposed method, 10 image samples from each category are collected into the test set.

According to the outstanding performance of our experiment, we took the best parameter as the final performance and compared it with the model consisting of original images. We performed various experiments by altering the image size. For example, we drafted the size of each picture after pre-processing as 350x350x3, 500x500x3, and 800x800x3. Among the data set, 90% of the data are the training set and 10% are the test set. 10,000 steps were applied for CNN training process ten times. Based on Top-1 error calculation, the result of the proposed system was acquired. According to whether the prediction is the same as the target category, the Top-1 error rate returns a Boolean value.

Table 2: Performance Comparison with Various Parameters

Input Image size	Numbers steps	Learning rate	Batch size	(VGG-16 System) Top1 Error	(Proposed FCN-8s-VGG-16 system) Top1 Error
350x350x3	10000	0.001	10	82.69%	92.15%
350x350x3	10000	0.001	20	84.34%	93.21%
500x500x3	10000	0.001	10	92.43%	97.25%
500x500x3	10000	0.001	20	93.54%	98.33%
800x800x3	10000	0.001	10	99.84%	99.57%
800x800x3	10000	0.001	20	99.88%	99.69%

Table 2 shows the comparison performance of various parameters. As seen from Table 2, the input size of 350x350x3 and 500x500x3 of the FCN-8s-VGG-16 system obtained a higher accuracy rate than the VGG-16 System which is trained without segmentation technology. Meanwhile, the accuracy rate of 800x800x3 input size achieved the closest result with the VGG-16 System. Furthermore, the running time has reduced on average 4.52% than the VGG-16 System in all three experiments owing to reduced pixel with segmentation technology. Figure 6 demonstrates the relationship between the batch size, image size and the system. We can conclude that as the batch size increased, both accuracy rate of VGG-16 System and FCN-8s-VGG-16 have increased slightly.

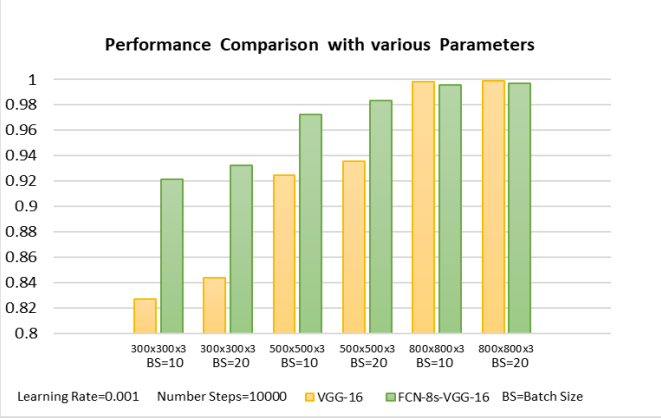


Figure 6: Performance Comparison with various Parameters

However, this will expand running time as well. Therefore, the batch size and running time could be a future trade off work for us. Above all, we could gather that image input size with high pixels will have a higher accuracy rate for recognition.

5. Conclusion and Future Work

In this paper, we proposed the FCN-8s-VGG-16 system to segment face images and recognize segmented images with high precision. Our face recognition system, FCN-8s-VGG-16 is composed of face image segmentation with FCN-8s [16] and

VGG-Face fine-tuning model recognition. Experiments were conducted on Celebrity-Face-Recognition dataset [40] for training and testing purposes. Compared with the original face recognition without segmentation accuracy, the effective classification rate using different parameters has been significantly improved from 92.15% to 99.69%, reaching a level of 82.69% to 99.88%. It also demonstrates that increasing the size of the input and batch number, while keeping the learning rate and number of steps constant, will improve the accuracy rate in certain situations. However, such observation is not completely ensured for other systems.

Accuracies on Top-1 in face recognition results means that some limitations happened in our algorithm, especially when applied to a wild data sets: First, even though FCN-8s has successfully segmented most face images, there were still a few images that did not achieve ideal results. For example, some small black areas occur after segmentation, which will affect the accuracy of performance. There are a few things that have triggered these failures. First, it takes three training processes to get the FCN-8s, which is not sensitive enough to combine the details of the image. This is because when the decoding is performed, that is, when the original image is restored, the label map of the input upsampling layer is too sparse. Secondly, FCN-8s does not consider the relationship between pixels during classification, and it lacks spatial consistency. Additionally, we only computed limited image input size, batch sizes, and learning rate, so it is very possible that we will receive a better accuracy rate if there are more parameters to choose from, with time allowed. Lastly, with the fine-tuning FCN model [16], the model will be more robust [43]. Only one recognition model processed in the experiment is limited. Some other CNN model has achieved an excellent performance in recent competition. For example, Google Net [44] is a winning architecture on ImageNet 2014. We will be applying it as our recognition system in future research for comparing with VGG-16. Besides, in reality, a human face is made of a 3D model; it's more realistic to segment and recognize images based on 3D faces. [45] has proposed a system of face recognition 3D based on segmentation by classifying fields of facial images before and after fusion of color and depth images. It brings us a potential research in face recognition study that we could investigate more about 3D image segmentation technologies instead of 2D.

Additionally, we will develop a mobile app such as [5], to complete the SFR-IoT architecture that could receive notification from our system though IoT devices when recognizing face in order to make contribution in real world application. In the future, data encoding and decoding will be applied to untrusted devices and servers to protect privacy data to make software more powerful and trustworthy [46].

Above all, the comparison reveals significant improvement in performance. Further experiments can be done by improving segmentation technology or by changing the recognition model.

References

[1] S. Zalke, P.M.-2015 I.C. On, U. 2015, Survey on implementation of enhanced uniform circular local binary pattern algorithm for sketch recognition, Oct. 2020.

[2] J. Cui, J. Li, Y.H.-2008 I.C. on, undefined 2008, "Research on Face Recognition Method based on Cognitive Mechanism," Ieeexplore.Ieee.Org, Oct. 2020.

[3] X. Dong, H. Huang, H.W.- ISCSCT, undefined 2010, "A comparative study of several face recognition algorithms based on pca."

[4] H. Yu, J. Yang, "A direct LDA algorithm for high-dimensional data — with application to face recognition," Pattern Recognition, **34**(10), 2067–2070, 2001, doi:10.1016/s0031-3203(00)00162-x.

[5] J. Bobulski, "Face recognition method with two-dimensional HMM," in Advances in Intelligent Systems and Computing, Springer Verlag: 317–325, 2016, doi:10.1007/978-3-319-26227-7_30.

[6] T. Ojala, M. Pietikäinen, D. Harwood, "A comparative study of texture measures with classification based on feature distributions," Pattern Recognition, **29**(1), 51–59, 1996, doi:10.1016/0031-3203(95)00067-4.

[7] S. Zalke, P.S. Mohod, "Survey on implementation of enhanced uniform circular local binary pattern algorithm for sketch recognition," in ICIIECS 2015 - 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems, Institute of Electrical and Electronics Engineers Inc., 2015, doi:10.1109/ICIIECS.2015.7192917.

[8] S. Karen, Z. Andrew, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Karen," American Journal of Health-System Pharmacy, **75**(6), 398–406, 2018.

[9] A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, **60**(6), 84–90, 2017, doi:10.1145/3065386.

[10] O.M. Parkhi, A. Vedaldi, A. Zisserman, "Deep Face Recognition," in Procedings of the British Machine Vision Conference 2015, British Machine Vision Association: 41.1-41.12, 2015, doi:10.5244/C.29.41.

[11] D. Virmani, T. Sharma, M. Garg, GAPER: Gender, Age, Pose and Emotion Recognition Using Deep Neural Networks, Springer, Singapore: 287–297, 2021, doi:10.1007/978-981-15-5463-6_26.

[12] F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June, 815–823, 2015, doi:10.1109/CVPR.2015.7298682.

[13] A. Aminian, M.S. Beni, "Face detection using color segmentation and RHT," in 3rd International Conference on Pattern Analysis and Image Analysis, IPRIA 2017, Institute of Electrical and Electronics Engineers Inc.: 128–132, 2017, doi:10.1109/PRIA.2017.7983032.

[14] S.D.F. Hilado, E.P. Dadios, R.C. Gustilo, "Face detection using neural networks with skin segmentation," in Proceedings of the 2011 IEEE 5th International Conference on Cybernetics and Intelligent Systems, CIS 2011, 261–265, 2011, doi:10.1109/ICCIS.2011.6070338.

[15] J.C. Klontz, A.K. Jain, "A case study of automated face recognition: The Boston marathon bombings suspects," Computer, **46**(11), 91–94, 2013, doi:10.1109/MC.2013.377.

[16] Y. Nirkin, I. Masi, A.T. Tuăn, T. Hassner, G. Medioni, "On face segmentation, face swapping, and face perception," in Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Institute of Electrical and Electronics Engineers Inc.: 98–105, 2018, doi:10.1109/FG.2018.00024.

[17] K. Okamoto, K. Yanai, "An automatic calorie estimation system of food images on a smartphone," in MADiMa 2016 - Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, co-located with ACM Multimedia 2016, Association for Computing Machinery, Inc: 63–70, 2016, doi:10.1145/2986035.2986040.

[18] C. Rother, V. Kolmogorov, A. Blake, "'GrabCut' - Interactive foreground extraction using iterated graph cuts," in ACM Transactions on Graphics, 309–314, 2004, doi:10.1145/1015706.1015720.

[19] Y. Ben Fadhel, S. Ktata, T. Kraiem, "Cardiac scintigraphic images segmentation techniques," in 2nd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2016, Institute of Electrical and Electronics Engineers Inc.: 364–369, 2016, doi:10.1109/ATSIP.2016.7523107.

[20] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society: 431–440, 2015, doi:10.1109/CVPR.2015.7298965.

[21] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, "One-Shot Video Object Segmentation," Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January, 5320–5329, 2016.

[22] S. Qin, S. Kim, R. Manduchi, "Automatic skin and hair masking using fully convolutional networks," in Proceedings - IEEE International Conference on Multimedia and Expo, IEEE Computer Society: 103–108, 2017, doi:10.1109/ICME.2017.8019339.

[23] LFW Face Database | Part Labels, Oct. 2020.

[24] D. Bussey, A. Glandon, L. Vidyaratne, M. Alam, K.M. Iftekharuddin, "Convolutional neural network transfer learning for robust face recognition in NAO humanoid robot," in 2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings, Institute of Electrical and Electronics Engineers Inc.: 1–7, 2018, doi:10.1109/SSCI.2017.8285347.
[25] D. Yi, Z. Lei, S. Liao, S.Z. Li, "Learning Face Representation from Scratch," 2014.
[26] A. Raj, S. Gupta, N.K. Verma, "Face detection and recognition based on skin segmentation and CNN," in 11th International Conference on Industrial and Information Systems, ICIIS 2016 - Conference Proceedings, Institute of Electrical and Electronics Engineers Inc.: 54–59, 2016, doi:10.1109/ICIINFS.2016.8262907.
[27] STL-10 dataset, Oct. 2020.
[28] M. Coskun, A. Ucar, O. Yildirim, Y. Demir, "Face recognition based on convolutional neural network," in Proceedings of the International Conference on Modern Electrical and Energy Systems, MEES 2017, Institute of Electrical and Electronics Engineers Inc.: 376–379, 2017, doi:10.1109/MEES.2017.8248937.
[29] F. Gao, J. Liu, "Face recognition using segmentation technology," in Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, Institute of Electrical and Electronics Engineers Inc.: 545–548, 2019, doi:10.1109/ICMLA.2019.00102.
[30] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, "Neural Codes for Image Retrieval," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8689 LNCS(PART 1), 584–599, 2014.
[31] The PASCAL Visual Object Classes Homepage, Oct. 2020.
[32] G. Csurka, D. Larlus, F. Perronnin, "What is a good evaluation measure for semantic segmentation?," BMVC 2013 - Electronic Proceedings of the British Machine Vision Conference 2013, 2013, doi:10.5244/C.27.32.
[33] B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, A.K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June, 1931–1939, 2015, doi:10.1109/CVPR.2015.7298803.
[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, **115**(3), 211–252, 2014.
[35] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, "Microsoft COCO: Common objects in context," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag: 740–755, 2014, doi:10.1007/978-3-319-10602-1_48.
[36] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Oct. 2020.
[37] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the Face Recognition Grand Challenge *, 2005.
[38] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," 2016.
[39] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matějka, J. Černocký, N. Poh, J. Kittler, A. Larcher, C. Lévy, D. Matrouf, J.F. Bonastre, P. Tresadern, T. Cootes, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2012, 635–640, 2012, doi:10.1109/ICMEW.2012.116.
[40] GitHub - prateekmehta59/Celebrity-Face-Recognition-Dataset: Dataset of around 800k images consisting of 1100 Famous Celebrities and an Unknown class to classify unknown faces, Oct. 2020.
[41] J.J. Lv, X.H. Shao, J.S. Huang, X.D. Zhou, X. Zhou, "Data augmentation for face recognition," Neurocomputing, **230**, 184–196, 2017, doi:10.1016/j.neucom.2016.12.025.
[42] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, 1995.
[43] Y. Sun, D. Liang, X. Wang, X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," 2015.
[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society: 1–9, 2015, doi:10.1109/CVPR.2015.7298594.
[45] M. Belahcene, A. Chouchane, M. Amin Benatia, M. Halitim, "3D and 2D face recognition based on image segmentation," in 2014 International Workshop on Computational Intelligence for Multimedia Understanding, IWCIM 2014, Institute of Electrical and Electronics Engineers Inc., 2014, doi:10.1109/IWCIM.2014.7008800.
[46] W. Xue, W. Hu, P. Gauranvaram, A. Seneviratne, S. Jha, "An Efficient Privacy-preserving IoT System for Face Recognition," in Proceedings - 2020 Workshop on Emerging Technologies for Security in IoT, ETSecIoT 2020, Institute of Electrical and Electronics Engineers Inc.: 7–11, 2020, doi:10.1109/ETSecIoT50046.2020.00006.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 600-611 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Real-Time Traffic Sign Detection and Recognition System for Assistive Driving

Adonis Santos[*,1,2], Patricia Angela Abu[3], Carlos Oppus[1], Rosula Reyes[1]

[1]*Electronics, Computer, and Communications Engineering Dept., Ateneo de Manila University, Quezon City, 1108, Philippines*

[2]*College of Engineering, First Asia Institute of Technology and Humanities, Tanauan City, 4232, Philippines*

[3]*Department of Information Systems and Computer Science, Ateneo de Manila University, Quezon City, 1108, Philippines*

ARTICLE INFO

Article history:
Received: 12 June, 2020
Accepted: 04 August, 2020
Online: 25 August, 2020

Keywords:
traffic sign detection and recognition
embedded system
computer vision
machine learning
convolutional neural network

ABSTRACT

Road traffic accidents are primarily caused by drivers error. Safer roads infrastructure and facilities like traffic signs and signals are built to aid drivers on the road. But several factors affect the awareness of drivers to traffic signs including visual complexity, environmental condition, and poor drivers education. This led to the development of different ADAs like TSDR that enhances vehicle system. More complex algorithms are implemented for improvement but this affects the performance of a real-time system. This study implements a real-time traffic sign detection and recognition system with voice alert using Python. It aims to establish the proper trade-off between accuracy and speed in the design of the system. Four pre-processing and object detection methods in different color spaces are evaluated for efficient, accurate, and fast segmentation of the region of interest. In the recognition phase, ten classification algorithms are implemented and evaluated to determine which will provide the best performance in both accuracy and processing speed for traffic sign recognition. This study has determined that Shadow and Highlight Invariant Method for the pre-processing and color segmentation stage provided the best trade-off between detection success rate (77.05%) and processing speed (31.2ms). Convolutional Neural Network for the recognition stage not only provided the best trade-off between classification accuracy (92.97%) and processing speed (7.81ms) but also has the best performance even with lesser number of training data. Embedded system implementation utilized Nvidia Jetson Nano with interface Waveshare IMX219-77 camera, Nvidia 7" LCD and generic speaker and programmed in Python with OpenCV, sci-kit learn and Pytorch libraries. It is capable of running at an adaptive frame rate from 8-12 frames per second with no detection and down to approximately 1 frame per second when there is traffic sign detected.

1 Introduction

The World Health Organization (WHO) in 2013 reported that road traffic accidents that result to loss of lives and damages to properties will continue to become a global challenge due to rapid motorization and insufficient action of national governments [1]. In the Philippines, traffic accidents are mainly due to driver's error [2] or violation [3] and this prompts government agencies to focus on regulation of driver by means of licensing and provide safer roads by building infrastructures like traffic lights and signs. Another solution includes the development of advanced driver-assistance systems (ADAs) that enhance vehicle systems to aid drivers on the road. But this system faces its own challenge, for instance, traffic sign detection and recognition (TSDR) system is often exposed to different weather phenomenon and road conditions [4] that degrades their effectivity [5]. This can be solved by employing more complex algorithms but this has an adverse effect to processing speed that is a must for smart vehicle systems. This trade-off between system accuracy and speed is still the biggest challenge in the implementation of ATSDR systems.

This paper "Real-Time Traffic Sign Detection and Recognition System for Assistive Driving" is an extension of work originally presented by the authors in International Symposium on Multimedia and Communication Technology (ISMAC) 2019 entitled "Traffic

[*]Corresponding Author: Adonis Santos, Ateneo de Manila University, adonis.santos@obf.ateneo.edu

Sign Detection and Recognition for Assistive Driving" [6]. The goal of this study is to help solve the problem of drivers neglect and lack of road education by implementing an ATSDR system that automatically detects and recognizes road traffic signs then provides information to the driver about the meaning of the signs. This work is the full implementation of a study that the previous work [6] is a part of. The differences between this work and the previous work include (1) gathering and utilization of new traffic road image dataset and road sign image datasets for training and testing the system, (2) evaluation of four pre-processing and detection methods for the detection stage, (3) addition of a Convolutional Neural Network (CNN) deep learning algorithm in the evaluation for the recognition stage aside from the original nine methods, (4) the detection and recognition methods are evaluated not only using the segmentation and classification success rate (accuracy) metric but also processing speed, (5) system-level optimization is done to improve accuracy by fine-tuning sub-processes and a negative class is added to reduce false detection, (6) the models are fully implemented not only in a personal computer but in an embedded system using microcomputer with interface camera, display and speaker, and (7) the system installed in a sedan car is tested in an actual road setup.

Traffic signs defined by the Department of Public Works and Highways to be recognized must be strategically positioned, clear and fully visible [7] and captured in good weather condition during daytime and will be limited only to six regulatory signs corresponding to speed limit, no overtaking and no turning shown in Figure 1. These are generally characterized by their red circular boundaries. The detection and recognition will only be limited to a single traffic sign in a single video frame. The capturing device will be installed to a sedan car with a target maximum speed of at least sixty (60) kilometers per hour. The system will be able to provide voice alert to the driver before the car completely passed by the traffic sign. Implementation of the system will make use of methods from previous studies using Python and open-source libraries.



Figure 1: Traffic signs to be recognized

Several implementations of traffic sign detection and recognition systems used different algorithms in pre-processing, detection and recognition phases. Pre-processing stage is necessary to suppress noise and improve image performance [8]. A study [6] showed the detection improvements in traffic sign recognition provided by the use of pre-processing and filtering methods by means of contrast stretching, color normalization and image enhancement before getting the regions of interest. Other techniques are applied depending on the choice of segmentation method in the detection stage which can be color-based, shape-based or combination of both, though color-based is widely implemented in the detection of road signs [9]. Shape based segmentation may have problem when there are similar objects to the traffic signs. It can be quite computationally expensive, and by nature, sensitive to noise. Many shape detectors are slow in computing over large and complex images [10] which will become an issue in a real-time system. The complexity of color based lies in the usage of three intensity values depending on the color space instead of just working in gray level image as in shape based. But still, color segmentation is used more often as color usually provides enough information on the meaning of the traffic sign to be recognized than shape segmentation. Difficulties in color-based detection include lack of standard colors among countries, and various phenomenons and different imaging conditions that cause color variations. In these cases, traffic sign images may appear with varying illumination, blurred, faded, highlighted, noisy, and physically damaged [5] affecting the performance of the color segmentation process. A study [11] solved this problem by instead of working on absolute RGB values, relative RGB values are used for they are almost unchanged with various illumination circumstances. Another study [8] combined four color spaces namely RGB, HSV, YIQ, and XYZ based on histogram to provide more accurate segmentation. Color is segmented using K-means clustering and effective robust kernel based Fuzzy C-means clustering resulting to a more accurate segmentation but has increased computational complexity. In [10], color segmentation in RGB color space followed by shape segmentation using joint transform correlator (JTC) template matching had been implemented for a more efficient system. A study [12] solved the unreliability issue of RGB due to sensitivity to illumination variation by working on HSV color space along with shape based filtering through template matching. HSV is a non-linear transform of RGB where any color can be represented by hue and the depth or purity of color by saturation [13] whose values are not affected by light intensity [12]. In [14], three methods of segmentation based on improved HSL color space were implemented that provided independence between chromatic and achromatic components resulting to robustness in changes in external light conditions. Based on success rate in segmentation, thresholding using global mean of luminance is the best method followed by segmentation using region growing then by improved HSL normalization. Another study [15] solved the effect of poor lighting by RGB channel enhancement using histogram equalization followed by color constancy algorithm before segmentation in HSV color space. Color constancy is the ability to correctly determine the color of objects in view irrespective of the illuminant [16].

Several studies evaluated different segmentation algorithms based on segmentation success rate and processing time. A study [5] evaluated four color segmentation methods namely, Dynamic Threshold Algorithm, Modified de la Escaleras Algorithm, Fuzzy Color Segmentation Algorithm, and Shadow and Highlight Invariant Algorithm, in red color segmentation of traffic sign in different road conditions and distances. Fuzzy Color Segmentation has the best segmentation success rate followed by Shadow and Highlight Invariant. In processing time, Shadow and Highlight Invariant has

the best average time, Fuzzy Algorithm has the best standard deviation and the most stable. Using the dataset with traffic image taken at different distances and conditions, Shadow and Highlight Invariant has the best performance. Another study [17] evaluated shape-based and color-based segmentation methods for traffic sign recognition using Matlab. Metrics in evaluation include number of recognized signs, lost signs, falsely recognized signs and speed of recognition. Among the evaluated methods are RGB Normalized Thresholding, Hue and Saturation Thresholding, Hue and Saturation Color Enhancement Thresholding, Ohta Space Thresholding, Grey-Scale Edge Removal, Canny Edge Removal, and Color Edge Removal. Results showed that color space thresholding methods are the best methods in segmentation success rate and speed. RGB normalized method provided best recognition score among the algorithms evaluated. Edge detection methods had the worst execution speed and excessive false detection percentage. Findings showed that edge-detection methods may be used as a complement to other color-segmentation methods, but they cannot be used alone, and normalization, as in the RGB Normalized method or Ohta Space Thresholding, improves performance and represents a low-cost operation.

In recognition stage, the use of neural networks in classification of road signs provides considerable saving in computational requirement as compared to template matching [9]. Several studies implemented deep learning algorithms for traffic sign recognition due to their high accuracy with faster processing speed [18]-[20]. In [21], detection and classification of circular traffic signs used CNN after pre-processing and Hough Transform using TensorFlow using accuracy metric attaining 98.2%. In [22], 24 traffic sign classes were used in recognition using deep CNN that achieved 100% accuracy. Another study [23] used masked R-CNN in the detection (Region Proposal Network) and recognition (Fast R-CNN) of over 200 traffic sign categories where data augmetation is implemented and resulted to less than 3% error rate. The disadvantage of neural network lies on the training overhead and architecture of multi-layer networks cannot be adapted for online application. In order to distinguish among different classes for classification, specific features of the input data must be chosen. In [4], histogram of oriented gradient is used as feature descriptor for classification of traffic signs as it captures color and shape as one feature.

2 Methodology

Shown in Figure 2 is the methodology for the proposed real-time traffic sign detection and recognition system starting with image acquisition for training the detection and recognition models followed by testing. The system will be designed in Python using the OpenCV, sci-kit learn, and Pytorch libraries and will be implemented in a microcomputer with interfaced camera and speaker.

2.1 Data Acquisition

Input traffic data will be acquired by the system using a camera interfaced to the microcomputer. Videos taken will be processed per frame by the system. Some testing process will make use of traffic images to be stored in the microcomputer.

Four sets of data will be used in the implementation and testing of the system. For the detection phase, a set of local traffic images from Google Images containing actual road and traffic images with different traffic signs to be detected will be used to test the models. Another set of data is composed of road traffic images from dataset of Napier University containing actual road with traffic signs to be detected captured in different distances. The recognition phase model will utilize one set of data containing images cropped to fully view the standard traffic sign from actual Philippine roads. The last set of data is actual video footage from local road to test the full performance of the traffic sign detection and recognition system.

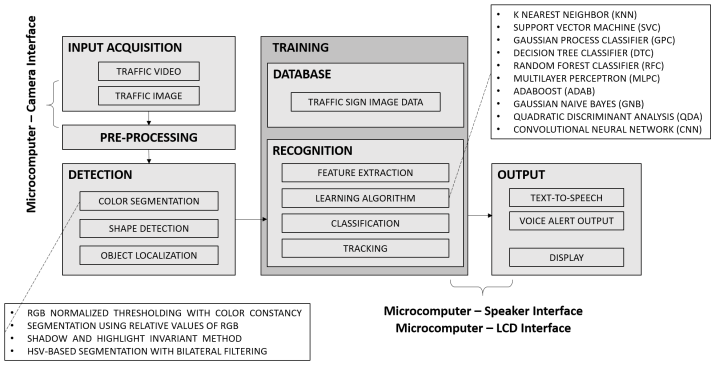


Figure 2: Overview of methodology

2.2 Pre-processing and Detection

The detection phase is composed of pre-processing, color-based segmentation, shape-based detection and object localization. This study will evaluate four different pre-processing and color-based segmentation methods that solved the problem of lighting variations affecting the detection phase. Methods to be evaluated that use different filtering techniques in different color spaces are RGB Normalized Thresholding with Color Constancy Algorithm, Segmentation using Relative Values of RGB, Shadow and Highlight Invariant Method, and HSV-based Segmentation with Bilateral Filtering.

2.2.1 RGB Normalized Thresholding with Color Constancy (CCM)

Histogram equalization of separate RGB channels is done followed by color constancy algorithm to extract the true color of the image [15]. The algorithm computes local space average color using a parallel grid of processing elements used to shift the color of the input pixel in the direction of the gray vector. Target color segmentation is done in HSV color space [16].

2.2.2 Segmentation using Relative Values of RGB (RRGB)

Segmentation that is based in three rules using the RGB color space. The pixel will be regarded as the target red color when (i) red component is greater than α, (ii) difference between red and green components is between β_1 and β_2 , and (iii) the difference between the red and blue components is between γ_1 and γ_2 . The values of $\alpha, \beta_1, \beta_2, \gamma_1$, and γ_2 are experimentally found to represent the red color perceived in different lighting conditions [11].

2.2.3 Shadow and Highlight Invariant Method (SHI)

Utilized the normalized HSV color space values. Segmentation happens in the hue image where hue is 255 if the pixel's hue is between 240 and 255, or hue is between 0 and 10. Hue will be set to 0 if pixel's saturation is less than 40, or pixel's value does not fall between 30 and 230. The hue image and seed image from the new hue is used, applying region growing algorithm to find target proper region [5].

2.2.4 HSV-based Segmentation with Bilateral Filtering (BFM)

Traffic images undergo pre-processing in the RGB color space before the detection stage. Bilateral filtering smoothens the image while preserving the edges by means of a nonlinear combination of nearby pixel values [24]. The weight assigned to each neighbor decreases with both the distance in the image plane and the distance on the intensity axis.

One of the four methods will be used for color-based segmentation to improve the efficiency and speed of detection of the next stage by reducing search areas that do not contain any possible traffic sign [9, 25] defined by pixels falling outside the target red color.

The next phase is the shape detection using Hough transform to detect the circular shape of the traffic sign. Hough transform has been recognized as a robust technique in curve detection that can detect objects even when polluted by noise [26]. It transforms a set of feature points in the image space into a set of accumulated votes in a parameter space. For each feature point, votes are accumulated in an accumulator array for all parameter combinations. Only one circle, by means of highest number of votes, will be detected as increasing the number of detection will also increase the time required to process multiple regions for the succeeding phases. The last stage in the detection phase is the object localization or the cropping of the candidate traffic sign from the traffic image to be used for the classification stage.

2.3 Feature Extraction

Histogram of Oriented Gradients (HOG) features are extracted from the candidate traffic sign images by the Python program to become the key feature for classification for the nine machine learning algorithms. The image from the dataset undergoes RGB to grayscale conversion and edge detection. The resulting image is then resized before using the built-in HOG extractor library in Python [6].

The steps above are important in ensuring optimum accuracy of the features when training the machine learning classifiers. HOG feature provides excellent performance relative to other existing feature sets [27].

2.4 Classifier

To classify the candidate traffic sign images, learning algorithms will be used. For this study, ten learning algorithms are implemented using sci-kit learn library [28] and Pytorch [29] in Python. The classifiers [30] to be evaluated based on accuracy and processing speed for traffic sign recognition are as follows.

2.4.1 K Nearest Neighbor (KNN)

KNN is a method where classification is computed from a simple majority vote of the nearest neighbors of each point [31]. Existing Python implementation of KNN uses model `KNeighborsClassifier(n_neighbors=1)` where a single neighbor is obtained for prediction.

2.4.2 Support Vector Machine (SVC)

An SVM is a discriminative classifier formally defined by a separating hyperplane. The algorithm outputs an optimal hyperplane using labeled training data which categorizes where the new examples are [32]. Implementation of the SVC will make use of the model `SVC(kernel="linear", C=0.025)` for linear kernel and with regularization parameter C of 0.025.

2.4.3 Gaussian Process Classifier (GPC)

Gaussian Processes predict by computing for empirical confidence interval using probabilistic Gaussian distribution [33]. For GPC implementation, the model `GaussianProcessClassifier(1.0 * RBF(1.0))` will be used utilizing the default squared-exponential kernel as radial basis function.

2.4.4 Decision Tree Classifier (DTC)

Decision Trees are non-parametric supervised learning method used for classification and regression with goal of creating a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [34]. The model to be used will be `DecisionTreeClassifier(max_depth=5)` with maximum depth of the decision tree equal to 5. The deeper the tree, the more complex the if-then-else decision rules and the fitter the model.

2.4.5 Random Forest Classifier (RFC)

Random Forest algorithm is a supervised classification algorithm which creates a forest with a number of trees. The higher the number of trees in the forest gives the high accuracy results [35]. The model to be used is `RandomForestClassifier(max_depth=5, n_estimators=10, max_features=1)` utilizing 10 decision trees, each with maximum depth of 5 and considering only 1 feature when looking for the best split.

2.4.6 Multilayer Perceptron (MLPC)

MLPC can be viewed as a logistic regression classifier where the input is first transformed using learnt non-linear transformation that projects the input data into a space where it becomes linearly separable [36]. Existing implementation model `MLPClassifier (alpha=1e-5)` will be used with L2 penalty parameter or regularization term equal to 0.00001. Smaller value of alpha may fix high bias which is a sign of underfitting by encouraging larger weights, potentially resulting in a more complicated decision boundary, while larger alpha value helps in avoiding overfitting by penalizing weights with large magnitudes [30]

2.4.7 AdaBoost (AdaB)

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases [37]. The model `AdaBoostClassifier()` will be used utilizing default hyper-parameter values where the base estimator is a decision tree classifier with a maximum depth of 1 and the maximum number of estimators is 50.

2.4.8 Gaussian Naive Bayes (GNB)

GNB is a special type of NB algorithm used when the features have continuous values and assumes that all the features are following a Gaussian distribution i.e, normal distribution. It scales cubically with the size of the dataset and might be considerably faster [38]. The model to be utilized calls `GaussianNB()`.

2.4.9 Quadratic Discriminant Analysis (QDA)

QDA is closely related to LDA but uses quadratic decision surface. It can be derived from simple probabilistic model where prediction is obtained by using Bayes' rule. The model fits a Gaussian density to each class [39]. In QDA, individual covariance matrix is estimated for every class of observations. QDA becomes useful when there is prior knowledge that individual classes exhibit distinct covariances. The model that will be utilized is `QuadraticDiscriminantAnalysis()`.

2.4.10 Convolutional Neural Network (CNN)

It is a neural network mathematical model based on connected artificial neurons similar to biological neural networks. Neurons are organized in layers and connections are established between neurons of adjacent layers. The output layer neurons amount is equal to the number of classifying classes with probabilities showing the possibility that the input vector belongs to a corresponding class [18]. The training process is to minimize the cost function with minimization methods based on the gradient decent also known as backpropagation [19]. The CNN model to be utilized is based on LeNet-5 architecture [40] composed of seven layer combination of convolution, pooling with rectified linear unit (ReLU) activation function and fully connected layers shown in Figure 3.

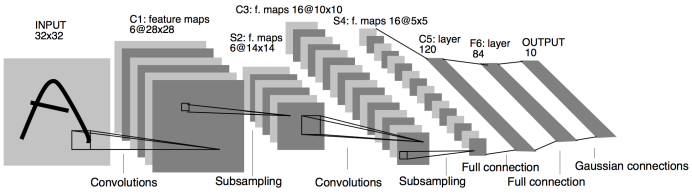


Figure 3: CNN model based on LeNet-5 architecture [40]

Table 1 summarizes the modification to hyper-parameter values done for the implementation of all the learning algorithms using scikit learn and Pytorch libraries in Python. All other hyper-parameters not modified are set to their default values.

Table 1: Summary of modified hyper-parameters for learning algorithms implementation

Learning Algorithm	Modified Hyper-parameter
KNN	n_neighbors=1
SVC	kernel="linear", C=0.025
GPC	None
DTC	max_depth=5
RFC	max_depth=5, n_estimators=10, max_features=1
MLPC	alpha=1e-5
AdaB	None
GNB	None
QDA	None
CNN	6 classes output

2.5 Evaluation

The evaluation phase of the designed model will be divided into three: detection, recognition, and system level testing. For this study, the detection and recognition models will be tested individually based on accuracy and speed using Lenovo U410 ultrabook before a system level testing is performed.

2.5.1 Detection Model Testing

After implementation of the four pre-processing and segmentation models, traffic images will be fed. Accuracy will be computed using Equation 1. TP is defined by correct detection on image with traffic sign, TN when there is no detection for traffic image without traffic sign, FP for detection when the image has no traffic sign or detection of other part of the image with traffic sign, and FN for non-detection of the traffic image with traffic sign. Processing speed of the sub-processes will also be measured for the entire detection process.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2.5.2 Recognition Model Testing

For the testing of the implemented recognition model, cross validation method will be used wherein portion of the dataset will be used for training while the remaining will be for testing. A confusion matrix will then be generated. Evaluation metric to be used is accuracy that determines whether a label predicted for a sample exactly matches the corresponding true label of the sample. Equation 1 will be used to compute for accuracy. Processing time of the entire recognition model and its sub-processes will also be recorded.

2.5.3 System Level Testing

For the system level testing, the best detection and recognition models based on the evaluation will be integrated to create the entire model of the system using Nvidia Jetson Nano with interface Waveshare 7" LCD, IMX219-77 camera and generic speaker. Accuracy and speed of detection will be tested using an actual video taken in local road. System level accuracy will be computed to be the product of detection accuracy and recognition accuracy using Equation 1.

A successful recognition means that the system has provided a voice alert about the detected traffic sign. A text-to-speech python module will be used for this function and the corresponding voice alert for each traffic sign detected is shown in Table 2. A text annotation will also be generated to the display.

Table 2: Voice alert and text annotation for target traffic signs

Traffic Sign	Voice Alert	Text Annotation
Left Turn Prohibition	No Left Turn	No Left Turn
Right Turn Prohibition	No Right Turn	No Right Turn
U-Turn Prohibition	No U-Turn	No U-Turn
Overtaking Prohibition	No Overtaking	No Overtaking
30kph Speed Limit	Thirty kph	30 kph
60kph Speed Limit	Sixty kph	60 kph

3 Results and Analysis

3.1 Dataset

A total of 2,194 images are used to test the detection phase of the models. Of these, 2,170 are traffic images from online sources with traffic sign having different viewing angle and position on image. These are evenly distributed to the six traffic sign classes and an additional class for images without traffic sign with 1366x768 pixels resolution. The remaining 24 images with 1936x1296 pixels resolution are foreign traffic images captured at different distances from 10 meters to 120 meters with 10-meter increment. For the recognition models, 12,120 images will be used to evaluate speed and accuracy. This dataset is augmented using different cropping methods including zooming, north-west cropping and south-east cropping to act as learning compensation when segmentation of the traffic sign is not centered. The dataset will be divided into train data (10%, 5%, 1%) and test data (90%, 95%, 99%) to determine the effect of the number of test data to models' accuracy. Another set contains six video footages with 360 frames of traffic images, 308 frames of which contain traffic sign. These are recorded using the actual system camera setup attached in a sedan car while approaching the traffic sign to be detected. Figure 4 shows a sample frame image from a video of a local road while Figure 5 shows sample images from the Napier University traffic dataset.



Figure 4: Sample traffic image from local road

Figure 6 shows the sample traffic signs from the dataset to be classified. This dataset contains a total of 2,020 images per class for the training and testing phases and is composed of different signs ranging from high resolution images up to 699x693 pixels down to low-resolution images of 12x12 pixels.



Figure 5: Sample traffic images from Napier University dataset



Figure 6: Sample traffic sign dataset

3.2 Traffic Sign Detection

The detection process is summarized by Figure 7 where the highlighted blocks are to be replaced by each of the four detection methods to be evaluated. From the original traffic image, red pixels are segmented using color thresholding. The resulting mask undergoes edge detection before Hough transform is applied to detect circular shape. The center and radius information is used to crop the candidate traffic sign for the next phase of classification. A traffic sign is considered detected if the cropped image fully shows the entirety of the traffic sign including the red circular boundary.

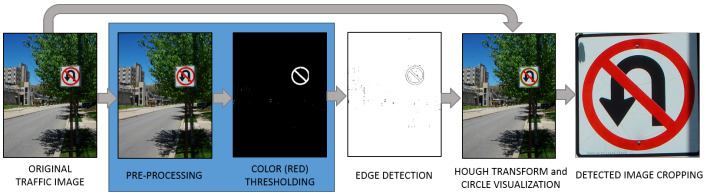


Figure 7: Traffic sign detection process

Experimentation for segmenting the red boundary of the traffic sign using HSV and relative RGB color spaces have resulted to the following threshold range summarized in Table 3 and Table 4. Figure 8 shows masks of segmented red color from an original traffic image using Relative RGB and HSV color segmentation.

Table 3: Range values for HSV color segmentation

Color Component/Channel	Range
Hue	$0 \leq h \leq 10$ or $160 \leq h \leq 179$
Saturation	$80 \leq s \leq 255$
Value	$60 \leq v \leq 255$

Table 4: Range values for relative RGB color segmentation

Color Component/Channel	Range
Red	$60 \leq r \leq 255$
Red minus Green	$35 \leq r\text{-}g \leq 135$
Red minus Blue	$8 \leq r\text{-}b \leq 204$



Figure 8: Sample relative RGB and HSV segmented red mask

Table 5: Accuracy of pre-processing and color segmentation methods

Method	Total Images	TP	TN	FP	FN	Accuracy
RRGB	2,170	535	299	11	1,462	32.12%
SHI	2,170	576	292	18	1,284	40.00%
BFM	2,170	503	304	6	1,357	37.19%
CCM	2,170	586	294	16	1,274	40.55%

Table 5 summarized the evaluation of the four pre-processing and color segmentation methods. In terms of accuracy, Color Constancy method provided the best performance successfully detecting 586 traffic signs from 2,170 traffic images. It is closely followed by the Shadow and Highlight Invariant method with 576 detection. Bilateral Filtering method and Color Constancy method have Shadow and Highlight Invariant method for color detection but results show improvement with Color Constancy method while performance degradation for Bilateral Filtering method due to (1) color improvement done by the Color Constancy method to the traffic image, and (2) images from the dataset are generally blurry due to vehicle vibration and environmental parameters that affect the performance of the Bilateral Filtering method. For Relative RGB implementation, color segmentation resulted to less false detection. Figure 9 shows the result of successful detection of traffic signs for six traffic images taken from local road using Shadow and Highlight Invariant method.



Figure 9: Sample detection of traffic sign

Table 6: Number of successful segmentation of pre-processing and color segmentation methods vs. traffic sign capturing distance

Distance	RRGB	SHI	BFM	CCM
120m	0	0	0	0
110m	0	0	0	0
100m	0	0	0	1
90m	0	0	0	0
80m	0	0	0	1
70m	0	2	1	1
60m	0	2	2	2
50m	0	2	2	2
40m	0	2	2	2
30m	0	2	2	2
20m	0	2	2	2
10m	2	2	2	2

Table 6 shows the evaluation of segmentation success of traffic sign with varying capturing distances. The traffic sign is captured twice per given distance. Color Constancy algorithm correctly segmented 15 out of 24 traffic images in the dataset with detection

distance up to 100 meters. Shadow and Highlight Invariant Method correctly segmented 14 out of 24 images with detection distance up to 70 meters. Bilateral Filtering Method correctly detected 13 out of 24 images with detection distance up to 70 meters. Relative RGB Method has the worse performance detecting only 2 out of 24 images with detection distance up to 10 meters and false detecting 13 images. This result validates the accuracy evaluation of the four methods previously discussed in Table 5 and suggests that traffic sign distance affects the image feature clarity of the sign on the image and causes significant decrease in detection success rate.

Table 7 summarizes the processing speed of the four pre-processing and color segmentation methods including execution speed of their sub-processes. This timing does not include shape detection and ROI localization as these processes are common to all methods. Relative RGB segmentation has the fastest processing time closely followed by the Shadow & Highlight Invariant method. The slight difference is due to color space conversion from RGB to HSV. Different processor runtime caused variable timing result for color segmentation process for Shadow & Highlight Invariant, Bilateral Filtering, and Color Constancy. These two employed simple pre-processing while Color Constancy and Bilateral Filtering methods are pre-processing methods themselves and perform multiple and complex image filtering using kernels.

Considering accuracy and processing speed, the pre-processing and segmentation method that provided the best trade-off is Shadow and Highlight Invariant method with 40.00% segmentation success rate and 31.2 ms processing speed.

Table 7: Processing speed of pre-processing and color segmentation methods and sub-processes

Method	Pre-processing	Color Segmentation	Total
RRGB	10.4 ms	15.6 ms	26.0 ms
SHI	5.2 ms	26.0 ms	31.2 ms
BFM	270.8 ms	18.2 ms	289.0 ms
CCM	494.8 ms	20.8 ms	515.6 ms

3.3 Feature Extraction

HOG feature extraction of a traffic image is shown in Figure 10. It starts from grayscale conversion followed by edge detection then by HOG feature extraction. HOG feature represents the general structure of the image using the dominant orientation of pixel groups. This process transforms the 2D image into 1D vector histogram representing the gradient direction of the image.

3.4 Classifier Evaluation

Table 8 shows the result of implementation of traffic sign classification using Multi-layer Perceptron Classifier with variable image input resolution. This shows that more feature pixels will not always provide the best accuracy for a given classifier and the optimum image resolution considering accuracy and processing speed is 50x50 pixels.

Table 9 summarizes the result of evaluating the 10 classification algorithms in terms of accuracy with the number training data varied. Most of the model showed downward trend in accuracy when training data is decreased. GPC, though has one of the best accuracy for higher number of training data, provided the worse performance for decreasing training data since optimization is being done using squared-exponential kernel to best fit few training data leading to misclassification. Convolutional Neural Network implementation obtained the best accuracy among the models. This is due to learning optimization using backpropagation decreasing the loss function resulting to higher classification success rate. With lesser training data, CNN still has the best accuracy. This is important to consider to ensure that the model will still provide better accuracy even when trained with fewer data. Figure 11 shows the normalized confusion matrix generated for CNN classifier.

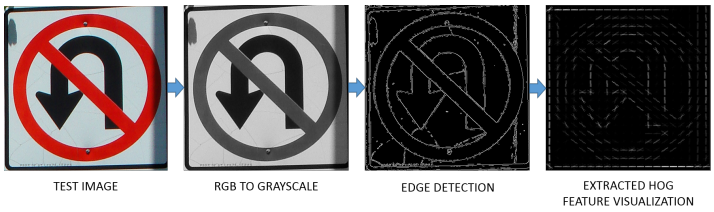


Figure 10: Test image feature extraction

Table 8: Effect of input image size to accuracy and processing speed for MLPC

Input Image Resolution	Accuracy	Processing Time
200x200 pixels	88.89%	454.5 ms
100x100 pixels	88.49%	331.2 ms
50x50 pixels	91.23%	283.2 ms
32x32pixels	85.32%	321.0 ms

Table 9: Effect of number of training data to accuracy

Classifier	10% Train	5% Train	1% Train	Average
KNN	94.63%	92.06%	86.70%	91.13%
SVM	89.72%	87.37%	57.21%	78.10%
GPC	93.87%	80.95%	16.63%	63.82%
DTC	72.45%	68.35%	57.90%	66.23%
RFC	76.59%	73.27%	60.17%	70.01%
MLPC	93.54%	91.19%	86.36%	90.36%
ADAB	62.50%	54.95%	35.64%	51.03%
GNB	85.44%	85.46%	77.50%	82.80%
QDA	54.07%	30.80%	26.28%	37.05%
CNN	96.64%	94.63%	87.64%	92.97%

K-Nearest Neighbor performed the second best in terms of accuracy closely followed by Multilayer Perceptron. This can be accounted to the characteristics of dataset. Each class of the dataset used is primarily composed of very similar images, and results suggest that KNN works better for data with close similarity.

The result of the evaluation of processing speed of the classification algorithms and their sub-processes for 90% of dataset is summarized in Table 10. CNN got the longest training time due to its nature of performing loss function optimization. Total recognition time is measured from color conversion until prediction and CNN obtained the fastest processing time. This is due to lesser

pre-processing being implemented by the method unlike the other algorithms that still perform color conversion and extract HOG features to every input traffic sign image to classify. Recognition time for all classifiers is below 20 ms except for GPC with 750.6 ms due to the use of more complex squared-exponential kernel.



Figure 11: Normalized confusion matrix for CNN classifier

Table 10: Processing speed of classifiers and sub-processes

Classifier	Feature Extract	Train Time	Color Convert	Image Resize	Predict	TOTAL TIME
KNN	16.04s	0.05s	1.56ms	10.42ms	5.21ms	17.19ms
SVC	16.04s	0.64s	1.56ms	10.42ms	0.00ms	11.98ms
GPC	16.04s	272.8s	1.56ms	10.42ms	738.6ms	750.6ms
DTC	16.04s	0.47s	1.56ms	10.42ms	0.00ms	11.98ms
RFC	16.04s	0.02s	1.56ms	10.42ms	2.08ms	14.06ms
MLPC	16.04s	6.62s	1.56ms	10.42ms	0.00ms	11.98ms
AdaB	16.04s	5.32s	1.56ms	10.42ms	6.77ms	18.75ms
GNB	16.04s	0.03s	1.56ms	10.42ms	2.60ms	14.58ms
QDA	16.04s	0.43s	1.56ms	10.42ms	4.16ms	16.15ms
CNN	-	581.4s	-	5.21ms	2.60ms	7.81ms

Table 11 shows result of optimization of the best detection and classification models based on the evaluation. These models are Shadow and Highlight Invariant method and Convolutional Neural Network classifier.

Table 11: Optimization of detection and recognition models

Parameter	Original Model	First Optimization	Second Optimization
Total Images	2,170	2,170	2,170
$TP_{detection}$	576	1,372	1,372
$TN_{detection}$	292	300	300
$FP_{detection}$	18	40	40
$FN_{detection}$	1,284	458	458
$TP_{recognition}$	547	1,240	1,257
$TN_{recognition}$	NA	NA	4
$FP_{recognition}$	47	172	147
$FN_{recognition}$	NA	NA	4

The first optimization resulted to improvement of correct detection and significant decrease in false detection. The process involved modification of threshold parameters of Hough circle transform for it to detect circle less strictly since traffic sign boundaries most often are elliptical in shape in the traffic image due to variable viewing angle. Hough transform is also adjusted to only detect circles with diameter less than one fifth of the image height. The limiting of circular detection is important since detecting larger circles will result to more pixel features thus will require longer processing time. Also in actual application, traffic signs do not often occupy the entire traffic image.

For the second optimization where a seventh class is added, results in Table 11 show further increase in correct recognition and decrease in false recognition. The additional class filters signs that are similar to any of the six classes, and once recognized as false class, no output will be generated by the model.

Results show that the optimized integrated model has a detection performance of 77.05% and recognition performance of 89.31%. The relatively low detection performance is due to the failure of the Hough transform to detect the circular boundary of the traffic sign since the appearance of the traffic signs are affected by the viewing angle and effect of environmental setting to image quality. Although adjusting the detector threshold to lower value resulted to detection improvement, further decrease in threshold resulted to more false detection. The overall accuracy of the model is 68.81%.

3.5 System Implementation

After implementing the system using Nvidia Jetson Nano platform, a timing comparison between the performance of the system using Lenovo U410 Ultrabook is presented in Table 12. Training time and accuracy of both implementation are relatively the same. The huge difference lies in processing time with Nvidia Jetson Nano almost doubled that of the Lenovo U410. This is due to the clocking difference of the processors of the two platforms.

Table 12: Performance comparison between Lenovo U410 and Nvidia Jetson Nano

Parameter	Lenovo U410	Nvidia Jetson Nano
Training Time	678.94 s	679.10 s
Processing Speed	131.3 ms	238.0 ms
Accuracy	95.58%	97.24%

Figure 12 shows some of the successful detection and recognition where traffic signs are annotated to each frame of actual local road video. Table 13 summarizes the timing analysis of the entire TSDR system with text annotation and voice alert output. The system has a total processing time of 1.0953 seconds. It takes about 0.73 ms to display the text annotation of the detected and recognized traffic sign while it takes 938.21 ms for the system to utter the entire phrase for voice alert. For voice alert system, the next frame can only be processed after it has finished all the required processing for current frame including uttering the meaning of the sign.

Given this timing values, the system without voice alert and only using text annotation as output can run at a maximum frame rate of 1/116.7ms or approximately 8 frames per second. For the system with voice alert output, the system requires a frame rate of 1/1.0953 seconds or less than 1 frame per second for it to run in real-time. Since voice alert must be implemented, having a frame rate of less than 1 fps will result to poor sampling causing traffic signs to not be detected especially for fast moving vehicles. A solution is to implement adaptive frame rate such that during non-detection,

the system will run at approximately 8 to 12 fps and when there is a traffic sign detected, frame rate shifts to approximately 1 fps to cater for the voice alert output. Figure 13 shows the hardware configuration block diagram of the system while Figure 14 shows the actual setup of the TSDR system for road testing. Figure 15 shows the testing of the TSDR system using actual footage from local road.



Figure 12: Sample detection and recognition of traffic frames

Table 13: Timing of sub-processes of the TSDR system

Process	Processing Time
Input Acquisition	13.69 ms
BGR to HSV Conversion	2.49 ms
Red Color Segmentation	5.22 ms
Gaussian Blurring	1.52 ms
Median Blurring	7.64 ms
Binary Thresholding	11.44 ms
Erosion	0.75 ms
Dilation	0.63 ms
Circle Detection	70.96 ms
ROI Cropping	7.09 ms
Image Transformation	4.79 ms
Prediction	30.13 ms
Text Annotation	0.73 ms
Voice Alert Output	938.21 ms
TOTAL TIME	1.0953 s

Table 14 shows the result of testing the TSDR system using a total of 360 frames from six different actual footage (F1-F6) from local roads. The footage contains 308 frames where traffic sign is fully visible with approaching distance and 52 frames where the traffic sign is not visible. The system had detection on 83 traffic images ($TP_{detection}$ + $FP_{detection}$), of which 30 were classified correctly to not contain traffic signs ($TN_{recognition}$), 20 were correctly classified ($TP_{recognition}$), and 33 were incorrectly recognized ($FP_{recognition}$). The system has a detection accuracy of 32.50% and a recognition accuracy of 60.24%. The overall accuracy of the system is 19.58%. The low accuracy is due to the poor image quality caused by several factors including (1) poor camera, (2) vehicle motion and vibration, and (3) environmental setting.

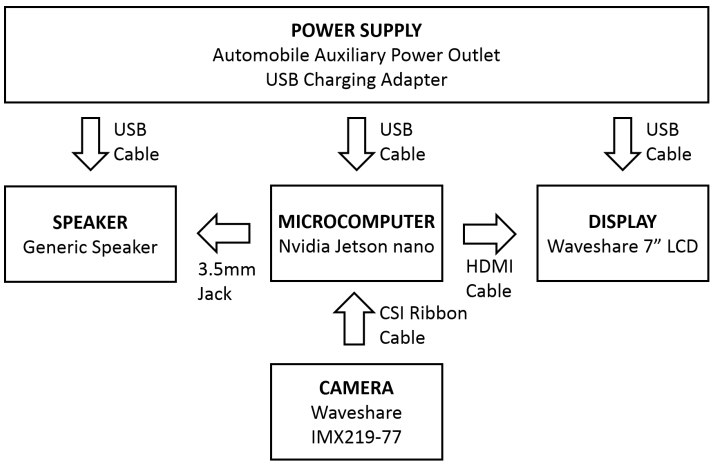


Figure 13: TSDR hardware configuration diagram

Detection and recognition of traffic sign in every frame is ideal for a ATSDR system but in actual application, one traffic sign correctly detected and recognized is enough to provide road information to the driver. Out of 6 sets of image sequences, 3 sets or 50% have at least one of the frames correctly detected and recognized.

Table 14: Test using frames from actual local road footage

Parameter	F1	F2	F3	F4	F5	F6	TOTAL
No. of Frames	87	73	64	32	48	56	360
$TP_{detection}$	12	29	2	0	0	23	66
$TN_{detection}$	12	7	8	8	8	8	51
$FP_{detection}$	3	8	0	0	0	6	17
$FN_{detection}$	60	29	54	24	40	19	226
$TP_{recognition}$	0	7	2	0	0	11	20
$TN_{recognition}$	9	15	0	0	0	6	30
$FP_{recognition}$	6	15	0	0	0	12	33
$FN_{recognition}$	0	0	0	0	0	0	0
Sign Detected?	No	Yes	Yes	No	No	Yes	50%



Figure 14: TSDR car setup for actual road testing



Figure 15: Sample frames from actual TSDR footage

4 Conclusion

Among the four evaluated pre-processing and detection methods, Shadow and Highlight Invariant method provided the best trade-off between segmentation success rate and processing time due to the less complex image filtering approach and use of HSV color space. Convolutional Neural Network classifier has the best performance in both accuracy and processing speed among the ten evaluated algorithms for traffic sign recognition due to the optimization method of the loss function using back propagation and the in-system feature extraction method of the algorithm. In classification, additional class to contain the not targeted traffic signs provides improvement of false detection performance.

An embedded system using Nvidia Jetson Nano platform with interface IMX219-77 camera, Waveshare 7" LCD and generic speaker using Python and OpenCV, sci-kit learn and Pytorch libraries has been developed. Implementation of adaptive frame rate improves sampling and provides an approach to handle variable frame processing time for a real-time system.

This study has not only provided comparative analysis on detection and recognition performance of several algorithms based on accuracy and processing speed but also able to implement a real-time embedded system. Further improvement to the system includes the use of better high definition camera with better image sensor and optical image stabilization. This can also be improved by the use of better detection and recognition algorithm like masked R-CNN. Also, further study can be done on the extension of application of the study to include other traffic signs and signals, and road markings.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to acknowledge Ateneo de Manila University, FAITH Colleges, and Fastech Synergy Philippines for supporting this research through their engineering scholarship program.

References

[1] World Health Organization, "Global status report on road safety", 2015, Retrieved from https://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

[2] M. R. De Leon, "Towards Safer Roads: Road safety initiatives of DOTr", ASEAN Automobile Safety Forum, May 2017.

[3] A. C. Mendoza, "Applications of Psychology to the understanding and modification of road user behavior", University of the Philippines National Center for Transportation Studies, August 2006.

[4] Y. MaryReeja, T. Latha, A. MaryAnsalinShalini, "Traffic sign detection and recognition in driver support system for safety precaution", ARPN J. Eng. Appl. Sci., **10**, 2015, 2209-2213.

[5] S. Feng, "Evaluation of colour segmentation algorithms in red colour of traffic sign detection", M.S. Thesis, Department of Computer Engineering, Hogskolan Dalarna, 2010.

[6] A. Santos, P. A. Abu, C. Oppus, R. Reyes, "Traffic sign detection and recognition for assistive driving", in 2019 International Symposium on Multimedia and Communication Technology (ISMAC), Quezon City, Philippines, 2019, pp. 1-6, doi: 10.1109/ISMAC.2019.8836161.

[7] Department of Public Works and Highways, "Highway safety design standards, road safety design manual", May 2012.

[8] C. Mythili, V. Kavitha, "Color image segmentation using ERKFCM", Int. J. Comput. Appl., **41**(20), 21-28, March 2012, doi: 10.5120/5809-8074.

[9] H. Fleyeh, M. Dougherty, "Road and traffic sign detection and recognition", Advances OR and AI Methods in Transportation.

[10] V. Deshmukh, G. Patnaik, and M. Patil, "Real-time traffic sign recognition system based on colour image segmentation", 2013, Int. J. Comput. Appl., **83**(3), 30-35, doi: 10.5120/14430-2575 10.5120/14430-2575.

[11] C. Lin, C. Su, H. Huang, K. Fan, "Colour image segmentation in various illumination circumstances" in Proceedings of the 9th WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 2010, 179184, doi: 10.13140/2.1.1045.3446.

[12] H. Rashmi, M. Shashidhar, G. Prashanth Kumar, "Automatic tracking of traffic signs based on HSV" International Journal of Engineering Research & Technology, **3**(5), May 2014.

[13] S. Sural, G. Qian, S. Pramanik, "Segmentation and histogram generation using the HSV color space for image retrieval" in Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 2002, pp. II-II, doi: 10.1109/ICIP.2002.1040019.

[14] H. Fleyeh, "Color detection and segmentation for road and traffic signs," IEEE Conference on Cybernetics and Intelligent Systems, 2004, Singapore, 2004, pp. 809-814, doi: 10.1109/ICCIS.2004.1460692.

[15] H. Fleyeh, "Traffic signs color detection and segmentation in poor light conditions", MVA2005 IAPR Conference on Machine Vision Applications, May 16-18, 2005 Tsukuba Science City, Japan.

[16] M. Ebner, "Color constancy using local color shifts", The 8th European Conf. on Computer Vision, Prague, Czech Republic, 2004, 276-287, doi: 10.1007/978-3-540-24672-5_22.

[17] H. Gomez-Moreno, S. Maldonado-Bascon, P. Gil-Jimenez, S. Lafuente-Arroyo, "Goal evaluation of segmentation algorithms for traffic sign recognition" in IEEE T. Intel. Transp., **11**(4), pp. 917-930, Dec. 2010, doi: 10.1109/TITS.2010.2054084.

[18] A. Shustanov, P. Yakimov, "CNN design for real-time traffic sign recognition", Procedia Eng., **201**, 718-725, 2017, doi: 10.1016/j.proeng.2017.09.594.

[19] S. Prasomphan, T. Tathong, P. Charoenprateepkit, "Traffic sign detection for panoramic images using convolution neural network technique", HPCCT 2019: Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference, 2019, 128-133, doi: 10.1145/3341069.3341090.

[20] Y. Wu, Y. Liu, J. Li, H. Liu, X. Hu, "Traffic sign detection based on convolutional neural networks", Proceedings of the International Joint Conference on Neural Networks, 2013, 1-7, doi: 10.1109/IJCNN.2013.6706811.

[21] Y. Sun, P. Ge, D. Liu, "Traffic sign detection and recognition based on convolutional neural network," 2019 Chinese Automation Congress (CAC), Hangzhou, China, 2019, pp. 2851-2854, doi: 10.1109/CAC48633.2019.8997240.

[22] D. Alghmgham, G. Latif, J. Alghazo, L. Alzubaidi, "Autonomous traffic sign (ATSR) detection and recognition using deep CNN", Procedia Comput. Sci., **163**, 2019, Pages 266-274, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.12.108.

[23] D. Tabernik, D. Skoaj, "Deep learning for large-scale traffic-sign detection and recognition" IEEE T. Intel. Transp., **21**(4), pp. 1427-1440, April 2020, doi: 10.1109/TITS.2019.2913588.

[24] C. Tomasi, R. Manduchi, "Bilateral filtering for gray and color images," Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), Bombay, India, 1998, pp. 839-846, doi: 10.1109/ICCV.1998.710815.

[25] N. Kehtarnavaz, A. Ahmad, "Traffic sign recognition in noisy outdoor scenes," Proceedings of the Intelligent Vehicles '95. Symposium, Detroit, MI, USA, 1995, pp. 460-465, doi: 10.1109/IVS.1995.528325.

[26] M. Rizon, H. Yazid, P. Saad, A. Shakaff, A. Saad, "Object detection using circular hough transform", American Journal of Applied Sciences, **2**, 2005, 10.3844/ajassp.2005.1606.1609.

[27] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), **1**, San Diego, CA, USA, 2005, pp. 886-893, doi: 10.1109/CVPR.2005.177.

[28] J. Lim, C. Oppus, "Identifying car logos using different classifiers in Python with HOG features".

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, "PyTorch: An imperative style, high-performance deep learning library", Adv. Neur. In., **32**, 2019.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: machine learning in python", J. Mach. Learn. Res. **12**, pp. 2825-2830, 2011.

[31] Nearest Neighbors Classification. Retrieved from https://scikit-learn.org/stable/modules/neighbors.html

[32] Introduction to Support Vector Machines. Retrieved from https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html

[33] Gaussian Processes. Retrieved from https://scikit-learn.org/stable/modules/gaussian_process.html

[34] Decision Trees. Retrieved from https://scikit-learn.org/stable/modules/tree.html

[35] How the Random Forest Algorithhm Works in Machine Learning. Retrieved from https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

[36] Multilayer Perceptron. Retrieved from http://deeplearning.net/tutorial/mlp.html

[37] An AdaBoost Classifier. Retrieved from https://scikit-learn.org/stable/ modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[38] Gaussian Nave Bayes Classifier Implementation in Python. Retrieved from https://dataaspirant.com/2017/02/20/gaussian-naive-bayes-classifier-implementation-python/

[39] Linear and Quadratic Discriminant Analysis. Retrieved from https://scikit-learn.org/stable/modules/lda_qda.html

[40] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," in P. IEEE, **86**(11), pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 883-893 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Evaluating the Effectiveness of Query-Document Clustering Using the QDSM Measure

Claudio Gutiérrez-Soto[*,1],Marco Palomino[2], Arturo Curiel[3], Héctor Riquelme Cerda[1], Fernando Bejar Rain[1]

[1]*Universidad del Bío-Bío, Department of Information Systems, Casilla, 5-C, Chile*

[2]*University of Plymouth, School of Engineering, Computing and Mathematics, PL4 8AA, England*

[3]*CONACYT-Universidad Veracruzana, Facultad de Estadística e Informática, 91020, México*

ARTICLE INFO

Article history:
Received: 07 July, 2020
Accepted: 19 November, 2020
Online: 08 December, 2020

Keywords:
Average link
Query logs
Query-sensitive similarity measure (QSSM)
Ward's method
Web search engines

ABSTRACT

It is well documented that the average length of the queries submitted to Web search engines is rather short, which negatively impacts the engines' performance, as measured by the precision metric. It is also well known that ambiguous keywords in a query make it hard to identify what exactly search engine users are looking for. One way to tackle this challenge is to consider the context in which the query is submitted, making use of query-sensitive similarity measures (QSSM). In this paper, a particular QSSM known as the query-document similarity measure (QDSM) is evaluated, QDSM is designed to determine the similarity between two queries based on their terms and their ranked lists of relevant documents. To this extent, F-measure and the nearest neighbor (NN) have been employed to assess this approach over a collection of AOL query logs. Final results reveal that both the Average Link Algorithm and Ward's method present better results using QDSM than cosine similarity.

1 Introduction

This paper is an extension of work originally presented in the International Conference of the Chilean Computer Science Society (SCCC) [1]. Nowadays, exist a large amount of information available on the Web; additionally, web search engines (WSE) daily index thousands of pages, by which finding relevant and timely information over this growth without constraints becomes quite a challenge. According to [2], this unrestricted growth has not been accompanied by corresponding technical advances in approaches to extract relevant information. Unsuccessful searches are common in WSEs and can be given by several reasons among which we can mention the following. First, the lengths of submitted queries by users are mostly short (e.g., the average size of a web search is 2.4 words [3]). Owing to queries are conformed by a few keywords, it is complicated to determine the specific topic in which the query is inserted, therefore, a result with a few keywords in the query can be formed by unrelated topics. Furthermore, users ignore how to formulate a correct query [4]. The problem is made more complex when the user does not have a specific idea about which should be the result that he/she is looking for (i.e., which turns in relevant information for him/her). In light of the foregoing, it is not easy for WSEs to interpret the meaning of what users are looking for. One way to tackle this issue, is to consider the context in which the queries are submitted at WSEs [5], [6]. To capture the context, WSEs should consider what queries are related among them (i.e., determining if the queries are similar) and how their results have been beneficial for users. A way to determine the relationship between similar queries and relevant documents for these queries is given by the cluster hypothesis, which establishes that all documents considered as relevant for a query are similar to each other (i.e., similar documents can be relevant for the same query)[7]. Accordingly, it can be assumed that relevant documents for a query q, are relevant for a query q', such as q' is similar to q. Thus, having methods that provide the similarity among documents and queries can bring a better characterization about the meaning of a new query, and as a consequence, it entails more effective results.

Whether the WSE is able to establish how similar is a new query regarding queries recently submitted, then the search engine

[*]Corresponding Author: Claudio Gutiérrez-Soto, Casilla 5-C Email: cogutier@ubiobio.cl

should provide documents, which were relevant in previous searches. Hence, the recent past queries along with the relevant documents provide a context, in which is feasible to improve the answers to new searches [8]. Nevertheless, measuring how similar are two queries using their documents is not easy, suitable metrics that allow representing the context are needed.

Aiming to capture the context in which queries are submitted, Tombros and Van Rijsbergen [9] present a pioneering approach introducing the measuring called query-sensitive similarity measure (QSSM). This measure establishes the following; two documents are more similar than others, whether both are more similar regarding a given query. Following this argumentation line, QSSMs can be used as a metric to measure the similarity between two queries considering the context. As such, a WSE using additional information can improve its effectiveness to answer a new query. To achieve this goal, queries alongside their documents should be stored in clusters. Currently, few approaches store the queries along with their documents [10]–[14] (these approaches recommend to the user a list of similar queries, which are related to the submitted query by the user). However, the approaches mentioned previously are not directly related to the approach presented in this paper.

1.1 Contribution

The main contribution of this paper is the effectiveness evaluation of QDSM. Roughly speaking, effectiveness is related to the quality of recovered documents. Better effectiveness occurs when more relevant documents are retrieved (from a total of N documents retrieved, there are more relevant documents than non-relevant documents). By contrast, worse effectiveness occurs when more non-relevant documents are retrieved. With this in mind, grouping similar queries alongside their relevant documents (in clusters) should directly affect the effectiveness. Therefore, improving clusters' effectiveness implies having more relevant documents by clusters, which aligns with the cluster hypothesis.

To evaluate QDSM effectiveness, the F-measure alongside the nearest-neighbor (NN) cluster hypothesis tests were used. Both tests were applied over the following algorithms; Single Link, Complete Link, Average Link, Bisection K-means, and Ward's method. Three relevance models were simulated with the aim to determine which documents are relevant for a specific query. Besides, a wide range of experiments was carried out over five sets of queries. Final results are presented as QDSM improvement percents regarding cosine measure (S_c) (or cosine similarity), which were contrasted with the values obtained applying the Student Paired T-test (two samples).

The remainder of this paper is organized as follows: In Section 2, a detailed review of articles related to the problem of capturing the context using clustering is presented. In Section 3, the methodological description, is exposed. Section 4 presents the experimental environment. Section 5 displays the empirical results, which are then discussed in Section 6. Finally, Section 7 gives some perspectives of future work along with the conclusions.

2 Related Work

There have been many works that deal with the use of clustering in Information Retrieval (IR). Clustering in IR has been employed to improve the effectiveness (i.e., quality of clusters). Overall, clustering-based approaches that intend to capture the context in which the queries are submitted can be classified into two categories, considering the underlying repositories (these are also known as collections or datasets). The first category involves using traditional IR datasets. Some of them use QSSM similarities such as [15]–[17]; meantime, the second category uses log-file data from search engines.

To improve effectiveness in the retrieval process, an approach relies on hierarchic query-specific clustering is presented in [15]. In pursuing this goal, a wide range of experiments was performed. According to the authors, given a specific query, the hierarchy should be adapted to increase the likelihood to situate relevant documents to the query in nearby clusters. Two characteristics stand out in this research. First, an analysis of optimal clusters variation considering the number of top-ranked documents allowing better effectiveness is exposed. Finally, a comparison between their results and inverted file search (IFS) is provided. Five traditional IR collections alongside four hierarchic agglomerative methods were employed in all experiments. Final results indicate that query-specific clustering outweighs static clustering in each of the experiments. On the other side, a framework based on probabilistic co-relevance, which gives a query-sensitive similarity, is presented in [17]. The similarity between two documents corresponds to the ratio between the co-relevance probability and a specific query. Two cases were considered to identify the co-relevance. First, the document's relevance is independent of the rest of the documents. Second, the document's relevance is dependent on the rest. Several experimental scenarios were studied using the nearest neighbor test on TREC collections. The final results reveal that the framework outperforms term-based similarity.

The approaches mentioned above expand the users' judgments grounded on the following assumption. All terms included in a relevant document for a specific query are relevant too. Consequently, it is assumed that all documents that include some of these terms are also relevant. Besides, these approaches do not deal with the similarity among queries, with the exception of [16],[18]. In [18], a method called Scatter/Gather, which explores clusters based on documents, is proposed. The method returns a ranked title's list for the organization and viewing of retrieval results. Scatter/Gather is used as a tool for retrieval of browsing results, which presents summaries to users. Towards that goal, documents are joined in similar topics. A fractional algorithm provides k clusters on TREC/Tipster dataset. As a result of experimentation, the authors assert that their method gives tailored clusters according to the query's characteristics. In such a way, their results corroborate the cluster hypothesis since relevant documents are more similar to each other than non-relevant documents. In [16], the authors introduce the Weighted Borda (WBorda) model, which determines the co-relevance of a document using different similarities' types. To this end, a Support Vector Machine (SVM) was trained to get the estimated co-relevance, fusing the induced rankings using several functions. Each function considers the similarity between documents and the query. Several similarity measures were considered in experiments such as cosine BM25, M1, and M3. The final results in tasks such as nearest-neighbor clustering, cluster-based, and graph-based document retrieval indicate that WBorda provides better results than several proposed

co-relevance models.

On the other hand, approaches such as [10],[19], and [20] belong to the second category. Users' log-based, an approach of query clustering is proposed in [19]. Towards that end, the documents previously read by users are employed to construct cross-references among documents and queries. According to the researchers exist a strong relationship between the selected documents and queries. This approach underlies two fundamental aspects: First, two queries are similar if users clicked on the same documents; Second, if a set of documents was selected for the same queries, then the documents' terms are related to the queries' terms. The empirical results were obtained using the DBSCAN algorithm and the Encarta encyclopedia dataset. The final results show that many similar queries are gathered in the same clusters utilizing this approach. A query-clustering classification, which compares various query similarity measures, is presented in [10]. Three groups: content-based approaches, feedback-based approaches, and results-based approaches are suggested in this classification. In content-based approaches, the common terms of queries are used to describe query clusters. Similarity functions such as Jaccard, Cosine, and Dice were employed to build the clusters. In that regard, the authors claim that this method is not convenient for search engines due to many queries have few terms. On the other side, in feedback-based approaches, the similarity measure is grounded users' selections over search results; therefore, two queries are similar whether they encourage the selection of similar documents. In turn, results-based approaches evaluate the similarity between queries through the overlap of returned documents. In this case, the researchers point out that this approach's principal drawback corresponds to high processing times. Notable results are obtained using the three approaches in parallel. In [20], a WSE provides a user with a list of similar queries regarding the user's submitted query. Semantically similar queries give support to the clustering process. Clusters are formed, taking into account the historical preferences of registered users in the WSE. To build the clusters, term-weight vector representation of queries considering the clicked URLs was employed. The method exhibits two benefits, (1) it discovers the related queries, and (2) sorts the queries rely on a relevance criterion. It is important to mention that the queries are sorted using the following criteria: (a) the similarity between the clusters' queries and the new query and (b) the support, which is related to how much the query answers capture the user interest. The experiments were conducted using the combination of (a) and (b). The results display improvements on average precision.

In summary, the first category is based on traditional IR datasets. A traditional IR dataset is formed by three sets, a set of documents (D), a group of queries (Q), and a set of users' judgments (JU). The user's judgments contain the relevant documents for a query in Q. Note that all works mentioned in this category include new relevant documents (if some document has some relevant term, then it is considered relevant), which are not part of the original JU. On the other hand, it should be noted that there are no JU in the second category (log-files from search engines). Consequently, subject matter experts evaluate the pertinence of a document given a query. Note that all approaches mentioned in this related work modified some documents' relevance, which directly impacts effectiveness. Contrary to these approaches, in this paper, three types of users' judgments are simulated without altering the documents' original-relevance.

The overall procedure and a discussion about the results are presented in the following sections.

3 Methodology

The methodology overview is as follows. Initially, a user submits a query to the WSE. Thereafter, the WSE returns the documents as a result of the query. These documents are ranked from the most similar to the least similar regarding the query. Once this is done, the documents are stored along with the query in clusters inside the WSE. Aiming to form the clusters considering documents and queries, QDSM is used. In this way, when a user submits a new query, it is contrasted with the past queries (these are the queries previously stored) in the query-document clusters. Accordingly, an effectiveness improvement should occur due to the clusters closest to the new query containing relevant documents for the new query.

Document relevances become a crucial factor in enhancing effectiveness. In a traditional IR dataset, document relevances are given by subject matter experts, who determine what documents are relevant given a query. These documents are reflected in the users' judgments. On the other hand, the relevance of documents in a WSE is given by ranking functions. Overall, ranking functions attempt to capture the relevance through users' clicks on documents, which are expressed in the ItemRanks. In this manner, a retrieved document (i.e., URL or web page) with an ItemRank high could be considered as relevant.

It is essential to keep in mind that most approach clustering-based extend or use subject matter experts to give relevance to the documents, due to none of these datasets have been designed to work with similar queries (past queries are part of clusters). According to [21], a good way to tackle this problem is by using simulation. In this paper, document relevances have been simulated. Two notable advantages are obtained with the simulation use. First, it is neither necessary to use subject matter experts nor extend the user's judgments. Second, several models of relevance (a model can be seen as a ranking function) can be used; for instance, given a query, a document can be relevant or non-relevant depending on the model. In this paper, this is given by a relevance function, which determines the relevance of a document considering both its corresponding ItemRank and relevance probability.

Aiming to shed light on how QDSM is evaluated using a relevance function, suppose the following example. Assume that three documents (d_5, d_{10}, and d_{12}) have been recovered for a query q, such as d_5 is the most similar document concerning the query. The respective ItemRank for each document is 20, 25, and 30. Likewise, the probabilities of relevance according to their respective ItemRanks are 90%, 40% 70%. Additionally, suppose a relevance function that only considers the last recovered document (in this case, d_{12}). As d_{12} has an ItemRank of 30, the probability of being relevant is 70%. To simulate the relevance probability of d_{12}, a binary array of 100 elements is used. Initially, this array is instantiated with 0 values; subsequently, 70 random positions with value 1 are assigned in the array using Uniform Distribution. In order to give the relevance to d_{12}, an array position is selected using Uniform Distribution; thus, if this value is 1, then d_{12} is relevant; in another case, d_{12} is

non-relevant (Note that d_5 and d_{10} are non-relevant).

Following the same example, suppose a relevance function that assigns the relevance individually, then the same procedure to give relevance is performed for each document (d_5, d_{10}, and d_{12}). Thus, a possible result could be that d_5 and d_{12} being relevant, meantime d_{10} could be non-relevant. Finally, suppose a relevance function that provides the average relevance, then the average of ItemRanks is obtained, and its relevance probability is used to give relevance to the three documents.

Note that different relevance functions could provide different results on QDSM, due to QDSM considers the relevant documents as part of its metric.

3.1 The Query-Document Similarity Measure

The Query-Document Similarity Measure (QDSM) is a Query-Sensitive Similarity Measure (QSSM), which has as a fundamental purpose to capture the semantic similarity between queries, taking into consideration terms that belong to the queries as well as the position in which appear the relevant documents in both lists. Indirectly, the terms associated with the relevant documents should contribute to providing context. Specifically, each list of documents is presented in descending order according to the similarity of documents regarding the query. From the semantic point of view, two queries are closer if they share more relevant documents in their lists. This can be appreciated by observing the number of relevant documents that form the intersection between the two lists. Therefore, while more relevant documents make up the intersection, the more similar the queries will be. Thus, this paper's primary assumption is that using similar queries alongside their relevant documents should provide clusters with better effectiveness than S_c, since additional information can be captured from documents, including the queries (i.e., information is not complete in each query individually). Specifically, this additional information is given by the union of queries and documents' terms but does not belong to the intersection among them. Using this rationale, QDSM is in line with the cluster hypothesis, which claims that relevant documents for a particular query tend to be close, whereby these relevant documents should tend to be in the same cluster for a specific query.

QDSM takes advantage from the place in which relevant documents appear on the list. As reported by [22], the most similar documents concerning the query tend to appear at the beginning of the list. On this basis, the order in which relevant documents appear in both lists gives information about the context (particularly the terms of relevant documents). QDSM deals with the order of relevant documents through the use of the Longest Common Subsequence (LCS) algorithm. LCS allows acquiring the relative similarity keeping the order in which simultaneously appears a relevant document in both queries. By doing so, the context capturing in which the queries are submitted is possible.

This measure is convenient in two situations:

- When terms of a query are few, as is currently happening in the WSEs.
- In a dynamic environment, where the documents' relevance could change (i.e., the position of a document in the list could change as well as its ItemRank), non-relevant documents could become relevant documents.

Accordingly, the queries are either short length (i.e., few keywords in the query) or ambiguous. Nevertheless, these can be enriched with more information associated with their relevant-documents retrieved.

Aiming to give formality, some definitions are detailed below.

Definition 1

Let D be a set of documents, such as every document in D, is formed by a set of terms (i.e., words contained in d). D is stored in a WSE W. Besides, let q be a single query, such as $q \in Q$, where Q is a set of queries interpretable by W.

Definition 2

The cosine measure between q and d_i, is defined as:

$$S_c(q, d_i) = \frac{\sum_{j=1}^{m} t_{qj} \cdot t_{ij}}{\sqrt{\sum_{j=1}^{m} t_{qj}^2 \cdot \sum_{j=1}^{m} t_{ij}^2}}$$

shuch as $d_i \in D \wedge q \in Q$, t_q are the query'terms and t_i are the document's terms.

Definition 3

Let $|d_i|$ be the number of terms in a document d_i, such as $d_i \in D$. Note that this definition can be applied to obtain the number of terms for a query q.

Lemma 1

$S_c(q, d_i) \neq 0 \iff (q \cap d_i) \neq \emptyset$

Proof by contradiction: Suppose $(q \cap d_i) = \emptyset$. Then, there are no terms in common between q and d_i. Hence, applying the *Definition 2* implies $S_c(q, d_i) = 0$. Likewise, if $S_c(q, d_i) = 0$, then $(q \cap d_i) = \emptyset$. □

Definition 4

Let $\mathcal{L}(q) = \{d_i | S_c(q, d_i) \neq 0 \ \forall d_i \in D\}$ be the set of documents whose similarity with $q \in Q$ is different to 0.

Definition 5

Let $\mathcal{L}_N(q)$ be a list of N retrieved documents from W, such as $\mathcal{L}_N(q)$ is ranked by decreasing order (i.e., they are ordered from highest to lowest according to S_c).

Definition 6

Let $F(ItemRank(d_i), M)$ be a binary function, which provides relevance to a document d_i, such as 1 is relevant and 0 is non-relevant. The function is defined as follows:

$$F(ItemRank(d_i), M) = \begin{cases} 1: & Pr(1) = \mathcal{M}(M, ItemRank(d_i)) \\ 0: & Pr(0) = 1 - Pr(1) \end{cases}$$

where M corresponds to the relevance model (i.e., PartialRel, AvRel or LastRel). *ItemRank* is a function that provides the rank for the document d_i, and $\mathcal{M}$ is the function that gives the probability considering M and *ItemRank*. A binary array formed by 100 elements is used to represent the probability in $\mathcal{M}$. The Uniform Distribution is employed to instantiate the values (the percentage is represented with values 1 in the array) and determine the relevance (if the array's selected position contains a 1, then the document is relevant).

Definition 7

Let $\mathcal{L}_{N,R}(q)$ be a list of retrieved documents along with their relevances, then:

$\mathcal{L}_{N,R}(q) = \{(d_i, F(ItemRank(d_i), M))|d_i \in \mathcal{L}_N(q)\}$.
Definition 8

Given two queries q and q' such as both queries are in Q, and their corresponding lists of documents $\mathcal{L}_{N,R}(q)$ and $\mathcal{L}_{N',R}(q')$. Then QDSM is defined as follow:

$QDSM(q,q') = \frac{S_c(q,q') + LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q'))}{1 + max(\|\mathcal{L}_{N,R}(q)\|, \|\mathcal{L}_{N',R}(q')\|)}$

where:

- $S_c(q,q')$ corresponds to the cosine measure between the queries q and q'.
- $LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q'))$ is the LCS algorithm applied over $\mathcal{L}_{N,R}(q)$ and $\mathcal{L}_{N',R}(q')$ [23].
- *max* gives the greatest number of relevant documents between the lists $\mathcal{L}_{N,R}(q)$ and $\mathcal{L}_{N',R}(q')$.

Theorem 1. $QDSM(q,q') = 0 \iff S_c(q,q') = 0 \wedge LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) = 0$.
Proof by contradiction.

- Suppose $S_c(q,q') \neq 0 \vee LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0$. If $S_c(q,q') \neq 0$ is enough for $QDSM(q,q') \neq 0$ (by *Definition 8*). Likewise, as $LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0$ then $QDSM(q,q') \neq 0$ (by *Definition 8*).
- Finally, as $QDSM(q,q') \neq 0$ then either $S_c(q,q') \neq 0 \vee LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0$ (by *Definition 8*). □

Lemma 2
$LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0 \iff \exists\, t \in (d \cap q \cap q')$
Proof.

- $LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0$, then $\exists\, d \in (\mathcal{L}_{N,R}(q) \cap \mathcal{L}_{N',R}(q'))$ (by the LCS invariant [23]). Thus, $d \in \mathcal{L}_{N,R}(q) \wedge d \in \mathcal{L}_{N',R}(q')$. Since $d \in \mathcal{L}(q) \wedge d \in \mathcal{L}(q')$, then $S_c(d,q) \neq 0 \wedge S_c(d,q') \neq 0$ (by *Definition 4*). Thus, $(d \cap q) \neq \emptyset \wedge (d \cap q') \neq \emptyset$ (by *Lemma 1.*). Hence, $(d \cap q \cap q') \neq \emptyset$. Therefore, $\exists\, t \in (d \cap q \cap q')$.
- Finally, as $\exists\, t \in (d \cap q \cap q')$ then $S_c(q,q') \neq \emptyset \wedge S_c(d,q) \neq \emptyset \wedge S_c(d,q') \neq \emptyset$ (by *Lemma 1.*). Thus $d \in \mathcal{L}_{N,R}(q) \wedge d \in \mathcal{L}_{N',R}(q')$ (by *Definition 7.*). Therefore, $LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0$ (by *Theorem 1*). □

Lemma 2 asserts that exists at least a common document in both lists, and therefore at least there is one term in common among queries and the document.

Lemma 3
$|d| \neq |q| \implies \exists t \in (d \triangle q)$
Proof. Suppose $|d| > |q|$, which implies that d has at least one term more than q. On the other hand, can ocurr that $|q| > |d|$, then q has at least one term more than d. Thus, $\exists t \in ((d-q) \cup (q-d)$. Therefore, $\exists t \in (d \triangle q)$. □

Lemma 3 points out that if different numbers of terms form the document and the query, then at least there is one term that does not belong to the intersection between them. Note that d and q are in $\mathcal{L}_{N,R}(q)$.
Theorem 2
$QDSM(q,q') \neq 0 \wedge |d| \neq |q| \neq |q'| \iff \exists\, t \in ((d \triangle q) \triangle q')$; such as d is in both $\mathcal{L}_{N,R}(q) \wedge \mathcal{L}_{N',R}(q')$.
Proof:

- $QDSM(q,q') \neq 0 \wedge |d| \neq |q| \neq |q'|$, then $LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0 \wedge S_c(q,q') \neq \emptyset \wedge |d| \neq |q| \neq |q'|$ (by *Definition 8*). As the first part, consider $LCS(\mathcal{L}_{N,R}(q), \mathcal{L}_{N',R}(q')) \neq 0 \wedge |d| \neq |q| \neq |q'|$ then, $\exists\, t' \in (d \cap q \cap q') \wedge |d| \neq |q| \neq |q'|$ (by *Lemma 2*). Thus $\exists\, t' \in (d \cap q \cap q') \wedge ((|d| \neq |q|) \neq |q'|)$. Subsequenlty, $\exists\, t' \in (d \cap q \cap q') \wedge \exists t \in ((d \triangle q) \neq q')$. (by *Lemma 3.*). Hence, $\exists\, t \in ((d \triangle q) \triangle q')$. The another case is $S_c(q,q') \neq \emptyset$, then $\exists\, t \in (q \triangle q')$, considering d in the hypothesis, $(q \triangle q') \subset ((d \triangle q) \triangle q')$. Therefore $\exists\, t \in ((d \triangle q) \triangle q')$.
- Finally and without loss of generality, if $\exists\, t \in ((d \triangle q) \triangle q')$ then $\exists\, t \in ((d-q) \cup (q-d)) \vee ((d-q') \cup (q'-d)) \vee ((q-q') \cup (q'-q)) \wedge ((S_c(q,q') \neq \emptyset \wedge \exists\, t \in (d \cap q \cap q'))$ (by *Lemma 1. and Lemma 2.)* then $QDSM(q,q') \neq 0 \wedge |d| \neq |q| \neq |q'|$. □

Definition 9

Let RQ be a set of queries, along with their retrieved documents and their corresponding relevances, then RQ is defined as follow:

$$RQ = \bigcup_{i=1}^{|Q|} \mathcal{L}_{N,R}(q_i)$$

Definition 10

Let $\widehat{D}$ be a benchmark query set, which is formed by Z subsets of queries, such as:

$$\widehat{D} = \bigcup_{i=1}^{Z} RQ$$

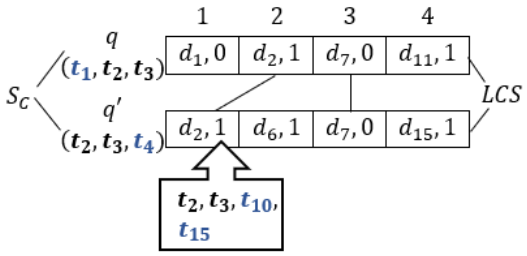


Figure 1: QDSM measure

Theorem 2 ensures that the context in which the queries are submitted in the WSE can be captured by the complementary terms to both queries and their relevant documents (i.e., the symmetric difference of sets d, q and q' ($(d \triangle q) \triangle q'$)). An example of the Theorem 2 essence and how QDSM is computed, is displayed in Figure 1. In Figure 1, the common terms for both queries are presented in bold (i.e., these are t_2 and t_3), note that both terms are common in d_2, which is a relevant document (i.e., all document d_i, 1 are relevant meanwhile d_i, 0 are non-relevant). LCS is applied over both lists of retrieved documents considering only the relevant documents in the lists (i.e., even though d_7 is in both lists, only d_2 is considered). Finally, all terms that give context are in blue color (i.e., t_1, t_4, t_{10} and t_{15}).

It is worth noting that QDSM takes the value 1 (see *Definition 8*) when q and q' are the same queries, and all retrieved documents are relevant. Specifically, this latter can be itemized in two parts. In the first part, $S_c(q,q')$ provides 1 because q and q' are the same. In the second part, both results lists are equal, and all retrieved documents

are relevant; therefore, both lists hold the relevant documents in the same positions.

In summary, QDSM provides a metric that captures the semantic relationship between two queries (context), considering the relevant documents' order in both lists. A wide range of experiments was conducted in order to compare the effectiveness between S_c and QDSM. The experimental setup is displayed in the following section.

4 Experimental Environment

A benchmark query set extracted from the well-known dataset of query log "AOL Query Logs Dataset (AOL) [24]", was used to carry out the experiments. This collection has more than 20 million web query logs stored, submitted by around 650 thousand users in more than 36 thousand lines of data. These queries were stored at an interval of three months in the year 2006. Broadly speaking, queries in AOL are depicted as rows in the database files, which contains five columns with the following fields:
{*AnonID*, *Query*, *QueryTime*, *ItemRank*, *ClickURL*}, where:

- AnonID: an anonymous user ID number.
- Query: the query submitted by the user in the WSE.
- QueryTime: the exact time at which the user submitted the query.
- ItemRank: if the user clicked on a result, it keeps the rank of the selected document; holds empty otherwise.
- ClickURL: The domain portion of the URL is showed as a result if the user clicked on a search result.

4.1 The Benchmark set of Related Queries

Aiming carrying out the clustering experiments, a benchmark set of related queries (*RQ*) (see *Definition 8.*) was processed randomly from AOL. To verify that the queries were partially related, each time a query was chosen, it was checked that at least existed another query, in such a way S_c was neither one nor zero. To achieve this goal, the queries with ClickURLs empty were removed due to these do not have answers associated with the queries. Furthermore, stopwords processing was previously performed before to apply S_c. The core insight is that ClickURLs allow depicting a list of retrieved documents for q (see Definition 5.). It should be noted that register with the same query q (i.e., the same terms), logged by the same user around the same time, corresponds to a single query, which was split into several registers. Providing the maximum amount of information implies to use the longest session, which at least contains one register (i.e., at least one result or document).

On the other side, it is important to highlight that AOL does not possess users' judgments. Note that the users' judgments play a fundamental role in order to know what documents are relevant for a specific query [25]. Furthermore, these relevant documents are necessary to evaluate precision, recall, and, therefore, effectiveness. To tackle this issue, users' judgments were simulated following the approach presented by [26]. Simulations of relevance judgments are presented in the following section.

4.2 Simulation of Relevance Judgments

Simulating document relevance regarding a query is not a trivial task. This task embraces a great variety of aspects, such as users' literacy [27], needed information at any one point of time, and the user's profile [28] among others. To address this problem, the approach proposed by [26], which provides the relevance probabilities for documents depending on their ItemRanks on AOL, is applied in this paper. Towards that end, $F(ItemRank(d_i), M)$ (see *Definition 6.*) is simulated using $\mathcal{M}(M, ItemRank(d_i))$ in *Definition 7.* In simple words, the relevance is assigned using a value 0 (non-relevant) or 1 (relevant), which is obtained considering the values presented in Table I (i.e, M in $F(ItemRank(d_i), M)$). The relevance probabilities were calculated using the ItemRanks, assuming the user clicks provide information about how the user interprets the query [30]. In Table I, two relevance models are presented by "AllRel" and "LastRel" columns. "AllRel" implies all clicked documents are considered as relevant; meantime "LastRel" reports that only the last clicked document is relevant. Regarding "AllRel", two variants were used for it. The first variant is named "PartialRel", which considers the individual ItemRank of each document obtained from Table 1. The second variant ("AvRel") corresponds to the average of ItemRanks of the query's recovered documents. For example, suppose three documents (d_5, d_{10}, and d_{12}) that have been recovered for a query q. The respective ItemRanks are 20, 30, and 40; therefore, the average is 30. Subsequently, the relevance probability is determined by the "ItemRank" (average) row and the "AllRel" column, so for this example, the relevance probability for each document is 0.5106. Although the three documents have the same relevance probability, the relevance for each document is individually obtained.

4.3 Clustering Experiments

Five algorithms were evaluated considering the three relevance models over the same $\widehat{D}$. Five sets of RQ were used; the smallest set of RQ contains 123 queries alongside their documents (i.e., for each document, the relevance has been assigned), meantime the biggest set comprises 2,141 queries. Aiming to compare the clusters' quality between S_C and QDSM, two well-known measures have been used, F-measure and the nearest neighbor (NN) cluster hypothesis test. F-measure was proposed by [29]; the idea behind this measure is to evaluate effectiveness in the post-processing step, in which each cluster is assigned to a class. The F-measure can be seen as a way of combining the precision and recall for a retrieval specific model, and it is defined as the harmonic mean of the model's precision and recall. In simple words, F-measure has as purpose to provide a binary classification as positive or negative according to the belonging of objects to determined classes in the clusters. F-measure allows giving more importance to precision, recall, or both. On the other hand, the nearest neighbor (NN) cluster test (which is also well-known as the (NN) test) was proposed by Voorhees ([30], [31]). In simple terms, the (NN) test reviews each of the retrieved documents for a specific query, identifying how many of its n close neighbors are relevant. The (NN) test is also used as a non-parametric classification and regression technique.

Table 1: Probability of a document being relevant in the AOL dataset if it has rank k, for two different click-based relevance interpretations, as calculated by [26]

ItemRank	AllRel	LastRel
≤ 20	0.4365	0.5702
≤ 120	0.5106	0.6278
≤ 300	0.5395	0.6493
> 300	0.4705	0.3507

Turning towards the cluster hypothesis, QDSM should provide better effectiveness than S_c if it is possible to find more relevant documents per cluster. Each experiment was executed ten times, and results are displayed as percentages of increasing or decreasing of QDSM regarding S_c. Specifically, F-measure was used giving the same weight for precision and recall; meantime, The (NN) test was instantiated with value three in all experiments. In addition, the Student's Paired t-Test (Two Samples test) was used to support the results. Five algorithms Single Link, Complete Link, Average Link, Bisection K-means, and Ward's Method, were used in each experiment.

All experiments were carried out on a server with: Intel Xeon Processor E3-1220 3.00 GHz; 16 GB Ram memory of 2133 MHz; 1 TB 7200 RPM Hard Drive; and Linux Operating System Debian Jessi 8.4.

5 Experimental Results

In this section, the quality of clusters (effectiveness) produced by QDSM and S_c is compared. To achieve this goal, the F-measure and the (NN) test were used considering the Single Link, Complete Link, Average Link, Bisection K-means, and Ward algorithms. Effectiveness was obtained using the relevance models PartialRel, AvRel, and LastRel. Note that the number of documents considered in the (NN) test corresponds to 3, it means that the relevances of the three closest documents with respect to the query were evaluated. To do that, the similarities between documents are checked alongside their relevance regarding the query. Overall, all results presented in each Table corresponds to QDSM, which are expressed in terms of percentages regarding the S_c. In Table 2, the QDSM effectiveness over the three relevance models was evaluated using the Single Link Algorithm. Note that the three relevance models were tested considering five sets of queries (# of q). From this Table, it is possible to appreciate that there is not an improvement of QDSM concerning S_c. Furthermore, Single Link presents better effectiveness for S_c than QDSM; this is consistent with the p-value obtained using The Student Paired T-test (two samples), which was 0.0029 for this Table. Generally speaking, Single Link exhibits the best results considering the "AvRel" relevance model, following by "LastRel" and finally "PartialRel" model. Continuing the same trend, the results for F-measure are exposed in Table 3. Here S_c shows again better results than QDSM, which is in line with the p-value: 0.00036, and notably the best results are presented by "AvRel" relevance model.

Regarding the Complete Link algorithm, similar results to the Single Link algorithm are presented in Tables 4 and 5, where S_c has better results than QDSM. In Table 4, the best results are provided by "LastRel" relevance model using the (NN) test, meantime that the best results using the F - measure are given by "PartialRel" model. The p-value for the (NN) test was 0.720, whilst the p-value for the F - measure was 0.00003.

On the other hand, the Average Link algorithm displays different results to Single Link and Complete Link algorithms, where QDSM is better than S_c. In Table 6, the best results are given by "PartialRel", followed by "LastRel" and "AvRel" respectively. Results showed in Table 6 are in line with the p-value: 0.00041. Following the same trend, in Table 7, QDSM presents better results than S_c for F-measure, which is in accordance with the p-value: 0.000002. It should be noted that there is no substantial difference between "AvRel" and "LastRel".

In turn, the results for the Bisection K-means algorithm are exposed in Tables 8 and 9. Both Tables provide conflicting results since Table 8 two relevance models ("PartialRel and AvRel") give good results for QDSM, meantime these relevance models provide opposing results in Table 9. These latter results are coherent with their respective p-values. The p-value for Table 8 corresponds to 0.124, whilst p-value for Table 9 is 0.904. These values are not significative due to they are not greater than 0.05.

Finally, in Tables 10 and 11, the results for the Ward's method are displayed. Both Tables provide excellent results for QDSM in contrast to S_c (except when the number of queries is 129, using the relevance model "PartialRel"). Furthermore, the p-values associated with both Tables present the more significative value among all algorithms. The p-value associated with Table 10 is 0.00007; meanwhile, the p-value for Table 11 corresponds to 0.00001.

To sum up, considering both measures, the three relevance models, and the p-values obtained, the best results are provided by Average Link algorithm and the Ward's method.

Table 2: The (NN) test over Single Link Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Single Link*		
	PartialRel	*AvRel*	*LastRel*
123	-30.92%	-33.73%	-42.36%
666	-11.76%	-106.36%	-69.62%
1,145	4.29%	-13.82%	4.36%
1,843	-14.70%	-22.68%	-32.44%
2,141	-40.20%	-4.74%	-11.20%

Table 3: F-measure test over Single Link Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Single Link*		
	PartialRel	*AvRel*	*LastRel*
123	-25.15%	-12.23%	-11.89%
666	-23.96%	-44.88%	-0.35%
1,145	-22.54%	-25.10%	-5.63%
1,843	-28.60%	-43.90%	-8.04%
2,141	-20.16%	-17.20%	-6.67%

Table 4: The (NN) test over Complete Link Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Complete Link*		
	PartialRel	*AvRel*	*LastRel*
123	-20.97%	26.76%	-41.72%
666	-14.90%	7.47%	-8.79%
1,145	-33.38%	28.44%	-29.72%
1,843	-30.80%	45.38%	-45.34%
2,141	-27.41%	41.98%	-55.13%

Table 5: F-measure test over Complete Link Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Complete Link*		
	PartialRel	*AvRel*	*LastRel*
123	-39.50%	-11.94%	-7.34%
666	-19.09%	7.48%	-33.90%
1,145	-32.67%	-47.29%	-22.01%
1,843	-15.06%	-14.54%	-14.28%
2,141	-8.71%	-15.88%	-17.39%

Table 6: The (NN) test over Average Link Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Average Link*		
	PartialRel	*AvRel*	*LastRel*
123	48.36%	4.50%	5.17%
666	38.47%	7.70%	18.89%
1,145	47.37%	23.69%	28.59%
1,843	47.22%	26.17%	37.29%
2,141	43.38%	32.89%	30%

Table 7: F-measure test over Average Link Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Average Link*		
	PartialRel	*AvRel*	*LastRel*
123	8.35%	6.76%	17.37%
666	7.83%	25.27%	12.14%
1,145	2.06%	20.44%	13.39%
1,843	3.17%	11.61%	13.19%
2,141	7.38%	12.85%	22.52%

Table 8: The (NN) test over Bisection K-means Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Bisection K − means*		
	PartialRel	*AvRel*	*LastRel*
123	15.81%	41.97%	-15.46%
666	31.73%	12.88%	-1.31%
1,145	24.22%	15.26%	-1.82%
1,843	29.48%	34.40%	-4.75%
2,141	37.98%	24.00%	-41.03%

Table 9: F-measure test over Bisection K-means Algorithm with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Bisection K − means*		
	PartialRel	*AvRel*	*LastRel*
123	-5.86%	-6.34%	5.25%
666	4.79%	-16.31%	-0.52%
1,145	-17.68%	-13.87%	2.83%
1,843	-2.50%	8.45%	7.77%
2,141	-7.82%	13.51%	12.96%

Table 10: The (NN) test over Ward's Method with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Ward′s Method*		
	PartialRel	*AvRel*	*LastRel*
123	36.47%	16.32%	39.44%
666	21.75%	8.09%	17.94%
1,145	30.61%	40.12%	28.12%
1,843	31.22%	29.21%	23.39%
2,141	21.62%	27.39%	18.80%

Table 11: F-measure test over Ward's Method with *PartialRel*, *AvRel*, and *LastRel*.

# of q	*Ward′s Method*		
	PartialRel	*AvRel*	*LastRel*
123	-4.38%	15.04%	15.37%
666	7.78%	9.34%	4.81%
1,145	8.22%	5.68%	6.85%
1,843	10.86%	7.07%	8.66%
2,141	13.08%	17.97%	8.55%

6 Discussion

The main reason traditional IR datasets were not used; it was because autonomous queries build them. As mentioned early, a traditional IR dataset is made up of a set of documents D, a set of queries Q, and a set of users' judgments JU. In a strict sense, evaluating the effectiveness considering similar queries (queries form part of

the clusters) means having a set of similar queries (Q') for Q. Note that Q' must consider a set of users' judgments (JU'). Indeed, the effectiveness of two similar queries should be different. Accordingly, these datasets are not suitable for evaluating approaches based on similar queries because they have neither Q' nor JU'. On the other hand, approaches based on log files employ subject matter experts to extend and evaluate whether a document is relevant or non-relevant given a query. In short, the extension of relevant documents is common in both types of collections. In this manner, aiming to avoid using subject matter experts or extending document relevances using relevant terms, three models of relevance have been simulated in all experiments.

Concerning the effectiveness evaluation for both measures (S_c and QDSM), it is noteworthy that F-measure has been widely used in several approaches, which deal with post-retrieval clustering. Nevertheless, the use of this measure provides two drawbacks. The first one is that the result associated with this measure comprises the number of relevant and non-relevant documents related to recall and precision in its mathematical formula. Thus, the initial effectiveness changes its value once new documents are considered relevant in the post-retrieval process. The second one refers to how the clusters are conformed taking into account the different classes of objects that these contain. Consider that objects can belong to predetermine classes, and the ideal situation is given when the clusters are formed only by objects of the same class. Two terms well-known in the cluster evaluation reflect this situation, homogeneity and completeness. The idea behind homogeneity is that each cluster has few classes; meantime, completeness intends each class to be contained in a few clusters. Thus, two-cluster forming using the same objects and the same classes can have the same F-measure, while their homogeneity and completeness are different. In turn, like F-measure, the NN-test has been extensively used in several approaches to assess effectiveness. Nevertheless, this measure is not sensitive to homogeneity and completeness, since it contemplates the direct search of the n-nearest neighbors. Hence, this test is more appropriate to corroborate the cluster hypothesis, which considers the relevant documents that form the clusters.

Regarding the results presented in section "Empirical Results", it is essential to point out that there is no significant difference between values provided by F-measure and the (NN) test, excepts for the Complete algorithm (Table 4 and 5), in particular for "AvRel" where for the (NN) test, Table 4 presents favorable results for QDSM in contrast to Table 5. Besides, the p-values for both Tables differ. On the other hand, exists a substantial difference with some results presented by [1]. In particular, regarding the relevance models used in that research. There, the relevance "AllRel" is used considering the proposed by [26], meantime in this research "AllRel" has been modified by "PartialRel" and "AvRel", it means that no all documents have been considered relevant such as occurs in [1]. It is important to point out that it is unlikely that all recovered documents are relevant, such as happens in the real world. Nevertheless, the Average Link algorithm presents interesting results in both works. Concerning the results provided by the algorithms in this research, the best results are provided by Average Link and Ward algorithms using both tests (F-measure and the (NN) test). The main Ward characteristic is that it minimizes the variance of the objects belonging to a particular cluster using the "error sum of squares". In this way, each cluster should tend to have objects of a few classes (relevant and non-relevant). Carried to the hypothesis cluster context should have a clear separation between clusters of relevant documents and clusters of non-relevant documents. Therefore, the nearest closest neighbor of a relevant document should be relevant too. On the other hand, the distance (S_c or QDSM) between two clusters for the Average Link Algorithm is determined as the average distance between each object in one cluster to every object in another cluster., by which it is feasible to avoid extreme measures obtaining more homogenous clusters. The latter is in contrast with the way to built clusters in Single and Complete algorithms. Finally, Bisection K-means is a hybrid approach between agglomerative and hierarchic clustering. This algorithm exhibits favorable results in the (NN) test except when the relevance is "LastRel", recall that in this case, only the last recovered document could be relevant.

Although the running times escape from the scope of this paper, it is worth noting that most time complexities are not high. To obtain the time complexities is necessary to consider visiting a distance matrix (i.e., one matrix for S_c and QDSM respectively) with the aim to find the n nearest-neighbor. Furthermore, calculating LCS implies to visit another matrix with M files and N rows. Note that M corresponds to a list of retrieved documents for a query q, whilst N is another list of retrieved documents for a query q'. Therefore, LCS takes $O(MN)$. Recall that LCS is used to evaluate QDSM. According to [32], the optimal implementation of Ward based on the algorithms, nearest neighbor chain, and reciprocal nearest neighbor, takes $O(N^2)$. In turn, the time complexity of Bisection K-means algorithm is $O(N^2 log_2 N)$. As for the implementation of Single Link algorithm takes $O(N^2)$ in time complexity [33]. On the other hand, Complete Link implies $O(N^2 log_2 N)$ [34]. The time complexity for the Average Link algorithm takes $O(N^2 log_2 N)$ [35]. It is important to mention that both matrixes of distances are previously built before using each clustering algorithm.

7 Conclusion and Future Work

This paper is intended to check the quality of clusters (effectiveness) built using Query-Document Similarity Measure (QDSM). To achieve this goal, the F-measure and the nearest neighbor (NN) test were used to evaluate clusters' quality. The clusters of documents were built using the AOL Query Logs Dataset. In order to provide relevance to the documents, three variants related to the ItemRanks over recovered documents were simulated. Extensive experimentation was carried out using the algorithms Single Link, Complete Link, Average Link, Bisection K-means, and Ward. According to results obtained, applying the nearest neighbor (NN) test, QDSM presents significant results using the Average Link, Ward, and Bisection K-means. On the other hand, in accordance with the results obtained by the F-measure; Ward and Average Link algorithms provide better results using QDSM than Cosine Similarity (S_C). The best results are provided by the Average Link algorithm, followed by Ward's method using QDSM, considering the three variants of relevance. Ideas for future research comprises the comparison between QDSM and other state-of-art measures.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The Group of Smart Industries and Complex Systems (gISCOM) under grant DIUBB 195212 GI/EF supported the work presented in this article. Marco Palomino acknowledges the funding provided by the Interreg 2 Seas Mers Zeeën AGE'IN project (2S05-014).

References

[1] C. Gutiérrez-Soto, A. C. Díaz, G. Hubert, "Comparing the Effectiveness of Query-Document Clusterings Using the QDSM and Cosine Similarity," in 2019 38th International Conference of the Chilean Computer Science Society (SCCC), 1–8, 2019, doi:10.1109/SCCC49216.2019.8966432.

[2] A. Mikroyannidis, "Toward a Social Semantic Web," Computer, **40**(11), 113–115, 2007, doi:10.1109/MC.2007.405.

[3] "Satista,Average number of search terms for online search queries in the united states as of august 2017," 2017.

[4] A. Kanavos, P. Kotoula, C. Makris, L. Iliadis, "Employing query disambiguation using clustering techniques," Evolving Systems, **11**(2), 305–315, 2020, doi:10.1007/s12530-019-09292-7.

[5] M. Alshomary, N. Düsterhus, H. Wachsmuth, "Extractive Snippet Generation for Arguments," in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, 1969–1972, Association for Computing Machinery, New York, NY, USA, 2020, doi:10.1145/3397271.3401186.

[6] Z. A. Merrouni, B. Frikh, B. Ouhbi, "Toward Contextual Information Retrieval: A Review And Trends," Procedia Computer Science, **148**, 191 – 200, 2019, doi:10.1016/j.procs.2019.01.036, the Second International Conference on Intelligent Computing in Data Science, ICDS2018.

[7] N. Jardine, C. van Rijsbergen, "The use of hierarchic clustering in information retrieval," Information Storage and Retrieval, **7**(5), 217 – 240, 1971, doi:10.1016/0020-0271(71)90051-9.

[8] C. Gutiérrez-Soto, A. C. Díaz, "Improving Precision in IR Considering Dynamic Environments," in Proceedings of the 21st International Conference on Information Integration and Web-Based Applications and Services, iiWAS2019, 458–462, Association for Computing Machinery, New York, NY, USA, 2019, doi:10.1145/3366030.3366101.

[9] A. Tombros, C. J. van Rijsbergen, "Query-sensitive Similarity Measures for Information Retrieval," Knowl. Inf. Syst., **6**(5), 617–642, 2004, doi:10.1007/s10115-003-0115-8.

[10] L. Fu, D. H.-L. Goh, S. S.-B. Foo, "The effect of similarity measures on the quality of query clusters," Journal of Information Science, **30**(5), 396–407, 2004, doi:10.1177/0165551504046722.

[11] K. W. Leung, W. Ng, D. L. Lee, "Personalized Concept-Based Clustering of Search Engine Queries," IEEE Transactions on Knowledge and Data Engineering, **20**(11), 1505–1518, 2008, doi:10.1109/TKDE.2008.84.

[12] D. Beeferman, A. Berger, "Agglomerative Clustering of a Search Engine Query Log," in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, 407–416, Association for Computing Machinery, New York, NY, USA, 2000, doi:10.1145/347090.347176.

[13] S. Chawla, "A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search," Applied Soft Computing, **46**, 90 – 103, 2016, doi:10.1016/j.asoc.2016.04.042.

[14] J. Sankhavara, "Feature Weighting in Finding Feedback Documents for Query Expansion in Biomedical Document Retrieval," SN Computer Science, **1**(5), 75–82, 2020, doi:10.1007/s42979-020-0069-x.

[15] A. Tombros, R. Villa, C. Van Rijsbergen, "The effectiveness of query-specific hierarchic clustering in information retrieval," Information Processing and Management, **38**(4), 559 – 582, 2002, doi:10.1016/S0306-4573(01)00048-6.

[16] F. Raiber, O. Kurland, F. Radlinski, M. Shokouhi, "Learning Asymmetric Co-Relevance," in Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR), ACM - Association for Computing Machinery, 2015.

[17] S.-H. Na, "Probabilistic co-relevance for query-sensitive similarity measurement in information retrieval," Information Processing and Management, **49**(2), 558 – 575, 2013, doi:10.1016/j.ipm.2012.10.002.

[18] M. A. Hearst, J. O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '96, 76–84, Association for Computing Machinery, New York, NY, USA, 1996, doi:10.1145/243199.243216.

[19] "Query Clustering Using User Logs," ACM Trans. Inf. Syst., **20**(1), 59–81, 2002, doi:10.1145/503104.503108.

[20] R. Baeza-Yates, C. Hurtado, M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," in W. Lindner, M. Mesiti, C. Türker, Y. Tzitzikas, A. I. Vakali, editors, Current Trends in Database Technology - EDBT 2004 Workshops, 588–596, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[21] C. Gutiérrez-Soto, G. Hubert, "Evaluating the Interest of Revamping Past Search Results," in H. Decker, L. Lhotská, S. Link, J. Basl, A. M. Tjoa, editors, Database and Expert Systems Applications, 73–80, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[22] C. Gutiérrez-Soto, G. Hubert, "On The Reuse of Past Searchesin Information Retrieval:Study of Two Probabilistic Algorithms," International Journal of Information System Modeling and Design, **6**(2), 71–90, 2015.

[23] L. Bergroth, H. Hakonen, T. Raita, "A survey of longest common subsequence algorithms," in Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000, 39–48, 2000, doi:10.1109/SPIRE.2000.878178.

[24] G. Pass, A. Chowdhury, C. Torgeson, "A Picture of Search," in Proceedings of the 1st International Conference on Scalable Information Systems, InfoScale '06, 1–es, Association for Computing Machinery, New York, NY, USA, 2006, doi:10.1145/1146847.1146848.

[25] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, "A Deep Look into Neural Ranking Models for Information Retrieval," CoRR, **abs/1903.06902**, 2019.

[26] S. Jiang, S. Zilles, R. Holte, "Empirical Analysis of the Rank Distribution of Relevant Documents in Web Search," in 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, volume 1, 208–213, 2008, doi:10.1109/WIIAT.2008.200.

[27] A. Crystal, J. Greenberg, "Relevance criteria identified by health information users during Web searches," Journal of the American Society for Information Science and Technology, **57**(10), 1368–1382, 2006, doi:10.1002/asi.20436.

[28] A. Spink, J. Bateman, J. B. J., "Searching the Web: a survey of EXCITE users," Internet Research, **9**(2), 117–128, 1999, doi:10.1108/10662249910264882.

[29] D. C. Blair, "Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50," Journal of the American Society for Information Science, **30**(6), 374–375, 1979, doi:10.1002/asi.4630300621.

[30] E. M. Voorhees, The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval, Ph.D. thesis, Cornell University, 1985.

[31] E. M. Voorhees, "The Cluster Hypothesis Revisited," in Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '85, 188–196, Association for Computing Machinery, New York, NY, USA, 1985, doi:10.1145/253495.253524.

[32] R. Amorim, V. Makarenkov, B. Mirkin, "A-Ward pβ : Effective hierarchical clustering using the Minkowski metric and a fast k-means initialisation," Information Sciences, **370**, 342–354, 2016, doi:10.1016/j.ins.2016.07.076.

[33] C. J. van Rijsbergen, "An algorithm for information structuring and retrieval," The Computer Journal, **14**(4), 407–412, 1971, doi:10.1093/comjnl/14.4.407.

[34] O. Aichholzer, F. Aurenhammer, "Classifying Hyperplanes in Hypercubes," SIAM Journal on Discrete Mathematics, **9**(2), 225–232, 1996, doi:10.1137/S089548019426348X.

[35] H. Schütze, C. Silverstein, "Projections for Efficient Document Clustering," in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '97, 74–81, Association for Computing Machinery, New York, NY, USA, 1997, doi:10.1145/258525.258539.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 46-54 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Using Leader Election and Blockchain in E-Health

Basem Assiri*

Computer Science Department, Jazan University, 45142, Jazan city, Saudi Arabia

ARTICLE INFO

Article history:
Received: 13 March, 2020
Accepted: 20 April, 2020
Online: 03 May, 2020

Keywords:
Electronic Personal Health Record
Parallelism
Distributed Systems
Leader Election Algorithms
Blockchain Technology

ABSTRACT

The development of electronic healthcare systems requires to adopt modern technologies and architectures. The use of Electronic Personal Health Record (E-PHR) should be supported by efficient storage such as cloud storage which enables more security, availability and accessibility of patients' records. Actually, the increase of availability of E-PHR enhances parallel access, which improves the performance and the throughput of the system. Using distributed systems, users are able to communicate and to share resources to achieve specific goals. Such kind of access needs to have more coordination to maintain parallelism, which can be provided through leader election algorithms. In leader election algorithms, users elect one of them as a leader to coordinate the work and to prevent conflicts. This paper introduces an adoptive leader election algorithm (ALEA) that considers medical and healthcare specifications, since it uses leader election algorithm for E-PHR in the cloud environment. The use of ALEA improves performance by allowing more parallelism and reducing the number of coordinating messages within the system, as well as facilitating the medical specifications such as having a primary doctor or handling emergency situations. Moreover, the paper highlights the strengths and weaknesses of using Blockchain technology in the field of healthcare. In fact, the paper investigates the implementation challenges of ALEA concepts using Blockchain technology.

1. Introduction

Within the last decades, the development of technologies, the Internet and digital applications makes them essential components in some other fields such as education and healthcare. This paper focuses on the development of e-health systems using some supportive algorithms and modern technologies. Actually, this paper is an extension of work originally presented in the 2ndInternationalConference on New Trends in Computing Sciences [1].

The competition among healthcare organizations encourages them to involve modern and advanced technologies to increase stakeholders' satisfaction. These technologies help to improve the provided services. For examples, a patient can schedule an appointment online; doctors (physicians) can access, maintain, and transfer E-PHRs electronically anytime and from anywhere; doctors would be able to diagnose, complete the required treatment and even participate in surgery remotely; prescriptions are sent to the corresponding pharmacy electronically. In addition, these technologies enable costs and managerial efforts to be reduced.

*Basem Assiri, 45142 - Jazan University - Jazan City – Saudi Arabia, +966599933185 & babumussmar@jazanu.edu.sa, bas0911@hotmail.com.

The services costs can be time, physical space, energy and infrastructure. Besides that, it gives deep and clear insight for better administration and decision making. The use of E-PHR as a digital version of PHR allows information to be accessed, updated and transferred in an electronic manner [2], which enhances information accessibility, availability, security, privacy, completeness and consistency. It also helps to avoid the risk of having traditional PHR in case of natural disasters such as Hurricane Katrina [2]-[6].

Moreover, an efficient pattern of storage such as cloud storage is required to support the use of E-PHR in distributed systems. Actually the E-PHRs are stored in servers and can be accessed securely on the Internet [7], [8].

The presence of E-PHR, servers, cloud storage and many connected devices creates a parallel and distributed system. In distributed systems, devices are connected through networks to perform specific tasks. Thus, it helps to improve efficiency and throughput of the process of sharing resources but it also requires more control and coordination. Therefore, leader election algorithms can be used to coordinate the parallel tasks and to preserve the consistency of E-PHRs [9]. Indeed, having

parallelism could result in conflicts, especially when a process tries to update an E-PHR, while another one is reading or updating it at the same time. In fact, having a leader allows for an exclusive access (token) to an E-PHR, which keeps it consistent.

In addition, Blockchain technology allows decentralized ledger to be managed within a distributed system. The ledger contains a chain of blocks (records). In Blockchain many nodes share and process distributed copies of the same ledger. In fact, when a node proposes a new block of transactions, the other nodes process it and vote to commit it if its valid or to abort to if it is not. Based of the votes of the majority (consensus), the block commits and every node updates its copy of the ledger or the block is ignored [10].

Nakamoto uses Blockchain to produce the Bitcoin as a first cryptocurrency, in which users execute electronic financial transactions without banks [10]-[12]. After that, Blockchain technology has been involved in many other areas such as judiciary, notary, copyrights, education and healthcare [12].

This paper introduces an adoptive leader election algorithm(ALEA) that is suitable for E-PHRs and healthcare systems. The paper proposes the principles of a primary and a secondary leader as well as having multiple tokens. ALEA allows the number of communication messages to be reduced in case of failures. Moreover, this work highlights the strengths and weaknesses of using Blockchain technology in the field of healthcare to implement the concepts of ALEA.

The rest of our paper is organized as follows: in Section 2, some related work is discussed. Sections 3 and 4, introduce our proposed system model and algorithm. Section 5 discusses many important issues and techniques such as algorithm correctness, consistency, synchronization, file sharing, traffic flow and replication. Finally, Sections 6 and 7 focus on the advantages and disadvantages of using Blockchain technology, while Section 8 provides conclusion.

2. Related Work

There are many techniques to preserve data consistency in cloud storage [13], [14]. Some research proposes strict consistency while the others relax this concept and accept weaker levels of consistency. Coppieters et al. provide an algorithm with strict consistency, where they order all concurrent processes on all replicas. Actually, the concurrent execution of processes should be matched with a correct sequential execution [15]. Zellag and Kemme introduce an efficient relaxed consistency model for cloud storage [16].

Some others use leader election algorithms for consistency, whereby a leader grants exclusive access to some memory objects to prevent conflict [9], [17]. For leader electing, a bully algorithm [9, 18], enables every user to have a unique identifier (Id) and every user sends its Id to all other users. So, they select the node with the maximum Id as a leader. The complexity of this approach reaches $O(n^2)$ messages, which is considered expensive. With a token ring algorithm [19], the users are structured in a linked-circle and every user sends its Id to the next one. After receiving the message, the user compares its own Id with the received one and sends the greater one to the next user. The complexity of this approach costs $O(n)$ messages. Numan et al. provide an algorithm that uses a centralized linked-list queue of all users. The leader is the head of the queue; and when it fails, another user acquires a lock and dequeues the old head. The complexity of this approach is $O(1)$ [20].

At the same time, many countries and institutions have started using E-PHRs. For example, at the beginning of 2014, the American Recovery and Reinvestment Act enforced all healthcare agencies (public and private) to use E-PHR. This facilitates accessibility, utilization and management. However, such a change requires specific technical and infrastructural support to be adopted [21].

In addition, Blockchain technology helps to allow decentralized management of E-PHRs, where there is no need for a third party such as hospitals or healthcare agencies. However, Blockchain has been enhanced with fairness and freedom [22]. Therefore, Blockchain provides many advantages for many areas such as health-care systems. First, it supports the availability, robustness and security of E-PHR. In fact, many projects and companies, such as Data Gateways, Gem Health Network, Deloitte and Guardtime have started using Blockchain to manage their E-PHR [11], [23]. Second, it supports all related financial operations such as funding, donations and insurance payments through cryptocurrencies. Third, Blockchain supports scientific clinical and biomedical research such as in the MedRec Health bank. Indeed, companies and organizations use Blockchain for data sharing and verification, ownership proofs and privacy of patients and organizations. In addition, it could apply the principle of "gain as you contribute" in scientific clinical and biomedical research. For example, while Bitcoin (which is the first cryptocurrency) is earned through solving puzzles, Gridcoin, is another cryptocurrency that is earned based on the contribution to scientific research [10], [11] and [24]. Fourth, Blockchain can be used to manage and process any kind of data such as the records of employees, healthcare facilities, medical instruments, medicines and pharmaceutical supply chains.

Laure and Martha use Blockchain as a control manager to manage the access of E-PHR. Actually, this work focuses on Blockchain's advantages such as scalability, security and privacy. Indeed, the work proposes using a Blockchain system as an access control manager, so it only has indexes of records, while the real records are stored in separate storage (out of the Blockchain). This would help to avoid the negative impact of data redundancy, where every node in the Blockchain has a copy of all indexes instead of having a copy of all records [21].

Kevin et al. suggest the use of Blockchain in healthcare to solve the issues of hardware and software heterogeneity. The work focuses on the quality of sharing E-PHR in an understandable and meaningful manner. So, they use the concept of Fast Healthcare Interoperability Resources with specific Application Programming Interfaces as standard for data formatting and presentation [25].

3. The System Model

Before you ALEA is designed according the concepts of in well-organized bully algorithm for leader election [20]. The well-organized bully leader election algorithm is implemented using a linked-list queue to minimize the cost of leader election to $O(1)$. ALEA is a modified algorithm that is efficiently applicable for medical and healthcare systems.

ALEA creates a linked-list queue *Queue* of size *Z*, where *Z* is the number of nodes in the *Queue*. The *Queue* is stored in shared memory, so it is always visible to all nodes. A node (processor/doctor) is denoted as *Nod* and is represented with a unique identifier (*doctorId*), a pointer that is pointing to the next node ,and a *token* flag. The new inserted node (doctor) can read the E-PHR but it has to get exclusive access to update it. The exclusive access is given by changing the *token* value from *0* to *1*. The queue *head* is the leader (*PL*). To change the leader, dequeue the current *head* node and move the leadership (*PL pointer*) to the next node.

For emergency cases where the *PL* is not accessible, a temporary queue *TQ* is created with a secondary leader called *SL*.

Figure *1* shows *Queue* where the list has three doctors, and *PL* points to the head of the queue. It also shows the emergency block which has another list of *TQ* and *SL*. In reality, *TQ* is not usually there.

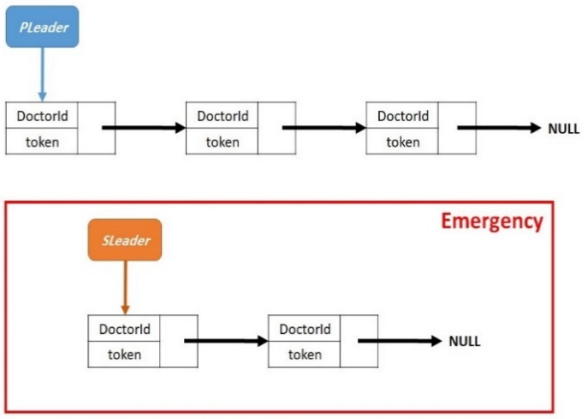


Figure 1: Leader election linked-list *Queue* with a *PL* pointing to the queue head; while the emergency linked-list *TQ* and a *SL* appear in the emergency block.

It is obvious that many doctors can read files in parallel, but conflict arises at the time of the update. For instance, the situation arises when one doctor is reading a file and another updating it. In this case, the data updated last should be visible to the reader. Also, when two doctors are writing to (updating) the same files, the two updating processes conflict with each others. To overcome this situation, it incorporates the idea of *TokenPtr*, which shows who holds the *token* flag to update the files.

4. Proposed Algorithm

- A hospital creates the patient's E-PHR with a primary doctor (leader) *PL*, (see Function 1). Actually, the node (doctor) has three attributes as follows: first, data to have the *doctorId*; second a *pointer* pointing to the next node; and third a *token* flag. When a new doctor is inserted into the *Queue*, the doctor can read the E-PHR directly, but for update permission, the *token* has to be changed to *1*. The *TokenPtr* is another pointer pointing to the node that has the *token* (*token=1*). Finally, it increases the size of *Queue*.
- If *PL* cannot be accessed for any reasons, except in the case of failure, it creates an emergence or temporary queue *TQ* with a secondary leader *SL*, (see Function 2). With the creation of a new node, the size of *TQ* is increased and the same procedures as in Function 1 are used.
- Now Function 3 shows the process of adding a new node to the *Queue*. Considering the medical needs, *PL* can add a new doctor to the team, by creating and enqueuing a new node to the *Queue*. Then it increases the *Queue* size. This is also applicable to *TQ*.
- Sometimes because of medical needs, the leader has to reorder doctors in the *Queue*, so it swaps the nodes as shown in Function 4. Based on the doctor's Id, *TempPtr1* starts searching from the head position, until it finds the first doctor. Then, *TempPtr2* starts searching from the head position, until it finds the second doctor. After that, it swaps the doctors by inserting *doctorId1* in the node of *TempPtr2* and *doctorId2* in the node of *TempPtr1*.
- When *PL* retires from leadership of the E-PHR's team, it follows the procedures for Function 5. If it is the only doctor who handles the E-PHR, which means *Queue* has one node only, then, the retirement is not allowed. Else, it uses *TPtr* to point to the *PL* node; moves the *PL* pointer to the next node; moves the *token* (if it is needed) and finally it dequeues the *TPtr* node and enqueues it again from the other end of the *Queue*. This is also applicable for *SL* and *TQ*.
- In Function 6, a doctor leaves the E-PHR's team (complete clearness). If it is the only doctor who handles the E-PHR, which means *Queue* has one node only, then, the clearness is not allowed. Else, it is dequeued from the *Queue* (as in Function 5), but there is no need to enqueue it again.
- Function 7, explains how to move the *token* among nodes. First, *PL* finds the targeted node according to its *doctorId*, then it gives the *token* (makes *token=1*), or gets it (makes *token=0*). In addition, *PL* allows parallelism by giving the *token* to many nodes simultaneously, which is explained in details later.
- Function 8, shows the case of a doctor requiring the *token*, so it sends an acquiring message to *PL* and waits for some time (*Timeout*). It should wait until the time becomes equal to *T* where *T=currenttime+Timeout*. In fact, it is supposed to receive a reply message (acknowledgment) from *PL*. However, if the timeout finishes without receiving the reply message, then *PL* fails, and it calls *PLFailure()*. Actually many nodes may detect *PL* failure simultaneously, so each node has to copy the *Id* of *PL* (failed leader's *doctorId*) (more details are given in Function 9).
- Function 9, shows the case of leader *PL* failing. Upon the detection of *PL* failure, the detector node calls *PLFailure(Id)*. It also passes *PL*'s *doctorId*. *PLFailure(Id)* moves the *PL* pointer to the next node and

1. ‖ **Algorithm 1: ALEA1.**

2. **1. Initialization()**
3. //To create the linked-list

```
Queue Z=0; //The queue size
Nod=newnode();
Nod→data=doctorId;
Nod→next=NULL;
Nod→token=0;
PL←node;
TokenPtr=PL;
TokenPtr→token=1;
Z++;
Return
```

```
2. Emergency()
//Adding a new doctor as SL & creating a temporary queue
Z=0; //The queue size
Nod=newnode();
Nod→data=doctorId;
Nod→next=NULL;
Nod→token=0;
SL←node;
TokenPtr1=SL;
TokenPtr1→token=1;
Z++;
Return
```

```
3. AddDoctor()
//Adding a new doctor to Queue
Nod=newnode();
Nod→data=doctorId;
Nod→next=NULL;
Nod→token=0;
Queue←enqueue();
Z++;
Return
```

```
4. SwapNodes(doctorId1, doctorId2)
//Swapping the doctors in Queue
TempPtr1=PL;
TempPtr2=PL;
i=1;
while(i ≤ Z) do
{
   if (TempPtr1→data, doctorId1) then
   {
   TempPtr1=TempPtr1→next;
   i++;
   }
   else
   {
   //First doctor is found, now find the other one
   Break;
   }
}
i=1;
while(i ≤ Z) do
{
   if (TempPtr2→data, doctorId2) then
   TempPtr2=TempPtr2→next;
   i++;
   }
   else
   {
   //Now Second doctor is found, so swap them
   TempPtr1→data=doctorId2;
   TempPtr1→token=0;
   TempPtr→data=doctorId1;
   TempPtr2→token=0;
   Break;
   }
}
Return
```

```
5. Rretirement()
//Retiring from the leadership
if (PL→next=NU LL) then
Return False;
else
{
TPtr=PL;
PL=PL→next;
   if (TokenPtr=TPtr) then
   {
   TokenPtr→token=0;
   TokenPtr=TokenPtr→next;
   TokenPtr→token=1;
   }
TPtr.dequeue();
T Ptr.enqueue();
}
Return
```

```
6. Clearness()
//Clearness
if (PL→next=NU LL) then
return False;
else
{
TPtr=PL;
PL=PL→next;
   if (TokenPtr=TPtr) then
   {
   TokenPtr→token=0;
   TokenPtr=TokenPtr→next;
   TokenPtr→token=1;
   }
TPtr.dequeue();
}
Return
```

dequeues the failed leader. If many nodes detect *PL* failure simultaneously, all of them invoke *PLFailure(Id*), which may cause multiple unnecessary dequeues. Thus, it is mandatory to use a Compare-and-Swap atomic operation (CAS statement), by which only one node changes the leader [26]. Using a CAS statement, one node checks if the *PL* is still in a failure (*PL's doctorId=Id*), and it invokes *Clearness()*. In *Clearness()*, *PL* (failed leader) is dequeued and another leader is elected. Therefore, the other nodes that detected the failure of *PL* also use CAS, but find PL's *doctorId*≠Id, since they find *doctorId* of the new *PL* that is not equal to the value of *Id*, and do nothing.

```
7. Leadership(doctorId1)
//Moving the token among nodes
//Getting the token
TokenPtr→token=0;
//Find the targeted node based on Id
i=1;
while(i ≤ Z) do
{
   if (TokenPtr→data, doctorId1) then
    {
    TokenpPtr=TokenPtr→next;
    i+ +;
    }
    else
    {
     //Give the token
    TokenPtr→token=1;
    Break;
    }
}
Return
```

```
8. Reminder()
//Node sends a message to acquire the token
Sendmsg(PL, "Acquire token");
//Waiting (Timeout)
T=CurTime()+Timeout;
while(receiveack() = false && CurTime() < T) do
wait();
//Timeout finishes and no response (PL fails)
if (receiveack()=false) then
{
Id=PL→data;
PLFailure(Id);
}
Return
```

```
9. PLFailure(Id)
//Leader still in failure
CAS (PL→data, Id, Clearness());
Return
```

Figure 2 is an example of a failed leader. Node 3 and 4 discover that the leader has failed. In Figure 2 (a), the two detector nodes elect a new leader in parallel, so both dequeue and move PL pointer. This causes one unnecessary dequeue. In Figure 2 (b) the two nodes elect a new leader in parallel using CAS (the lock can also be used), so both of them copy the doctorId of the PL in Id (Id=1). Both of them apply CAS such that one node will have successful CAS, and it dequeues the failed leader. However, since a new leader is already there, the doctorId of the new leader does not equal to Id anymore. Obviously, the doctorId is 2, while Id=1. As a result of this, the other node gets a failed CAS, so it does nothing.

5. Analysis

5.1. Correctness

For ALEA correctness, the correctness of concurrent operations, that are made by doctors, are proved by satisfying Linearizability [26]. This requires the concurrent operations to be ordered to match a correct (valid) sequential execution. Indeed, ALEA is considered an event-based model [9], where a doctor performs the operations(a read/update) on the E-PHR, and each operation is represented by two instantaneous events (*begin()* and *end()*). A complete execution is a sequence of operations where there is no pending operation and every operation has the two events. For the correctness and legality of all operations, it is not difficult to argue about the correctness and legality of a sequential execution where all operations are running in one processor and one after another. This helps to prove the memory consistency and to predict the situation of the file, before and after each operation. Since ALEA has two kinds of operations which are update and read, an update operation is legal if it appears instantaneously and all later reads read it, until another update takes place. A read operation is legal if it reads the last written data, and all later reads read the same data until another update takes place. A sequential execution is legal if all its operations are legal. Then, the concurrent execution is correct and the memory is consistent if the order of concurrent execution (including events of all concurrent operations) matches the order of a legal sequential one; this is known as Linearizability [26].

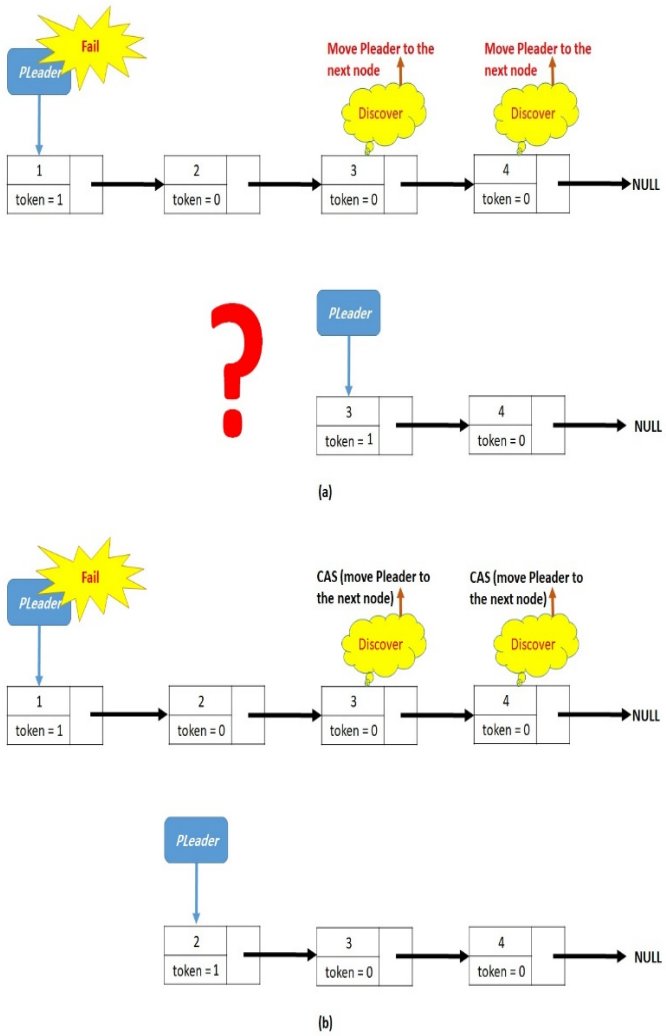


Figure 2: (a): As result of the failure of the leader, two nodes are electing a new leader in parallel, which causes one more unnecessary dequeue: (b) As result of the failure of the leader, two nodes use CAS to elect a new leader in parallel, so one node dequeues the failed leader and the other does nothing

In ALEA, there are two kinds of operations, some operations process the linked-list queue and the other operations process the E-PHR. The first kind is to modify the linked-list queue of doctors. To update the queue, there are operations such as enqueue, dequeue (even electing new leader is conducted through dequeuing) or move the token. In fact, ALEA maintains a queue of linked-list, so it follows the concepts of Michael and Scot [27], that is considered as one of the best lock-free algorithms in the field. Actually the enqueue and dequeue operations are conducted based on Michael and Scot's algorithm [27], which is linearizable. For the operation of moving the token, it is executed only by the leader, so it is serialized. The read operation simply tries to find the current leader which is always able to see the last change. In fact, the linked-queue is stored on a shared memory which makes all updates visible instantaneously. The second kind of operations is to update the E-PHR. Actually, some procedures that prove the correctness of the bully leader election algorithm [9, 18], and the well-organized bully leader election algorithm [20], are applicable in ALEA. To update the E-PHR it has to use the token which gives exclusive access, so it does not conflict with other operations and the operations can be ordered based on the token movement. On the other hand, all read operations are executed concurrently and are ordered with respect to the real-time order. Indeed, ALEA satisfies Linearizability.

5.2. Synchronization

Processes synchronization allows resources to be shared without causing any conflicts (it preserves consistency). Clearly, ALEA uses application level synchronization, but at some points it also relies on operating system primitives such as using locks and CAS atomic operation. Focusing on how to order events, Linearizability follows the real-time order of the concurrent operations (execution). In fact, when there is one thread, events are not interleaved on the same object. However, when there are many threads the synchronization satisfies a happens-before relationship. Therefore, the synchronization of events respects a well-formed clock. Actually, the order depends on the exact time in terms of where and when the operation takes effect. However, having multiple time zones of doctors and patients due to remote access or travel; challenges the use of physical clock. Therefore, a logical clock such as Lamport's logical clock [9, 23], is preferred. Lamport's logical clock is one of the famous approaches that uses the happens-before (|) relationship such that, for any two operations *a* and *b*, it is said *a* happens-before *b* (*a*|*b*) if and only if the *end()* of operation *a* occurred before the *begin()* of *b*. However, if there is an overlap between operations then the two possibilities are considered (*a*|*b*) or (*b*|*a*). Using ALEA, it orders the operations even if they are executed in multiple processors since all of them are executed on a single version of E-PHR. The read operation isordered easily, while the update operation gets exclusive access(token), so it is also ordered based on the token movements.

5.3. Parallel Access of E-PHR

In this part, it is suggested the E-PHR to divided into multiple sections *s*, and doctors are able to access different sections concurrently. Therefore, ALEA has to use multiple tokens, let say *k* tokens, where *k=s* (a token for each section). Now, the leader gives a suitable token to a doctor according to the needed section. Thus, *TokenPtr* of the original ALEA is replaced with a two-dimensional array of pointers (with *k* rows and 2 columns). Each row of this array shows the doctor's id and the corresponding section.

5.4. Traffic Flow

Processes It is known that the election of a new leader requires a number of messages to passed and that may costn2messages for *n* nodes [17, 18]. On the other hand, having one leader has severe negative impacts on the system in case of leader failure. To handle this situation, decentralized leader election algorithms enable many replicas for each file and more than one leader. To access any replica, voting is conducted and access is allowed according to the majority (consensus). Such type of permission requires approximately *2n* messages [17, 18]. However, with ALEA the number of messages is reduced to *0*, as it dequeues the head node (leader) and the new head will be the new leader (no traffic). In addition, other messages are sent to acquire the token. In ALEA the leader can hold a token or move it to the targeted node as needed with no messages. Rarely, if a doctor needs to get the token, the doctor sends an acquire message (as shown in Function 8), and it receives a reply message from the leader, which causes no traffic.

5.5. Fairness and Starvation (timeout)

This part focuses on fairness of the leader election process and fairness of token movements. For the fairness of leader election process, the leadership appointment has to follow the queue property first-in-first-leader. However, this is relaxed to approximate-first-in-first-leader to handle medical and healthcare requirements. Indeed, in ALEA, the leader is able to swap the doctors in the queue based on the medical needs, which is completely fair from the medical point of view. On the other hand, ALEA allows the leader to move token among nodes as needed, which is also fair from the medical point of view.

5.6. Replication

This part It would never be advisable to use a single centralized copy of the E-PHR since there is no way for recovery in case anything goes wrong. In this regard, the idea of having multiple replicas of the same file is integrated, which is very important for reliability and performance. Firstly, many replicas allow recover of files if the reis an issue with a server, security problem, file corruption or failed operation (read/update). Secondly, the replicas allow improved performance as they can be distributed based on system capabilities, competences, load balancing or geographical distribution. However, having many replicas requires more effort to keep them consistent. For such an issue, it is suggested that the number of replicas be reduced to three copies. The main (permanent) replica is stored in a suitable place bearing in mind the geographical location of the patient and all doctors accessing this replica. The second replica (backup1) should be stored in the same system or in a very close one, so it can be used easily for recovery. The third replica (backup2) must be stored in a different server that is geographically located far away from the main one; so it can be used in case of natural disasters, for example. In order to maintain consistency in replicas, it is suggested to use the eventual consistency criteria where the consistency of file is relaxed [9, 15]. Eventual consistency is very suitable to E-PHRs and cloud storage. Using eventual consistency, read operations do not cause any harm, as the replicas remain consistent, but this scenario will be different in the update operation because the update operation changes the content of the replica. After

confirmation of an update, the change should be reflected in all other replicas in an asynchronous manner (over the network). In ALEA, the token can be given to more than one doctor working on different parts of the file, which causes a write-write conflict. Therefore, every update is labeled with two timestamps, which are the timestamp when the update takes effect, and the timestamp of the last update before this one took place. This helps all replicas to order the changes and maintain consistency among all of them. Another way is to use Primary-Based Protocol to conduct remote write/update; in this scheme, the user holds a local copy of a file and after performing the update operation locally, it sends are quest to the server for final approval [9]. This way, the main server takes responsibility for preventing conflicts. To avoid any type of conflict, it updates the other replicas instantly. This protocol (of maintaining the local replica) is an actual support for mobile applications, where the user works on those selected files from different locations, which may affect the connection with the server. Besides this, a huge number of update operations on the system and the processes of maintaining replica consistency result in very high contention on the network. To solve this difficulty, the techniques below is effective:

- Directly, send the modified part of a file to replicas. This can be used in case of a considerable number of update operations.
- Only send a notification to invalidate the other replicas. The invalidation notification has a smaller size compared to the update messages. Indeed the invalidation can be more specific to tell which part of the replica is not valid anymore. Then, the system denies access to that part until it finds a suitable time to update it.
- For some kind of update, send the computation itself, so the others process it locally and make the update themselves.

6. The Use of Blockchain

Blockchain technology is a distributed ledger that is shared and processed by nodes according to consensus [10, 11]. In fact, Blockchain algorithms have three major phases which are proposing a new block, voting on the new block, committing or aborting it based the consensus [10, 11, 22]. The implementation of ALEA concepts using Blockchain technology requires some modifications. The creation of a new queue in Functions 1 and 2, will be based on the consensus of users, rather than hospitals or healthcare agencies. Moreover, adding, removing and swapping doctors (in Functions 3, 4, 5 and 6) will not be executed by the leader; instead, they have to make a proposal, vote and then take a decision according to the majority. The same thing is applicable to token movements in Functions 7 and 8, as well as in case of the failure of the leader in Function 9.

6.1. Transaction's Validation

The Blockchain consists of a number of nodes (processors). Each node can access the patients' E-PHRs and has a copy of the ledger. The ledger contains blocks and each block contains a number of transactions. The transaction contains some operations that are executed on E-PHRs. The operations on the E-PHR are either read or update (the update includes creating, editing and deleting E-PHR). After the execution of transactions, the miners validate the transactions by validating the output of the operations of those transactions. Indeed, the miners also consider the concurrency of transaction and can use some standard property for such issues such as Opacity [28]. According to this correctness property, a valid transaction commits and is placed on the blocks, while an invalid transaction aborts [29]. Note that more than one miner may validate the same transaction and every one places it in a different block. or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

6.2. Block's Validation

When the block is full (based on the size of the block), it is proposed to the validators (who are the team of doctors), so they validate the correctness of the block, considering the concurrent blocks as well. In fact, the content of the block is hashed to secure it. The hashing can be produced in many ways such as Proof-of-Work (PoW) or Proof-of-Stack (PoS) [30]. In addition, validators validate the signature of the proposer node to verify that the block has been created by a legitimate node. Actually, the validators validate the identity of the proposer through its digital signature, the PoW and content of the block. According to this, validators vote to commit the block or abort it.

6.3. Consensus

Blockchain uses consensus to consider the validation processes of all validators. If the majority votes to commit the block, then it is added to the ledger, otherwise it is ignored (and then there is no need for a leader like in ALEA). The majority of votes means more than 50% is needed, and in some systems they increase it to 67% [31, 22]. In addition, the order of the committed blocks in the ledger is decided based on who gets PoW first. Some work uses a unique timestamp for each block, so the one with a smaller timestamp is added first [22]. The blocks are chained in the ledger using the hash number, as every block has the hash of the previous one; thereforethe ledger is unchangeable.

7. Discussion on the Use of Blockchain

The implementation of ALEA's concept using Blockchain has some strengths and weaknesses. Thus, this section discusses the issues of security, privacy, immutability, decentralization, robustness, availability, ownership, computational costs and system traffic flow [10, 11, 24, 21].

7.1. Security, Privacy and Immutability

Using Blockchain technology the identities of doctors are hidden. The E-PHRs and the operations on them are hashed and encrypted. Such security and privacy are positive points that encourage everyone to use Blockchain. On the other hand, users will be untraceable and that is an issue for healthcare systems, especially in terms of tracking suspicious and illegal behaviors.

In addition, using Blockchain guarantees immutability, since the ledger (that has the records of all operations on E-PHRs) is unchangeable. Obviously, this is a positive point; however, such a system does not allow to rollback.

7.2. Decentralization

To avoid the presence and the control of the centralized third party such as a hospital, decentralization enables queue to be

created or emergency to be handled using consensus. The negative impact of consensus is the voting delays; as well as having non-trusted nodes or misleading votes.

7.3. Robustness

To The distribution of Blockchain technology helps to prevent a single point of failure, especially in the case of leader failure. Instead, the work will continue as long as the majority of nodes is still working.

7.4. Availability

Blockchain provides high availability, since all nodes have replicas of the files. The large redundancy of replicas requires more space (memory), communication and processing (to preserve the consistency of the replicas). To avoid such an issue, it is proposed to use Blockchain as an access control manager that shares copies of indexes rather than the real records, and the records are stored in separate centralized storage [21], as shown in Figure 3.

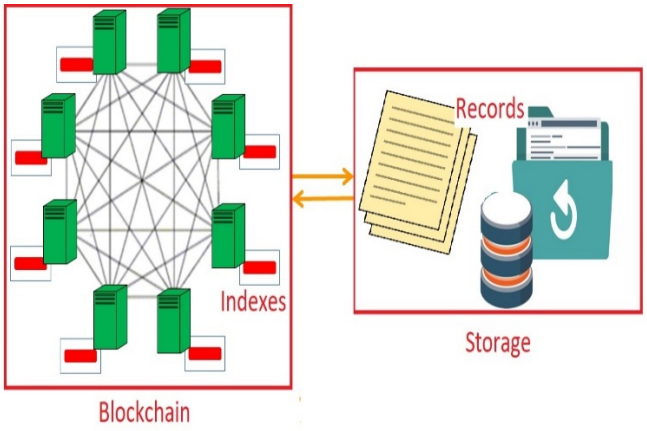


Figure 3: Blockchain with indexes and separated centralized storage of records.

7.5. Performance Cost

The use of Blockchain technology results in huge increases in computational and communication costs. In ALEA, the computation is executed by one node, while in Blockchain all validators execute the computation to confirm its correctness. In addition, when a node proposes a new block, it broadcasts to all validators, so if there are n validators, it broadcasts n messages. After the validators execute the computation, they send votes to all nodes, let's say m nodes, which costs m^2 messages. Actually, the set of the validators is a subset of all nodes in the Blockchain. Then, every node calculates the majority and broadcasts messages of the decision (commit/abort), which also costs m^2 messages. Finally, in case of commit, all nodes update their own copy of the ledger, rather than a limited number of replications. Thus, the use of Blockchain negatively affects the speed of the system and its traffic flow.

7.6. Fault Tolerance

The consensus is a fault tolerant correctness property, where some validators do not confirm the validity of the block. In medical cases, it is not suitable to ignore the votes of 49% or even 33% of doctors because they are not the majority. This means, the Blockchain may still allow for some errors, which is very critical for healthcare specifications.

8. Conclusion

This paper proposes an adaptive algorithm, for E-PHRs in a cloud environment, so it can be easily used with minimal infrastructure. ALEA enhances parallelism using an alternative and modified leader election technique that suits medical and healthcare systems. This work investigates the performance and the characteristics of the ALEA in comparison to Blockchain technology, which shows that the use of Blockchain technology may result in some negative impacts

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Basem Assiri. Leader election and blockchain algorithm in cloud environment for e-health. In2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), pages 1–6. IEEE, 2019.

[2] Paul C Tang, Joan S Ash, David W Bates, J Marc Overhage, and Daniel Z Sands. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. Journal of the American Medical Informatics Association, 13(2):121–126, 2006.

[3] Selena Davis, A Roudsari, and Karen L Courtney. Designing personal health record technology for shared decision making. Studies in health technology and informatics, 234:75–80, 2017.

[4] Janet Woollen, Jennifer Prey, Lauren Wilcox, Alexander Sackeim, Susan Restaino, Syed T Raza, Suzanne Bakken, Steven Feiner, George Hripcsak, and David Vawdrey. Patient experiences using an inpatient personal health record. Applied clinical informatics, 7(02):446–460, 2016.

[5] Arloc Sherman and Isaac Shapiro. Essential facts about the victims of hurricanekatrina.Center on Budget and Policy Priorities, 1:16, 2005.

[6] Shayne Sebold Taylor and Jesse M Ehrenfeld. Electronic health records and preparedness: lessons from hurricanes katrina and harvey. Journal of medical systems, 41(11):173, 2017.

[7] Mu-Hsing Kuo. Opportunities and challenges of cloud computing to improve health care services. Journal of medical Internet research, 13(3):e67, 2011.

[8] Hoang T Dinh, Chonho Lee, Dusit Niyato, and Ping Wang. A survey of mobile cloud computing: architecture, applications, and approaches. Wireless communications and mobile computing, 13(18):1587–1611, 2013.

[9] Andrew S Tanenbaum and Maarten Van Steen. Distributed systems: principles and paradigms. Prentice-Hall, 2007.

[10] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Technical report, Manubot, 2019.

[11] Tsung-Ting Kuo, Hyeon-Eui Kim, and Lucila Ohno-Machado. Blockchain distributed ledger technologies for biomedical and health care applications. Journal of the American Medical Informatics Association, 24(6):1211–1220, 2017.

[12] Michael Crosby, Pradan Pattanayak, Sanjeev Verma, Vignesh Kalyanaraman,et al. Blockchain technology: Beyond bitcoin. Applied Innovation, 2(6-10):71, 2016.

[13] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. Information systems, 47:98–115, 2015.

[14] Divyakant Agrawal, Sudipto Das, and Amr El Abbadi. Big data and cloud computing: current state and future opportunities. In Proceedings of the 14th International Conference on Extending Database Technology, pages 530–533, 2011.

[15] Tim Coppieters, Wolfgang De Meuter, and Sebastian Burckhardt. Serializable eventual consistency: consistency through object method replay. In Proceedings of the 2nd Workshop on the Principles and Practice of Consistency for Distributed Data, pages 1–3, 2016.

[16] Kamal Zellag and Bettina Kemme. How consistent is your cloud application? In Proceedings of the Third ACM Symposium on Cloud Computing, pages 1–14, 2012.

[17] Gerard Tel. Introduction to distributed algorithms. Cambridge university press, 2000.
[18] George Coulouris, Jean Dollimore, and Tim Kindberg. Distributed systems: Concepts and design edition 3.System, 2(11):15.
[19] P Beaulah Soundarabai, J Thriveni, HC Manjunatha, KR Venugopal, and LM Patnaik. Message efficient ring leader election in distributed systems. In Computer Networks & Communications (NetCom), pages 835–843. Springer, 2013.
[20] Muhammad Numan, Fazli Subhan, Wazir Zada Khan, Basem Assiri, and Nasrullah Armi. Well-organized bully leader election algorithm for distributed system. In 2018 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), pages 5–10. IEEE, 2018.
[21] Laure A Linn and Martha B Koo. Blockchain for health data and its potential use in health it and health care related research. In ONC/NIST Use of Blockchain for Healthcare and Research Workshop. Gaithersburg, Maryland, United States: ONC/NIST, pages 1–10, 2016.
[22] Basem Assiri and Wazir Zada Khan. Enhanced and lock-free tendermint blockchain protocol. In 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), pages 220–226. IEEE, 2019.
[23] Marc Pilkington. Blockchain technology: principles and applications. In Research handbook on digital transformations. Edward Elgar Publishing, 2016.
[24] Usman W Chohan. Environmentalism in cryptoanarchism: Gridcoin case study. Available at SSRN 3131232, 2018.
[25] Kevin Peterson, Rammohan Deeduvanu, Pradip Kanjamala, and Kelly Boles. A blockchain-based approach to health information exchange networks. InProc. NIST Workshop Blockchain Healthcare, volume 1, pages 1–10, 2016.
[26] Maurice Herlihy and Nir Shavit.The art of multiprocessor programming. Morgan Kaufmann, 2011.
[27] Maged M Michael and Michael L Scott. Simple, fast, and practical non-blocking and blocking concurrent queue algorithms. In Proceedings of the fifteenth annual ACM symposium on Principles of distributed computing, pages267–275, 1996.
[28] Rachid Guerraoui and Michal Kapalka. On the correctness of transactional memory. In Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming, pages 175–184, 2008.
[29] Nir Shavit and Dan Touitou. Software transactional memory. Distributed Computing, 10(2):99–116, 1997.
[30] Iuon-Chang Lin and Tzu-Chun Liao. A survey of blockchain security issues and challenges. IJ Network Security, 19(5):653–659, 2017.
[31] Basem Assiri and Costas Busch. Approximate consistency in transactional memory. International Journal of Networking and Computing, 8(1):93–123, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 55-61 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Organizational, Social and Individual Aspect on Acceptance of Computerized Audit in Financial Audit Work

Bambang Leo Handoko*, Nada Ayuanda, Ari Tihar Marpaung

Accounting Department, Faculty of Economics and Communication, Bina Nusantara University, Jakarta, Indonesia, 11480

ARTICLE INFO

Article history:
Received: 26 February, 2020
Accepted: 14 April, 2020
Online: 03 May, 2020

Keywords:
Organizational
Social
Individual
Audit
Software

ABSTRACT

Auditors now can no longer rely on the old-fashioned way of auditing manually. More and more jobs, increasingly complex work environment, the demands of the times, accuracy and speed of work require auditors inevitably must adopt technology. This research began with our success as academics in the audit family. Related to Indonesia, a large country and has several hundred public accounting firms and thousands of auditors, but computerized use of audits using software is still very little. The public accounting firm still uses a manual system, using normally typed paperwork. We want to find out what can boost the use of software among auditors. Our results are useful for auditors in Indonesia. From the results of statistical tests we found that auditors use compilation software by individual auditors themselves rather than organizations and individuals

1. Introduction

This paper is an extension of work originally presented in 2019 International Conference on Information Management and Technology (ICIMTech) [1]. The development of the information system used by the client has an impact on the expertise that must be mastered by the auditor who was originally approached manually, so with these changes the auditor is required to master the information system process used by the client and computerized audit technique for adjusting the audit process and the procedures used when carrying out field work such as changes in the manual accounting system environment into a computer-based accounting information system causes the auditor to study a system. Purpose auditing for carried out effectively and efficiently, the auditor should adjust his audit techniques to the client's information system [2].

General Audit Software or commonly abbreviated as GAS is an auditing with computer technique, or better known as Computer Assisted Audit Techniques. In Western countries, this GAS has been widely applied by audit companies. [3] Revealed that "Adoption of General Audit Software (GAS) can improve audit quality, and this is the reason why its use is applied by US Auditing Standards".

In terms of regulation (environment), the Indonesian Audit Standards suggest GAS in conducting works [4], but its use is not mandated / charged. Therefore, the use of technology in audit companies in Indonesia is not fully regulated. One of the Public Accounting Firm's Big Four states they use Generalized Audit Software because they have the software and own the resources.

In terms technology, information technology proficiency including English as foreign language is very influential for public accounting firms medium and small. In terms of technological skills, the Big Four is not in doubt. Obviously, become problem in domain of public accounting firms that smaller. On the contrary, is nor serious problem; the elite office believe language matters won't hamper their work, it used in operation every day. In addition, auditors must understand compatible software that is compatible to check/audit clients for accuracy in testing data integrity.

Associate terms of audit companies (organizations), IT Capital Budget is very influential on the implementation of General Audit Software. Big companies tend having customized audit tools, while medium and small audit companies usually rely on commercially available and cheaper software. Besides that, the IT skills of auditors in the company are also very influential.

On the other hand, the attitudes and intentions of auditors are also very influential in adopting General Audit Software (GAS). One senior auditor from the intermediate audit commented "Software is to help auditor work, nor to make it more complicated". Difficult to operate software make hamper usage.

*Bambang Leo Handoko, Email: bambang.handoko@binus.edu

As a result, the auditor will need even more effort to be able to use it. Interestingly, the intermediate audit office found young employee pay more interest in Generalized Audit Software (GAS) [5]. These premises also supported by research conducted by [6] which resulted that people willing to adopt, when it simplify audit process times and do their jobs more efficiently. Auditors who are young and have IT insights will usually be more interested if given the opportunity to enhance their skills.

2. Theoretical Framework and Hypothesis

2.1. Technology Acceptance Model

The Technology Acceptance Model (TAM) developed by [7] a successful and highly acceptable model for predicting willingness to adopt current information approach. Until now this theory was a pioneer and the foremost in uncovering the reasons people want to change to keep abreast of technological developments. Many studies have re-examined, expanded, and used TAM.

The TAM model originated from theory of reason action [8], that is grounded from study of reaction perception for something, effect on person's will and behavior. Reaction and perception affect in acceptance of technology. One factor that can influence is the user's will, so that the reason someone sees the actions / behavior as a benchmark in the acceptance of a technology. The TAM model that is grounded with a premise reaction or perception for something which outcome in result behavior. Reaction and perception affect in acceptance of technology. One factor that can influence it is the user's will his or herself.

According to [9] people willing to leave old method and change to new modern one, because they think it was useful and easy. Both of these components when associated with TRA are part of belief.

Basically it will be very strong if someone already has a perception, perception will be a suggestion, it will make someone want to do it [10]. The main perception here is the perception of usefulness, namely that this technology is very useful. But in addition to usability, it must also be supported by perceptions of ease, because we find many useful but difficult, it will also be left by people. Based on this premises and previous study by [11], [12] we formulated our premise as follows:

Hypothesis 1: PU has positive influence on System Usage (SU).

Hypothesis 2: PEU has positive influence on System Usage (SU).

Hypothesis3: PEU has positive influence on Perceived Usefulness (PU).

2.2. Auditing Software

CAATs according to [13] software designed to enable auditors with less sophisticated computer skills to carry out audits related to data processing functions. These packages can carry out certain analytical calculations, thus detect anomalies. Audit software is also interpreted as a computer program that allows automatic decision. Conventional tools such as system use programs, information reappearance programs, or high-level programming languages can be used for this audit. Real data processing goes through an audit program. Outputs are simulated and compared with regular outputs for monitoring purposes. Parallel simulation, redundant processing of all input data by conducting a separate program test, allows comprehensive and very precise validation to be carried out on important transactions that require 100% audit [14]. The audit program used in parallel simulations is usually a type of general audit program that processes data and produces output that is identical to the program being audited. Audit software is one of the software that adapted in educational work, along with other software used in college/university [15].

2.3. Social, Individual, and Organizational

Reflecting on the preliminary research, we find that there are recognizable 3 factors that are triggers for the perception of ease and perception of use. these three factors are individual, social and organizational [16]. The organization in our research context is the office where the auditor works. The policies taken by partners, which are the highest leaders in the public accounting firm, determine whether software is used or not. In accordance with [16] organization impact both Perceived Usefulness, Perceived Ease of Use (PEU).

The second factor to be investigated is social. Socials in our study are peers or peers of the auditor. When working of course the auditor works in a team, where in the team consists of several people. This is where social interactions were influenced [16]. [11] said people tend to be the same and not too different from their peers or social community. If one uses software easily, the others don't want to be left behind to use it. this opinion is reinforced by the results of the study [17], [18].

The third factor is individual. These factors originated from within the auditor himself. Where they have their own intentions, have their own perceptions before being influenced by the surrounding environment [16]. Individual factors are suspected to have strong potential, because it involves beliefs from within [13], Based on this premises, we form these several hypotheses related to the social, individual and organizational factors to PU and PEU.

H4: Organizational factor has positive influence on Perceived Usefulness

H5: Social factor has positive influence on Perceived Usefulness

H6: Individual factor has positive influence on Perceived Usefulness

H7: Organizational factor has positive influence on PEU

H8: Social factor has positive influence on PEU

H9: Individual factor has positive influence on PEU

3. Research Methodology

Our research was an associative research using path analysis test. We use a total of six factors. The six variables contain three independents, namely: individual, social, organizational, also two intervening variables namely perceived usefulness and perceived ease of use and one dependent variable, namely system usage. Hypothesis testing performed using statistical software.

Hypothesis testing is done after the precondition namely validity and reliability.

3.1. Population and Samples

Respondents in our study were people who worked as independent auditors or financial auditors in public accounting firms. We do not limit the accounting firm, whether the big four, ten or others. The population size is difficult to know, because the number of auditors is very large and changing, based on the approach taken by Chassan [19]. it was concluded that the respondents were at least 30, so we decided to use 100 respondents similar like [20]. .

3.2. Measurements of Factors

Factors used within our research are adjectives, which are basically abstract. For this reason, in order to be concrete and can be explored with certainty, then we made the operation of variables to make measurements. Variable measurements are made to make abstract variables (derived from adjectives) more tangible and can be calculated quantitatively. Variable measurement is based on preliminary research and grand theory. Here in table 1 presented the operation of variables.

Table 1: Table Operation of Variables

Operation of Variables		
Variable	**Main indicator**	**Source**
Organizational Factor (X1)	*1. Support* *2. Training* *3. Management Support*	[21]
Social Factor (X2)	*1. Internalization* *2. Image*	[21]
Individual Factor (X3)	*1. Job relevance* *2. Output quality* *3. Result demonstration*	[21]
Perceived Usefulness (Z1)	*1. Improve job performance* *2. Useful*	[9]
Perceived Ease of Use (Z2)	*1. Easy to Understand* *2. Easy to Use*	[9]
System Usage (Y)	*1. Frequency of Use* *2. Use anytime*	[9]

4. Research Result

We use path analysis to examine whether there is influence and how strong is the influence between variables in our study. However, before we conduct hypothesis testing, the data must first be tested for validity, reliability and classic assumptions. This is done to ensure that the data is valid, reliable and worth testing.

4.1. Reliability and Validity Testing

A variable is said to be reliable when the Cronbach alpha value is above 0.7 [22], reliable means that even if tested over and over again, the results will remain consistent.

Table 2: Cronbach's Alpha

No.	Variable	Reliability indicator
	List of variables	Cronbach's Alpha
1.	X1 Organizational factor	0.950
2.	X2 Social factor	0.830
3.	X3 Individual factor	0.879
4.	Z1 Perceived usefullness	0.820
5.	Z2 Perceived ease of use	0.861
6.	Y System Usage	0.947

Different from reliability, validity is intended to know whether if a question is asked to many parties then the answer remains the same. Validity can be seen from the r count greater than r table.

Table 3: Validity Testing

Variable			Variable		
Indicator	r count	r table	Indicator	r count	r table
OF			PU		
OF1	0.842	0.195	PU1	0.697	0.195
OF2	0.932	0.195	PU2	0.697	0.195
OF3	0.917	0.195	PEU		
SF			PEU1	0.756	0.195
SF1	0.712	0.195	PEU2	0.756	0.195
SF2	0.712	0.195	SU		
IF			SU1	0.900	0.195
IF1	0.697	0.195	SU2	0.900	0.195
IF2	0.862	0.195			
IF3	0.751	0.195			

4.2. Classical Assumption

This must be performed as a data quality test in testing the linear regression model. Quality multiple regression analysis in this study are free from deviations of assumptions. The classical assumption test, there are several tests including normality, multicollinearity, heteroscedasticity, and autocorrelation.

4.2.1. Normality Test

In this study normality testing is performed using graph analysis and statistical analysis.

The statistical test used in the study is Kolmogorov-Smirnov (K-S) non-parametric statistical test. Data requirements are

normal if the probability or p> 0.05 in the Kolmogorov-Smirnov normality test. Kolmogorov-smirnov test value > 0.05 means that the data is normally distributed [22]. If the Kolmogorov-Smirnov test value <0.05 then the data is not normally distributed. In Table 4 the K-S test value is 0.200 which means it is greater than 0.05, which means that the data are normally distributed.

Table 4: Normality Test

		Unstandardized Residual
N		100
Normal Parameters[ab]	Mean	.0000000
	Std. Deviation	1.95132147
Most extreme differences	Absolute	.088
	Positive	.088
	Negative	-.082
Test Statistic		.088
Asymp. Sig. (2 tailed)		.200[c]

The criteria used in chart analysis are, if the data spreads around a diagonal line and follows the direction of the diagonal line or the histogram graph shows a normal distribution pattern, then the regression model meets the normality assumption.

Based on Figure 1 it can be concluded that the data is spread around the diagonal line, so it can be concluded that the regression model meets the normality assumption

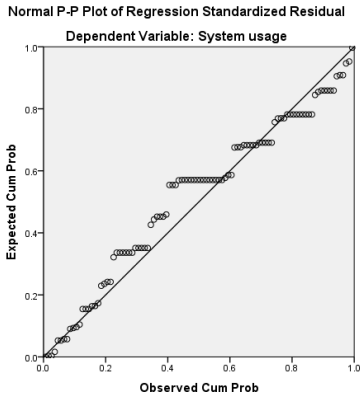


Figure 1: Graphic Normality

4.2.2. Multicollinearity Test

Multicollinearity test aims to test whether the regression model found a correlation between independent variables. A good regression model should not accept a correlation between independent variables. To find out the presence / absence of multicollinearity is to use Variance Inflation Factor (VIF) and tolerance. A low tolerance value is the same as a high VIF value (because VIF = 1 / Tolerance). If the tolerance value ≤ 0.10 or VIF value ≥ 10, it means that there is multicollinearity, while the tolerance value ≥ 0.10 or VIF value ≤ 10, it means there is no multicollinearity.

Based on Table 5, the tolerance values for the OF, SF, and IF variables are 0.597, 0.338 and 0.486, also for the PU and PEU variables, each at 0.532, which is greater than 0.100. In addition, the VIF values of the OF, SF, and IF variables of 1,674, 2,956 and 2,059, also for the PU and PEU variables, respectively 1,880, which are all smaller than 10. Based on these results, we can conclude that the research data free from multicollinearity

Table 5: Multicollinearity Test

Model		Collinearity Statistics	
		Tolerance	VIF
1.	(Constant)		
	PU	.532	1.880
	PEU	.532	1.880
	Organizational	.597	1.674
	Social	.338	2.956
	Individual	.486	2.059

4.2.3. Heteroscedasticity Test

Heteroscedasticity test aims to determine whether in the residual regression model there is an imbalance of variance from one observation to another. If the variance from one observation to another is the same, it can be called homoscedasticity and if it is different it can be called heteroscedasticity. Testing heteroscedasticity in this study is using statistical tests and plot graph tests.

The statistics used in this study to determine the presence or absence of heteroscedasticity is the Goldfeld-Quandt Test. The testing steps are as follows:

1. Sort the independent variable X from the smallest largest
2. Then make two separate regressions, first for the smallest X value. Second for large X values and omit some data in the middle.
3. Make the ratio of RSS (Residual Sum of Square = error sum if square) from the second regression to the first regression (RSS2 / RSS1) to get the calculated F value.
4. Perform the F test using degrees of freedom of (n-d-2k) / 2, where

n = number of observations,

d = the amount of data or observation values lost

k = estimated number of parameters.

F test criteria if:

F arithmetic > F table, then there is heteroskedasticity

F arithmetic < F table, then there is no heteroskedaticity

Results of group I regression with RSS1 = 381.974. The results of group II regression with RSS2 = 814.614. F-stat = RSS2 / RSS1 = 814.614 / 381.974 = 2.132. F-table is 4.85. It can be concluded that F-statistic < F-table it means that no heteroscedasticity

Heteroscedasticity test plot graph to detect the presence or absence of heteroscedasticity by looking at the plot graph, if there are certain patterns such as points that form certain patterns and orderly, then heteroscedasticity has occurred. If the pattern image spreads above and below the number 0 on the Y axis indicates the absence of heteroscedasticity.

Figure 2 illustrates the spread of scattered data so that it can be concluded that the data in this study is free from heteroscedasticity.

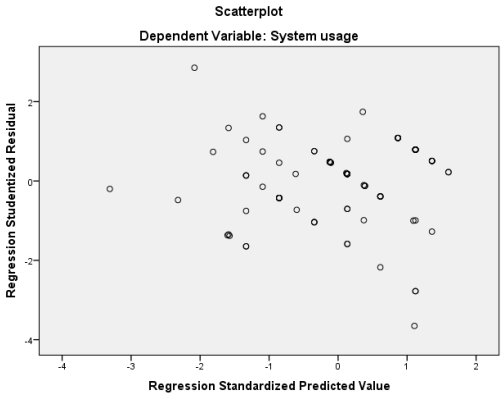


Figure 2: Graphic Heteroscedasticity

4.2.4. Autocorrelation Test

Autocorrelation test was performed using Durbin Watson. If the Durbin Watson value ranges between the upper limit values (dU), then an autocorrelation violation is not expected. The following is a table of autocorrelation test results:

Table 6: Autocorrelation Test

dL	dU	4-dU	4-dL	DW
1.61	1.74	2.26	2.39	1.740

Based on Table 6 it is known that the Durbin Watson (DW) value is 1,740. In the table, we see DW to obtain dL value of 1.61 and dU of 1.74. So, that in the regression equation, the DW value is in the dU <d <4-dU region. Then H_0 is accepted meaning that the DW value is in the criteria of no autocorrelation. Thus, the assumptions on autocorrelation in the regression equation model have been fulfilled.

We also used Breusch-Godfrey or often called the LM test. In order to detect the presence of autocorrelation, the following are things that can be done:

1. Pay attention to the t-statistic value, R^2, F test, and Durbin Watson (DW) statistics.
2. Perform LM test (Breusch Godfrey method). This method is based on the values of F and Obs*R-squared, where if the probability value of Obs*R-squared exceeds the level of confidence, then H_0 is accepted. This means that there is no autocorrelation problem.

   Testing the autocorrelation hypothesis:

   a. H_0: autocorrelation does not occur
   b. H_a: autocorrelation occurs
3. If the p-value Obs*R-square < α, H_0 is rejected.

Following are the results of autocorrelation testing with the Breusch-Godfrey test:

Table 7: Breusch Godfrey Test

F-Statistic	0.316	Probability	0.726
Obs*R-Squared	0.835	Probability	0.657

Based on the results of calculations using the Breusch-Godfrey test, Obtained probability value Obs*R-square is equal to 0.657. This matter means probability > α = 0.01, then the conclusion is the level of confidence 99% of the regression models are free from autocorrelation problems.

4.3. Hypothesis Testing

The data has passed the prerequisite test, which means it is time to test the initial assumption / hypothesis. The test is presented in Tables 8 and 9. If the p-value is significant below 0.05 we find that the variable has a significant effect. After having an effect, we also see whether the effect is positive or negative. The direction is seen from the value of the beta coefficient, whether positive or negative. If positive means positive and vice versa. When referring to tables 8 and 9 we can conclude that hypotheses 1, 2 and 3 are all accepted, have a significant effect and are in a positive direction, this result supported previous result in preliminary study by [23].

Table 8: Hypothesis Testing 1

No.	Variable	Unstandardized B	t	Sig.
1.	PU	-0.319	3.439	0.001
2.	PEU	0.977	10.458	0.000
Dependent Variable: SU				

Table 9: Hypothesis Testing 2

No.	Variable	Unstandardized B	t	Sig.
1.	PEU	0.689	9.285	0.000
Dependent Variable: PU				

H4 can't be accepted and H5 and H6 can't be rejected, this resulted in line with [24], in other word not supported [16] and [25].

Table 10: Hypothesis Testing 3

No	Variable	Unstandardized B	t	Sig.
1.	OF	0.009	0.135	0.893
2.	SF	0.297	2.074	0.041

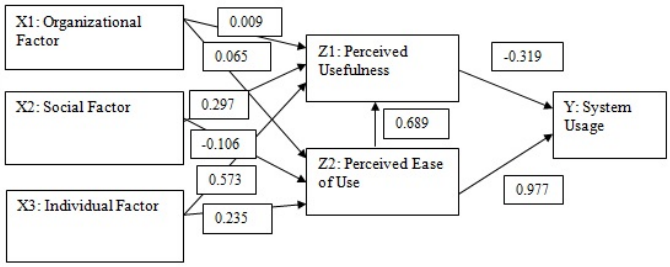
3.	IF	0.573	7.219	0.000
Dependent Variable: PU				

While H7 is can' be accepted, H8 also can't be accepted and H9 positively influenced. We have result that support [26],[25], [27].

Table 11: *Hypothesis Testing 4*

No.	Variable	Unstandardized B	t	Sig.
1.	OF	0.065	0.830	0.408
2.	SF	-0.106	-0.599	0.551
3.	IF	0.235	2.398	0.018
Dependent Variable: PEU				

The research path coefficient is presented below.



5. Conclusion and Suggestion

After getting the test results statistically, we followed up. The follow up that we do is to conduct interviews and observations of respondents. Our results have found that audit judgment expectancy and expectancy incentives can influence auditors' interest in using GAS. This is because the auditor's performance will increase, when performance increases, then incentives such as salary increases, benefits, bonuses, which are material and also praise/acknowledgement, non-material recognition will accompany their careers. Defers case with age similarity influence; It is true that auditors in their daily life prefer to associate with the same age. However, this association only affects behavior outside of work obligations. For example, it only affects hobbies, games and favorites foods. Meanwhile, to use audit software, it still has to be mandatory from the lead leader.

While other results state that the availability of software and the desire to adopt influence people to actually use it. This is reasonable, because how is it possible for an auditor to use, if the software is not provided by the relevant accounting firm where he works. It is impossible for the auditor to install himself on a personal laptop and use it without the company's approval. For the influence between intention and behavior clearly follows psychological rules, that someone who already wants, will usually continue to use.

Future studies can examine the actual process, on internal auditors and small accounting firms outside the big four and big ten. Future studies can test the extent to which the ability of the software. Future studies examine affect partners or owners can increase utilization by providing facilities and infrastructure as well as conducting training and socialization of its use.

References

[1] B. L. Handoko, Meiryani, S. Sabrina, and N. Ayuanda, "Admission of Information Technology in External Audit Profession: Impact of Organizational, Social and Individual Factors," in *Proceedings of 2019 International Conference on Information Management and Technology, ICIMTech 2019*, 2019, pp. 36–41.

[2] N. Mahzan and A. Lymer, "Examining the adoption of computer-assisted audit tools and techniques," *Manag. Audit. J.*, vol. 29, no. 4, pp. 327–349, 2014.

[3] J. J. Schultz, J. L. Bierstaker, and E. O'Donnell, "Integrating business risk into auditor judgment about the risk of material misstatement: The influence of a strategic-systems-audit approach," *Accounting, Organ. Soc.*, 2010.

[4] S. M. Glover, M. H. Taylor, and C. Western, "Mind the Gap: Why Do Experts Have Differences of Opinion Regarding the Sufficiency of Audit Evidence Supporting Complex Fair Value Measurements?," *Contemp. Account. Res.*, 2019.

[5] N. K.-G. J. of and U. 2017, "'Attitudes and Perceptions Towards Incorporating Computer Assisted Audit Techniques in an Undergraduate Auditing …," *Theibfr.Com*, vol. 11, no. 3, pp. 55–71, 2017.

[6] M. Mustapha and S. J. Lai, "Information Technology in Audit Processes: An Empirical Evidence from Malaysian Audit Firms," *Int. Rev. Manag. Mark.*, vol. 7, no. 2, pp. 53–59, 2017.

[7] F. D. Davis, "Perceived Usefulness , Perceived Ease Of Use , And User Acceptance," *MIS Q.*, vol. 13, no. 3, pp. 319–339, 1989.

[8] M. Fishbein and I. Ajzen, "Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research," *Read. MA AddisonWesley*, no. August, p. 480, 1975.

[9] V. Venkatesh, F. D. Davis, and S. M. W. College, "Theoretical Acceptance Extension Model : Field Four Studies of the Technology Longitudinal," vol. 46, no. 2, pp. 186–204, 2012.

[10] A. Rogers, "Examining Small Business Adoption of Computerized Accounting Systems Using the Technology Acceptance Model.," *Walden Diss. Dr. Stud.*, p. 126, 2016.

[11] Y. L. Chen, H. T. Chih, and C. C. Wan, "The relationship between attitude toward using and customer satisfaction with mobile application services: An empirical study from the life insurance industry," *J. Enterp. Inf. Manag.*, vol. 28, no. 5, pp. 680–697, 2015.

[12] H.-J. Kim, A. Kotb, and M. K. Eldaly, "The use of generalized audit software by Egyptian external auditors," *J. Appl. Account. Res.*, vol. 17, no. 4, pp. 456–478, 2016.

[13] R. Widuri, B. L. Handoko, and I. E. Riantono, "Perception of Accounting Student on Learning of Generalized Audit Software," in *Proceedings of 2019 International Conference on Information Management and Technology, ICIMTech 2019*, 2019.

[14] H. J. Kim, A. Kotb, and M. K. Eldaly, "The use of generalized audit software by Egyptian external auditors: The effect of audit software features," *J. Appl. Account. Res.*, 2016.

[15] E. Symeonaki, M. Papoutsidakis, D. Tseles, and M. Sigala, "Post-Implementation Evaluation of a University Management Information System (UMIS)," in *Third International Conference on Mathematics and Computers in Sciences and in Industry (MCSI)*, 2016, pp. 14–19.

[16] K. Rosli, P. H. P. Yeow, and E.-G. Siew, "Adoption of Audit Technology in Audit Firms," *Proc. 24th Australas. Conf. Inf. Syst.*, pp. 1–12, 2013.

[17] A. S. Shammari, "An Examination Factors Influencing The Intention To Use Instructional Technologies: An Extention Technlogy Acceptance Model (TAM)," *Int. J. Inf. Res. Rev.*, vol. 04, no. 02, pp. 3637–3641, 2017.

[18] A. Ahmi and S. Kent, "The utilisation of generalized audit software (GAS) by external auditors," *Manag. Audit. J.*, vol. 28, no. 2, pp. 88–113, 2013.

[19] J. B. Chassan, *Research Design in Clinical Psychology and Psychiatry*. New York: Irvington Publishers Inc, 1979.

[20] J. B. Chassan, "Intensive design in medical research," *Pharmacol. Ther. Part B Gen. Syst.*, vol. 1, no. 1, pp. 139–148, 1975.

[21] H. J. Kim, A. Kotb, and M. K. Eldaly, "The use of generalized audit software by Egyptian external auditors: The effect of audit software features," *J. Appl. Account. Res.*, vol. 17, no. 4, pp. 456–478, 2016.

[22] U. Sekaran and R. Bougie, "Research Methods For Business. A Skill Builing Approch. 7th Edition," *Book*, 2016.

[23] A. Tarhini, K. Hone, and X. Liu, "A cross-cultural examination of the impact of social, organisational and individual factors on educational technology

acceptance between British and Lebanese university students," *Br. J. Educ. Technol.*, vol. 46, no. 4, pp. 739–755, 2015.

[24] M. Sharif Abbasi, A. Tarhini, M. Hassouna, and F. Shah, "Social, Organizational, Demography and Individuals' Technology Acceptance Behaviour: a Conceptual Model," *Eur. Sci. J.*, vol. 11, no. 9, pp. 1857–7881, 2015.

[25] K. Rosli, P. Yeow, and E.-G. Siew, "Factors Influencing Audit Technology Acceptance by Audit Firms: A New I-TOE Adoption Framework," *J. Account. Audit. Res. Pract.*, vol. 2012, pp. 1–11, 2012.

[26] C. Chiu, S. Chen, and C. Chen, "An Integrated Perspective of TOE Framework and Innovation Diffusion in Broadband Mobile Applications Adoption by Enterprises," *Int. J. Manag. Econ. Soc. Sci.*, vol. 6, no. 1, pp. 14–39, 2017.

[27] R. Widuri, B. L. Handoko, and I. C. Prabowo, "Adoption of Information Technology in Public Accounting Firm," 2019.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 63-68 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Fast Stream Cipher based Chaos Neural Network for Data Security in CAN Bus

Zhongda Liu[*,1], Takeshi Murakami[2], Satoshi Kawamura[3], Hitoaki Yoshida[4]

[1]*Faculty of Science and Engineering, Ishinomaki Senshu University, Ishinomaki, 986-8580, Japan*

[2]*Technical Division, Iwate University, Morioka, 020-8550, Japan*

[3]*Faculty of Humanities, Morioka University, Takizawa,020-8550, Japan*

[4]*Faculty of Education, Iwate University, Morioka, 020-8550, Japan*

ARTICLE INFO

Article history:
Received: 02 July, 2020
Accepted: 24 August, 2020
Online: 09 September, 2020

Keywords:
Stream cipher
CAN
Chaos
Encryption and decryption

ABSTRACT

Vehicle systems are controlled by embedded electronic devices called electronic control units (ECUs). These ECUs are connected together with network protocols. The Controller Area Network (CAN) protocol is widely implemented due to its high fault tolerance. However, the CAN is a serial broadcast bus, and it has no protection against security threats. In this paper, we propose a fast stream cipher based on a chaos neural network (CNN) that is able to generate pseudo-random numbers at a high speed, faster than that of the Advanced Encryption Standard, to protect ECUs on the CAN bus by encrypting CAN messages. We discuss the chaotic orbit of the CNN and statistical properties of pseudo-random numbers from the CNN. For a stream cipher, it is very important to share the symmetric key. We designed a symmetric key that is shared with ID-based encryption without the need to use digital certificates. We evaluated our method's performance with embedded boards and showed that the stream cipher is efficient for the embedded software of the ECU. Further, it does not need a hardware security module to accelerate the encryption.

1 Introduction

This paper is an extension of work originally presented at the IEEE 10th International Conference on Awareness Science and Technology [1]. In that work, we found that a chaos neural network (CNN) is able to generate pseudo-random numbers (PRNs) at high speed, 49% faster than that produced with the Advanced Encryption Standard (AES) [2], [3], and it can be easily implemented even for embedded devices.

Generally, electronic devices embedded in vehicles to control vehicle systems are called electronic control units (ECUs). A modern vehicle is usually equipped with more than 70 ECUs [4]. To share information and control the subsystems, those ECUs are connected together with network protocols, such as a Controller Area Network (CAN), Local Interconnect Network (LIN), Media Oriented Systems Transport (MOST), or FlexRay.

The CAN is a broadcast serial communications bus that is widely introduced because of its fault tolerance. The CAN identifier (ID) (see Sec. 2 and Fig. 1) is used for prioritizing messages on the bus and avoids collisions by design. However, security issues were ignored during designing since people took it for granted that a vehicle would be a closed system [5], [6]. Unfortunately, messages are broadcast on the CAN bus, and external devices, such as on-board diagnostics readers, are able to access the CAN bus in modern vehicles.

A pseudo-random number generation (PRNG) is crucial to a stream cipher in information security field. We have reported various PRNG methods [7]–[11] and the property of PRNs from a CNN [9] has been confirmed [10] by National Institute of Standards and Technology (NIST) Special Publication 800-22 [12]. An ultra-long period PRNs that has reached 10^{22432} [11] can be generated with the chaotic time series from the CNNs. The chaotic time series is hard to predict because it is sensitive to tiny change of the initial status.

In this paper, we propose a fast stream cipher based on a CNN to

[*]Corresponding Author: Zhongda Liu, Faculty of Science and Engineering, Ishinomaki Senshu University, Ishinomaki, 986-8580, Japan
Email: liuzd@isenshu-u.ac.jp

protect CAN messages by encrypting them. The remainder of this paper is organized as follows: Section 2 introduces CAN and security issues and surveys some related work. Section 3 describes our CNN and discusses some of its characteristics. Section 4 presents the CNN stream cipher, including sharing of the symmetric key and the procedure for encryption and decryption of the stream cipher. A performance evaluation of the proposed CNN stream cipher is given in Section 5. Finally, Section 6 concludes this paper.

2 Related Work

A CAN is a serial communications bus defined by the International Organization for Standardization (ISO) and originally developed for the automotive industry to replace the complex wiring harness with a two-wire bus [13]. Balanced differential signaling reduces noise coupling and enables high noise immunity in the CAN bus.

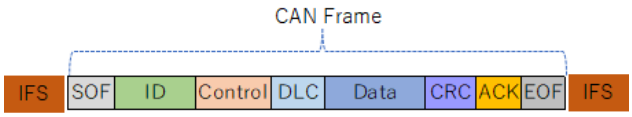


Figure 1: Structure of the CAN data frame

The CAN communication protocol is a carrier-sense multiple access protocol with collision detection and arbitration on message priority. A CAN message contains a unique ID field that represents the priority and function of the message. The CAN protocol supports four different message types: overload, error, remote, and data frame. The CAN data frame begins with a start-of-frame (SOF) bit and is followed by the ID, a control field (6 bits), 4-bit data length code (DLC), 0-64 bits of data, a cyclic redundancy check (CRC) sequence (15 bits), a 2-bit acknowledgment (ACK), and a 7-bit end of frame (EOF) sequence that marks the end of the frame. Between CAN frames, a 7-bit inter-frame space (IFS) is required by the CAN controller to provide time for moving a received frame to a message buffer area (see Fig. 1).

The CAN was subsequently adopted as ISO standards. ISO 11519 (low-speed CAN) is for applications up to 125 kbps with a standard 11-bit ID, while ISO 11898 (high-speed CAN) provides for signaling rates from 125 kbps to 1 Mbps. Furthermore, high-speed CAN supports two data frame formats, where the standard frame consists of an 11-bit ID, while the extended format contains a 29-bit ID.

Unfortunately, security issues were ignored during designing because people took it for granted that CANs would be a closed system in automobiles [5], [6]. Security issues with CANs relate mainly to authentication and encryption at the present time.

Authentication: To identify whether an ECU is authorized, several authentication proposals based on message authentication codes (MACs) have been released. Key sharing is a matter of grave concern. CANAuth [14] implements a backward-compatible message authentication protocol on the CAN bus. One or more pre-shared 128-bit MAC keys are to be available on each CANAuth node. CANAuth assumes that the keys are intended to be stored in tamper-proof storage that cannot be queried by anything but the node itself. LiBrA-CAN [15] splits authentication keys between groups of multiple nodes, rather than achieving authentication independently for each node.

Encryption: A CAN frame is broadcast over the bus. In modern vehicles, external devices, such as on-board diagnostics readers, are able to access the CAN bus, making it is easy to intercept a CAN message. Cryptographic approaches based on the AES have been proposed to guard against such interception. The problem is the computation load of the AES, which might have an undue influence on the response of the ECU. Wolf and Gendrullis [16] and the EVITA Project [17], [18] implemented a hardware security module (HSM) to accelerate the AES measures. However, even if a HSM is used, the cryptographic measure requires 60 clock cycles (at 100 MHz) for the encryption of one AES block [18]. This is insufficient for dealing with the real-time response required of an ECU. Also, the additional hardware increases the cost of the ECU.

In this study, we focused on the encryption issues in CANs. We propose a fast stream cipher based on a CNN that does not need the additional HSM hardware.

3 Chaos Neural Network

As a chaos generator the CNN consisted of 4 neurons in a discrete time system (see Fig. 2).

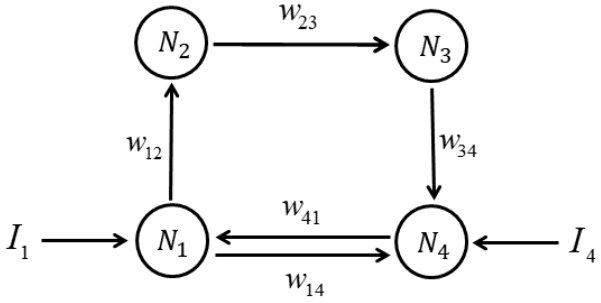


Figure 2: CNN consisting of four artificial neurons

An output from the jth neuron at time $t+1$ is defined as:

$$x_j(t+1) = f[u_j(t)] \tag{1}$$

Here, An activation-function f (see Fig. 3) is an asymmetric piecewise-linear function (APLF).

For the jth neuron, the total value of inputs at time t is defined as :

$$u_j(t) = \sum_{i=1}^{n} w_{ij} x_i(t) + I_j \tag{2}$$

I_j is an external input of the jth neuron. $x_i(t)$ is an input from the ith neuron at time t, and w_{ij} is a synaptic. Generally, the start value of x_i is set as 0, and the synaptic weights are set as follows: $w_{12} = -12.60001$, $w_{14} = 4.511$, $w_{23} = 5.951$, $w_{34} = -4.7004$ and $w_{41} = -7.345007$. The synaptic weights adjust the input values

from other neurons. If extreme synaptic weights were set, the output range of neurons would be limited [19]. The external inputs I_1 and I_4 share a common value ($I_1 = I_4$), and I_2 and I_3 are set as 0 ($I_2 = I_3 = 0$). Thus, a different CNN would be obtained if a different value for I_1 and I_4 were set.

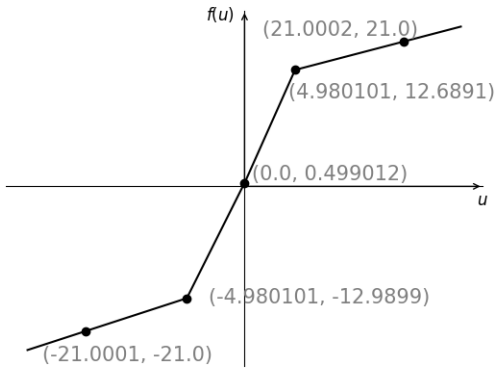


Figure 3: An activation-function f (APLF)

An activation-function APLF can avoid a periodic window corresponding to a non-chaotic periodic orbit. The activation-function APLF composed of linear functions by connecting five points. Those points can be changed as independent parameters. In a cipher system, the points of APLF can be selected as secret keys [9], [10], [20].

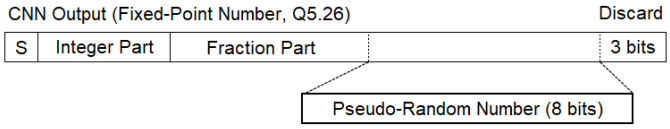


Figure 4: The extraction method of a PRN from a CNN output

Generally, discrete time system of the CNN is implemented with floating-point arithmetic [9]. But many embedded devices do not support 64-bit floating-point arithmetic. In this paper, the CNN is computed by 32-bit fixed-point arithmetic (Q5.26) and it allows overflow and underflow of variables. Comparing to 32-bit floating-point arithmetic, the fractional part of Q5.26 has enough long bit length. PRNs are extracted by the method presented in Figure 4. With regard to the CNN output, the lowest 3 bits of the fraction are discarded [8] and the lower 8 bits of the fraction are extracted as a PRN. Those PRNs from the CNN are applied to the proposed stream cipher .

3.1 Chaotic Orbit

A chaotic orbit is hard to predict because it is sensitive to tiny changes of the initial status. Here, Figure 5 shows time series from a CNN (Q5.26). Corresponding to all external inputs, output time series in the diagram present no bifurcation pattern but chaotic characteristics. It suggests that all external inputs can be used for chaos generation. In fact, the CNN generates the same time series on the ARM CPU and X86 CPU when the same parameters are set. Therefore the CNN is portable between different machines. Moreover, the Lyapunov exponents λ are computed per time series. The maximum Lyapunov exponents is about 2.5 and all of value is $\lambda > 0$. The Kolmogorov-Sinai entropy [21] is also computed by use of Lyapunov exponents, it is about 4.2. Those results demonstrate that the time series from the CNN has chaotic orbit and a high degree of randomness.

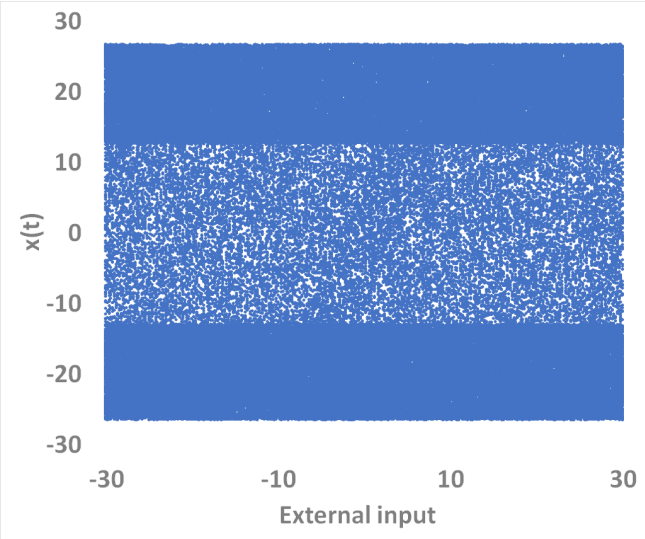


Figure 5: The input-output characteristics of time series from a CNN

3.2 Randomness

Randomness of the PRNs that were extracted from the CNN was confirmed by NIST Special Publication 800-22 statistical test suite [12]. Since NIST test has a couple of trouble (asymptotic approximation, etc.) even the test suite is updated [22], [23], NIST test method presented in the literature [24] was adopted.

We performed the NIST test for 1,000 times. And $10^6 \times 1,000$ bits of PRNs were generated by the method shown in Fig. 4 for per test. NIST test results are presented in Table 1. The failure ratio for the proportion is under 1%, and the failure ratio for the P-values that check for uniformity of distribution is less than 0.1%. All of those results suggest that all of tests passed these criteria, and the tested PRNs from the CNN have good statistical properties.

4 CNN Stream Cipher

The proposed stream cipher has two phases: an ID-based encryption (IBE) phase and a stream phase (see Fig. 6). Each phase uses different CAN IDs; that is, the CAN ID associates a CAN frame with a specific phase. In the advance IBE phase, the symmetric key is shared with IBE [25], [26] among authorized ECUs. Subsequently, an authorized ECU sends encrypted data frames to other authorized ECUs and those ECUs can decrypt data using the symmetric key in the stream phase.

Table 1: NIST SP800-22 statistical test results

	FR[a]	BF[a]	CS[a]	RU[a]	LR[a]	RK[a]	FF[a]	NT[a]
Proportion	0.3	0.2	0.3	0.5	0.4	0.4	0.9	0.2
P-value	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0
	OT[a]	UN[a]	AE[a]	RE[a]	RV[a]	SE[a]	LC[a]	
Proportion	0.3	0.7	0.2	0.4	0.2	0.2	0.1	
P-value	0.0	0.0	0.0	0.0	0.0	0.0	0.1	

FR: Frequency, BF: Block Frequency, CS: Cumulative Sums, RU: Runs, LR: Longest Run, RK: Rank, FF: FFT, NT: Non-overlapping Template, OT: Overlapping Template, UN: Universal, AE: Approximate Entropy, RE: Random Excursions, RV: Random Excursions Variant, SE: Serial, LC: Linear Complexity

[a] Numbers are average ratio of failed tests (%).

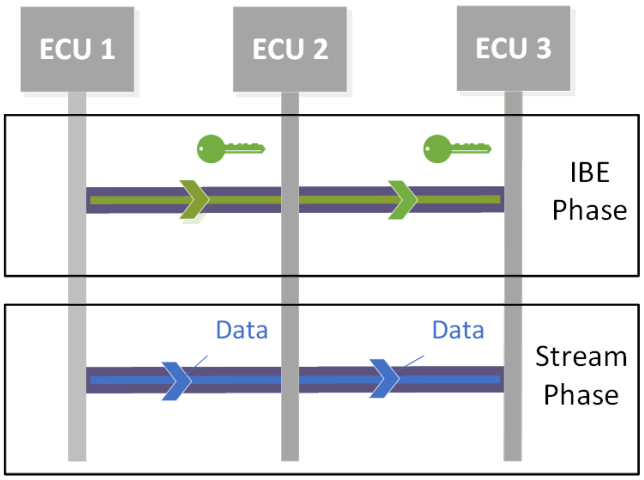


Figure 6: Overview of CNN stream cipher

4.1 IBE phase

It is an important step for the stream cipher to create, manage, and share the symmetric key. A public key infrastructure is used in the Internet to ensure secure communication generally. With this infrastructure, an on-line certificate authority (CA) is necessary, and the cost for issuing digital certificates may become prohibitive [27].

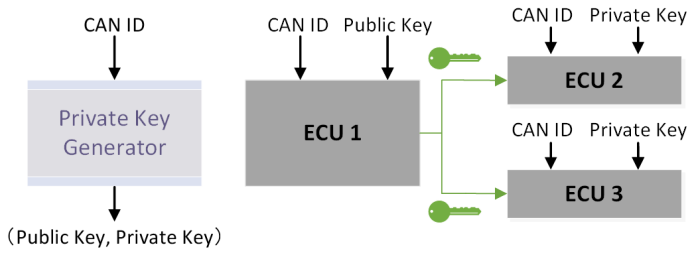


Figure 7: IBE phase

In the proposed stream cipher, we use IBE [26] to create, manage, and share the symmetric key. Fig. 7 shows how the symmetric key is shared among authorized ECUs. The private key generator (PKG) can be offline and does not need to use digital certificates instead of a CA. The PKG initially defines which CAN IDs are used in the IBE phase. For each valid CAN ID, the PKG outputs a public and private key pair that is issued to authorized ECUs. Thus, an ECU can use the public key to encrypt the symmetric key that is used in the stream phase and send it over the CAN bus. When an ECU receives a CAN frame, it checks the CAN ID to confirm the phase and decrypts the data in the CAN frame to obtain the symmetric key in the IBE phase.

4.2 Stream phase

In the stream phase, authorized ECUs use the symmetric key that was obtained during the IBE phase to encrypt and decrypt CAN frames (see Fig. 8). The CNN implemented in authorized ECUs generates a stream of pseudo-random bits: R_1, R_2, R_3, ..., R_i with the symmetric key. This stream is XORed with a stream of bits, D_1, D_2, D_3, ..., D_i, which are from the data in a CAN frame, to produce the stream of cipher text bits. Then each cipher text character is given by $C_i = D_i \oplus R_i$, which is loaded into a CAN frame and translated with the CAN bus. The procedure of decryption is almost the same: when an authorized ECU has received a CAN data frame, the CNN in the ECU generates the same stream of pseudo-random bits R_i and the original data is obtained by $D_i = C_i \oplus R_i$.

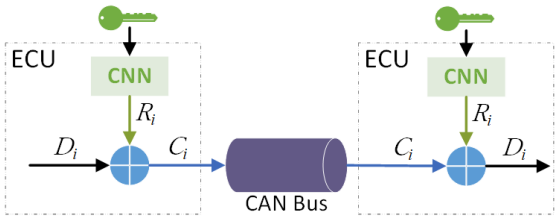


Figure 8: Stream phase

5 Evaluation

It is most important to ensure the safety of the vehicle and its passengers. Therefore, the embedded software of the ECU must run quickly to deal with the constraints of a real-time response. This section describes the performance evaluation of the CNN stream cipher with two embedded CAN boards (listed as Board A and B in Table 2) that were provided by P&A Technologies Inc. These CPUs have a different architecture, where Board A was implemented with a SH2A CPU, while Board B used an ARM CPU. Those boards are connected by a length of about 80*cm* twisted pair cable with D-sub connector.

Table 2: Specifications of experimental CAN boards

Component	Board A	Board B
CPU	R5S72630P200FP SH2A-FPU core (196 MHz)	SAMA5D27C-D1G ARM Cortex-A5 (492 MHz)
RAM	SDR SDRAM (64 MHz)	DDR2-SDRAM (120 MHz)
CAN Controller	Built-in CPU	FPGA IP

5.1 Experimental setup

In our experiments, we tested only the high-speed CAN whose bit rate is typically 500 Kbps, up to 1 Mbps. In fact, another CAN standard specifies low-speed CAN (see Sec. 2) at transmission rates above 40 Kbps up to 125 Kbps. It is more difficult to deal with real-time constraints at the high-speed CAN bit rate. Thus, we assumed that our stream cipher would work well at the low-speed CAN bit rate if it successfully ran with the high-speed CAN.

The PKG can be performed offline. Thus, we assume that Boards A and B are two authorized ECUs and they have already gained the symmetric key. Then we implemented the CNN on Boards A and B. According to the symmetric key, the same stream of pseudo-random numbers was generated in both boards and used to encrypt the data part of a CAN frame in one board and decrypt it in the other board.

5.2 Experimental Results

One thousand CAN message frames were sent between Boards A and B to confirm the validity of the CNN stream cipher and measure the encryption and decryption time. We tested with 500-Kbps and 1-Mbps bit rates. The CAN bus was loaded with over 60% higher-priority traffic.

Table 3: Results of performance testing with CAN boards

	average time
Board A	44 μs
Board B	4 μs

We confirmed the CAN log data of Boards A and B, which showed that each board encrypted and decrypted CAN data frames successfully. With Board A, the procedure for encryption or decryption was performed within 44 μs on average. With Board B, the procedure only took 4 μs on average (see Table 3). The results suggest that the performance of the CNN stream cipher is adequate for real-time requirements of an ECU without additional HSM hardware.

6 Conclusions

In this paper, we have proposed and evaluated a fast stream cipher based on a CNN to provide security for the ECUs on a CAN bus. We have shown that the CNN is chaotic and have strong randomness, and that PRNs with a high degree of randomness can be generated from a CNN. In the proposed stream cipher, IBE is used to create, manage, and share the symmetric key. The PKG can be performed offline and does not need to use digital certificates. The stream cipher was evaluated with embedded CAN boards. The performance test results suggested that our method is efficient for software embedded in an ECU and has no need for a HSM to accelerate the encryption process.

As future work, we will design a new activation-function APLF to extend randomness of the CNN and improve the performance of the stream cipher based on the CNN.

References

[1] Z. Liu, T. Murakami, S. Kawamura, H. Yoshida, "Parallel Implementation of Chaos Neural Networks for an Embedded GPU," in 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), 1–6, IEEE, 2019, doi:10.1109/ICAwST.2019.8923383.

[2] NIST FIPS PUB, "197, Advanced Encryption Standard (AES)," Federal information processing standards publication, **197**(441), 0311, 2001, doi:10.6028/NIST.FIPS.197.

[3] J. Daemen, V. Rijmen, Rijndael/AES, 520–524, Springer US, 2005, doi:10.1007/0-387-23483-7_358.

[4] A. Albert, et al., "Comparison of event-triggered and time-triggered concepts with regard to distributed control systems," Embedded world, **2004**, 235–252, 2004.

[5] P. Carsten, T. R. Andel, M. Yampolskiy, J. T. McDonald, S. Russ, "A system to recognize intruders in controller area network (CAN)," in 3rd International Symposium for ICS & SCADA Cyber Security Research 2015 (ICS-CSR 2015) 3, 111–114, 2015, doi:10.14236/ewic/ICS2015.15.

[6] R. Buttigieg, M. Farrugia, C. Meli, "Security issues in controller area networks in automobiles," in 2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), 93–98, IEEE, 2017, doi:10.1109/STA.2017.8314877.

[7] H. Yoshida, Y. Nihei, T. Nakanishi, "Comparative study on structurally different chaos neural network," in Proceedings of Papers, International Symposium on Information Theory and its Applications, ISITA 2004, 1046–1050, 2004.

[8] S. Kawamura, H. Yoshida, M. Miura, M. Abe, "Implementation of Uniform Pseudo Random Number Generator and Application to Stream Cipher based on Chaos Neural Network," in The International Conference on Fundamentals of Electronics, Communications and Computer Sciences, 2002, 4–9, 2002.

[9] H. Yoshida, T. Murakami, Z. Liu, "High-speed and highly secure pseudo-random number generator based on chaos neural network," New Trends on System Science and Engineering: Proceedings of ICSSE, **276**, 224–237, 2015, doi:10.3233/978-1-61499-522-7-224.

[10] H. Yoshida, T. Murakami, T. Inao, S. Kawamura, "Origin of Randomness on Chaos Neural Network," Trends in Applied Knowledge-Based Systems and Data Science, **9799**, 587–598, 2016, doi:10.1007/978-3-319-42007-3_51.

[11] H. Yoshida, Y. Akatsuka, T. Murakami, "Implementation of High-Performance Pseudo-Random Number Generator by Choas Neural Networks using Fix-Point Arithmetic with Perturbation," in Proceedings of Papers, The 2018 International Symposium on Nonlinear Theory and Its Applications, NOLTA2018, 46–49, 2018.

[12] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, editors, A statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST Special Publication 800-22, 2001.

[13] S. C. HPL, "Introduction to the controller area network (CAN)," Appl. Rep. SLOA101, 1–17, 2002.

[14] A. Van Herrewege, D. Singelee, I. Verbauwhede, "CANAuth - a simple, backward compatible broadcast authentication protocol for CAN bus," in ECRYPT Workshop on Lightweight Cryptography, volume 2011, 2011.

[15] B. Groza, S. Murvay, A. Van Herrewege, I. Verbauwhede, "LiBrA-CAN: a lightweight broadcast authentication protocol for controller area networks," in International Conference on Cryptology and Network Security, 185–200, Springer, 2012, doi:10.1007/978-3-642-35404-5_15.

[16] M. Wolf, T. Gendrullis, "Design, implementation, and evaluation of a vehicular hardware security module," in International Conference on Information Security and Cryptology, 302–318, Springer, 2011, doi:10.1007/978-3-642-31912-9_20.

[17] O. Henniger, A. Ruddle, H. Seudié, B. Weyl, M. Wolf, T. Wollinger, "Securing vehicular on-board IT systems: The evita project," in VDI/VW Automotive Security Conference, 41, 2009.

[18] H. Schweppe, Y. Roudier, B. Weyl, L. Apvrille, D. Scheuermann, "Car2x communication: securing the last meter-a cost-effective approach for ensuring trust in car2x applications using in-vehicle symmetric cryptography," in 2011 IEEE Vehicular Technology Conference (VTC Fall), 1–5, IEEE, 2011, doi:10.1109/VETECF.2011.6093081.

[19] H. Yoshida, H. Fukuchi, T. Murakami, "Implementation of High-Speed Pseudo-Random-Number Generator with Chaotic and Random Neural Networks," in Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020, doi:10.24251/HICSS.2020.786.

[20] H. Yoshida, T. Murakami, Japan patent JP5504501B, 2014.

[21] T. S. Parker, L. O. Chua, "Chaos: A tutorial for engineers," Proceedings of the IEEE, **75**(8), 982–1008, 1987, doi:10.1109/PROC.1987.13845.

[22] K. Hamano, T. Kaneko, "Correction of overlapping template matching test included in NIST randomness test suite," IEICE transactions on fundamentals of electronics, communications and computer sciences, **90**(9), 1788–1792, 2007, doi:10.1093/ietfec/e90-a.9.1788.

[23] H. Okutomi, K. Nakamura, "A study on rational judgement method of randomness property using NIST randomness test (NIST SP. 800-22)," IEICE Trans. A, **93**(1), 11–22, 2010.

[24] H. Yoshida, T. Murakami, S. Kawamura, "Study on testing for randomness of pseudo-random number sequence with NIST SP800-22 rev. 1a," Technical report, IEICE Technical Report, 2012.

[25] A. Shamir, "Identity-based cryptosystems and signature schemes," in Workshop on the theory and application of cryptographic techniques, 47–53, Springer, 1984, doi:10.1007/3-540-39568-7_5.

[26] D. Boneh, M. Franklin, "Identity-based encryption from the Weil pairing," in Annual international cryptology conference, 213–229, Springer, 2001, doi:10.1007/3-540-44647-8_13.

[27] A. Nash, W. Duane, C. Joseph, PKI: Implementing and Managing E-security, McGraw-Hill, Inc., 2001.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 69-73 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Low Power Bulk Driven Series Parallel OTA for Low Frequency Applications

Sushma Padubidri Shivaprasad[1,*], Sreemannarayanay Kulkarni[2]

[1]*Department of Electronics and Communication Engineering, NMAM Institute of Technology, Nitte, 574110, India*

[2]*School of Electronics and Communication Engineering, Reva University, Bangalore, 560064, India*

ARTICLE INFO

Article history:
Received: 14 June, 2020
Accepted: 16 August, 2020
Online: 09 September, 2020

Keywords:
Low power OTA
Low cutoff frequency
Multifunction filter
Bulk driven
Series parallel current mirror
Biquad

ABSTRACT

Low power OTAs are the most preferred circuits in the realization of continuous time filters of analog front end of wearable healthcare devices. A low transconductance OTA with series parallel current mirror to realize large time constant of the filter is designed. The differential pair of the OTA uses bulk driven PMOSFETs and the subthreshold operation of the circuit achieves 44 nW power with supply voltage of ±0.4 V. The designed OTA has DC gain of 29.59 dB and UGBW of 34.28 KHz. Using the proposed OTA, a multifunction filter which can operate as low pass and high pass filter, having cut-off frequency in the range 25 Hz - 225 Hz is designed in gpdk 180 nm CMOS technology. The simulation is performed using Cadence virtuoso design environment.

1. Introduction

Energy efficient wearable biomedical devices are most widely used in modern healthcare services. In such systems, the design of analog front end (AFE) circuits using existing CMOS technology has multiple design constraints having tradeoffs. As the AFE circuits need to remain on all the time to sense the bioelectric signals continuously, they consume largest share of the overall power budget. So, the energy efficiency of these circuits is major concern along with other performance parameters.

As shown in Figure 1, the analog front end (AFE) consists of low noise amplifier followed by continuous time active filters. Using the sensors or electrodes, biomedical signals such as ECG, EEG, EMG etc. are captured. These signals are having amplitude in the range of μV to few mV and frequencies in the range of dc to few Hz. Therefore, these circuits have to be designed with good low frequency performance. As the signal from the electrodes has dc component, large off chip capacitors are used. However, use of subhertz cutoff frequency high pass filter eliminates these large capacitors. The low pass and band pass filters are used to select the required signal.

The analog filters were designed using opamps. The bandwidth of opamp-RC filters is limited by the excessive gain-bandwidth product requirement for the opamp, to obtain ideal filter response. The high gain-bandwidth product of opamp based topologies consume more power. Therefore, in recent designs, Operational Transconductance Amplifier (OTA) are used in place of opamps. OTAs are also used as the low noise amplifier (LNA), where they need to have high gain and less noise. Therefore, OTA is an important circuit in the analog front end (AFE) of low power systems. OTA is a voltage controlled current source and it converts differential input voltage to output current. The circuit symbol for OTA is shown in Figure 2.

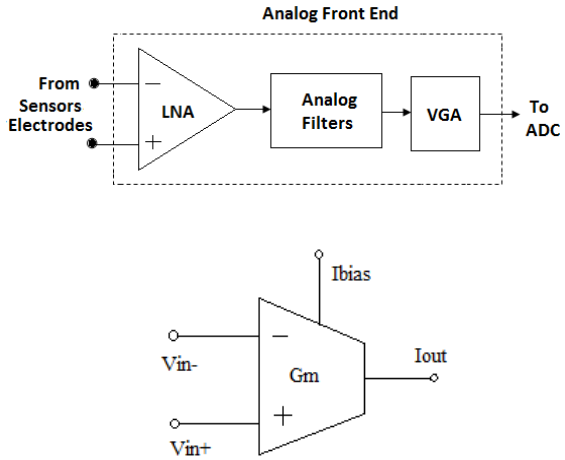


Figure 2: Circuit symbol for OTA

*Corresponding Author: Sushma Padubidri Shivaprasad, pssushma@nitte.edu.in

Figure 3 shows the circuit diagram of single ended current mirror OTA with bulk driven PMOS differential pair, where all nodes have either drain-gate connected or source connected MOSFETs except for input devices. The diffential input (V_{in+} - V_{in-}) is applied to the bulk terminal of PMOSFET pair M1 and M2 that results in output current given by (1). The transconductance of the OTA is constant for the specifies range of differential input and is controlled by bias current I_{bias}.

$$I_{out} = G_m \, (V_{in+} - V_{in-}) \tag{1}$$

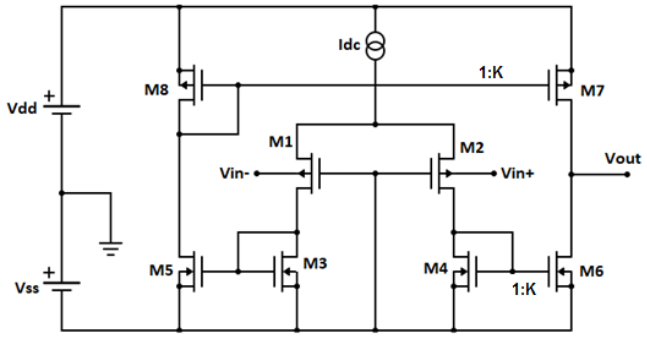


Figure 3: Schematic of current mirror OTA with bulk driven PMOS differential pair

The low voltage operation of OTA is necessary for low power operation. In strong inversion region operating voltage cannot be lowered below the threshold voltage of the MOSFETs. The state of the art digital circuits are operating at 0.3 V to 0.5 V range. To make analog circuits work with same supply voltage range in a chip, they are operated in subthreshold or weak inversion region by choosing appropriate inversion level i_f[1][2]. To lower the operating current level bulk driven MOSFETs are used, resulting in low transconductane and increased linearity. The bulk driven MOSFETs have low cutoff frequency leading to poor frequency response characteristics. The low frequency operation requires large time constants which can be achieved by designing OTAs with small transconductance. In weak inversion gm/Id is high. In order to reduce gm series parallel current mirror [2] is used wherein small bias current and large division factor is used to reduce the output current. Choice of small bias current leads to poor input linear range. Compared to low transconductor design techniques like current division [3], bulk driven [4], attenuator based modular reduction of transconductance [5], current steering [6], source degeneration [7], current cancellation and series parallel current division technique gives lowest transconductance, low offset and better linearity. However, there is trade off with low power.

The proposed series parallel bulk driven OTA (SPOTA) is explained in section 2. The simulation results of the OTA is discussed in section 3. In Section 4, second order multifunction Gm-C filter with tunable cut-off frequency as an application of the proposed SPOTA is explained followed by conclusion in Section 5.

2. Proposed Bulk Driven Series Parallel OTA

In low power analog designs, choice of circuit topology significantly matters for power reduction. In the design of analog filters, OTAs are suitable alternative for opamps. The transconductance of the OTA is an important parameter in the design of active filters. In analog filter applications, the transconductance value is chosen depending on the cutoff frequency range. For low cut off frequency analog filters, large capacitors and resistors are required. As integration of the large resistor is not area efficient, small transconductance cells are designed. In series parallel current division technique, output current is divided by high ratio to achieve very small transconductance [2]. The bulk driven PMOSFET differential pair further reduces overall transconductance.

The OTA shown in Figure 3 is modified to obtain bulk driven series parallel (SP) OTA as shown in Figure 4. The differential input voltage is connected to the substrate of PMOSFETs M1 and M2 and the gates are connected to V_{ss} . Therefore, the transconductance of the OTA depends on g_{mb} , the body conductance which is 0.2 to 0.4 times g_m, the transconductance of the MOSFET. The bulk driven PMOS differential pair at the input with series-parallel current mirror load reduces the transconductance and increases the linearity of the OTA. All devices in the circuit are biased to operate in weak inversion region, by maintaining gate voltage $V_{GS} < V_{th}$, the threshold voltage and drain to source voltages $V_{DS} \geq 4V_T$. The weak inversion drain current I_D under these condition is as given in (2),

$$I_D = 2 \, n \, K \, V_T^2 \, (W / L) \, exp((V_{GS} - V_{th}) / n \, V_T) \tag{2}$$

where $K = \mu_n C_{ox}$, the process dependent parameter which is the product of mobility of charge carrier and the unit gate capacitance, V_T is the volt equivalent of temperature approximately equal to 26 mV at room temperature and n is given by (3),

$$n = (g_m + g_{mb}) / g_m \tag{3}$$

The transconductance g_m in subthreshold region is as given in (4),

$$g_m = I_D / n \, V_T \tag{4}$$

Also, the drain conductance g_o is expressed as in (5),

$$g_o \approx \lambda \, I_D \tag{5}$$

The current mirror load of the differential pair has series and parallel placed MOSFETs so that output current is small. The overall transconductance of the OTA is calculated using (6),

$$G_m = g_{mb1} / N^2 \tag{6}$$

where g_{mb1} is the body transconductance of PMOS M1 and N is the number of series-parallel connected transistors of current mirror load of the differential pair. The current source I_{dc} is realized using PMOS current mirror to provide bias current.

The voltage gain of the circuit is given by (7), with K=1.

$$A_v = g_{mb1} / (g_{o7} + g_{on}) \tag{7}$$

where g_{on} is the overall output conductance of series connected NMOSFETs at the output node. The unity gain cutoff frequency is given by (8),

$$f_T = g_{mb1} / 2 \, \pi \, C_L \tag{8}$$

The transconductance in pS to few nS can be realized when current is divided at the load instead of at the input side. The advantages of series parallel current mirror OTA are low input noise, wide linear input range, small transconductance and nano

watt power. This OTA is used as Gm-cells of a second order Gm-C filter.

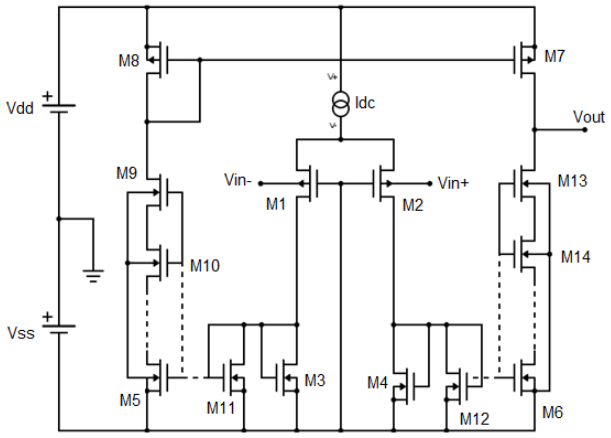


Figure 4: Schematic of proposed bulk driven series-parallel OTA

3. Discussion on Simulation Results of SPOTA

The series parallel current division technique is applied to the current mirror OTA for gate driven and bulk driven input PMOS differential pair. Using the spectre simulation in Cadence Virtuoso tool power dissipation, DC gain, unity gain bandwidth, transconductance, gain and phase margin of circuit are determined and compared. The CMOS technology used is gpdk 180nm. The gate driven SPOTA has power supply of 1.6 V while bulk driven circuit is operated at supply voltage of ± 0.4 V. The input and output waveforms of the bulk driven SPOTA is shown in Figure 5. For ± 1mV input, output voltage of 128 mVpp is obtained. The DC gain is 29.59 dB and 44 nW of power is consumed as can be seen in Figure 6 and in Figure 7 respectively. The SPOTA is stable with 53o of phase margin and has gain margin of 28.2 dB. The OTA has unity gain cutoff frequency of 34. 28 kHz and CMRR of 74.68 dB. The output is linear for ± 50 mV differential input voltage, which is a good range for biomedical applications.

Table 1 gives the comparison of bulk driven SPOTA with gate driven circuit. It is observed that the bulk driven MOSFETs can be operated with supply voltage near the threshold voltage, while for gate driven MOSFETs overdrive voltage requirement has to be met. Therefore, gate driven SPOTA with 1.6 V supply voltage has higher power consumption of 194.336 nW whereas bulk SPOTA power is reduced to 44 nW. The transconductance of the OTA is function of bias current. For bulk driven circuits,

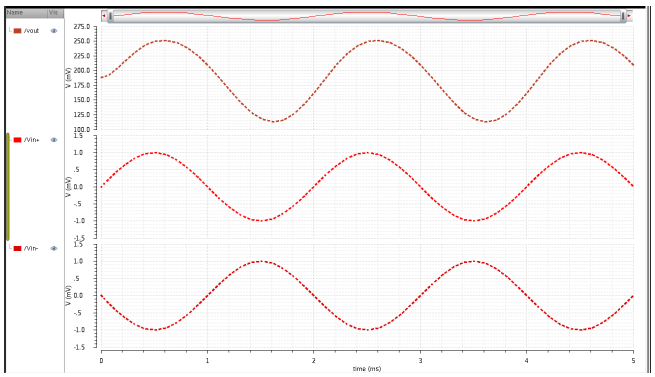


Figure 5: Input and output waveforms of bulk driven SPOTA

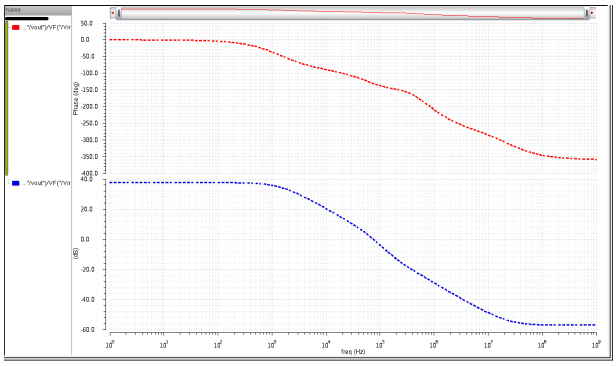


Figure 6: Plot of DC gain and phase of bulk driven SPOTA

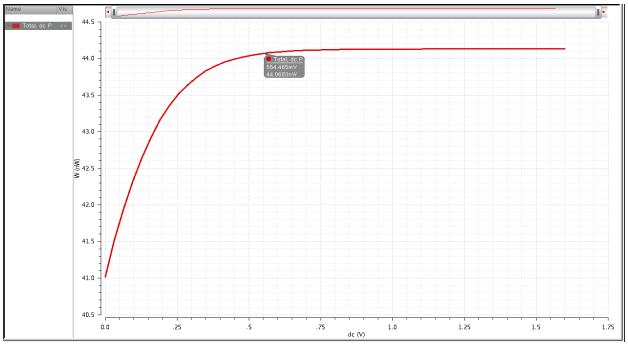


Figure 7: Power dissipation curve of SPOTA

the operating current can be very small. A transconductance of 6.03 nS for bulk driven and 24.63 nS for gate driven is achieved using series parallel technique for N=7 and can be reduced to pS by choosing large value for N. Hence the transconductance of the bulk driven SPOTA is small compared to that of gate driven at the cost of reduced bandwidth, gain margin and phase margin. The CMRR is improved in bulk SPOTA.

4. Gm-C Filter Design

The SPOTA proposed has small transconductance, hence its use in analog filter design for low frequency applications is demonstrated. The circuit structures in CMOS for active filters include Gm-C, OTA-C, switched capacitor integrator and Opamp-RC structures. The Gm-C filter is preferred because of its simplicity, ease of cascading, less area and minimum power consumption.

Table 1: Comparison of Bulk Driven SPOTA with Gate Driven SPOTA

Design Parameter	Gate SPOTA	Bulk SPOTA
Technology (nm)	180 nm	180 nm
V_{DD}(V)	1.6	0.4
Bandwidth (kHz)	74.93	3.504
DC gain (dB)	21.15	29.59
Phase Margin	62.6º	53º
Gain Margin (dB)	38.1	28.2
CMRR (dB)	62.96	74.68
Transconductance	24.63 nS	6.03 nS
Power (nW)	194.336	44

Gm-C filters rely on OTAs and capacitors to realise the transfer function of the filter. Since the transconductance of proposed SPOTA is less than 10 nS and can be even pS if N is chosen large, the integrated capacitor value can be limited in the order of fF to few pF. A second order Gm-C filter realized using biquad structure [8] [9] is shown in Figure 8. It can be set as low pass or high pass filter using the inputs $V_{in\,LP}$ and $V_{in\,HP}$. The two OTAs are used in unity gain mode. For low pass filter, the input is connected to $V_{in\,LP}$ with $V_{in\,HP}$ grounded, whereas it is connected to V_{inHP} with V_{inLP} grounded for high pass filter. The filter is synthesized by neglecting the nonidealities. In Figure 8, voltage $V1$ is as given by (9),

$$V1 = G_{m1}(V_{out} - V_{in\,LP}) / sC_1 \quad (9)$$

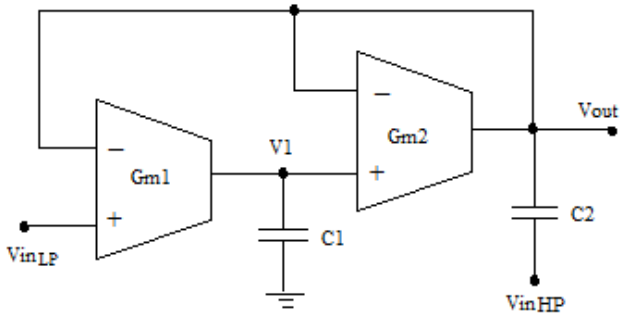


Figure 8: Block schematic of biquad Gm-C filter

Also, the output voltage is as given by (10),

$$V_{out} = (G_{m2}V1 + sC_2V_{in\,HP}) / (sC_2 + G_{m2}) \quad (10)$$

Substituting (9) in (10) we obtain the output voltage V_{out} as in (11),

$$V_{out}(s) = ((G_{m1}\,G_{m2} / C_1\,C_2)\,V_{in\,LP} + S_2\,V_{in\,HP}) / (s^2 + s\,(G_{m2} / C_2) + G_{m1}\,G_{m2} / C_1\,C_2) \quad (11)$$

Choosing $G_{m1} = G_{m2}$ the transfer function of biquad Gm-C low pass filter is obtained by setting $V_{in\,HP} = 0$ and is given by (12),

$$H_{LP}(s) = (G_m^2 / C_1\,C_2) / (s^2 + s\,(G_m / C_2) + G_m^2 / C_1\,C_2) \quad (12)$$

Similarly, the transfer function of high pass filter is obtained by setting $V_{in\,LP} = 0$ and is given by (13),

$$H_{HP}(s) = s^2 / (s^2 + s\,(G_{m2} / C_2) + G_{m1}\,G_{m2} / C_1\,C_2) \quad (13)$$

The 3 dB cutoff frequency and the quality factor of the filter is as given in (14),

$$f_c = (1/2\pi)\,sqrt\,(G_m^2 / C_1\,C_2) \text{ and } Q = sqrt\,(C_2 / C_1) \quad (14)$$

5. Simulation Results of Biquad Gm-C Filter

The simulation of biquad Gm-C filter is carried out using Cadence Virtuoso Tool with gpdk180nm CMOS technology. The magnitude response of the low pass and high pass filter are shown in Figure 9 and Figure 10. The proposed OTA connected as voltage follower is used as Gm cell. The filters have butterworth magnitude response with cutoff frequency in the range 25 Hz to 225 Hz for C_1 = 10 pF. Gain and power consumption are around -7.4 dB to -7.8 dB and 91.7 nW respectively.

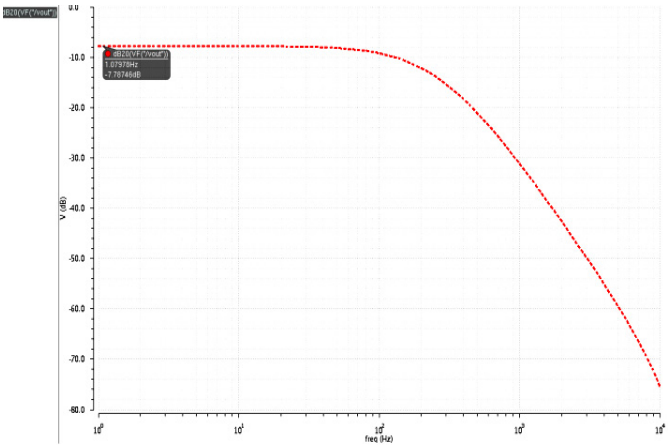


Figure 9: Magnitude response of low pass filter using bulk driven SPOTA

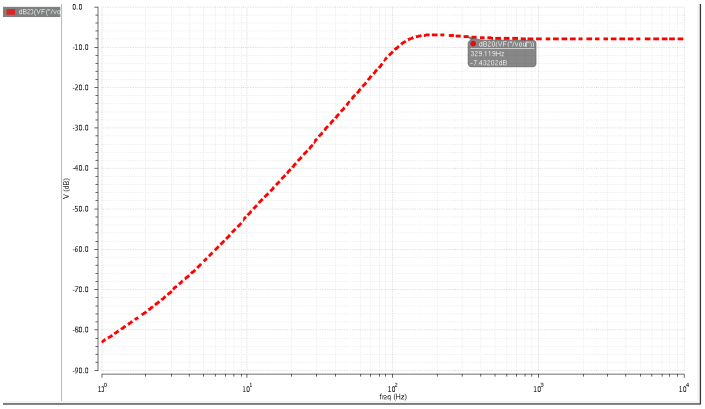


Figure10: Magnitude response of high pass filter using bulk driven SPOTA

By varying capacitor values or by varying the bias current to SPOTA, required cutoff frequency for the filter can be set. When the input signal is applied to $V_{in\,LP}$ with $V_{in\,HP}$ connected to ground, biquad filter is configured as low pass filter, whose cutoff frequency can be tuned by varying C_2 from 25 Hz to 225 Hz. Therefore it can be used to select biomedical signals such as ECG. When the signal is applied to $V_{in\,HP}$ with $V_{in\,LP}$ connected to ground, it is configured as high pass filter, which can be used to allow signals starting from DC to 225 Hz. The series parallel technique reduces the transconductance of the filter to very minimum value. Implementation of low cut off frequency filters using bulk driven SPOTA, results in small capacitors for the filter because of low transconductance. Also the bulk driven SPOTA consumes low power, hence it is useful in active filter section of the low power biomedical AFE.

Table 2 gives comparison of Gm-C filter implementations with different types of OTA. It can be seen from the table that the series parallel current division technique gives smallest transconductance value. The series and parallel connected unit transistors are used to realize the circuit. The designed Gm-C filter is a multifunction filter as low pass and high pass operations can be realized in the same circuit. Also filter can be electronically tunable if the dc current to the OTA is varied. The range of 3 dB frequency of the filter is suitable for selection of biomedical signals. The circuit operating voltage is less than 0.5V hence can be used in the analog section of low voltage, low power system on chip (SoC).

Table 2: Comparison of Gm-C filter Implementations

Parameters	Ref [3] 2002	Ref [1] 2006	Ref [5] 2014	Ref [6] 2018	This work
Technology	1.2 µm	0.8 µm	0.35 µm	180 nm	180 nm
V_{DD} (V)	±1.35	2	1	1	±0.4
Transconuctance	10 nS	33pS-2.35nS	33ps	50pF/pole	6.03 nS
Power	8.2µ (OTA+bias circuit)	100 nW	82 n	1.75µW	91.7 nW
Type of OTA used in the filter	Current division, bulk driven current mirror OTA	Series parallel current mirror OTA	Preattenuated input, current mirror OTA with bump transistor	Current steering current mirror OTA	Series parallel current mirror Bulk driven OTA
Cut-off frequency	0.17Hz	10 Hz-1.2 kHz	0.69 -72 Hz	0.225- 16 kHz	25-225 Hz

6. Conclusion

An OTA operating at ±0.4 V supply, consuming around 44 nW power with transconductance as low as 6.03 pS is designed. Very small transconductance is obtained by series parallel current division technique. Further reduction is transconductance is obtained using bulk driven input PMOS differential pair. The bias current and the number of series-parallel connected NMOS current mirror determine the transconductance of the OTA.

A low cutoff frequency Gm-C biquad filter is then realized using the SPOTA proposed, which can be configured as low pass or high pass filter. The second order filter thus designed consumes 91.7 nW of power. The small value of transconductance of the SPOTA has led to small capacitors in the design. So this multifunction filter can have on-chip capacitors. The filter has Butterworth response with low power consumption, low cutoff frequency and hence suitable for AFE of biomedical devices.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Arnaud et al., "Nanowatt Sub-nS OTAs With Sub-10 mV Input Offset Using Series-Parallel Current Mirrors," IEEE Journal of Solid-State Circuits, **41**(9), 2009–2018, 2006, doi:10.1109/JSSC.2006.880606.

[2] A. Arnaud, C. Galup-Montoro, "Pico-A/V range CMOS transconductors using series-parallel current division," Electronics Letters, **39**(18), 1295–1296, 2003, doi: 10.1049/el:20030840.

[3] A. Veeravalli et al., "A CMOS transconductance amplifier architecture with wide tuning range for very low frequency applications," IEEE Journal of Solid-State Circuits, **37**(6), 776–781, 2002, doi:10.1109/JSSC.2002.1004583.

[4] D. Arbet, M. Rak, V. Stopjakov,"Comparison of Gate-Driven and Bulk-Driven Current Mirror Topologies," in 2016 IEEE 19th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS), 1–4, 2016, doi::10.1109/DDECS.2016.7482457.

[5] S. Thanapitak, "An 1-V, 74-dB, sub-Hz Gm-C filter based on a modular transconductance reduction technique," in 2014 IEEE 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 1–5, 2014, doi.org/10.1109/ECTICon. 2014.6839763.

[6] J. Perez-Bailon, Marquez, A., Calvo, B., Medrano, N., Sanz-Pascual, M.T., "A 1V–1.75µW Gm-C low pass filter for bio-sensing applications," in 2018 IEEE 9th Latin American Symposium on Circuits & Systems (LASCAS), 1–4, 2018, doi10.1109/LASCAS.2018.8399960.

[7] K. Garradhi, "Low-voltage and low-power OTA using source-degeneration technique and its application in Gm-C filter," in 2016 IEEE 11th International Conference on Integration, Design & Test Symposium (IDT), 221-226, 2016, doi: 10.1109/IDT.2016.7843044.

[8] M. Kumngern, Dejhan K., "Voltage-mode low-pass, high-pass, band-pass biquad filter using simple CMOS OTAs," in 2009 IEEE International Conference on Instrumentation and Measurement Technology (I2MTC'09), 924-927, 2009, doi: 10.1109/IMTC.2009.5168583.

[9] C. Chanapromma et al., "High even-order fully-differential biquadratic continuous-time low-pass Gm-C filter," in 2014 IEEE International Electrical Engineering Congress (IEECON), 1–4, 2014, doi:10.1109/iEECON.2014.6925853.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 72-95 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Automated Extraction of Heavyweight and Lightweight Models of Urban Features from LiDAR Point Clouds by Specialized Web-Software

Sergiy Kostrikov[1,2,*], Rostyslav Pudlo[2], Dmytro Bubnov[2], Vladimir Vasiliev[2], Yury Fedyay[2]

[1]*Department of Human Geography and Regional Studies, School of Geology, Geography, Recreation and Tourism, V. N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv, 61022, Ukraine*

[2]*EOS Data Analytics, 31 Alchevskyh St., Kharkiv, 61002, Ukraine*

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 08 October, 2020
Online: 08 November, 2020

Keywords:
LiDAR
AFE
Phased Methodological Flowchart
High Polyhedral Modeling
Low Polyhedral Modeling
Heavyweight Models
Lightweight Models
Web-Software
ELiT Server
ELiT Geoportal

ABSTRACT

3D city modeling may be considered as one of the key applications, that are provided by the Automated Feature Extraction (AFE) techniques from LiDAR data. The authors attempt to prove that with growing availability of LiDAR surveying methods the resulted 3D city models become the most significant modeled features for any urban environment. Our paper represents the conceptual multifunctional approach within the AFE frameworks, that has been introduced through consequent steps of the phased methodological flowchart with its two branches: High Polyhedral Modeling (HPM) and the Low Polyhedral Modeling (LPM) of buildings. Both branches result in the heavyweight models, and in the lightweight ones, correspondingly. The research purpose of this paper is to outline our multifunctional approach (functionalities of Building Extraction, Building Extraction in Rural Areas, Change Detection, and Digital Elevation Model Generation) to the fully automated extraction of urban features, and present our original contributions to the relevant algorithmic solutions within both HPM, and LPM pipelines, as well as represent desktop, web-, and cloud-based software elaborated for these intentions. Original enhancements and optimizations of the adopted AFE-techniques have been bounded to the phases of the methodological flowchart, while some derivative results have been presented not only as the software description, but also in the discussion chapter. Joint implementation of various functionalities in a web-based application (the Server) is presented with several interface samples of research in urban block and district scopes, while a cloud-based application (the Geoportal) is an Internet-toolbox for solutions in the scope of a whole city.

1. Introduction

This text is a significantly changed extension of the work firstly presented as a conference paper [1].

The global world has been already transferring into the information society for several recent decades. It is quite acceptable to consider *a Geographic Information System – GIS* as one of the core tools of such transfer together with the relevant technologies of remote sensing, including *LiDAR (Light Detection and Ranging)* technique. It has been remarkable for few recent decades that exactly this period also has been featured by the continuing urbanization process, that still takes place nowadays in many developing countries. Numerous facts and phenomena indicate, that we may face the largest urban growth wave within the whole human history, which also concurs with prompt development of information technologies, remote sensing, computer sciences, geoinformatics, and *GIS-platforms / modules*. All listed entities could and should be involved in resolving of those drastic problems arisen within the urbanized areas, for example, by considering these territories as the *hierarchical urban geosystems* and making the corresponding decision support recommendations [2, 3]. Thus, even several highly challenging issues relevant to the Smart City development can be met in this way [4]. Moreover, the rapid growth of both remote sensing technique and GIS-technological involvement in surveying environmental consequences of the urbanization process has been

*Corresponding Author: Sergiy Kostrikov, sergiy.kostrikov@karazin.ua

clear evident over the past few decades [5], while quite a few both approaches, and methods, as well as user interfaces have been developed [6-15].

Urban features can be accepted as the core constituents of any virtual presentation for each digitally simulated city. A *3D City Model* is the key formalized entity of the mentioned city content, and it enables effective and accurate 3D modeling of the urbanized environment with respects to housing sets and infrastructures [16-23]. Within the urban studies perspectives it can be possible to accept a 3D object / feature model as a work-flow key issue of the GIS output completed from the urban remote sensing input. This assumption may be even more evident, while the discrete objects are considered as those modeled results, which are intended for a customized solution within the framework of one only, or several urban applications. In this meaning 3D modeling appears to become a key component within the common geoinformation paradigm [24]. It is considered be possible to accept a 3D city model of urban area as that entity, which corresponding environmental analog is placed within 3D urban environment described by routine urban features and structures with buildings as the dominant objects among them.

Exactly for a couple of latest decades LiDAR (Lidar) data collecting processing technique has become an alternative (to areal imageries) data source for generating a 3D representation of natural landscapes and housing environments as well as the overall human-environment intercourse [25, 26]. Being able to collect directly the accurate *3D point clouds* of various density over urban territories, the LiDAR technique proposes an effective and beneficial data source in this meaning. By the way, this is illustrated further in this text.

There are Airborne (ALS), Terrestrial (Mobile - MLS), and UAV-LS (Unmanned Aerial Vehicle Laser Scanning) LiDAR techniques within the common remote sensing technology, that measures distances on the base of the time intervals between the laser signal transmitting / receiving. All ALS / MLS /UAV-LS mapping are surveying and mapping tools based on hardware platforms and software solutions [27-30]. The ALS lidar normally completes low-average density measurements of the underlying topographic surface, that result in the georeferenced, but unstructured set of points – lidar Point Clouds, while the MLS technique usually captures the walls – with a high resolution of façade details for buildings of different sizes. The UAV-LS lidar accomplishes *high density measurements* and can provide surveys that may be considered as hybrid ones, since it combines those output data, which can be provided by both ALS, and MLS Lidars.

The *automated feature extraction* (AFE) methods are the derivatives of ALS / MLS / UAV-LS Lidar remote technique. AFE-procedures produce 3D city models mentioned above, that are positioned in three dimensions just because of the AFE-method implementation. Since a common three-dimensional urban model may be a subject of more, than one hundred applications in various industrial domains [24], the AFE-methodology itself can be hardly overvalued. Municipal management, urban emergency services, noise and other hazard mapping, visibility analysis city infrastructure inventory, population and energy demand estimation with building models are those key urban applications, in which 3D models obtained from LiDAR data through AFE procedures are in an evident growing marketing demand according to the obvious reasons. To meet this demand requirements an AFE-model should be of high accuracy [31], what in some cases implies *a hybrid data source*. The latter means either a "LiDAR + areal image" hybrid [32], or additional involvement of MLS / UAV-LS data processing procedures for detailed extractions of building façades [33].

The *main research purpose* of this text is to outline with the phased methodological flowchart our conceptual multifunctional approach to the fully automated extraction of urban features, and present the original contribution to relevant algorithmic solutions within the modeling pipeline, as well as represent the web- and cloud-based software elaborated for these intentions.

2. Previous Works Done due to Building Detection, Extraction, and 3D Reconstruction from LiDAR Data

Fully Automated Feature Extraction, building Point Cloud segmentation, rooftop modeling, and 3D reconstruction of buildings, all have been among the main topics of thematic discussions in the remote sensing community through the relevant papers and on forums of various scales especially for the latest decade [26, 32, 34-42].

Now, AFE is still a vitally crucial part of what is being done and what researchers and other professionals are attempting to do better within the LiDAR surveying / processing domain. How has it further been progressed with its technique recently? We may consider as a key issue for efficient AFE-results a provision of a bridge between MLS / UAV-LS lidars, from one side, and ALS lidar, from another, one and vice-versa. Without any further explanations just in this paper section, we may only express a somewhat trivial idea, according to which *the composite models* of urban features extracted may be considered as the most effective ones [43].

2.1. Overall AFE issues

Because of ALS / MLS / UAV-LS joint surveying methods the automated feature extraction as a generating procedure for 3D urban models has become an evident alternative to the urban photogrammetry. It is a well-known fact, that besides various direct processing of aerial images, the Urban Remote Sensing (URS) (e.g., multispectral, hyperspectral ones [44]) traditionally provides the 3D building model generation from airborne-mobile-drone photogrammetric point clouds, but the lidar surveying / processing approach deals with similar dataset structures. Moreover, the novel sensor technology with its lower cost in comparison with previous hardware has expedited Lidar involvement in urban studies. Significant pros for such solution may be that fact, according to which ALS / MLS / UAV-LS sensors can deliver point datasets with densities of huge range (from one up to several thousand points per sq. meter). Even with the lower edge of this density range, it is feasible to detect urban structures, their approximate boundaries, and other various man-made features. Those models can be generated, which correctly resemble both wall, and roof structures. Many relevant methods for ALS / MLS / UAV-LS surveyed data processing have been proposed due to building extraction, detection, and 3D reconstruction [45-54].

Once we already commonly classified the automated feature extraction approaches on the base of an input data source [1]. The first approach means processing the high-resolution airborne imageries with additional including of *digital elevation models* (DEM) into an AFE-pipeline [7, 55]. Although some significant results have been obtained, the "only aerial imagery" approach may be considered as that one, which does not perform well enough in the densely built-up urban areas, primarily because of shadows, landscape gaps, and unsatisfactory contrasts in various urban configurations. According to this, the AFE procedures based exclusively on the first approach may not be fully reliable for the robust practical usage [33, 56]. The second approach straightforwardly employs LiDAR data and produces improved AFE-results, if compared to the imagery-only methods [31, 35, 37, 56]. What is more, this second approach implies an employment of very different techniques for processing ALS point clouds [57-59]. Methods applied within the third approach combine in a common case both aerial imageries, and all kinds of LIDAR data (ALS / MLS / UAV-LS) so that to employ the complementary information from all data sources [33, 60].

The AFE complete algorithmic content implies the ground and vegetation detection, while the man-made feature AFE-technique has to use either single-, or multi-return ALS / MSL / UAV-LS range and intensity information with application of the various thematic algorithms, e.g., as: neural networks [61]; *RANSAC* (the *RANdom SAmple Consensus* algorithm) approach for extraction of feature plains [62] with its key modifications [63, 64], that have been successfully employed by the Polygonal Surface Reconstruction method (the *PolyFit*) for feature reconstruction [65]; 3D Standard / Randomized Hough Transform (SHT / RHT) methodology that generally consists of three main steps: building points' detection, detection of building planes, and these planes' refinement [66-68]; implementation of knowledge-based entities [69]; the multi-scale approach [70].

Besides all algorithmic solutions mentioned above it may be reasonable to emphasize the hierarchical terrain recovery algorithm, that may robustly provide distinguishing between ground and non-ground points within an input point cloud by implementing the "adaptive and robust filtering" method [50]. Within this approach it is necessary to consider the whole range of input data to evaluate a DEM of high resolution for further feature extraction steps using this relevant hierarchical strategy. Thus, road linear features can be identified by classifying signal intensity and elevation data. Not only discrete building boxes, but also network features can be detected and extracted then. For example, man-made features of the road networks can be derived using a customized transformation technique, and then validated with road lines and topographic shapes obtained from an initial Lidar point cloud. Further one can obtain the attributes of road segments such as their widths, lengths, and slopes by computing some derivative information and enhancing existing metadata in this way. Other man-made features, firstly, building models are created with the higher level of accuracy, which would correspond to various Levels of Detail (LODs), beginning from LOD1 simplified box models and up to the models with internal partitions [71]. City Geography Markup Language (*CityGML*) is employed as a geoinformation data standard for presentation of the geometry and geographical data in digital models related to city buildings.

We have already presented *the High Polyhedral* models *(HPM)* of buildings and *the Low Polyhedral* ones *(LPM)* [1]. These two categories not only are the components of the authors' Lidar data processing methodology but also they are mentioned in the surveying section of this text only because HPM / LPM definitions reflect two significant mainstreams in the existing automated feature extraction. Obviously, these two dominant trends can be named with other titles by different authors.

Mostly a literature survey already completed above concerns exactly the HPM issues. The latter imply the generation of building models consisting of numerous polyhedrons, and because of this the relevant modeled entities can be accepted as "heavy ones" – *the heavyweight (HW) models*, which are produced by *the High Polyhedral Modeling approach*. It means, an HPM-building model may be generated from up to more, than one hundred thousand of Lidar points. It may be reasonable to state that the HPM-procedures are primarily based on the *Lidar point classification* that one, which is not directly associated with clustering, while the LPM-operations – *on the Lidar point cloud segmentation* through *clustering.*

The common work-flow of the HPM building model generation can be outlined as follows on the base of those literature sources that we have already referred to.

At its *first stage*, the building footprints (building base boundaries) are detected by segmenting DEM data obtained from LiDAR for two general classes: *ground* class and *non-ground* ones. The bare ground as a grid is delineated upon this step. A well-known, so called "sequential linking technique" is often suggested to reconstruct building footprints into regular polygons. These polygons then are improved so that to reach the cartographical standards [52, 72].

The so-called *prismatic models* are generated for those urban structures, which roofs are flat, and *polyhedral models* are created for those structures, which roofs are non-flat, at the *second stage.* At last, at the *third stage*, the vertical wall rectification operations should be applied, if there are sufficient MSL / UAV-LS or other correcting data in a relevant geodatabase for processing of this *area-of-interest* (AOI).

These three introduced stages may overlap almost any LiDAR data processing workflow. Most urban attributes of these building models are obtained from ASL / MSL / UAV-LS data. All corresponding HPM AFE-algorithms, that conclude the three stages workflow referred to above, should be tested using several geodatabases of varying earth surface type, vegetation coverage type, urban area type and LiDAR point density. Afterwards, normally the most effective algorithm should be chosen.

If we complete the general summary even for several overviewed HPM AFE-algorithmic results, this summary may demonstrate that in many urban territories the derivative DEMs accumulate most topographic details and remove non-ground features reliably enough. The transportation infrastructural network features are also depicted mainly satisfactorily even within densely built-up city districts. The extracted building footprints demonstrate to have enough positioning accuracy. Their estimated values may be equal to the accuracy obtained from data surveyed in field monitoring, while this traditional

surveying technique in many cases is a routine procedure of Lidar processed result accuracy evaluation [73, 74].

2.2. Detection, Segmentation, and Reconstruction of Building Roofs

The relevant modeling software tools can provide by a Point Cloud segmentation and clustering procedures of building detection and extraction the models of low-rise buildings preferably through rural areas and suburbs [1]. Simulating procedures stay within the Low Polyhedral Modeling approach, which is based on procedures of planar segmentation of Lidar point clouds rather, than on their classification (the case of HPM). The LPM building models produced are composed of not many polyhedral facets, and the number of points intended for a single model generation is limited approximately by a range from five and up to thirty thousand points maximum. Thus, the low polyhedral models can be considered as *the lightweight (LW) ones*. Overall low polyhedral modeling frameworks adapted to our LPM methodological approach proceed from the series of seminal papers in roof segmentation and reconstruction, therefore we have named this algorithmic technique as the *SAS*-methodology (the title has been abbreviated according to the authors of the initial approach) [35, 37, 75-77].

Methodologically we can reasonably define HPM as the automated extraction of *a whole building*, while LPM – the automated extraction of *this building roof.* A *segmentation* procedure of roof plains may be a key one within a corresponding AFE-pipeline titled for this section of our paper. This data-driven procedure was initially adapted for Lidar data from imagery processing [78]. Segmentation normally starts with applying *clustering methods* to Lidar point clouds [37, 76, 77].

The automated extraction of a whole building can be normally fulfilled by three similar sub-procedures outlined above for roofs, i.e., *building detection*, *building segmentation*, and *building reconstruction* [50, 59, 70, 79]. Although, contrary to roof extraction, all three sub-procedures, which are related to a whole building, may not be evidently less distinguishable. The case is a completely automated process of a whole building extraction may not yet be reliable enough from a practical point of view, because of the great complexity of actual urban environment with tremendous variety of its configurations. Thus, a whole building fully automated processing may need to pass a longer distance to provide wider usage of available supplementary data sources, for instance, city ground plans or municipal architecture schemes, so that to significantly enhance an ultimate processed result.

Somewhat more simplified methods, which provide roof extraction as equal to roof detection, are based on a *Digital Surface Model (DSM).* This model contrary to a DEM includes not only the ground, but the discrete features also. According to existing references, the DSM is calculated using an imagery and feature pyramids [80]. The finalized surface is refined then on the base of local adaptive regularizations. The roof detection step is grounded on the fact that a roof should be higher, than the neighboring ground – the topographic surface. This is normally estimated applying tools of the mathematical morphology to this DSM. The sliding window size technique involves the input information about the maximum roof size in its existing geographical extent.

This "only DSM involved" method has been further specified, when the building roofs are segmented depending on their estimated complexity, and in the same way finally reconstructed [81]. Usually, two types of parametric models are used for simple building roofs in those cases, when a building possesses either a flat, or a symmetrically sloped roof [50, 82]. Within other variants prismatic models are applied for complex separate roof structures or to connected building roof sets.

Two general classes of the roof extraction approaches are either the *data-driven technique* (e.g., *SaS*-methodology already mentioned above, and this technique also known as *a generic of polyhedral technique*), or the *model-driven one* (also known as *a parametric technique*). The latter implies some assumptions about topological and geometrical properties of a whole building model, generally, and due to a roof model, particularly. Two methodologies can contrast one to another in a categorical perspective: the data-driven method can be accepted as the *geometrical approach* (due to creating the roof geometry from the points of a given cloud), and the *topological one*. It is necessary to emphasize, that there is no a clear boundary between these two approaches [83]. Some researchers report the Hough transform technique within the model-driven segmentation methods [37], while other scientists recall it as one from the key data driven approaches together with region growing and *RANSAC* [66]. By the way, in the opposite case some fully automated approach has been clearly defined as a model-driven one, which applied for the 3D model reconstruction with prototypical roof templates (*CityGML LOD2*) [84].

The invariant moment technique has been applied for generation of the roof template library according to various building classes [46]. It is a quite known fact, that the model-driven approach may fail, when the inhomogeneity of feature distribution within a point cloud leads to biased parameters. Therefore, extracted models of complicated, not ordinary building roofs are produced by using data-driven algorithms. These algorithms are normally based on segmented intersecting planes and operate with triangulated point clouds [35-38, 62-68, 85].

It is a commonly known fact for the data driven approach, that most critical errors occur upon the determination of a total building roof outline, when trees are placed near a building box [62, 64, 66-68]. To avoid this issue various algorithmic approaches have been modified independently based on the detection and outlining of planar facets, e.g. [46]. A facet plane is normally determined by the point clustering procedure. A roof scheme is outlined by a connected component analysis. A key related assumption is that all the geometric model boundaries for roof segments are either parallel, or perpendicular to the main building disposition.

Thus, it is necessary to emphasize once more, that it may be not reasonable at all to distinguish whole building modeling from its roof reconstruction, while only ALS data are involved. It can be the same or, at least, very similar procedure, if only façades are not generated from MSL / UAV-LS data sources. Also, this procedure may be both within the HPM, and the LPM frameworks. In the mentioned context various researches suggested a boundary-based building / its roof extraction and corresponding TIN-based modeling and reconstruction [50, 86]. The buildings and their roofs

are reconstructed within this approach by triangulating each point cluster of identified facet candidates and clustering those fragmentary triangles into small patches, concluding piecewise planar fragments. Finally, the reciprocal intersections of the summarized planar facets are employed for extraction of building corners and for definitions of corresponding extracted feature orientations. A quite resembling technique have been used in our high polyhedral modeling for the building extraction [1].

It may sound reasonable to conclude this literature review by mentioning again, slightly more in details, the Hough transform and *RANSAC* algorithms and making references to so-called "global solutions to building extraction and reconstruction" [87]. Most earlier reports of the Hough transform (HT) applications mainly concerned 2D mapping of point clouds, until the "extending generalized HT" was proposed that provided detection of 3D features in a point cloud [88]. Standard and Randomized Hough Transform methods are ones among the most popular techniques of plane detection [68], and it can be employed for the determination of more, or less precise roof plane parameters – the triplets selected are accepted as possible roof planes [89]. Comparison of one of the Hough methods – RHT with RANSAC demonstrates a significant advantage of the RHT in the processing accuracy versus computing time tradeoff [68].

The classic RANSAC (FB81 according to [65]), as it was already mentioned above extracts primitives from point datasets [62]. The high-quality geometric instances extracted are a subject for further polygonal surface reconstruction of various features with 3D geometry. Nonetheless there are still no proved evidence that either Hough, or RANSAC optimizations have been successfully applied for modeling urban areas [68], but even optimized RANSAC methods may lead to the numerous false plains [37]. Probably it would be right to affirm that RANSAC is that approach, which has faced almost the greatest number of its optimizing solutions among other point cloud detecting / segmenting methods. Even early optimizations of this algorithms provided the removal of up to a half the outliers resulted from a given data set [90]. Probably the best overview of this algorithm improvements together with one more original optimization has been given in [63].

Despite all optimizations followed by rising algorithmic efficiencies, both HT, and RANSAC still hardly could be applied to more, or less significant urbanized areas, e.g. to the city district scope. To meet such challenge among number of other purposes sophisticated methods have been elaborated within the paradigm of "global solutions to building segmentation and reconstruction" [87, 91]. Roof segmentation and reconstruction have been consequently developed within the frameworks of this methodology. Both these two stages should be recognized as a solution of the appropriate energy functions' minimization problem. Firstly, after an initial segmentation completed, every Lidar point has been accurately assigned to its optimal plane by minimization of a global energy function [87]. It was named "a global solution", because the method could define multiple roof plains concurrently. After it, on the base of the segmented in this way roof planes another "global solution" has been developed and applied already for the roof reconstruction stage. The building box has been partitioned into volumetric cells, what allows to construct the roofs of the sustainable topology and the correct geometry [91].

2.3. Generalizing LiDAR-Based Solutions in 3D Building Modelling

We have finalized with building reconstruction reviewing AFE steps in the previous subsection of the text, while shortly mentioning in subsection 2.1 processing integrated data sources. According to number of references, reconstruction of buildings can be accepted as a complicated procedure of the digital presentation generation for those physical urban features, that can be extracted from Point Clouds and transformed into effectively structured 3D models with various attributes, while the quality of these models should be evaluated further [58, 86, 92, 93]. Generalizing or *hybrid* solutions in the mentioned extent mean not only the data fusion involvement [94, 95], but also using this basis for *footprint extraction* in the automated mode and its *boundary regularization*, since both operations are the basis for a roof reconstruction, and even for wall raising, when MSL data are not available.

Four following groups of authors almost independently developed some interesting *hybrid* approach lied within the following workflow of five major algorithmic steps [37, 38, 96-98]: A density-based algorithm of clustering for the individual building segmentation, which begins with footprint delineation; A rectified boundary-tracing algorithm is applied; A hybrid method for footprint planar patches segmentation developed; these methods select so-called "seed points" in the parametric space and generate the regions in usual (spatial) space; A boundary regularization approach that examines point outliers; Finally, reconstructing procedures are accomplished due to the topological and geometrical information about building roofs and using the intersections of footprint planar patches.

This hybrid approach proceeds from that fact, according to which there are two main classes of methods applied to trace building footprint boundaries: *raster-based* method and *vector-based* one. By the raster-based technique, the point clouds are usually transformed into a regular grid, after it the image processing methods are employed to indicate, trace, delineate, and regularize the boundary and footprint edges [99, 100]. Upon the vector-based approach, that was developed quite long before, a simple string delimits the exterior boundary, while other parameters, representing the inner boundaries, are extracted from raw LiDAR data [101].

3. Original Methodological Approach

3.1. Phased Methodological Flowchart

While introducing the methodological basis of our complete R&D cycle (from raw ALS data processing to solutions of use cases with building models), we stand on that broadly accepted point of view, according to which LiDAR remote sensing can be defined as the research and technological approach primarily used to obtain for further processing the information about the topographic surface, vegetation, and various features of the human infrastructure at a certain distance from an observer's point (buildings, bridges, roads, powerlines, etc.) [25-27, 31].

Due to the options and the necessities to cover large areas in the cities the ALS surveys obtain data from very different urban configurations, that can be distinguished, initially, by building

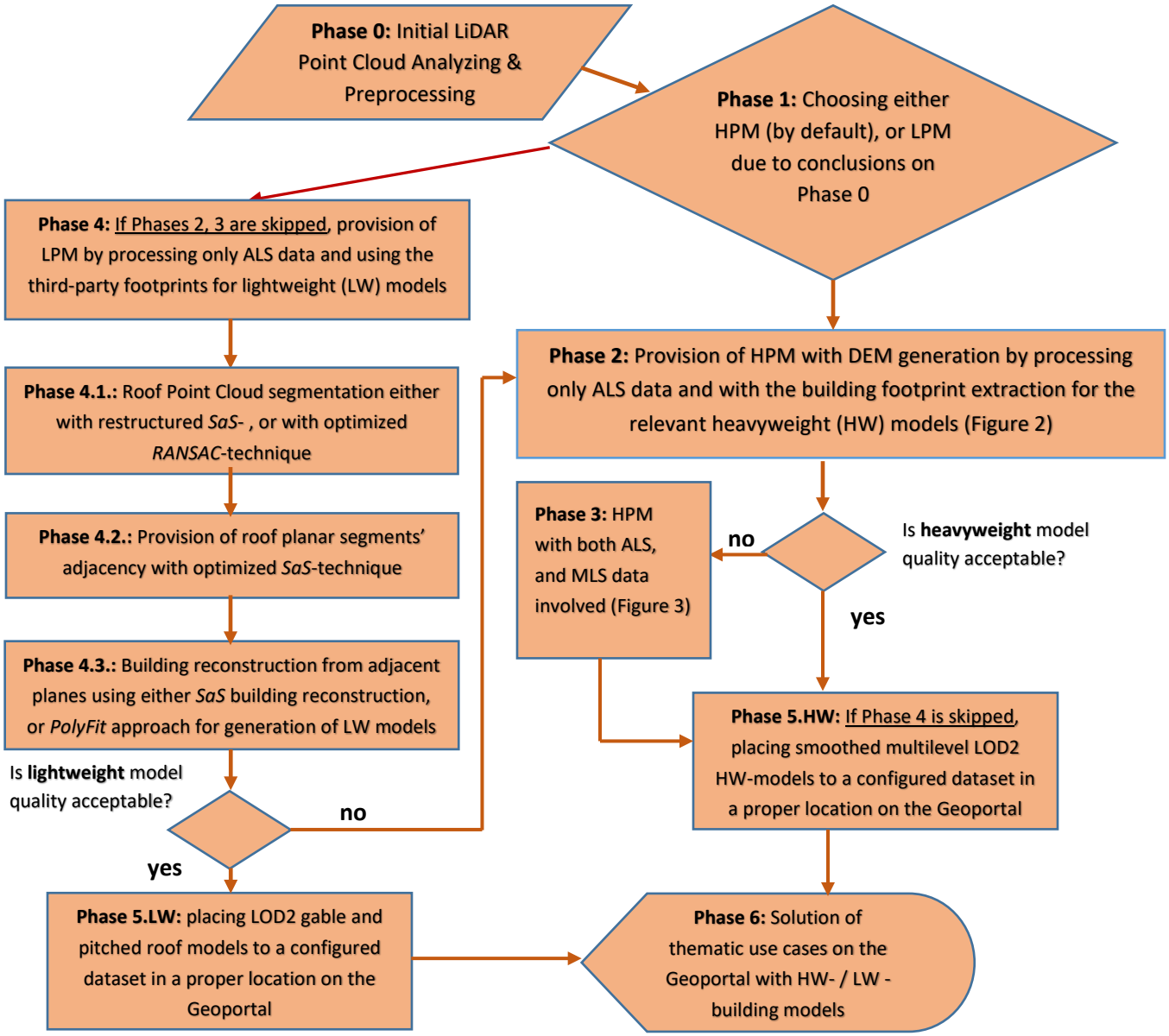


Figure 1: Phased flowchart of the overall methodological approach

geometries and by densities of sets of buildings. Probably two key types of the mentioned urban configurations are *high-rise buildings* of city downtowns and other central districts, on the one hand, and *low-rise buildings* of city suburbs / outskirts and neighboring rural areas, on the other hand. Two methodologically different AFE-techniques outlined in the previous chapter of literature review – the High Polyhedral Modeling of buildings and the Low Polyhedral Modeling – should be applied to these two urban configuration types exclusively – HPM to high-rise building sets, and LPM – to low-rise ones. We can emphasize as a strong point of our overall R&D approach just this joint employment of these two technically different methods within the united building detection, extraction, and modeling methodology, what is resulted further in the relevant software elaboration with two different tools – *Building Extraction (BE*, for a case of HPM), and *Building Extraction Rural Area* (*BERA*, for a case of LPM). Some common features of both approaches and relevant software tools have been already presented in some of our previous publications, although we have not made yet an emphasis on an allocation of both HPM, and LPM within a unified workflow [1, 4, 43, 102]. Since in many cases Airborne LiDAR survey relies on existing territorial regularities of urban areas, it is often provided consequently – from city central parts to outskirts. According to this, if presenting our general approach as *a phased methodological flowchart*, it would be reasonable to accept the HPM as *Previous AFE-Phase* and the LPM - as *Subsequent AFE-Phase* (Figure 1):

It is shown in Figure 1, that an activity diagram combines both HPM, and LPM issues, while an input is a raw LiDAR Point Cloud (*Phase 0). Analyzing & Preprocessing* phase implies evaluating point densities, survey induced and filtering induced errors, as well as checking a for a point cloud proper georeferencing. Two key

factors, that determine a solution to be made on *Phase 1* (either HPM, or LPM chosen) are the territories of specific urban configurations, that are overlapped by .LAS files, and if the third-party (*OpenStreetMap, Microsoft*, etc.) footprints are available for a given area.

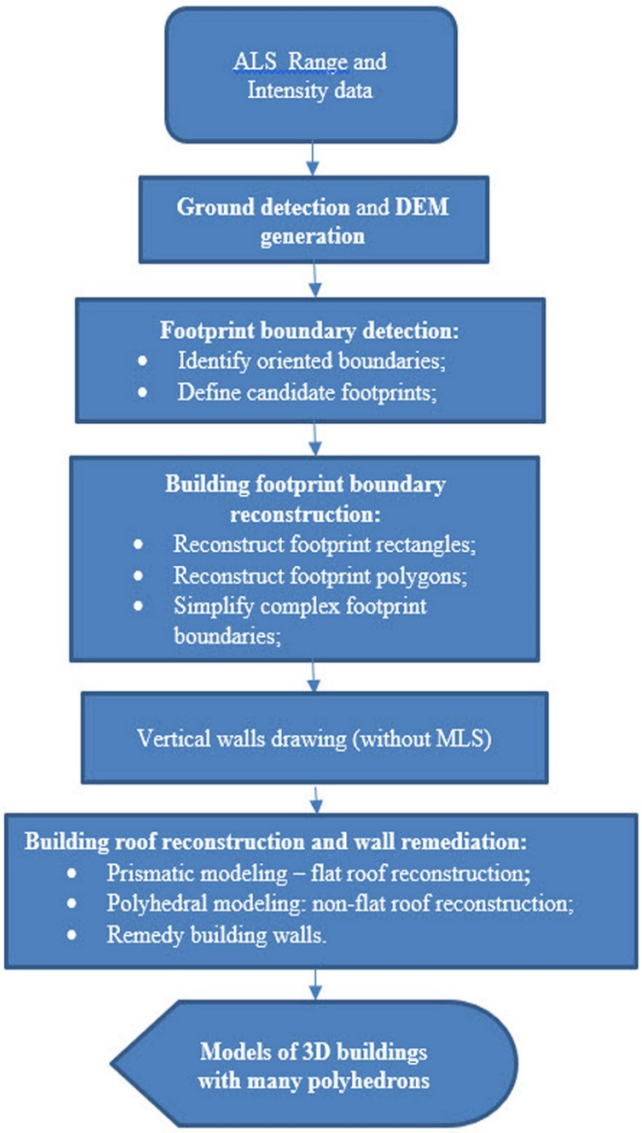
Phase 2 implies a provision of the phased flowchart HPM-branch by processing only ALS data with extraction of the original footprints and with generation of HW-building models, which surfaces consist of numerous polyhedral facets. The output model quality is estimated by several algorithmic evaluating parameters, and in the web-software UI (a user interface of the *Detailed Viewer* tool) by comparison of building surfaces' allocation in 3D space with their neighboring Lidar points. A procedure of such evaluation is illustrated further in this text. If model quality is accepted as either good, or satisfactory, it may be necessary to go directly to *Phase 5.HW*, since both *Phase 3* (involvement of MSL data in addition to ALS ones), and *Phase 4* (a whole LPM-branch) have to be skipped in this phased flowchart scenario. *Phase 3* is accomplished, if MLS data employment can substantially contribute to the quality of HPM-building models with not only precise reconstruction of their facades, but also with increasing of overall sustainability of the high polyhedral model of a building by adding supplementary facets of its minor surfaces. *Phase 5.HW* is also a further target, if acceptable model quality is obtained on *Phase 3*, and, in general, if the most of HW-models are of acceptable quality. Then on *Phase 5.HW* these models are place to a configured dataset to a proper location on our Geoportal. A digital elevation model is generated within the HPM branch only.

Phase 4 provides the LPM-workflow, which starts from *Phase 4.1*, within which surface normal determination as an initial step of roof surface reconstruction within the *SaS*-approach, is accomplished. It is followed by a roof point cloud segmentation either in the updated *SaS*-methodological frameworks, or by a segmentation provided with the optimized *RANSAC*-technique. Both these choices are described further in the text. *Phase 4.2* represents preliminary solutions for the LPM-building reconstruction stage with our original contribution to building a matrix of roof planar segment adjacency. *Phase 4.3* finalizes the building reconstruction stage by combining a building from geometric primitives extracted from a point cloud on the previous stages. If the most of LW-models of buildings with gable and pitched roofs are of acceptable quality, upon *Phase 5.LW* they are placed to a configured dataset of models to the Geoportal (it is defined further in this text). If the quality of the LPM output results is not acceptable, the HPM-pipeline is attempted as an alternative, and a repeated workflow starts again from *Phase 2.* The LPM branch of our overall methodological approach does not support the DEM generation functionality.

An accomplishment of the overall approach is normally finalized on *Stage 6* by solutions with HW- / LW- models of the thematic use cases on the Geoportal, what is briefly illustrated further in our paper.

Optimal LiDAR point density values that were empirically defined for processing techniques involved on *Phases 2* and *4* are between 5 and 140 ALS points per sq. m. If MSL / UAV-LS data are added to our HPM AFE-pipeline, these data are thinned out to acceptable values by sophisticated thinning algorithms, although *Phase 3* deals with up to several hundred of MSL points per sq. m.

A sketch of our core algorithmic workflow of High Polyhedral Modeling is depicted in two flowcharts below (Figures 2, 3). It has been elaborated within the frameworks of the integrated *Building Extraction (BE) / Change Detection (CD) / Digital Elevation Model-Generation (DEM-G)* pipeline of ALS / MLS / UAV-LS data processing. The complete algorithmic workflow depicted on the first flowchart (Figure 2) intends to extract both topographic surface, and urban features from ALS data arrays only, what corresponds to *Phase 2* of the overall methodological flowchart (Figure 1).



3.2. High Polyhedral Modeling for Building Extraction – Heavyweight Models

Thus, a flowchart depicted in Figure 2 does explain the thematic content of the *Phase 2* block of the phased

methodological diagram (Figure 1). Detailed consideration of this phase shows that initially the necessary preprocessing is accomplished within the first algorithmic block (*ALS Range…*). Then all Lidar points are separated for those, that belong to ground, and for other ones, that belong to non-ground features (block *Ground detection…*).

Delineation of the original building footprints (as a valuable alternative to the third-party ones) is a crucially important component of the HPM pipeline (*Phases 1-3, 5* Figure 1), and it is completed within the third block of the HPM ALS algorithmic flowchart (*Footprint boundary detection*) – Figure 2. The footprints extracted are a self-sufficient entity for a whole HPM-ALS pipeline of *Phase 2*, and the following block (*Building footprint boundary reconstruction*) predefines modeling of a whole building, but the preliminary to this reconstruction procedures of footprint boundary optimization and more precise allocation have to be accomplished according to three sub-blocks of this block indicated by the bulleted records (*Reconstruct footprint rectangles, Reconstruct footprint polygons, Simplify complex footprint boundaries*). Thus, through processing upon this fourth algorithmic block building footprints are determined as quadrangles, rectangles, or routine polygons. Then "pretended" walls are arisen from the defined boundaries of footprints (*Vertical walls…*block).

With the next algorithmic block (*Building roof reconstruction….*) the rooftops of buildings may be raised from these delineated boundaries and then corrected by the data of the same ALS point cloud used upon all five previous blocks of this flowchart (Figure 2). A building which has a flat roof is modeled in prismatic geometries (a bulleted record for a subblock *Prismatic modeling – flat roof reconstruction*). If a building possesses some complicated shape of its rooftop, it is modeled as a polyhedron (a bulleted record for a subblock *Polyhedral modeling: non-flat roof reconstruction*). As it was already mentioned above "HPM" means that initially reconstructed building facets consist of many polyhedrons and represent HW-models contrary to that solution ("LPM"), while LW-models can consist of few polygons only. According to the understandable reasons, HW-models are normally constructed and visualized as comparably heavier entities (from 20 to 150 thousand of points are processed per model). Thus, mandatory smoothing and noise removing should be evidently provided. For these aims our original contribution to a Delaunay refinement algorithm [100] has been employed. The corresponding "covering Delaunay TIN" is involved to algorithmic sub-blocks *Polyhedral modeling: non-flat roof reconstruction*, then – to a subblock *Remedy building walls.*

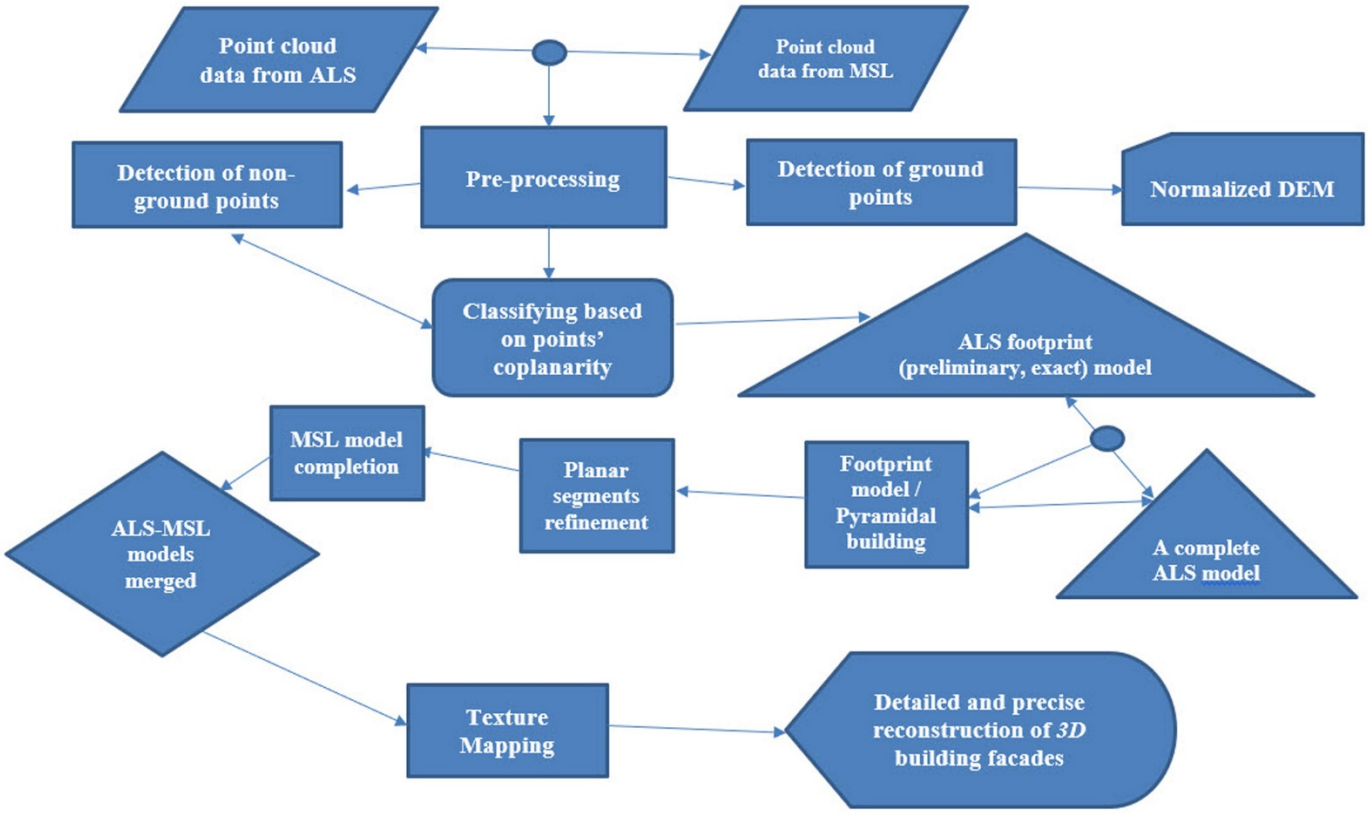
In the presented way we have just introduced the thematic content of *Phase 2*, which is in our methodological flowchart (Figure 1). As it has been emphasized above, if the output model quality (the finalized algorithmic block *Models of 3D buildings with many polyhedrons*) for the results based on the ALS data only is not acceptable *Phase 3* of the phased methodological diagram should be involved.

The initial input data for the HPM ALS / MLS / UAV-LS algorithm, which is core content of *Phase 3*, and which flowchart is presented in Figure 3, are as follows:

- The points that have been received by ALS (**x**, **y**, **z** coordinates and RGB color attributes) Lidar survey. It is geoprocessed as a set of 3D point layers;
- The points that have been received by either MSL, or by UAV-LS scanning, or by both (**x**, **y**, **z** coordinates and RGB color attributes). It is also geoprocessed as a set of 3D points layers;
- The regular earth surface (a grid layer - a DEM), that has been generated upon the first algorithmic steps with existing gaps within those locations, where ground data are absent, is also accepted as an initial input for further processing;
- The smoothed regular earth surface (a refined grid layer) as a DEM with gaps removed by chosen algorithmic procedures, therefore this surface is a continuous one.

On the base of points' distance to a smoothed DEM the HPM ALS / MLS algorithm of *Phase 3* (Figure 3) separates raw Lidar points into two categories, as it is done in those procedures completed by the HPM ALS algorithm of *Phase 2* (Figure 2). The first one contains the ground points that form "a ground level of this building footprint". The second category contains non-ground points that are clustered with this building footprint. The relevant point cluster normally represents a single building or a tree. The upper blocks of the ALS / MLS algorithmic flowchart display all preliminary pre-processing / classifying steps that are provided (*Detection of non-ground* and *ground points*; *Pre-processing; Classifying based on points' coplanarity*; *Delineation of preliminary footprints*) (Figure 3). On further classifying steps any Lidar point is assigned to be either a building point, or a non-building one. According to this binary labeling the primary footprint extraction is implemented as a technique of obtaining building footprint polygons from ALS data exclusively, thus this procedure is completely the same in both algorithmic workflows (Figures 2 and 3). This operation is normally performed in two steps: generating *preliminary footprints* from *existing grid gaps* and extracting the *exact finalized footprints* from these preliminary ones. The preliminary footprints are extracted as *No Data* holes in a grid. It is reasonable to examine this extraction slightly more in details in comparison with other components of both algorithmic workflows. It is the content of the flowchart block *ALS footprint (preliminary, exact) model* (Figure 1). The following modeling-extracting steps are performed for preliminary footprints:

1) An initial grid is divided into the blocks of a size not more than 4'000'000 cells. Neighboring blocks possess an intersecting part of the size, that is specified by the *Max building size* parameter. After this step each building block is processed detachedly;
2) Each cell either is marked as a ground (where the surface grid contains some value), or not classified at all (where a grid has no data);
3) Thus, assuming *expandCount = (params-> MaxCorridorSize / (2 * cellSize) + 1*; We expand the ground class on *expandCount* units in the urban metrics. Afterwards we expand *No Data* class on *expandCount***;** This step separates single long buildings in their ribbon sets connected by urban corridors;
4) Assuming *expandCount = params-> MaxCorniceSize / cellSize + 1*, we expand *No Data* class on *expandCount* to join semi-attached buildings;



5) Consequently, we find gaps in a grid through 1)-4) and each of these gaps retrieves its own index. In this way, only those holes that are completely isolated, i.e. surrounded by the bare ground, are expected to be found;
6) Afterwards, we should check each of gaps and store only those ones, that possess an area larger, than a certain threshold value specified in the preferences;

Thus, we obtain the areal boundary for each gap considered to be a preliminary footprint. The *exact footprints* are generated proceeding from the preliminary ones and from some supplementary information of this point cloud. For every "exact footprint" several following steps are provided.

1) All the points delineated by a geographical extent of some boundary (its size is the algorithmic parameter, which is equal to *MaxCorniceSize*) are selected for processing;
2) All Lidar points, that are above the ground surface, but lower than *2 m* above are classified as the *ground points*;
3) After this, through all points the triangulated network is being built. Within this network all edges, which possess at least one junction, that has been classified as the ground, should be deleted.
4) This triangulated network is divided for several connected segments by removing all the edges, that are longer than *MinDistanceBetweenBuildings.* All network segments that have an area more, than a minimal building footprint value (that is the algorithmic parameter), are assumed to be *the footprints* or *building parcels*;
5) The next algorithmic step is *filtering trees*. There are three possible customized options in this procedure:
a) Tree filtration is *turned off*. In such a case this step is skipped;
b) Applying *preliminary filtration*. The information about *trees* or *non/trees* is obtained from the classified source Lidar file;
c) Applying to a point cloud our own classifying technique.

Upon this fifth algorithmic step, the triangulated network is normally built through all points that are located inside those previously created footprints (preliminary ones). Then all vertical edges associated with this extracted from a point cloud entity are removed. The edge is supposed to be vertical, if its horizontal slope is more, than a certain threshold – a minimal angle, that is a parameter of the algorithm. After this procedure completed, the triangulated network is partitioned for several connected components. Then the points from components, which area is less than *MinBuildingPartArea*, are designated as non-building points, and are removed from the footprint entity.

6) At the next algorithmic step, a new triangulated network is built through the points that belong to building parcels only; moreover, those vertical edges, which 2D length is longer than *MinWallSize* (the key algorithmic parameter) should be removed. The jointed polyhedrons with their aggregated areas, which are more than minimal building parcel area, are accepted as *the footprints*.
7) The final step of the exact footprints extracting is expanding of their preliminary templates obtained through 1)-6). This step is necessary, since some building and infrastructural units, that are above the ground or above other constructive parts, are filtered out on the previous algorithmic step. To

expand a footprint template the iterative algorithm is applied. Upon every step of this algorithm those points, that have not been classified yet as a constructive building or the ground, are classified as *a building* in a case, when this point is nearer to a footprint, than some assigned threshold distance. Afterwards, all these points are added to a footprint, and the new expanded footprints are built as an external boundary, that crosses all points that were previously classified as building points.

After all *exact footprints* are extracted from the preliminary ones (footprint templates), each of them should be checked for intersection with a source preliminary footprint so that to avoid topology errors. All those ones that do no intersect have to be filtered out. At the end of the block *ALS footprint (preliminary*, *exact) model* all Lidar points, that are located inside the examined parcels, and that do intersect, should be marked as the *building points* (in this way avoiding an extraction of other footprints that may be located within the same area). Finally, we accept the exact footprints as some *planar parcels*. The algorithmic step of their refinement is completed at the next algorithmic block *A complete ALS model*. The procedure of the planar segment refinement is quite lengthy therefore it may be a subject of description in another text. Thus, summarizing all stated above, we should emphasize that a preliminary footprint is an entity, which is extrapolated through the ground points without filling corresponding gaps at perspective footprint locations. The exact footprints are built proceeding from preliminary ones and by providing outlined algorithmic steps.

Just as in the HPM ALS algorithmic pipeline of *Phase 2* (Figure 2) our update of a Delaunay refinement algorithm has been employed. The relevant "covering Delaunay TIN" creation is completed in the flowchart blocks of *Planar segment* refinement, *Footprint model / Pyramidal model*, *A complete ALS-model* (Figure 3). All algorithmic blocks mentioned above are provided for processing ALS data only and reconstructing only *HPM-building roofs* and some other supplementary constructive components.

The HPM AFE results of both *Phase 2*, and *Phase 3* obtained by the *Building Extraction* software tool is a set of building models, each of them, as a rule, consists of few façades. The textures may be placed on these façades. Each of these heavyweight models is stored by *.OBJ /.B3DM* ("Batched 3D Model") /*.KML* formats, but in the relevant software environment the main inner format is *.OBJ.* Each HW-model has its six mandatory components:

- *A footprint*. It represents a smoothed *2D* polygon, which allocates a building footprint;
- *A pyramidal model* that is *not draped*. This model consists of several horizontal polygons at different levels. From every layer a vertical building wall is dropped to the previous level. From the lowest level the walls are dropped to the ground level;
- *A draped pyramidal model*. This model is the previous one, but each of its polygons has *a texture*. ALS points are only used for creating textures due to quasi-horizontal polygons ("roofs"). As far as creating textures for vertical polygons ("walls") is concerned, preferably the MSL points should be applied to support a desirable level of details (*LOD*);
- *A complete model*. This model is generated either by using of the detailed TIN techniques on the Airborne Lidar points only (*A complete ALS model* block – Figure 3), or by combining ALS results with the wall segments reconstructed on the base of MSL data and obtaining a combined model (*ALS-MLS models merged*);
- *Texture Mapping:* draping textures on an ALS-MLS merged model;
- *Detailed and precise* reconstruction and visualization of the finalized model with 3D building façades.

The reconstructing operation for building façades is exposed by the *MLS model completion* block of the flowchart in Figure 3. The latter provides a necessary "noise clipping" procedure for building wall surfaces.Since if the distance threshold value, that means a metric length from a given Lidar point to an extracted façade, is either lower, or approximately equal to the points' density value, the reconstructed planar surface may be a set of outliers, e.g., small peaks. Exactly in this case a procedure of noise clipping should be provided.

As it has been aforementioned already, the ALS models, on the one hand, and the MLS models, on the other hand, are merged at a certain point of the flowchart in Figure 3, just after the *Planar segments refinement* and the *MSL model completion* blocks, and before the *Texture mapping* block. This block means the penultimate algorithmic step – texture draping on a merged model. In this way, for each modely component, every polygon in the pyramidal model, the textures are being built. All those Lidar points that are located near the texture extent can be found. These points are projected on the roof / façade plane, and the coloring of the texture pixel is initiated. All predeccor algorithmic blocks in the flowchart of are ended by the *Detailed 3D reconstruction of building façades* one. Finally, the results are being delivered into *.OBJ* and *.B3DM* formats, and a composite high polyhedral model can be displayed by the relevant visualizing software tools developed by the authors (Figure 4). It can be seen, that the HPM AFE results in numerous polygonal segments, which represent quite a rough surface, only while zoomed in. Even a model of numerous polyhedrons can be impressively displayed by our provided visualizing technique.

After completing the HPM-algorithmic workflow, a user obtains building models for display either in the *ElitCore* desktop application, or in the web-based *ELiT* (EOS LiDAR Tool) *Viewer* with three levels of detail (*LODs*): *LOD 0* represents a *model footprint* as its projection on the plane; *LOD 1* is a 3D object that exposes a building as a set of prisms - *a pyramidal model*;

LOD 2 displays smoothed multilevel models in minute details (Figure 5). We have expressed an idea in the introduction to this text, that the composite (ALS / MLS-UAV-LS) model of extracted features may be the most effective one, primarily, with respect to user' various applications in many industrial domains. Such a typically smoothed multilevel model of *LOD2*, a HW-model, may be like follows in the *ElitCore* desktop UI (a user interface) (Figure 5). Thus, despite visualizing rough building surfaces, if significantly zoomed in, while applying to the LiDAR data intentionally selected and refined by the ALS / MLS algorithmic workflow on *Phase 3* of the overall methodological flowchart (Figure 3).

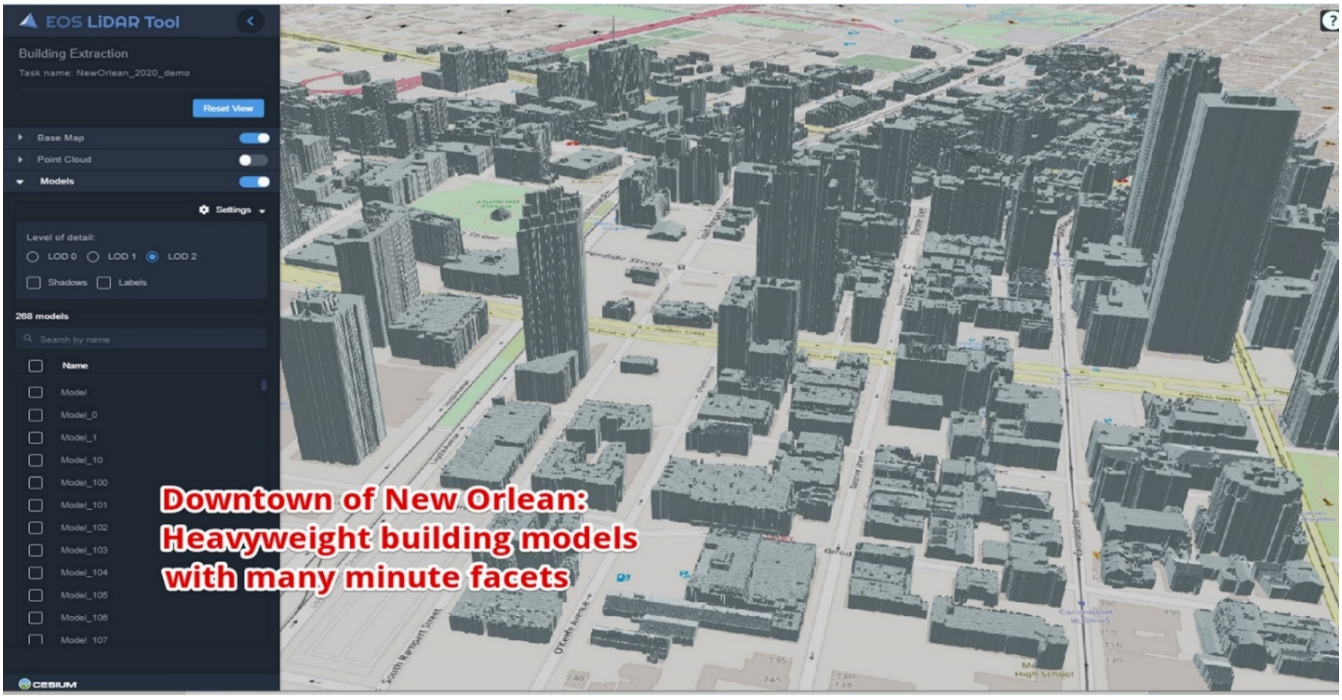


Figure 4: Results of the HPM ALS / MLS workflow displayed in the viewer of the web-based applications – building roofs of many polygonal segments. *The downtown of New Orlean, USA* (a dataset from the *EOS LIDAR Tool* landing page: https://eos.com/eos-lidar)

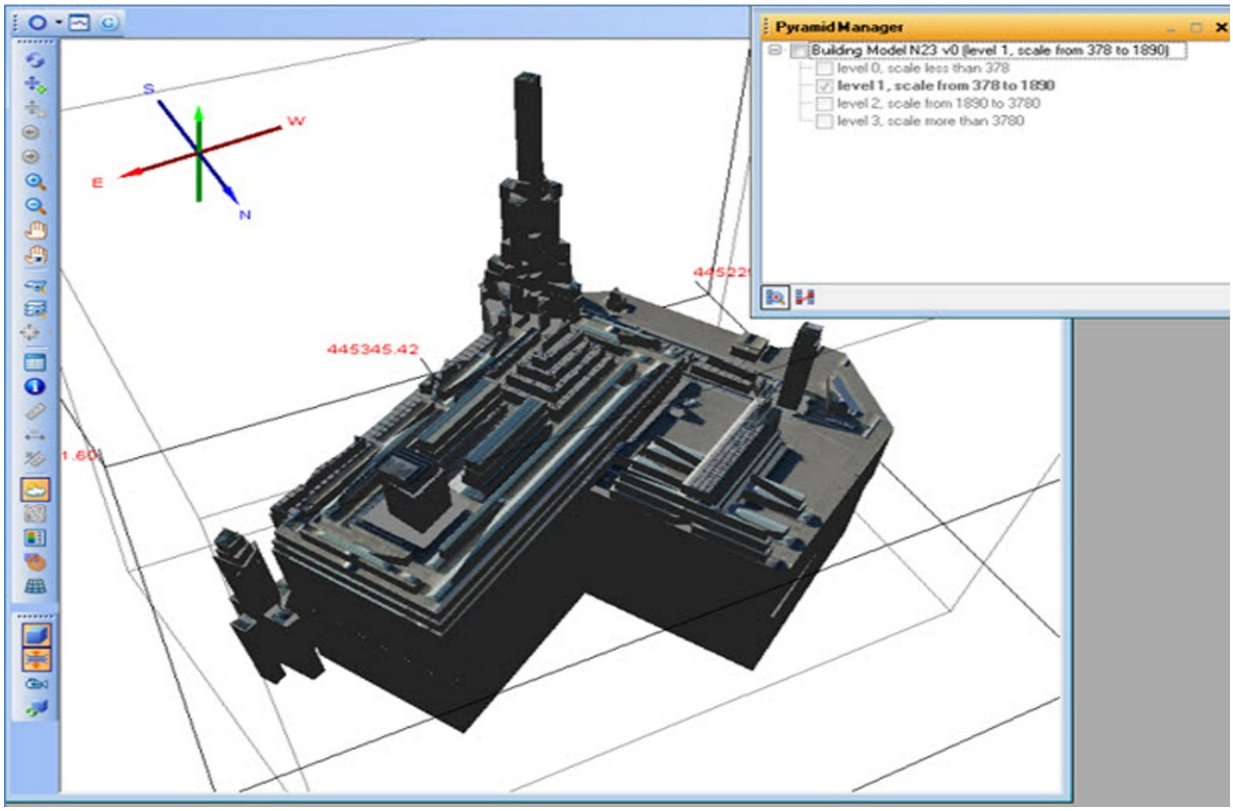


Figure 5: Model of one the historical buildings from the Ottawa downtown in the *ElitCore* desktop user interface

The presented methods of the HPM on *Phases 2, 3* is a completely original AFE-methodology elaborated by the authors. It has been implemented in *Building Extraction (BE)* (Figures 4, 5) and *Change Detection (CD)* (Figure 6) functionalities of our both desktop, and web-based applications.

3.3. Low Polyhedral Modeling for Building Extraction in Rural Areas – Lightweight Models

3.3.1. Common issues

Oppositely to the High Polyhedral Modeling technique our another AFE-approach is the Low Polyhedral Modeling method,

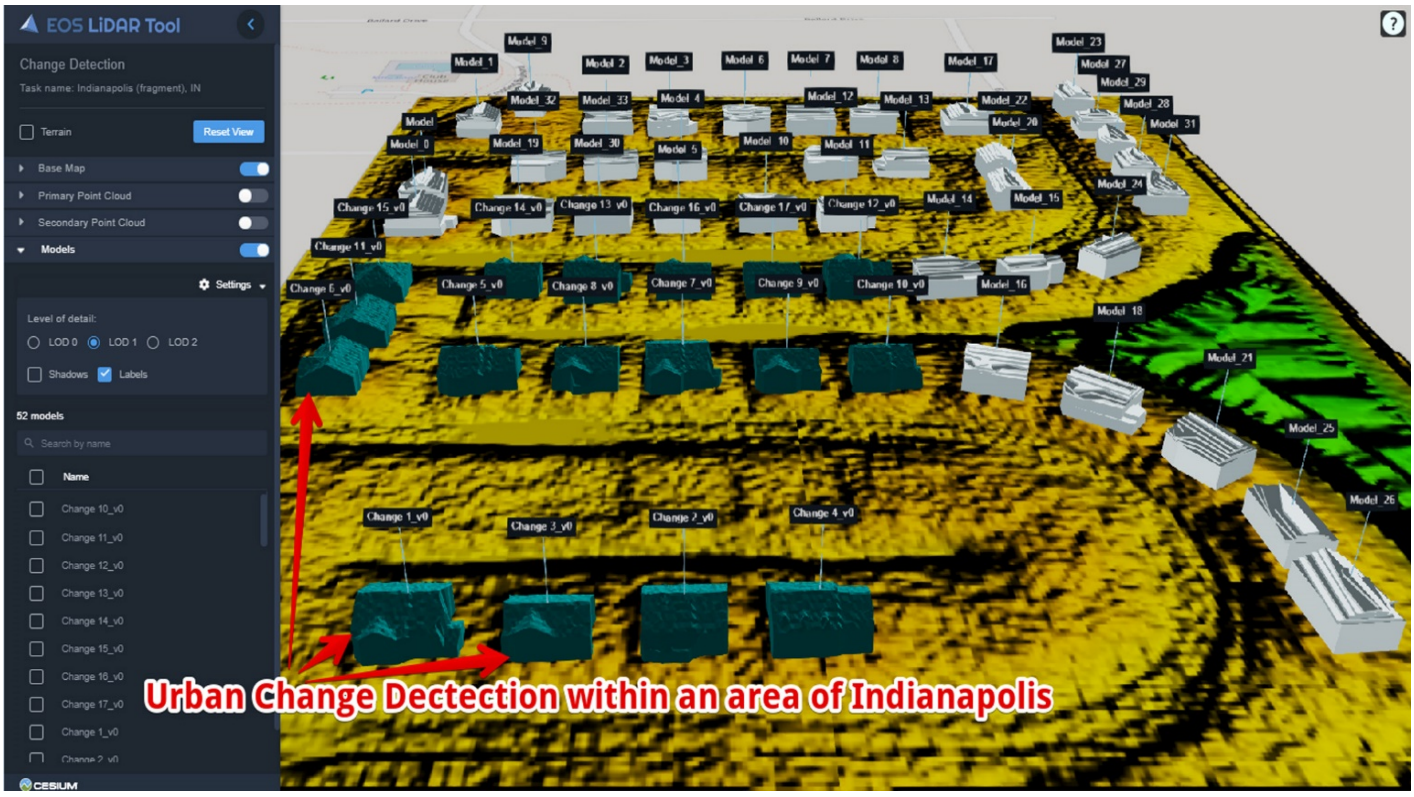


Figure 6: Results of the CD functionality application: new buildings (labeled as *Changes* in dark green color) as urban changes that appeared within one year's time. Models of both new, and old buildings (labeled as *Models* in light grey color) are placed on a DEM generated. *An area of Indianapolis, Indiana, USA* (a dataset from the *EOS LIDAR Tool* landing page: https://eos.com/eos-lidar)

On the finalizing building reconstruction stage theoretical basics of the *PolyFit* [65] have been effectively implemented in the applied solutions with web-software (*Phase 4.3*).

We have mentioned already, that initially the LPM is intended to extract low-rise buildings in either rural areas, or city suburbs. Nonetheless, it has been successfully applied to various urban configuration even within central parcels of the large city areas. Because of the segmenting / clustering procedures that drastically decrease a number of polyhedrons as constituents of a model extracted, such model is titled as "low polyhedral" one. We emphasized already, that these models are extracted, reconstructed, and visualized as the *LW-models* (approximately from 5 to 40 thousand of Lidar points processed per one entity). After completion these models are composed of only few polygonal planar segments. Models are reconstructed in this way, and each of them represents a lightweight modeled feature as a final solution.

Building reconstruction of low-rise constructive features is implemented as a finalizing procedure of building modeling (*Phase 4.3*). This algorithmic stage begins with the adjacency matrix of roof plains generation, that exposes the connectivity of the delineated planar segments. Our original contribution to the formation of the plane adjacency matrix is described in the one of the following subsections of this text.

Both roof interior, and exterior vertexes are determined. Topologically consistent and geometrically correct building models are obtained through implementation of the extended boundary regularization approach. The latter is based on multiple parallel and perpendicular line pairs, and just this approach expedites an achievement of the reliable building models. The model precision can be effectively tuned, despite it understandably still depends on the data precision. In any case the model precision may meet the strict customer requirements. The low polyhedral model implemented corresponds to *LOD 2* of the City GML standards [17, 19, 22]. The following visual presents the main content and some attribute information of this LPM algorithmic result in the interface of our cloud-based software (Figure 5). A number of acceptable models among all LW-models generated is a criterion of going either to *Phase 5.LW* with model output to the Geoportal, or *Phase 2* so that to repeat processing and modeling by shifting to the HPM branch (Figure 1).

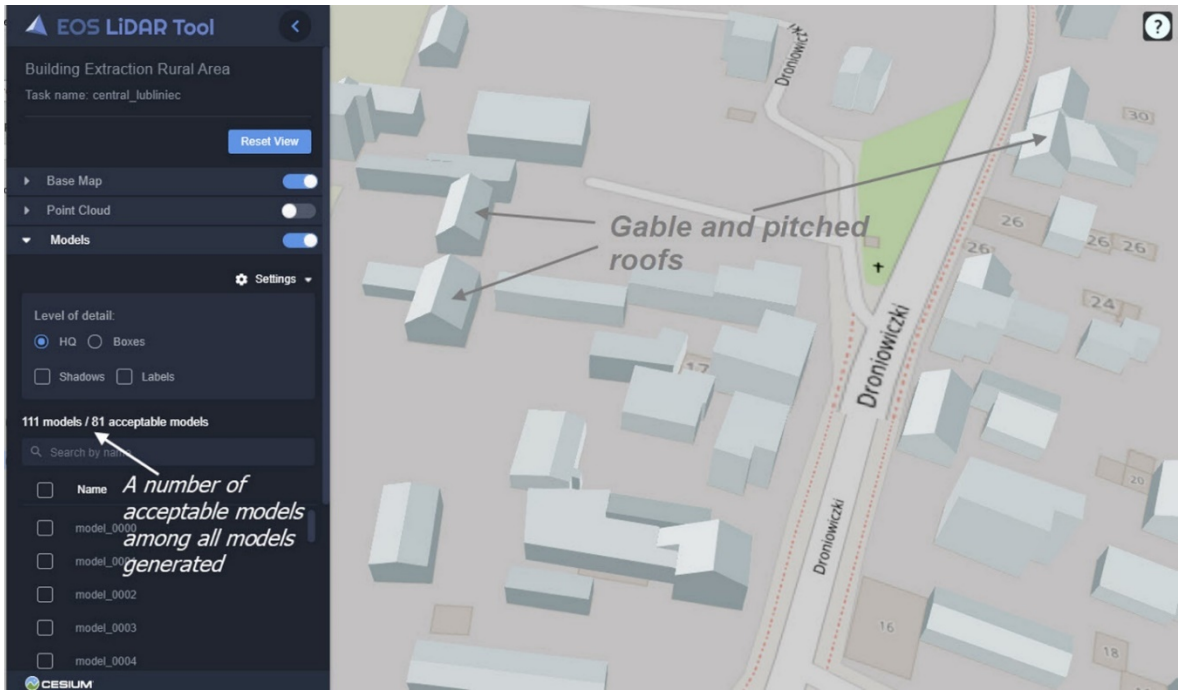


Figure 7: The LW-models of buildings with gable and pitched roofs (pointed to by grey arrows) in a small town. *The city of Lubliniec, Poland* (a dataset from the *EOS LIDAR Tool* landing page: https://eos.com/eos-lidar)

3.3.2. Enhancing SaS-segmentation by restructuring its pipeline

It is an initial content of *Phase 4.1* of the phased methodological flowchart (Figure 1).

2. To Assign unclassified points to the clusters. An addition of formerly unassigned points to a proper cluster might be a key procedure for various configurations of roof individual segment delineation. Some of these points might not be assigned to any cluster due to noise impact (a roof segment under a tree), or because of neighborhood nonplanarity on a seam between two or more roof slopes.

3. To Separate parallel planes inside each cluster. Surface normal vectors are being clustered initially. A clustered entity with the same normal vector is created. Thus, primarily a cluster is a set of planes, and a distribution of parallel planes into separate clusters inside this primary cluster takes place just on this step. The parameter ***D*** (from Item 1) is being computed for all points, and the latter are clustered. The output is a set of planes with the same normal vector, but on the different distance from the center of coordinates.

4. To Split coplanar clusters using the Voronoi neighborhood. A segmented plane may combine those roof segments, that are not connected spatially, but other segments are between them. Such coplanar planes are separated on this step. Commonly, these Items 5 and 6 correspond to the Separation of Parallel and Coplanar Planes step in [37].

5. To Remove small clusters (optionally – using the Voronoi neighborhood). Small area segments can be left after separation of parallel and coplanar clusters. These features should be eliminated as the noise and clustering errors.

6. To Remove the near-vertical clusters, because the modeling technique does not use them due to their unreliability.

It is reasonable to emphasize intentionally that the *Phase 4.1* initial content steps 1-6 can be completed only for a point cloud of some robust density, that may lie within the earlier mentioned range of 5-140 ALS points *per sq. m.*

3.3.3. Providing RANSAC enhancements

It is another content of *Phase 4.1*, and the optimized *RANSAC*-technique can be employed for a roof plane segmentation as an alternative to the *SaS*-approach, which has been somewhat restructured in comparison with an original issue in Item 1-6 in the previous subsection. We have already mentioned in the reviewing section of this text, that the RANSAC algorithm is able to extract a manifold of geometric primitives with different types of their shapes. RANSAC can deal with large number of outliers in the initial data, since this resampling method uses the smallest number of points necessary for estimation a given geometric primitive [62-64, 89, 103, 104]. Thus, the corresponding geometric primitives are obtained, if they approximate definite majority of points.

Some debatable issues present in the initially attempted *SaS*-segmentation, which appeared to be drastically sensitive to various outliers in the point clouds like, e.g., overhanging (above a building) trees. Because of this and due to the Lidar induced errors usually caused by sharp changes in the heights for the points belonging to the same feature, we have finally employed *RANSAC* with some editing of its general scheme. It may assist in meeting the challenges mentioned. For example, when the most points do not have the planar neighborhood, while the normal surface vectors, that are nonetheless found for the point minority, are located randomly, and this does not allow to delineate a corresponding point cluster for plane extraction.

In the outlined case, the *RANSAC* shape extraction strategy gives an opportunity to delineate a planar segment even facing the

challenges mentioned, because the sufficient condition is a randomly drawn point set, that can be quite feasible for this drawing, and employed then for constructing planar segment primitives. 2D planar points and 3D non-planar points can be separated ones from others using this technique. Figure 8 shows the example of a building model, which planes have been segmented with the optimized *RANSAC* from a point cloud with numerous outliers. Then this model has been refined and efficiently reconstructed, and even an overhanging tree, that is comparable in its size with a building, has been removed.

We have selected a following number of the basic *RANSAC* enhancements practicable for the ALS point cloud segmentation and implemented it in the relevant algorithmic pipeline embedded in our web-based software:

- *Number of iterations* upon the planar facet-candidate selection *is not a constant*, but it depends on a qualitative value that is the *Best Current Candidate Plane Index (BCCPI)*, a total number of Lidar points involved, and number of attempts completed yet. The purpose of the BCCPI value introduction is somewhat resembling to the RANSAC "score of the shape" [63], that also provides formalized estimation for the candidates of planar facets.
- Any BCCPI value is defined as *number of points within the delta-neighborhood of a plane minus the "penalty charges" for the point dispersion.* Thus, the less is point scattering from a detected plane, the higher is a qualitative

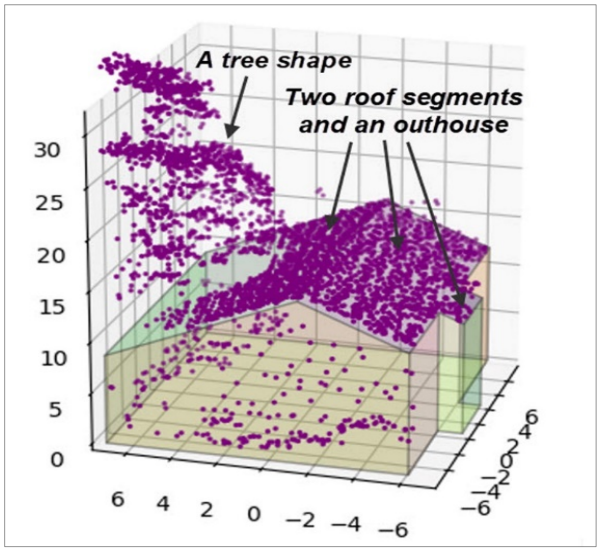


Figure 8: A building model reconstructed from the plains detected and segmented by the optimized *RANSAC*. A model is visualized by the *Matplotlib* python library. A building located in *the city of Lubliniec, Poland*

value of a candidate plane; the bigger is the scattering, the lower is a value mentioned. This technique reasonably allows to prefer those candidates, which approximate well even a relatively small point cluster, but not those random plains, that cross a whole extent of a model and overlap a big number of points in its delta-neighborhood. It is quite significant especially for the *gable roofs* with minor angles between its two slopes, or for *pitched roofs* with small angles within each pair of this roof slopes.

in point clouds with low data noise, but instead, a threshold for the BCCPI values may be intentionally decreased for overlapped strips resulted upon the airborne laser scanning, when a flight changes surveying direction.

- According to the improved *RANSAC* basics the point sampling for a plane-candidate may proceed from a whole point cloud with associated normal vectors, and the technique output is a set of planar primitives with corresponding point sets [105]. Such procedural content may cause some significant inaccuracy itself, while attempting to detect a plane, which may cross a whole cloud as explained above. Instead, we introduce *the point sampling from a randomly chosen limited neighborhood.* On the one hand, with this selection we still have the chances to approximate a big planar facet, if one does exist, on the other hand, we are expected to approximate some relatively small planar segment precisely enough, if a big one does not actually exist. If sampling neighborhood reduction is not provided, the probability that all three points, *a point triple*, would approximate (a key criterion of plane detection) one small planar facet has a low value.
- A point triple should be checked for *the degeneracy*: if an area of their triangle is too small, then the probability of processing error is high enough, and the whole computation cycle may fail. Therefore, *such triple is rejected*, and another one is sampled.

The overall structure of our optimizations of *RANSAC* is presented in the following *Algorithm Pseudo-Code* (Figure 9):

1. Main cycle:

```
working_set := input_points
result_planes := []
WHILE len(working_set) < threshold:
  plane_parameters, remaining_points := fit_plane_ransac(working_set)
   result_planes.append(plane_parameters)
   working_set := remaining_points
```

2. Selecting candidate plane function (fit_plane_ransac)

```
best_sample_quality := 0
best_sample_size := 0
best_sample := None
iteration := 0
WHILE probability(best_sample_size, points, iteration) <
quality_threshold:
  IF iteration >= iteration_threshold:
    reduce_dispersion_influence_on_quality_estimation()
```

```
  IF check_quality_limit_reached()
    BREAK
  iteration := 0
 p1, p2, p3 := select_candidate_points(points)
 iteration := iteration + 1
 IF NOT check_candidate_area(p1, p2, p3):
  CONTINUE
 candidate_sample, candidate_quality := get_best_candidate(p1, p2,
p3, points)
 IF candidate_quality > best_sample_quality:
  best_sample_quality := candidate_quality
  best_sample_size := size(candidate_sample)
  best_sample := candidate_sample
```

```
3. Assessing candidate quality function – BCCPI value
(get_best_candidate)
plane_parameters, sample_points := fit_plane(p1, p2, p3, points)
points_distance := get_distance_to_plane(plane_parameters,
sample_points)
distance_dispersion := mean(points_distance)
quality := quality_influence * (len(sample_points) / len(points)) + (1 -
quality_influence) * (1 - dispersion / plane capture_distance)
```

Figure 9: The Pseudocode of the optimized *RANSAC*

It has been determined that the optimal initial value of quality_influence may be 0.4, while the iteration step in reduce_dispersion_influence_on_quality_estimation is 0.1.

3.3.4. Optimizing SaS-reconstruction by Voronoi Neighborhood

One of our key contributions to the SAS-methodology corresponds to *Phase 4.2* of the phased methodological diagram (Figure 1). It consists in the refinement of the optimized adjacency matrix obtainment on the preliminary stage of building reconstruction. This matrix is the most significant issue for the delineation of the adjacent planar segments in the model generated. Thus, we have enhanced the *SaS*-workflow by *the extensive use of the Voronoi neighborhood* for computation of the roof segment cluster adjacency on the building *reconstruction stage.* Although the authors of the original workflow referred to the Voronoi diagram only on the point cloud *segmentation step* – the Voronoi neighborhood V_p has been employed for the surface normal computation only [37, 76]. Applying to the Voronoi diagram on the building reconstruction stage, we remove both horizontal, and vertical gaps in the processed data, as well as mitigate the nonhomogeneous point density of a primary point cloud.

The Voronoi diagram has been applied for the roof cluster adjacency determination and for separation of coplanar clusters, while the limited Voronoi diagram has been used for avoiding the side effects of the cluster adjacency determination. Also, the Voronoi diagram has been applied for the reliable identification of some traditional architectural constituents such as building awnings and or building overhands [102].

The Voronoi neighborhood for the roof cluster adjacency determination means the obtainment of the planar segment optimized adjacency, on the condition that these segments are preliminary delineated. The authors of the *SaS*-methodology use the Voronoi diagram while providing eigenvalue analysis, and each point of neighbors are being delineated. Nonetheless upon computing the cluster adjacency these authors apply a routine distance between all pairs of points using the following formula [37, P. 1562]:

$$d\,(P,Q) = min\,(d\,(p_i,q_j))\ \forall p_i \in P;\ \forall q_j \in Q, \qquad (1)$$

where: $\boldsymbol{d(p_i,q_j)}$ is a distance between any pair of points $\boldsymbol{p}_i$ and $\boldsymbol{q}_j$, which belongs to two different clusters $\boldsymbol{P}$ and $\boldsymbol{Q}$ correspondingly. The problematic issue of the *SaS*-approach like presented with (1) proceeds from a case of nonhomogeneous point cloud density that is a subject for clustering and segmenting. E.g., a sparse point cloud, which is also worsened by the faults of surveying technique, reasoned the case, according to which clustered points appear far from the boundary of a cluster. In such a case a distance value (from (1)) fails to be proved, if it is checked by a threshold parameter. Consequently, this may be a reason for the errored adjacency determination and for the consequent wrong model reconstruction. Nonetheless, it may be not so wrong to conclude, that because there are no Lidar points between two delineated planar segments, which would belong to other clusters, these two segments are sooner adjacent, than not and their seeming "non-connectivity" might be caused by the gaps and outliers in Lidar data only. This problem may be resolved by the adjacency determination with the Voronoi neighborhood.

The possible solution can be based on the following assumption. Even in a case, when two points lie far from each other, and there are no those points between them, which may belong to the third cluster (besides a given pair of clusters), and the Voronoi cells of these two points have common edges, then this pair of points can be determined by their Voronoi neighbors. In another case, when there are points of one more cluster between a given pair of points, these points are not considered as the adjacent ones even, if they pass a comparative test by applying a threshold parameter large enough.

On the base of all mentioned above and referring to an existing relevant example [106], we assume to consider as *adjacent ones* only those *clusters*, which points are *neighbors* within the Voronoi neighborhood. The only criterion is a case, when their Voronoi cells possess a *common edge*. Thus, we both solve a problem of data scarcity, gaps / outliers, and find a solution for a sparse point cloud. What is more, we get rid of a necessity of the enlarged threshold value introduction, but implement its definition by interpolating technique instead, since this value should be both big enough (not to remove the actual adjacent clusters), and small enough (not to define non-adjacent clusters as adjacent ones). In this way, we can significantly increase the applicability of our approach, and decrease its narrowness, e.g., when it strongly depends on point density and on the equitability or at least on the similarity of spatial distribution of the points that belong to two clusters. Moreover, the search of the Voronoi neighbors can be completed faster with such approach, because each point has a computed list of neighbors to be checked for their spatial acceptability. Thus, the overall algorithmic efficiency can be increased. It is only the first aspect of the overall LPM optimizing. The second one, that is not a less important solution, is introduced in the following paragraph.

While arranging the matrix adjacency in the frameworks of the *SaS*-technique, we define the distance between two clusters of planar points ***d* (*P*, *Q*)**, as the minimum of all point possible combinations between two clusters distance (1). Then, $m = |P|$ is a point number in the first cluster, and $n = |Q|$ - a point number in the second cluster. Hence, a whole number of possible combinations to be computed in such a case is the $m \cdot n$ value, the product of the point numbers.

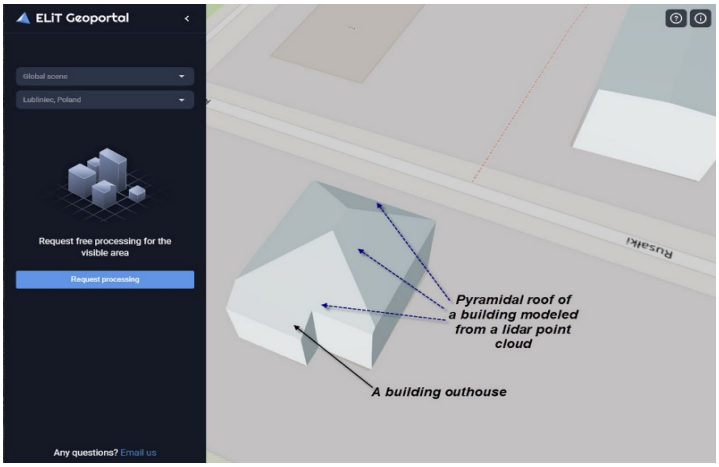


Figure 10: A lightweight model of a building with the pyramidal roof and an outhouse reconstructed at the *ElIT Geoportal* in the location of *the city of Lubliniec, Poland*

At this algorithmic step we suggest an alternative solution: to extend each of two clusters of planar points up to their mutual intersection, while a relevant linear segment of the cluster intersection is generated. Then we compute the distance between two clusters as the minimal distance of all possible measured combinations between points belonging to each of two clusters, from one side, and this linear segment of a cluster intersection, from another one. The statistical significance of difference between two mentioned values can be estimated. A combination number in this case becomes substantially fewer: the $m + n$ value only, the sum, but not the product. It defines the significant simplification of the overall LPM algorithmic complexity from the quadratic complexity $O(n^2)$ (a case of the *point product* that defines number of combinations) to the linear complexity $O(n)$ (a case of the *point sum*, that defines number of combinations). Decreased complexity provides the overall enhancement of the LPM-algorithmic efficiency.

It is evident, that the better a roof segment matrix is optimized, the more sustainable number of geometric primitives (e.g., vertices) is necessary for the robust building roof reconstruction. The introduced update of the adjacency matrix computation provides the definition of some threshold levels between two clusters of coplanar points within a footprint boundary, which is accepted as a geometric analogue of the roof edge. These clusters should never intersect, and this circumstance causes an indefinite exit of computing guaranteed, if the traditional approach is attempted to provide. Otherwise, the updated LPM technique, in its turn, completes the modeled topological sustainability and geometric correctness of a building footprint and of its roof as shown in Figure 10. It displays a model example from a fragment of one from modeled *CityGML LOD2* locations placed at our internet-resource – *ELiT (EOS LiDAR Tool) Geoportal* (http://elit-portal.eos.com/). The presented lightweight model has been selected for display, just because the relevant point cloud possesses quite a few non-intersecting planar segments, that is why the routine *SaS*-algorithm of reconstruction has failed while processing this data, but its update has succeeded – Figure10.

Phase 4.1 and partially *Phase 4.2* examined above deal with the point cloud segmentation stage in an overall processing workflow. Correspondingly, our prime LPM-concern has been a set of the roof segmentation operations, as it has been previously discussed for these two phases of the methodological diagram. *Phase 4.3* provides the finalized stage of building modeling – its reconstruction, while *Phase 4.2* grounds some substantial premises for it. We consider the urban feature reconstruction stage as presentation of building geometry and topology within a certain *LOD* with a defined number of feature segments. That is why we have attempted to implement in *Phase 4.3* one of the most interesting among other relevant theoretical technique – the polyhedral (polygonal) surface reconstruction (the *PolyFit* approach already mentioned in the literature review) [65, 107].

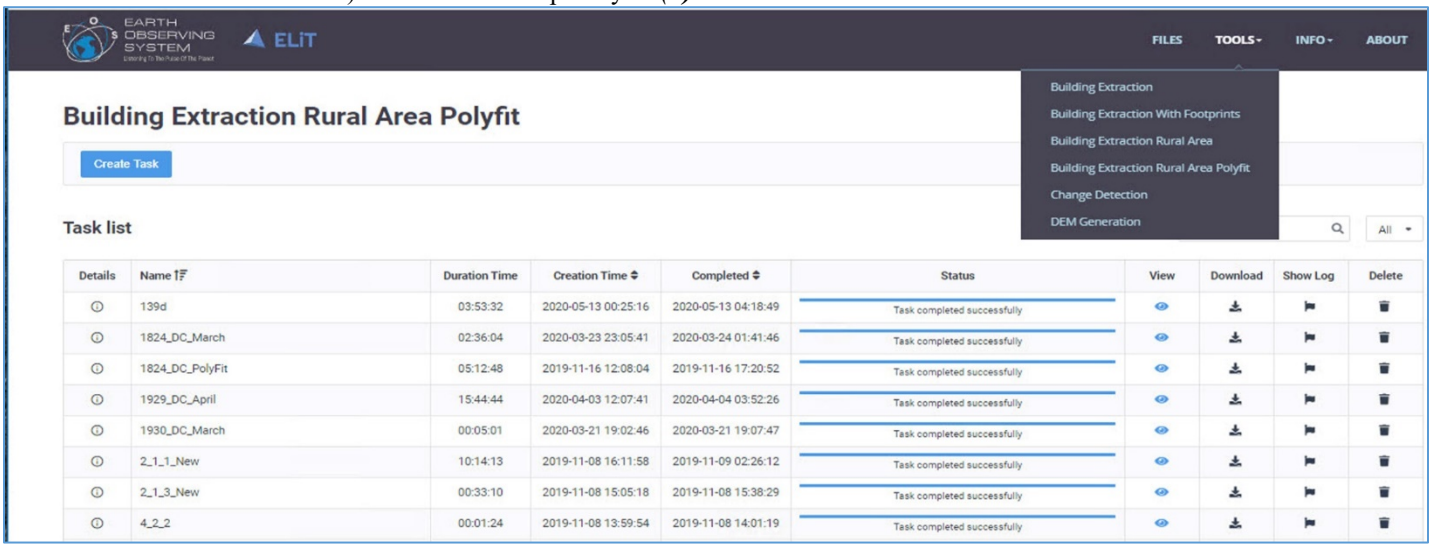


Figure 11: The *PolyFit* page selected among a list of other AFE-techniques in the menu *TOOLS* of the web-based software interface

This approach demonstrates, that reconstruction of geometric primitives from a point cloud can be treated as a binary labeling problem (what is, by the way, close enough to "global solutions" in the point cloud processing [87, 91]). The method proceeds from the hypothesizing and selecting strategy and provides intersecting the building planes already segmented, then seeking for a necessary combination of intersected planes and getting a manifold polygonal surface model. Since in its segmentation stage the method is strongly based exactly on the *RANSAC* algorithm for plane detection, then by enhancing *RANSAC* plane primitive extraction and refinement, what has been introduced in this text above, we understandably contribute to making the *PolyFit* solution more robust.

Therefore, constructing a whole pipeline for the *PolyFit* implementation in our web-based software, we attempted to employ on a relevant segmentation step both the original advances of the *SaS*-approach, and possibilities of enhanced *RANSAC*. Thus, the *PolyFit* appears to become one more AFE-option in the relevant menu option of our web-software package (Figure 11).

4. Results: Software Elaborated for AFE Purposes

The applied software solutions, that can be elaborated on the base of the presented multifunctional methodological approach to the generation of heavyweight and lightweight AFE-models (Figure 1) can be introduced as the robust results of this research. These software solutions are completed with the functionalities, which can provide the efficient support and realization of all three groups of advantages referred to above.

4.1. Desktop and Web-Based Applications

The ***ELiT*** (**E**OS **Li**dar **T**ool) Web-Based software is both a separate web-based (network) generator - *ELiT Server* (a landing demo-page available at https://eos.com/eos-lidar/), and an integrated component of EOS Platform-as-a-Service software - *ELiT Cloud.* The latter is the basis for *ELIT Geoportal* - a type of web portal used to find, access, and process LiDAR geospatial information, both primary, and resulted one, as well as provide the associated geographic services (display, editing, analysis, etc.) though the Web (http://elit-portal.eos.com/). Few examples of the *ELiT* Server interface have been already presented above (Figures 4, 6,7, 10, 11).

The congeneric to these two products is our desktop *ElitCore* package, which possesses even somewhat broader functionality, but can produce only the heavyweight polyhedral models (Figure 5). The *Building Extraction* functionality implemented in the standalone *ElitCore* provides detection, extraction, and reconstruction of heavyweight models according to our general methodological workflow of the HPM AFE (Figure 1) and its algorithmic content of *Phases 2* and *3* (Figures 2 and 3). It normally results in a DEM ("a bare ground class" delineated from an initial point cloud) and a set of urban features which are the discontinuous objects.

All three mentioned software products are based on the sophisticated HPM / LPM solutions based on those complicated algorithms for Lidar data processing that have been introduced in the previous section of this text. Both *ELiT Server*, which has been implemented as a distributed information system, and *ELiT Geoportal* usually operate from the Web browser installed on a user's workstation.

ELIT 3D city models as many other 3D models appear to be predominantly used for display; however, they are being increasingly employed in number of subject areas and for a large range of tasks beyond the visualization. In our previous publications, we sought to understand and present several use cases regarding their presentation not only on the *ELiT* Server, but also on the *ELiT* Geoportal with further utilization of these City GML LOD1 models for some tasks within and beyond model display [4, 102].

Our web-based software implies provision of the following *ELiT* tools as the functionalities of the distributed information system (these tools are shown in Figure 11 above):

- The *Building Extraction (BE)* tool (a sub-page *Building Extraction* of the *Tools* page – Figure 11) provides the HPM building detection, extraction, and reconstruction through various sub-procedures pursuing the final sub-procedure of building reconstruction for high-rise buildings, preferably in city downtowns. Thus, the *BE*-functionality accomplishes the generation of heavyweight models, which have been already grounded and presented above (Figures 2-4, 6). The building extraction functionality is that option, according to which any single building construction may be detected, extracted, and reconstructed as a heavyweight model with further visualization by the open source *Cesium 3DTiles* library with a certain *LOD* and with relevant spatial, geometric, and semantic attributes. Key components of the massive urban environment can be simulated as the heavyweight models with minor details (Figure 12);

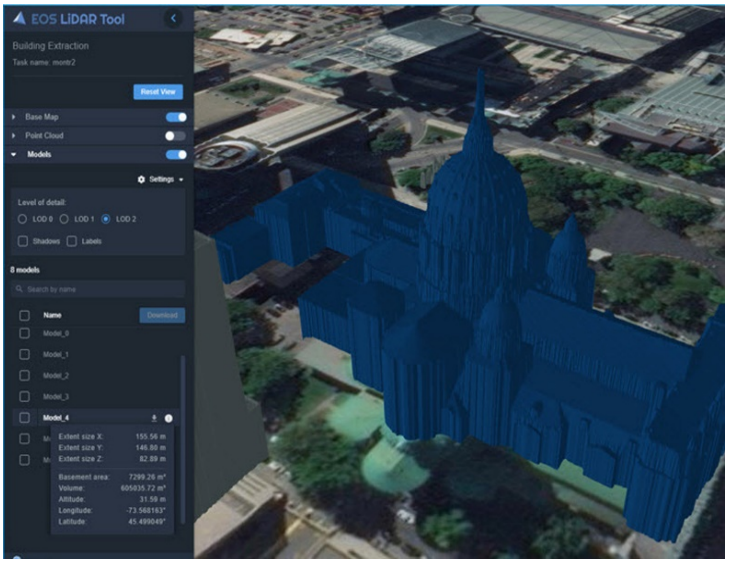


Figure 12: A historical building in *a downtown of Ottawa, Canada*, visualized in the *ELiT Server* interface

- The *Building Extraction with Footprints (BEwFP)* tool (a sub-page *Building Extraction with Footprints* of the *Tools* page – Figure 11) accomplishes the same HPM building extraction as the *BE*-tool producing the heavyweight building models. Oppositely to the previous tool, the *BEwFP*-functionality involves the third-party data – building footprints (global maps as *Open Street Maps*, *ESRI Imageries,*

Microsoft footprints, etc.) instead of accomplishing its own *ELiT* footprint delineation described in 3.2 Section of this text. Normally the footprints input is an archived ESRI *SHP* vector file;

- The *Building Extraction Rural Area* tool (*BERA* - a sub-page *Building Extraction Rural Area* located on the *Tools* page – Figure 11) provides "a classical pipeline", by which we implement the *SaS* point cloud segmentation / building reconstruction methods [37] of low polyhedral modeling introduced in *Phase 4.3* of the methodological flowchart (Figure 1). According to the *SaS*-pipeline and to its optimization described in subsections 3.3.2 and 3.3.4 the *BERA* tool accomplishes the hierarchical segmentation of a Lidar point cloud, refinement, and separation of extracted planes with further building reconstruction in rural areas and urban suburbs. As it has been already mentioned above (Subsection 2.2), the generated lightweight building models consist of not so many facets, and number of points involved in a single model generation is limited by a number from five to thirty thousand. Necessary limitation may be met by adaptive thinning at the cost of modeled details, if an initial value ranks over this quantity. The correspondence of a resulted model to an initial point cloud can be visually assessed for a conclusion, if it is modeled accurately or not, and if it is allowed to go to *Phase 5.LW* , or it is necessary to go back to *Phase 2*, and provide HPM instead of with LPM (Figure 1). This quality estimation is provided with the *ELiT Detailed Viewer* (Figure 13). In case of acceptable quality, the point locations should not deviate on more than a certain distance threshold value from the modeled surfaces. An exact threshold is introduced, basing on the point cloud characteristics, e.g., on its density.
- The *Building Extraction Rural Area Polyfit* tool (*BERA-P* - a sub-page *Building Extraction Rural Area Polyfit* located on the *Tools* page – Figure 7) implements somewhat another algorithmic pipeline in comparison with the *BERA* tool. The *RANSAC* optimized clustering and segmentation is used on the initial building modeling (subsection 3.3.3). The hypothesizing and sampling strategy of the *PolyFit* basics [65] together with its finalized reconstruction of polyhedral models is applied to concluding building modeling. Our optimization of the *RANSAC* segmentation described above is embedded in the relevant workflow, that fulfils the *BERA-P* tool. Its processing may take somewhat longer time, than the "ordinary" *BERA* tool does, but usually provides an increase of number of topologically correct and geometrically precise models among all output models in a dataset.
- The *Change Detection* tool (*CD* - a sub-page *Change Detection*, which is on the *Tools* page – Figure 11) automatically detects block-, district-, and city- scope alterations in an AOI selected. Changes in urban architectural morphology usually happen through city areas over some significant period, if only it is not any catastrophic event of either environmental, or social demolition. The *ELiT* CD tool detects locations of changes in positions and shapes of buildings and infrastructures as 3D models with additional spatial information. Usually two Point Clouds (the temporally primary, and the temporally secondary ones) are compared. It is not *BERA*, but the *BE*-functionality is the only one that is involved in preparing two resulted modeled sets. The difference between two input point clouds is computed as the *BE*-output of the high polyhedral modeling (Figure 6).

All those already listed *ELiT* functionalities are supported by the relevant web-API and by the Front-end web-interface for routine procedures of uploading, processing, and analyzing the initial data and for further downloading derivative data. The sets of modeled *.OBJ* and. *B3DM* entities (the latter are *Cesium 3D*

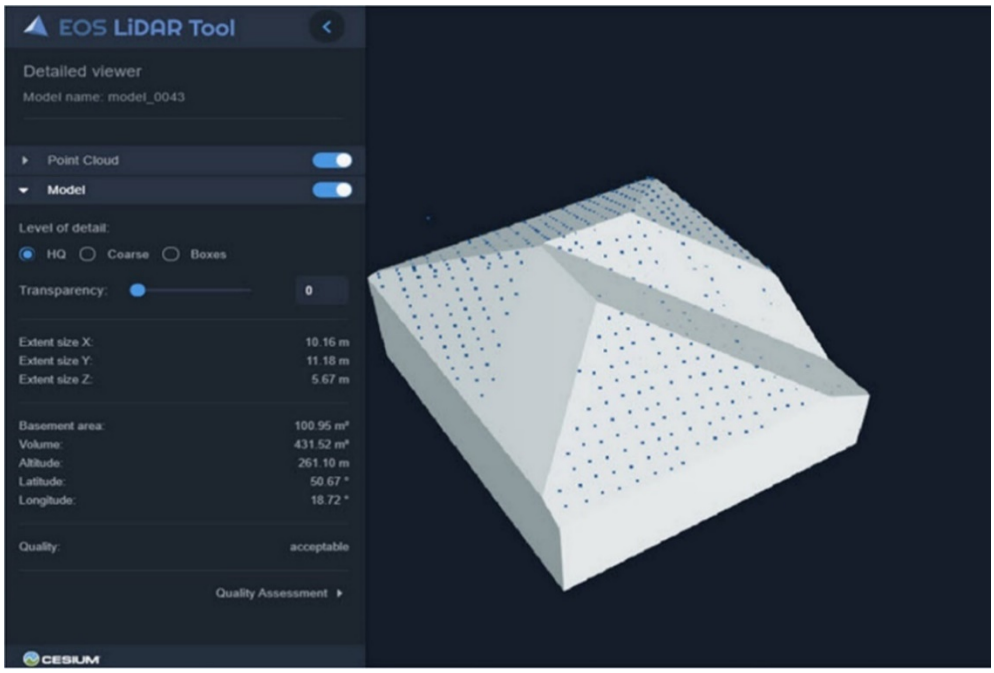


Figure 13: The *Detailed Viewer* Tool for examining accuracy of a lightweight model according to point cloud content

Tiles component) have to be delivered to 3D models, and can be stored and visualized in *.COLLADA, .KML,* and *glTF* formats.

According to the content of *3D Tiles* specification a *.glTF* is employed by *.B3DM* as *its payload* for delivering 3D building geometry. If data processed by *ELiT Server*, the results of all functionalities, *BE, BEwFP, BERA, BERA-P*, and *Change Detection,* are displayed by the *ELiT Viewer.*

- The *DEM Generation* tool (*DEM-G* - a sub-page of the *Tools* page Figure 11) runs a generation of a field of topographic heights making *a DEM / DSM grid* (Figure 17). It means we can create a gridded surface from sample data, what is commonly referred to as the interpolation. According to the contemporary basics of topographic modeling the most accurate uniformly spaced DEM matrix can be compiled directly either from photogrammetric processing of remote sensing imageries, or by processing results of LiDAR surveys [8]. In other words, a series of irregularly spaced height points are acquired, from which uniformly spaced elevation benchmarks are interpolated. A created DEM we understand as a synonym of a digital terrain model, which simulates the bare earth surface with uniformly spaced *z*-values of heights within a DEM neighborhood. Our automated processing for generation of a DEM of high precision provides modeling topography of various genetic types, e.g., sand-eolian (as shown in Figure 14), glacial, fluvial, volcanic, topography of human impact, etc. Results of the *DEM-G* functionality are visualized in *ELiT Viewer* by the *TerrainProvider* option of the open source *Cesium JavaScript* library (Figure 14).

4.2. Cloud-Based Application – ELiT Geoportal

A common definition of the *EliT Geoportal* (*EGP*) as a *cloud-based* application has been already done in the previous subsection of this text. We already explained in one of our previous publication the necessity of a geoportal as a robust solution for the web-GIS software promotion [102]. *EGP* is a kind of web portal used to transform occasional visitors of the EOS LiDAR web-site into the warm leads. The *EGP* site can be used to find and access the derivative Lidar information and relevant services for number of thematic use cases, that can be resolved through geospatial modeling, visualization, and GIS-analysis.

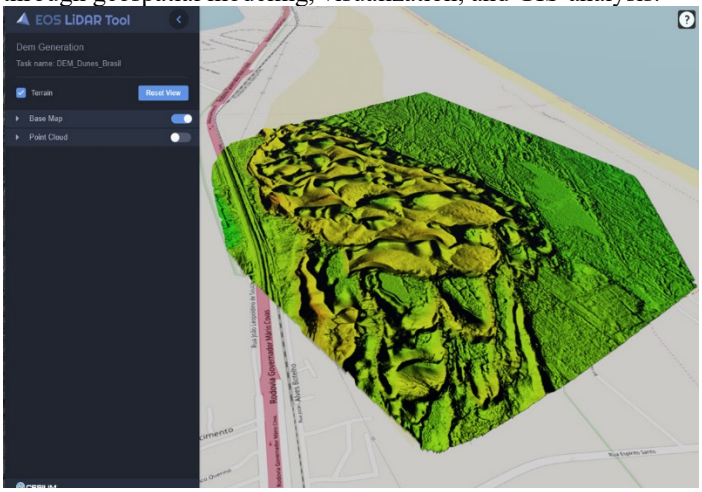


Figure 14: Topography of sand dunes in Brazil visualized in the *ELiT Viewer* interface (a dataset from the *EOS LIDAR Tool* landing page: https://eos.com/eos-lidar)

Thus, the Geoportal can be also defined as the Internet-toolbox for resolving local, regional, national, and even globalized urban problems, i.e., by performing at very different scales. A routine Web-based map-server software usually allows only backups of mapping results. Nonetheless, for various domains of Urban Studies, which can be provided with our research and modeling technique [1], we do need a Geoportal Internet-toolbox associated with a web-based Geodatabase, that would allow us to sort, query, update, and delete both different features and data associated with these features. Such architectural and structural web-software scheme differs towards its more efficiency if compared with solutions of a common case. Within these solutions some web-based GIS application generator is developed, and it provides on the base of the routine database scheme the number of web scripts. Then it deals with a map interface and with a database, as well as with provision of interactive maps defined by a user [19, 108, 109]. Otherwise, the EGP may support an effective integration of a Web-based GIS interface with a Geodatabase chosen within a web-architecture. Spatial references and 3D regional layers are automatically chosen for a given territory. What is more, semi-automatic generation of Web 3D Scenes according to a selected AOI can be provided. Owing to this architectural scheme, series of thematic EGP use cases are provided (http://elit-portal.eos.com/): *Population estimation in an AOI based on 3D building models; Energy consumption estimation based on 3D building models; Visibility analysis in urban environment for various city development scenarios' evaluation and comparison.*

5. Discussion

Choosing either the high polyhedral modeling, or the low polyhedral one with corresponding generation of either HW-, or LW-models may be selected as introductory topic of the discussion, which summarizes our research.

As explained above, in a common case, *Phase 1* of the methodological flowchart (Figure 1) suggests either one, or another option according to environmental and urban configurations of a selected AOI.

The LW-models are built exclusively on the base of ALS surveys together with input of the third-party footprints (Figure 7). The lightweight model accuracy exclusively depends on the segmentation quality and the reconstruction robustness. What is more, exactly clustering / segmenting procedures significantly define consequent LM-results. Application of the originally optimized *RANSAC* (Figure 9) has resulted not only in removal of

a point cloud outlier caused by a tree (Figure 8), but also in predetermined appearance of a building outhouse on the reconstruction stage in a finalized model, what cannot be guaranteed with the routine approach of segmentation (Figure 10).

The comparative evaluation of optimized clustering/segmenting technique with both restructuring *SaS*-pipeline (subsection 3.3.2), and with enhancing *RANSAC* (subsection 3.3.3) has been provided on the base of several ALS data surveys accomplished over the territory of the city of Lubliniec, Poland, in 2017-2018 (Figures 7, 8, 9). Table 1 is produced to compare the modeled results of the methods mentioned. Three different values of the key *RANSAC* parameter from Figure 9, *plane capture distance (PCD)*, are used in *meters*, evaluating the position uncertainty for each Lidar point. Consequent iterations are completed for *PCD=0.1*, *0.2*, and *0.3*. The model quality is evaluated both by a visual comparison with a processed point cloud (as in Figure 13), and by customized estimating of algorithmic parameters. Modeling warnings designate partially correct models that have some flaws.

Table 1: Comparison of the modeled results depending on clustering and segmenting method (a LiDAR point dataset on Lubliniec, Poland)

#	Method of clustering / segmentation	Pro-cessing time	Total number of models	Correct models	In-correct models	Modeling warnings
1	Original *SaS*	8 h 45 min	934	534	289	111
2	Original *RANSAC*	12 h 11 min	934	577	248	109
3	*SaS* with restructured pipeline	11 h 24 min	934	588	243	103
5	Optimized *RANSAC* with *PCD=**0.1***	15 h 57 min	934	662	167	105
6	Optimized *RANSAC* with *PCD=**0.2***	10 h 42 min	934	677	147	111
6	Optimized *RANSAC* with *PCD=**0.3***	7 h 42 min	934	699	133	102

Shown in Table 1 are the comparative statistics, which has stipulated the necessity development of the *BERA PolyFit* building modeling tool described above. *The total number of models* column has the same value for all records, since it is a number of building footprints obtained from a data provider for this AOI. Another almost constant column for all methods is number of partially correct models (*Modeling warnings*). The table shows, that for all other records the evidently best results are demonstrated by the optimized *RANSAC* with *PCD=0.3* (Method *# 6*). The accuracy of the segmented roof planes follows from a selected clustering-segmenting method thus it does predefine either a success, or not on the stage of reconstruction. Therefore, this technique has been selected for the *BERA-P* tool. Briefly the most efficient *RANSAC* method according to Table 1 can be illustrated in the following way, if examining only one model (*# 43*) from a dataset, what from Table (1) (Figure 15).

Two different iterations of Method # 6, which present resembling results (the same number of up to seven face candidates), are illustrated, because it is the only method, that produces an acceptable number of similar results through various iterations. Lidar point as color-coded segments have ratios for each coded class. Ratio numerator is the number of break-line (non-planar points that do not belong to this segment due to the *PCD* value) points, while ratio denominator – planar points (they satisfy the *PCD* threshold). Methods *##1-6* (Table 1) have been compared with respect to similarity of various iterations like in Figure 15. Method *# 6* meets criterion of the best efficiency in this respect. The following illustration shows that this technique is the only tool (among all other from Table 1), that can provide the clustering-segmenting premises for reconstruction of a problematic roof slope, that belongs to model # 21 in the considered Lubliniec dataset (Figure 16). All other five techniques missed this roof slope.

The finalized illustration of this text demonstrates an increase for up to 27% (*41 acceptable among 62 all models* versus *30 / 62 ratio*) in a number of acceptable models for a small urban area in Washington, D.C. by providing Method *# 6* as a software solution (Figure 17).

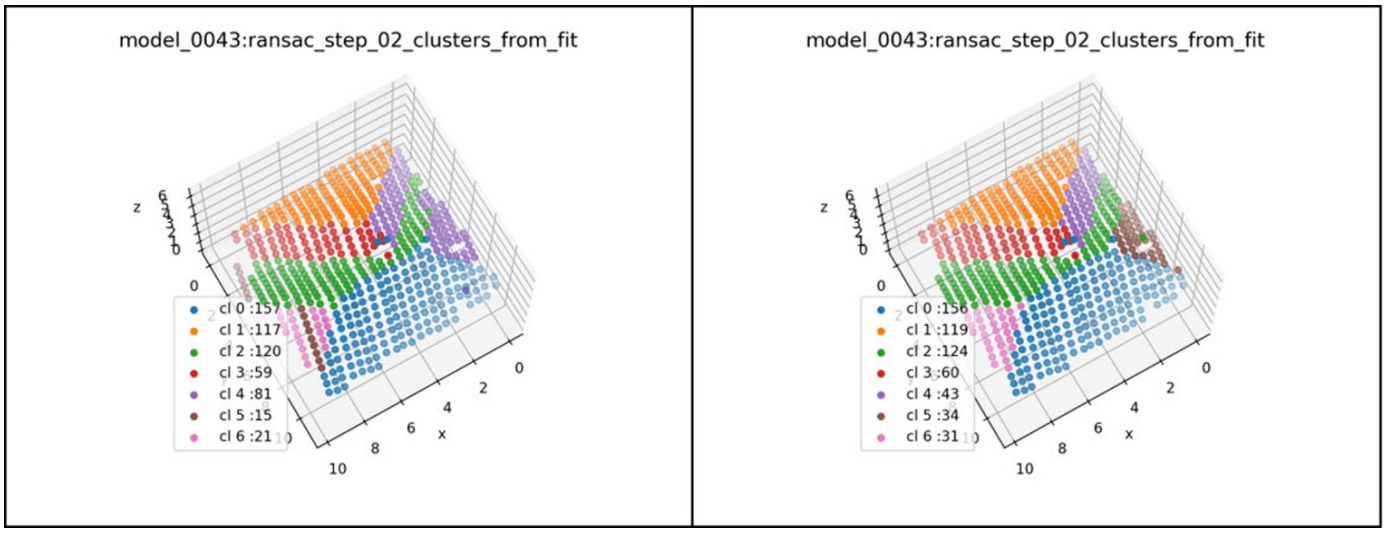


Figure 15: Two different clustering / segmenting iterations of Method # 6 (Table 1) to a point cloud of model # 43, that demonstrate resembling results

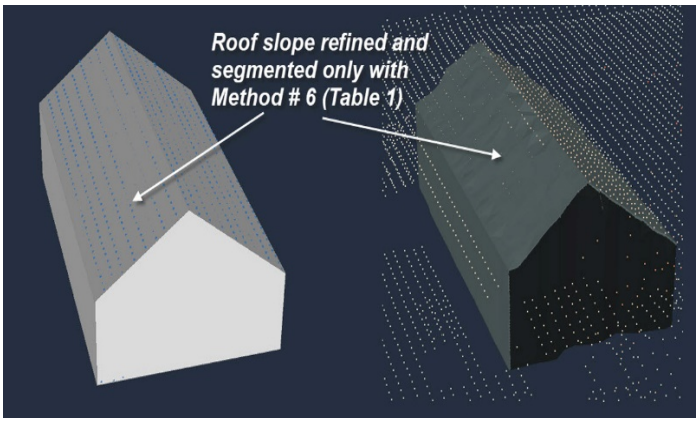


Figure 16: The roof slope pointed by arrows segmented by Method # *6* for further reconstruction. A CAD model built from a Google imagery (to the right) is compared with a point cloud with quite weak correlation for this building segment, while *ELiT* LW-model reconstructed is to the left

6. Conclusion

Our paper focuses on the conceptual multifunctional (functionalities of *Building Extraction*, *Building Extraction* in *Rural Areas*, *Change Detection*, and *DEM Generation*) approach within the AFE frameworks, that has been introduced through consequent steps of the phased methodological flowchart. In general, a workflow presented that has been completely automated for building detection, extraction and modeling by processing ALS and MLS point clouds. Two approaches of the high polyhedral modeling, and the low polyhedral one, have been proposed as two phased branches of a united overall workflow. These two modeling mainstreams have been compared on the base of their resulted models: the heavyweight models generated together with their original footprints within the HPM frameworks, and the lightweight models that employ the third-party footprints – with the LPM technique. With modeled results illustrated we have attempted to prove that the HPM branch can be appropriate for high-rise building reconstruction with either flat, or pitched roofs in densely build-up urban areas. The LPM branch – for modeling low-rise buildings with gable, or pitched roofs in either rural areas, or in city suburbs, as well as in small towns, i.e., in areas with thinned build-up.

A plot of our overall AFE workflow has been presented as the content of the methodological flowchart phases, while its *high polyhedral modeling* branch has been described by the relevant core algorithmic pipeline separately for phases of ALS and MLS data processing. The ALS block of footprint extraction and delineation has been presented in details. Visual samples of HPM-results processed by the *Building Extraction* and the *Change Detection* desktop and web-software tools have been provided.

The *low polyhedral modeling* has been explained as the technique consisting of two consequent stages: the point cloud clustering / segmenting stage, and the stage of building reconstruction. Our methodological flowchart has presented detailed phases of the LPM implementation. The *SaS*, *RANSAC*, and *PloyFit* approaches have been adopted, and selectively enhanced for implementation in the web-software as *Building Extraction Rural Areas* and *BERA-PolyFit* tools:

SaS-segmentation has been enhanced by restructuring its pipeline; *RANSAC*-segmentation has been optimized by both the introduction of the key BCCPI parameter, and by the adaptive alterations of rules due to the plane candidate "penalty charges", as well as by introducing the point sampling from a randomly chosen limited neighborhood; the optimized *RANSAC*-segmentation method with a certain parameter has been selected as the most efficient one among six compared methods on the base of results obtained for our key dataset;

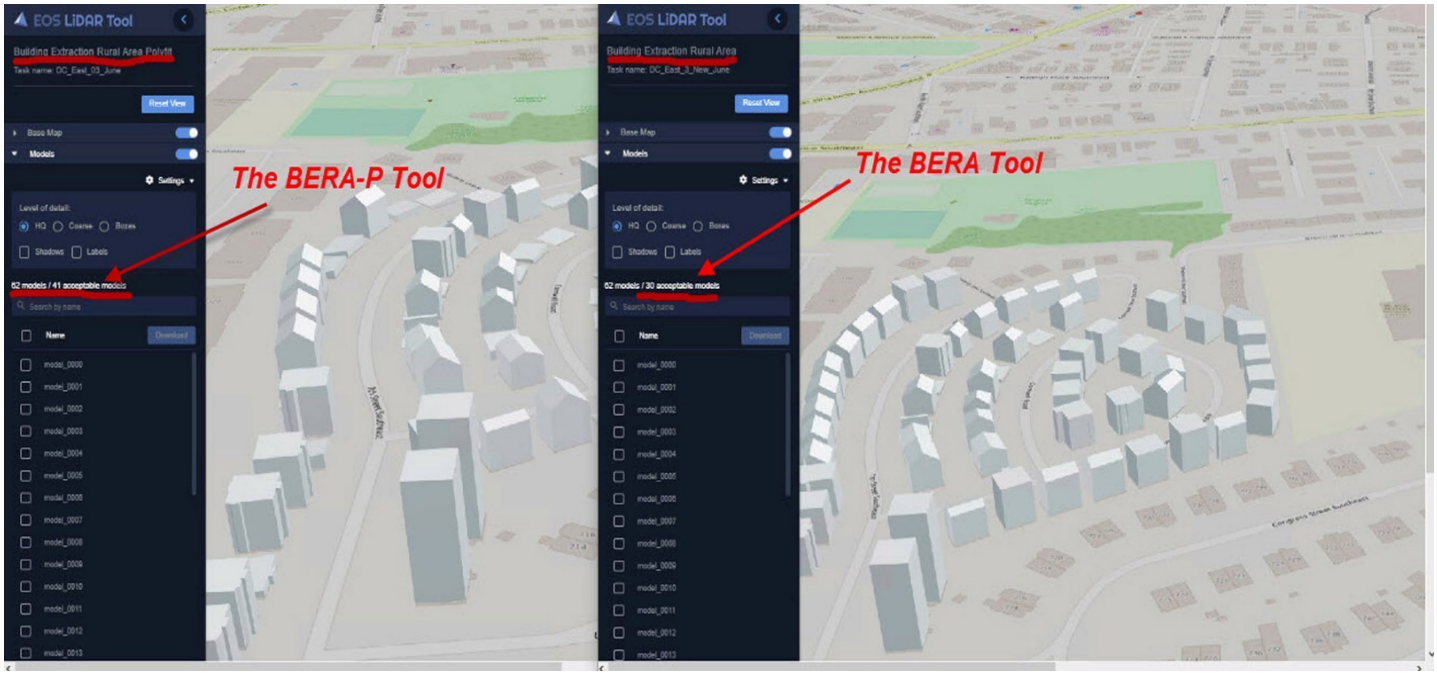


Figure 17: Comparison of modeled results, the number of acceptable models, produced by the *BERA Polyfit* Tool (with Method # *6* in clustering / segmenting algorithmic core - the left visual), on one hand, and by the *BERA* Tool (the right visual), on the other hand. *A small urban parcel in Washington, D.C., USA*

SaS-reconstruction has been optimized by the Voronoi neighborhood introduction, and thus obtaining the enhancement of overall algorithmic efficiency and the better optimization of the roof plane adjacency matrix. The Voronoi diagram can be accepted as a natural neighborhood without the necessity to specify its size. This property can be crucial for employing Voronoi neighbors within the reconstruction stage, while in the original methodology its usage is strictly limited to the stage of segmentation;

PolyFit approach has been implemented in the *BERA-P* web-software tool and proved to be as the most efficient one on the reconstruction stage if based on the optimized *RANSAC*. This issue has finalized our contributions to the low polyhedral modeling.

Besides all mentioned above functionalities the graphic user interface of the original family of software products for LiDAR data processing has been introduced as desktop, web-based, and cloud-based solutions: *ElitCore*, *ELiT Sever*, and *ELiT Geoportal*, correspondingly.

As far as future works are concerned the suggested algorithmic optimization and enhancements should be exercised through much more urban configurations, e.g., on those locations already placed on the *ELiT Geoportal.* Routine GIS-data consumers may employ the *ELiT Server* to search, access, and process their own Lidar data by the multifunctional tools in both block, and district scopes, and the data delivered to them by a provider. In such case, the *ELiT Geoportal*, as an Internet-toolbox, may serve an increasingly significant role in sharing derivative LiDAR information on the structure and spatiotemporal development of large human settlements, i.e., provide solution in a whole city scope.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research and development have been sponsored by the Noosphere Ventures that is an international asset management Company.

References

[1] S. Kostrikov, R. Pudlo, D. Bubnov, A. Kostrikova, "Studying of urban features by the multifunctional approach to LiDAR data processing," in Joint Urban Remote Sensing Event JURSE 2019: New Methodologies for urban investigation through remote sensing. UBS, France, 2019. https://ieeexplore.ieee.org/document/8809063. IEEE Xplore Digital Library, 2019. http://dx.doi.org/10.1109/JURSE.2019.8809063

[2] M. Batty, B. Hutchinson, Systems Analysis in Urban Policy Making and Planning, Plenum Press, 1982.

[3] S.V. Kostrikov, L.M. Niemets, K.Y.Sehida, K.A. Sehida, C. Morar, "Geoinformation approach to the urban geographic system research (cases studies of Kharkiv region," Visnyk of V. N. Karazin Kharkiv National University, series "Geology. Geography. Ecology", **49**(1), 107-124, 2018. https://doi.org/10.26565/2410-7360-2018-49-09

[4] S.V. Kostrikov, "Urban remote sensing with LiDAR for the Smart City concept implementation," Visnyk of V. N. Karazin Kharkiv National University, series "Geology. Geography. Ecology", **50** (2), 101-124, 2019. https://doi.org/10.26565/2410-7360-2019-50-08

[5] T. Esch, W. Heldens, A. Hirner, "The Global Urban Footprint," in Urban Remote Sensing. CRC Press, 34-43, 2018.

[6] M. Batty, D. Chapman, S. Evans, M. Haklay, S. Kueppers, N. Shiode, A. Smith, P.M. Torrens, "Visualizing the city: communicating urban design to planners and decision-makers," Center for Advanced Spatial Analysis, Working Paper Series, 2000. [Online]. Available: http://www.casa.ucl.ac.uk/visualcities.pdf

[7] C.J. Bellman, M.R. Shortis, "A classification approach to finding buildings in large scale aerial photographs," ASPRS Archives, **XXXV**(B3), 337–342, 2004.

[8] D.F. Maune (Editor), Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd Edition, ASPRS, 2007.

[9] F. Leberl, "Human Habitat Data in 3D for the Internet," in Computer Vision, Imaging and Computer Graphics. Theory and Applications. VISIGRAPP 2009. Communications in Computer and Information Science, **68**, Springer, Berlin, 2010. DOI https://doi.org/10.1007/978-3-642-11840-11

[10] T. Esch, M. Marconcini, D. Marmanis, J. Zeidler, S. Elsayed, A Metz, A. Müller, S. Dech, "Dimensioning urbanization—An advanced procedure for characterizing human settlement properties and patterns using spatial network analysis," Appl. Geogr., **55**, 212–228, 2014. http://dx.doi.org/10.1016/j.apgeog.2014.09.009

[11] D. Potere, A. Schneider, S. Angel, D.L Civco, "Mapping urban areas on a global scale: Which of the eight maps now available is more accurate?," Int. J. Remote Sens. **30**(24), 6531–6558, 2009. http://dx.doi.org/10.1080/0143116090312113 4

[12] J. Congote, A. Moreno, L. Kabongo, J.L. Pérez, R. San José, O. Ruiz, "Web based hybrid volumetric visualization of urban GIS data," in Usage, Usability, and Utility of 3D City Models - European COST Action TU0801, EDP Sciences, Nantes, France, 2012. http://dx.doi.org/10.1051/3u3d/201203001

[13] J. Richards, Remote Sensing Digital Image Analysis, Springer, Heidelberg, Germany, 2013.

[14] S. Yasumoto, A. Jones, K. Yano, T. Nakaya, "Virtual city models for assessing environmental equity of access to sunlight: a case study of Kyoto, Japan," Int. J. Geogr. Inf. Sci., **26**, 1-13, 2012. http://dx.doi.org/10.1080/13658816.2011.570268

[15] I. Delikostidis, J. Engel, B. Retsios, C.P.J.M. van Elzakker, M.J. Kraak, J. Döllner, "Increasing the usability of pedestrian navigation interfaces by means of Landmark visibility analysis," J. Navig. **66**, 523–537, 2013. http://dx.doi.org/10.1017/S0373463313000209

[16] W. Förstner, "3D City Models: Automatic and Semiautomatic Acquisition Methods," in Proceedings Photogrammetric Week, University of Stuttgart, 291-303, 1999.

[17] J. Döllner, T.H. Kolbe, F. Liecke, T. Sgouros, K. Teichmann, "The virtual 3D city model of Berlin – managing, integrating and communicating complex urban environment," in Proceedings of the 25th International Symposium on Urban Data Management, 15-17 May 2006, UDMS, Alborg, Denmark, 2006.

[18] W. Ribarsky, T. Wasilewski, N. Faust, "From Urban Terrain Models to Visible Cities," IEEE Comput. Graph. Appl., **22**(4), 10-15, 2002. http://dx.doi.org/10.1109/MCG.2002.1016692

[19] A. Stadler, C. Nagel, G. König, T.H. Kolbe, "Making interoperability persistent: A 3D geodatabase based on CityGML," in Proceedings of the 3rd International Workshop on 3D Geo-Information, Seoul, Korea. Lecture Notes in Geoinformation & Cartography, Springer Verlag, 2009. http://dx.doi.org/10.1007/978-3-540-87395-2_11

[20] D. Fisher-Gewirtzman, "3D models as a platform for urban analysis and studies on human perception of space," in Usage, Usability, and Utility of 3D City Models - - European COST Action TU0801, EDP Sciences: Nantes, France, 2012. [Online]. Available: http://3u3d.edpsciences.org. https://doi.org/10.1051/3u3d/201201001

[21] K. Sanjaya, F. Henning, K.R. Purba, "3D LIDAR City Model Application and Marketing Plan Development," in 2017 International Conference on Soft Computing, Intelligent Systems and Information Technology (ICSIIT), 238-242, 2012. https://doi.org/10.1109/ICSIIT.2017.50

[22] T.H. Kolbe, "Representing and Exchanging 3D City Models with CityGML," in Proceedings of the 3rd International Workshop on 3D Geo-Information, Seoul, Korea. Lecture Notes in Geoinformation & Cartography, Springer Verlag, 2009. http://dx.doi.org/10.1007/978-3-540-87395-2_2

[23] S. Zlatanova, J. Beetz, "3D spatial information infrastructure: The case of Port Rotterdam," In Usage, Usability, and Utility of 3D City Models - European COST Action TU0801, EDP Sciences, Nantes, France, 2012. http://dx.doi.org/10.1051/3u3d/201203010

[24] F. Biljecki, J. Stoter, H. Ledoux, S. Zlatanova, and A. Çöltekin, "Applications of 3D city models: state of the art review," ISPRS Int. J. of Geo-Inf., **4**, 2842–2889, 2015. http://dx.doi.org/10.3390/ijgi4042842

[25] E. Baltsavias, "A comparison between photogrammetry and laser scanning,", ISPRS J. Photogramm. Remote Sens., **54**(2-3), 83-94, 1999. http://dx.doi.org/10.1016/S0924-2716(99)00014-3

[26] P. Dong, Q. Chen (Editors), LiDAR Remote Sensing and Applications, Boca Raton, FL, CRC Press, 2018. http://dx.doi.org/10.4324/9781351233354

[27] E.P. Baltsavias, "Airborne laser scanning: Existing systems and firms and other resources," ISPRS J. Photogramm. **54**, 164–198, 1999. http://dx.doi.org/10.1016/S0924-2716(99)00016-7

[28] Y. Lin, J. Hyyppa, A. Jaakkola, "Mini-UAV-LS-Borne LIDAR for Fine-Scale Mapping," IEEE Geosci. Remote Sens. **8**, 426–430, 2011. http://dx.doi.org/10.1109/LGRS.2010.2079913

[29] B. S. Yang, Z. Wei, Q. Q. Li, J. Li, "Automated extraction of street-scene objects from mobile LiDAR point clouds," Int. J. Remote Sens., **33**(18), 5839-5861, 2012. http://dx.doi.org/10.1080/01431161.2012.674229

[30] Y. Wang, L. Cheng, Y. Chen, Y. Wu, M. Li, "Building Point Detection from Vehicle-Borne LiDAR Data Based on Voxel Group and Horizontal Hollow Analysis," Remote Sens., **8**, 419, 2016. http://dx.doi.org/10.3390/rs8050419

[31] R. Wang, "3D building modeling using images and lidar: a review," Int. J. of Image and Data Fusion, **4**(4), 273–292, 2013. http://dx.doi.org/10.1080/19479832.2013.811124

[32] M. Awrangjeb, C. Zhang, C. Fraser, "Automatic extraction of building roofs using LIDAR data and multispectral imagery," ISPRS J. Photogramm. and Remote Sens., **83**, 1-18, 2013. http://dx.doi.org/10.1016/j.isprsjprs.2013.05.006

[33] B. Yaung, Z. Wei, Q. Li, J. Li, "Semiautomated building façade footprint extraction from mobile LiDAR point clouds," IEEE Geoscince and Remote Sens. L., **10**(4), 766-770, 2013. https://doi.org/10.1109/LGRS.2012.2222342

[34] M. Awrangjeb, C. Fraser, "Automatic segmentation of raw LIDAR data for extraction of building roofs," Int. J. Remote Sens., **6**, 3716–3751, 2014. http://dx.doi.org/10.3390/rs6053716

[35] J. Shan, A. Sampath, "Urban DEM generation from raw LiDAR data: a labeling algorithm and its performance," J. Photogramm. Eng. Remote Sens., **75**, 427-442, 2009. https://doi.org/10.14358/PERS.71.2.217

[36] N. Haala, M. Kada, "An update on automatic 3D building reconstruction," ISPRS J. of Photogram. Remote Sens., **65**, 570-580, 2010. http://dx.doi.org/10.1016/j.isprsjprs.2010.09.006

[37] A. Sampath, J. Shan, "Segmentation and reconstruction of polyhedral building roofs from aerial LIDAR point clouds," IEEE T, Geosci. Remote Sensing, **3**, 1554–1567, 2010. http://dx.doi.org/10.1109/TGRS.2009.2030180

[38] A. Habib, R. Zhai, C. Kim, "Generation of complex polyhedral building models by integrating stereo-aerial imagery and LiDAR data," J. Photogram. Eng. Remote Sens. 75, 609-623, 2010. http://dx.doi.org/10.1016/j.isprsjprs.2010.09.006

[39] P. Kumari, W.E. Carter, R.L. Shrestha, "Adjustment of systematic errors in ALS data through surface matching," Adv. Space Res., **47**, 1851–1864, 2011. http://dx.doi.org/10.1016/j.asr.2010.12.015

[40] M.Y. Belkhouche, B. Buckles, "Iterative TIN-based automatic filtering of sparse LiDAR data," IEEE Geosci. Remote. Sens. Lett., **2**(3), 231-240, 2012. http://dx.doi.org/10.1080/01431161.2010.515266

[41] D. Chen, L. Zhang, R. Liu, "Urban building roof segmentation from airborne LiDAR point clouds," Int. J. Remote Sens., **33**, 6497–6515, 2012. http://dx.doi.org/10.1080/01431161.2012.690083

[42] S. Sun, C. Salvaggio, "Complex Building Roof Detection and Strict Description From LIDAR Data and Orthorectified Aerial Imagery," in 2012 IEEE Geoscience and Remote Sensing Symposium, 5466- 5469, 2012. http://dx.doi.org/10.1109/IGARSS.2012.6352369

[43] S. Kostrikov, R. Pudlo, A. Kostrikova, "Three key EOS LiDAR Tool functionalities for Urban Studies," in 39th Asian Conference on Remote Sensing (ACRS 2018): Remote Sensing Enabling Prosperity, Proceedings of a meeting held 15-19 October 2018, Kuala Lumpur, Malaysia, AARS – Curran Associates, Inc., **3**, 1676-1685, 2019.

[44] J.R. Jensen, "Remote Sensing of the Environment: An Earth Resource Perspective (2nd edition)," Prentice Hall, Upper Saddle River, NJ, 2006.

[45] A. Gruen, "TOBAGO—A semi-automated approach for the generation of 3-D building models," ISPRS J. Photogramm. Remote Sens., **53**, 108–118, 1998. http://dx.doi.org/10.1016/S0924-2716(97)00034-8

[46] H.-G. Maas, G. Vosselman, "Two algorithms for extracting building models from raw laser altimetry data," ISPRS J. Photogramm. Remote Sens., **54**, 153–163, 1999. http://dx.doi.org/10.1016/S0924-2716(99)00004-0

[47] N. Haala, C. Brenner, "Extraction of buildings and trees in urban environments" ISPRS J. Photogramm. Remote Sens., **54**, 130–137, 1999. http://dx.doi.org/10.1016/S0924-2716(99)00010-6

[48] G. Priestnall, J. Jaafar, A. Duncan, "Extracting urban features from LiDAR digital surface models", **24**(2), 65-78, 2000. https://doi.org/10.1016/S0198-9715(99)00047-2

[49] U. Stilla, U. Soerge, U. Thoennessen, "Potential and limits of InSAR data for building reconstruction in built-up areas," ISPRS J. Photogramm. Remote Sens. **58**, 113–123, 2003. http://dx.doi.org/10.1016/S0924-2716(03)00021-2

[50] Y. Hu, "Automated Extraction of DTM, roads, and buildings using Airborne lidar," PhD Thesis, University of Calgary, 2007.

[51] C. Alexander, S. Smith-Voysey, C. Jarvis, K. Tansey, "Integrating building footprints and LiDAR elevation data to classify roof structures and visualize buildings," Comput. Environ. Urban Syst., **33**, 285-292, 2009. http://dx.doi.org/10.1016/j.compenvurbsys.2009.01.009

[52] A.K. Aijazi, P. Checchin, L. Trassoudaine, "Automatic detection and feature estimation of windows in 3D urban point clouds exploiting façade symmetry and temporal correspondences," Int. J. Remote Sens., **35**, 7726–7748, 2014. http://dx.doi.org/10.1080/01431161.2014.975420

[53] E. Orthuber, E., J. Avbelj, "3D building reconstruction from Lidar point clouds by adaptive dual contouring," ISPRS Ann. Photogram. Remote Sens., II-3/W4, PIA15+HRIGI15, Joint ISPRS conference 2015, 25–27 March, Munich, Germany, 2015.

[54] T. Freeland, B. Heung, D.V. Burley, G. Clark, A. Knudby, "Automated feature extraction for prospection and analysis of monumental earthworks from aerial LiDAR in the Kingdom of Tonga," J. Archaeol. Sci., **69**, 64-74, 2016. https://doi.org/10.1016/j.jas.2016.04.011

[55] C. Brenner, "Towards fully automatic generation of city models," Int. Arch. Photogram., Remote Sens., **33**, 1-8, 2000.

[56] M. Romano, "Innovation in LiDAR processing technology," Photogramm. Eng. Remote Sens., **70**, 1202-1212, 2004.

[57] G. Sithole, G. Vosselman, "Filtering of airborne laser scanner data based on segmented point clouds," Int. Arch. Photogram. Remote Sens. , **34,** 66-71, 2005.

[58] M. Kada, L. McKinley, "3D building reconstruction from LiDAR based on a cell decomposition approach," Int. Arch. Photogramm., Remote Sens. Sensing, **38**, 47-52, 2009.

[59] J.X. Zhang, X.G. Lin, X.G. Ning, "SVM-based classification of segmented airborne LIDAR point clouds in urban areas," Remote Sens., **5**, 3749–377. 2013. http://dx.doi.org/10.3390/rs5083749

[60] C. Flener, M. Vaaja, A. Jaakkola, A. Krooks, H. Kaartinen, A. Kukko, E. Kasvi, H. Hyyppä, J. Hyyppä, P. Alho, "Seamless Mapping of River Channels at High Resolution Using Mobile LiDAR and UAV-LS-Photography," Int. J. Remote Sens., **5**, 6382-6407, 2013. http://dx.doi.org/10.3390/rs5126382

[61] C. Liu, B. Shi, X. Yang, N. Li, H. Wu, "Automatic buildings extraction from LiDAR data in urban area by neural oscillator network of visual cortex," IEEE J-STARS, **6**, 2008–2019, 2013. http://dx.doi.org/10.1109/JSTARS.2012.2234726

[62] M.A. Fischler, R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, **24**(6), 381–395, 1981. http://dx.doi.org/10.1145/358669.358692

[63] R. Schnabel, R. Wahl, R. Klein, "Efficient RANSAC for point-cloud shape detection," in Computer graphics forum, volume 26, 214–226, 2007. http://dx.doi.org/10.1111/j.1467-8659.2007.01016.x

[64] B. Li, R. Schnabel, J. Shiyao, J., R. Klein, "Variational surface approximation and model selection," Computer Graphics Forum (Proc. Pacific Graphics) 28, 7, 1985—1994, 2009. https://doi.org/10.1111/j.1467-8659.2009.01577.x

[65] L. Nan, P. Wonka, "PolyFit: Polygonal surface reconstruction from point clouds," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017. http://dx.doi.org/10.1109/ICCV.2017.258

[66] L. Xu, E. Oja, Kultanen, "A new Curve Detection Method: Randomized Hough Transform (RHT)," Pattern Recognit. Lett., **11**, 331–338, 1990. https://doi.org/10.1016/0167-8655(90)90042-Z

[67] D. Borrmann, J. Elseberg, K. Lingemann, A. Niichter, "The 3D hough transform for plane detection in point clouds: a review and a new accumulator design," 3D Res., **2**, 1-13, 2011. http://dx.doi.org/10.1007/3DRes.02(2011)3

[68] E. Maltezos, C. Ioannids, "Automatic extraction of building roofs from Airborne LiDAR data applying and extended 3D randomized Hough transform," ISPRS Ann. Photogramm. Remote Sens., **III-3**, 209-216, 2016.http://dx.doi.org/10.5194/isprsannals-III-3-209-2016

[69] J. Susaki, "Knowledge-based modeling of buildings in dense urban areas by combining airborne LiDAR data and aerial images," Remote Sens., **5**, 5944–5963, 2013. http://dx.doi.org/10.3390/rs5115944

[70] L. Zhang, T. Xu, J. Zhang, "Building extraction based on multiscale segmentation," in 5th International Congress on Image and Signal Processing (CISP), October 16–18, 2012. http://dx.doi.org/10.1109/CISP.2012.6470035

[71] J. Cheng, Y. Deng, Q. Du, "Mapping between BIM models and 3D City models of different levels of detail," in Proceedings of the 13th International Conference on Construction Applications of Virtual Reality, 30-31 October 2013, London, UK, 2013.

[72] N.H. Arachchige, S.N. Perera, H.G. Maas, "Automatic processing of mobile laser scanner point clouds for building façade detection," ISPRS International Archive of Photogrammetry and Remote Sensing in Spatial Information Science, **XXXIX-B5**, 187–192, 2012. http://dx.doi.org/10.5194/isprsarchives-XXXIX-B5-187-2012

[73] R. Brinkman, C. O'Neill, "LiDAR and photogrammetric mapping," The Military Engineer, **5**, 56–57, 2000.
[74] M. Hodgson, P. Bresnahan, "Accuracy of airborne lidar-derived elevation: Empirical assessment and error budget," J. Photogram. Eng. Remote Sens., **70**, 331–339, 2004. http://dx.doi.org/10.14358%2FPERS.70.3.331
[75] A. Sampath, J. Shan, "Building boundary tracing and regularization from airborne LiDAR point clouds," J. Photogramm. Eng. Remote Sens., **73**(7), 2007, 805–812. http://dx.doi.org/10.14358/PERS.73.7.805
[76] J. Shan, A. Sampath, "Building extraction from LiDAR point clouds based on clustering techniques," in Topographic Laser Ranging and Scanning: Principles and Processing, Boca Raton, FL: CRC Press, 423–446, 2008. http://dx.doi.org/10.1201/9781420051438.ch15
[77] A. Sampath and J. Shan, "Building roof segmentation and reconstruction from lidar point clouds using clustering techniques," Int. Arch. Photogramm., Remote Sens., **XXXVII**(B3a), 279-284, 2008.
[78] R. C. Gonzalez, R.E. Woods, Digital Image Processing, Englewood Cliffs, NJ: Prentice-Hall, 2002
[79] K. Zhang, D. Whitman, "Comparison of three algorithms for filtering airborne LiDAR data," Photogramm. Eng. & Remote Sens., **71**, 313–324, 2005. http://dx.doi.org/10.14358/PERS.71.3.313
[80] F. Rottensteiner, "Using the Dempster-Shafer method for the fusion of LiDAR data and multispectral images for building detection," Inf. Fusion, **6**(4), 283–300, 2005. https://doi.org/10.1016/j.inffus.2004.06.004
[81] F. Rottensteiner, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," ISPRS J. Photogram. Remote Sens., **93**(7), 256–271, 2014. http://dx.doi.org/10.1016/j.isprsjprs.2013.10.004
[82] L. Xu, D. Kong, X. Li, "On-the-fly extraction of polyhedral buildings from airborne LiDAR data," IEEE Geosci. Remote Sens. Lett., 1946–1950, **11**(11), 2014. https://doi.org/10.1109/LGRS.2014.2314458
[83] F. Tarsha-Kurdi, P. Landes, M. Grussenmeyer, M. Koehl, "Model-driven and data-driven approaches using LiDAR data: Analysis and comparison," in ISPRS Workshop, Photogrammetric Image Analysis (PIA07), Sep 2007, Munich, Germany, 87-92, 2007.
[84] A. Henn, G. Groger, V. Stroh, L. Plumer, "Model driven reconstruction of roofs from sparse LIDAR point clouds," ISPRS J. Photogram. Remote Sens., **76**, 17-29, 2013. http://dx.doi.org/10.1016/j.isprsjprs.2012.11.004+
[85] K. Kim, J. Shan, "Building footprints extraction of dense residential areas from LiDAR data local planarity," in 2011 Annual Conference of the American Society for Photogrammetry and Remote Sensing, Milwaukee, Wisconsin, 193-198, 2011.
[86] V. Verma, R. Kumar, S. Hsu, "3D building detection and modeling from aerial LIDAR data," in 2006 IEEE computer society conference on computer vision and pattern recognition, 17-22 June 2006, New York, NY, 2006. http://dx.doi.org/10.1109/CVPR.2006.12
[87] J. Yan, J. Shan, W. Jiang, "A global optimization approach to roof segmentation from airborne LiDAR point clouds," ISPRS J. Photogramm. Remote Sens., **94**, 183–193, 2014. http://dx.doi.org/10.1016/j.isprsjprs.2014.04.022
[88] K. Khoshelham, "Extending generalized Hough transform to detect 3D objects in laser range data," IAPRS, **XXXVI,** Part 3 / W52, 407-412, 2007.
[89] F. Tarsha-Kurdi, T. Landes, P. Grussenmeyer, "Hough-transform and extended RANSAC algorithms for automatic detection of 3D building roof planes from LiDAR data," IAPRS, **XXXVI,** Part 3 / W52, 206-210, 2007.
[90] G. Roth, M.D. Levine, "Extracting geometric primitives," Comput. Vis. Image Underst., **58**(1), 1-22, 1993. http://dx.doi.org/10.1006/cviu.1993.1029
[91] J. Yan, W. Jiang, J. Shan, "A global solution to topological reconstruction of building roof models from airborne LiDAR point clouds," ISPRS Ann. Photogram., Remote Sens. **3**, 379–386, 2016. http://dx.doi.org/10.5194/isprsannals-III-3-379-2016
[92] H. Huang, C. Brenner, M. Sester, "A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data," ISPRS J. Photogramm. Remote Sens., **79**:29–43. 2013. http://dx.doi.org/10.1016/j.isprsjprs.2013.02.004
[93] R.J. You, B.C. Lin, "A quality prediction method for building model reconstruction using LiDAR data and topographic maps," IEEE Trans. Geosc., **49**(9), 2011. https://doi.org/10.1109/TGRS.2011.2128326
[94] F. Rottensteiner, C. Briese, "Automatic generation of building models from LiDAR data and the integration of aerial images," Int. Arch. Photogram. Remote Sens., **34**(3/W13), 2003.
[95] L. Cheng, L. Tong, Y. Chen, W. Zhang, J. Shan, Y. Liu, M. Li, "Integration of LiDAR data and optical multi-view images for 3D reconstruction of building roofs," Opt. Laser. Eng, **51**(4), 493–502, 2013. http://dx.doi.org/10.1016/j.optlaseng.2012.10.010
[96] I. Stroud, Boundary representation modelling techniques, London: Springer-Verlag London, 2006.
[97] Z. Lari, A. Habib, "An adaptive approach for the segmentation and extraction of planar and linear/cylindrical features from laser scanning data," ISPRS J. Photogram. Remote Sens., **93**(7), 192–212, 2014. http://dx.doi.org/10.1016/j.isprsjprs.2013.12.001
[98] R. Cao, Y. Zhang, X. Liu, Z. Zhao, "3D building roof reconstruction from airborne LiDAR point clouds: a framework based on a spatial database," Int. J. Geog. Inf. Sci., **16**, 1-22, 2017. http://dx.doi.org/10.1080/13658816.2017.1301456
[99] P. Gamba, B. Houshmand, "Joint analysis of SAR, LIDAR and aerial imagery for simultaneous extraction of land cover, DTM and 3D shape of buildings," Int. J. Remote Sens., **23**(20), 4439-4450, 2002. http://dx.doi.org/10.1080/01431160110114952
[100] J. Shewchuk, "Delaunay refinement algorithms for triangular mesh generation," *Comput. Geom.*, **47**(1), 741-778, 2014. http://dx.doi.org/10.1016/S0925-7721(01)00047-5
[101] R. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," Remote Sens. Env., 37(1), 35-46, 1991. http://dx.doi.org/10.1016/0034-4257(91)90048-B
[102] S.V. Kostrikov, D.Y. Bubnov, R.A. Pudlo, "Urban environment 3D studies by automated feature extraction from LiDAR point clouds," Visnyk of V.N. Karazin Kharkiv National University, series "Geology. Geography. Ecology", **52,** 156-181, 2020. https://doi.org/10.26565/2410-7360-2020-52-12
[103] J. Matas, O. Chum, "Randomized RANSAC with $T_{d,d}$ test," Image and Vision Computing, **22**(10), 837-842, 2004. https://doi.org/10.1016/j.imavis.2004.02.009
[104] M. Li, L. Nan, N. Smith, P. Wonka, "Reconstructing building mass models from UAV-LS images," Comput. Graph., **54**(C), 84–93, 2016. http://dx.doi.org/10.1016/j.cag.2015.07.004
[105] B. Xu, W. Jiang, J. Shan, J. Zhang, L. Li, "Investigation on the Weighted RANSAC Approaches for Building Roof Plane Segmentation from LiDAR Point Clouds," Remote Sens., **8**(1), 5, 2016. https://doi.org/10.3390/rs8010005
[106] R. Tse, Ch. Gold, D. Kidner, "Using the Delaunay Triangulation/ Voronoi Diagram to extract Building Information from Raw LIDAR Data," in 4th International Symposium on Voronoi Diagrams in Science and Engineering (ISVD 2007). IEEE Xplore Digital Library, 2007. http://dx.doi.org/10.1109/ISVD.2007.40
[107] L. Nan, C. Jiang, B. Ghanem, P. Wonka, "Template assembly for detailed urban reconstruction," Computer Graphics Forum, **34**(2), 217–228, 2015. http://dx.doi.org/10.1111/cgf.12554
[108] E. Aoidh, M. Bertolotto, D.C. Wilson, "Understand geospatial interests by visualising map interaction behavior," Inf. Vis., **7**(3), 275-286, 2008. https://doi.org/10.1057%2FIVS.2008.24
[109] D. Eum, T. Minoura, "WebSiteGen: Web-based database application generator," IEICE Trans. Inf. Syst., **E86-D**(6), 112-126, 2003.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 74-79 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Power Loss Minimization using the Integration of DGs and Reconfiguration of Distribution System: Applied on Real Distribution Feeder of Urbain Areas of Kenitra City in Morroco

Ismail Moufid*, Soukaina Naciri, Hassan EL Moussaoui, Tijani Lamhamdi, Hassane El Markhi

Intelligent Systems, Georessources and Renewable Energies laboratory (ISGREL), Sidi Mohamed Ben Abdelah University; FST,Fez, 2202, Morocco

ARTICLE INFO

Article history:
Received: 29 June, 2020
Accepted: 21 August, 2020
Online: 09 September, 2020

Keywords:
NOP
Distributed generation
Power loss
Distribution network

ABSTRACT

Optimal integration of distributed generation (DG) into the distribution system results in reduced power losses and improved bus voltages. In this article, a combination of two techniques has been analyzed:
The integration of DG and reconfiguration of the distribution system by removing the Normally Open Point NOP in different places of the system.
These two techniques are applied to a real distribution network " distribution network of Kenitra city in Morroco", considering as key objectives the reduction of power loss and improvement of voltage profile.
To investigate the effectiveness and robustness of our system a model was performed using ETAP. The simulation results improve that we can minimize greatly the power losses in the distribution network by the implementation of DGs and reconfiguring our distribution network.

1. Introduction

The role of the distribution network transfers the electrical energy directly from the substations to consumers. The distribution network administrator currently faces several challenges resulting from increased penetration of renewable energy resources and the demand of customers. The Distribution Network Operators (DNO) has noted that the traditional solution of the network may not be sufficient. The increasing load demand in the distribution network creates an important challenge for research in phases of studying the management of the distribution network to satisfy the need with the actual infrastructure.

The determination of the grid concept of the distribution system depends on special specifications and side conditions [1]. distribution of electrical energy in a grid network is done in two ways: closed-ring networks and radial networks as presented in Fig.1.

Radial grid: The open ring is generally the structure of the distribution network as presented in Figure 1, the advantage of this structure is that during a fault, the currents propagate in one unidirectional direction, so the detection and the position of faults became more flexible than in closed rings arrangement. Moreover, the inconvenience of this architecture is that when a fault befalls, the part of the network after the fault is detached from the substation, this situation makes the management of the distribution system more complicated.

Closed-ring grid: The closed ring grid has the same structure as the radial structure, just that a part of the rings is performed as closed rings as we can see in Figure 1 b. the problem of this kind of structure that the protection of the grid became more complicated than in the radial case because when a fault happens the currents propagate in different ways. The disadvantage of a radial configuration can be skipped by adopting the closed-ring structure [1].

The benefit of the closed-ring structure is that more than one charger feeds the system, so if a feeder is inoperative or in maintenance, the grid is stay powered by the other feeders. In this way, customers are not concerned, even if one of the feeders is down or becomes inoperative. Moreover, the ring dispensing of the network is provided with many isolated parts. Furthermore, in the case of a fault in a section, this section is isolated by opening the corresponding breakers.

Usually, the structure of distribution networks is carried out according to the network shown in Figure 1.the closed-ring distribution network is created with a NOP, dividing the circle into two outlets from an open electrical point [2].

*Corresponding Author: Ismail Moufid, ismail.moufid@usmba.ac.ma

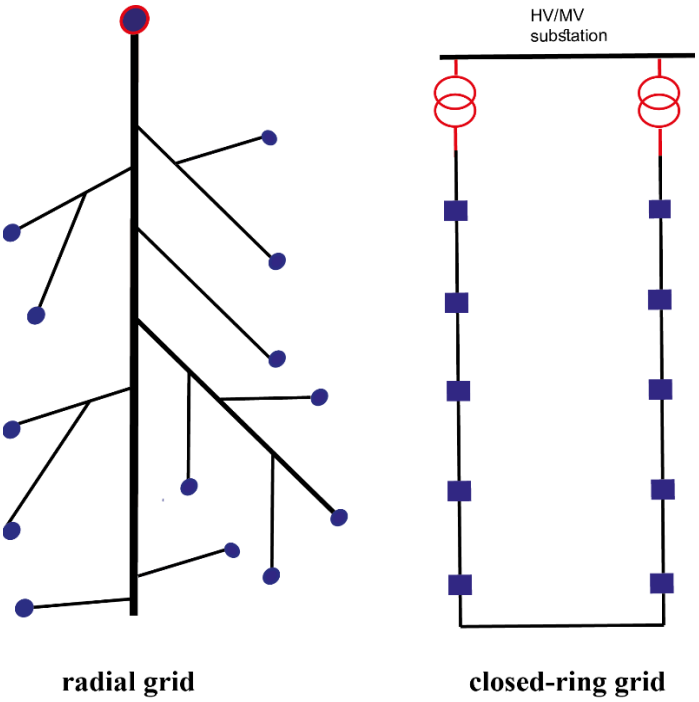


Figure 1: The structure of the closed-ring grid and radial grid.

By using NOP in the distribution network we assure the selectivity of the protection systems and minimize the consequence of faults; so the number of customers affected in the event of failure can be reduced in this case.

The power flow can be affected by the position of NOP. Throughout the day, the number of customers changes, so the power flow also changes.

Consequently, the optimal position of the NOP would change, so the current in each power line became nearly equal. we can reduce power losses and improve the voltage profile of the distribution network by adjusting the loads between the adjacents feeders.

Authors in [3] have proposed a strategy to determine the suitable best location for DG "Distributed Generations" to reduce distribution power losses. The approach is based on a specific loss equation. The strategy has been examined with different size of DG. The authors in [4] have defined an objective function of cost to place solar, wind, and fuel cell. An analytic hierarchy method has been utilized to decide on finding the optimal positions for various sorts of DGs.

The improvement of voltages profile placing DGs in different positions has been discussed in [5]. They have suggested voltage sensitivity to install DGs. Authors in [6] have defined the position of DG to minimize the power losses, the algorithm of Kalma is used to determine the optimal size of DG.

Authors in [7], proposes an approach to determine the Hosting Capacity of distributed generation (DG). The approach proposes a model of the lines utilizing the π type. The capacitance of different lines is not neglected, and we obtained a higher value of HC. we are concentrate on this study on line overloading.

An optimal power flow method and hybrid PSO for optimal position and sizing of DG considering the power loss reduction and DG cost minimization was proposed in [8]. The optimization problem satisfies some constraints like the thermal limit of the transformer and lines and the limit of the voltage profile. Authors in [9] proposed a method for the optimal position of DGs. They have also discussed the sensibility factor to choose the best place of DG. In [10] a GA methodology respecting indices of power loss and voltage profile was discussed. Authors in [11]have suggested a strategy to choose the position and the size of the DG regarding the constraints of minimizing power losses and harmonic distortion. They adopted a multi-objective PSO approach based on conventional weighted aggregation which may lead to a suboptimal solution in some cases. The optimal DG placement is determined based on voltage stability study as a security estimate by authors in [12].A method to evaluate the Sizing and sitting of various DGs based on the reduction of power loss was proposed by authors in [13].

In [14], authors have presented a technique to choose the optimal DG placement Based on Power Stability Index (PSI) based DG placement. In this method, the value of PSI is calculated for each row and sorted from the most important value to the most inferior value. DG was placed at the bus terminal of the most important PSI branch. The PSI value has been recalculated and by the same method location for the next DG. Authors in [15] have discussed a loss sensitivity based DG placement strategy. They used the bus ranking approach based on loss sensitivity and for the placement of DG they are placed in the buses who have the highest loss-sensitive. A simple strategy for the optimal position and sizing of generators in different buses was presented in [16]. The most important aim of this strategy is to minimize the power loss and cost of the distribution network. The authors also discuss the optimization of the weighting factor, which equilibrate loss factors and cost in order to achieve the desired aims with the highest advantage possible. In [17] an optimal placement and sizing of CBs and DGs in distribution systems the based on Spring Search Algorithm (SSA) are proposed. The authors in [18] have studied the impact of using SOP in the distribution system to reduce power losses and improving the voltage profile. Authors in [19] have proposed a new strategy for distribution systems including DG. The principal objects were the reduction of the risk factor and the cost reduction of total installation and operational. The optimal size and position of the DG are fixed in the planning step. The optimization of real power loss and the DG injection index are the two objective functions studied in this DG planning. Authors in [20] have analyzed the optimal placement of DG in order to assure voltage stability and power loss minimization of the distribution network. They have also adopted the PSO technique to find the optimal size and the position of DG. The optimal size and location of a single DG unit in different distribution systems are studied in the paper.

This paper proposes a comparison between two approaches in order to minimize power loss and improving the voltage profile of the distribution network. The paper is organized as follows.

In the first section, the formulation of the problem of power loss minimization and voltage improvement is discussed. In the second part, a description of our test network was discussed.

The third section is divided into two parts. In the first part, the efficiency of the suggested approach is tested to choose the best point of the open point in our distribution systems.

In the second part, DG is placed in a different place to choose the best point of the minimum of power loss, and the achieved results are presented. Lastly, some important conclusions are drawn.

2. Problem formulation

2.1. Power and current in the branches

we consider a distribution network with a radial configuration formed of a set of branches as presented in fig.2. Each branch of this network is modeled as a resistance in series with a pure inductance. The impedance of any branch "i" of this network is written as follows:

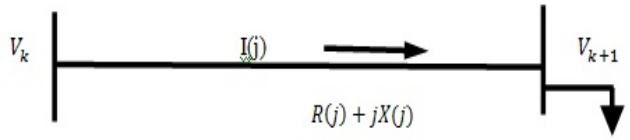


Figure 2: Distribution network with a radial configuration formed of two buses.

The load flow in this type of network is done using a BIBC matrix (**b**us **i**njection to **b**ranch **c**urrent) to calculate the currents flowing through all branches of the network.

For a busbar to which a load is connected, the apparent power S is represented by:

$$S_i = P_i + jQ_i \tag{1}$$

where $i = 1,2,3, \dots, n$

The equivalent current charge corresponding to the kth iteration is represented by:

$$I_i^k = I_i^r + jI_i^l = \left(\frac{P_i + jQ_i}{V_i^k}\right)^* \tag{2}$$

where:

V_i^k and I_i^k : are the voltage load and current of the busbar for the kth iteration.

$I_i^l and\ I_i^r$: are the imaginary and real parts of the busbar load.

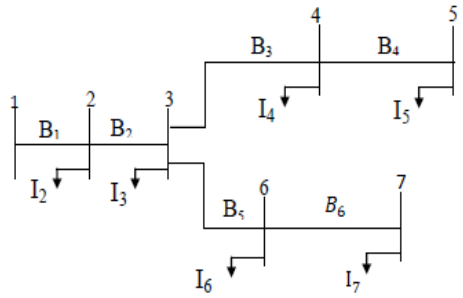


Figure 3: A simple distribution network of 7 busbars and 6 branches.

2.2. The BIBC Matrix

The load currents at the busbar are obtained using equation (2), as well as the currents circulating through the branches are determined by utilizing Kirchhoff's law to the studied distribution network. A simple distribution network constituted of 7 busbars and 6 branches presented in Figure 3 is used as an example to facilitate the description of the method for determining the BIBC matrix.

The branches currents can be represented as a function of load currents as results:

$$\begin{aligned} B_1 &= I_2 + I_3 + I_4 + I_5 + I_6 + I_7 \\ B_2 &= I_3 + I_4 + I_5 + I_6 + I_7 \\ B_3 &= I_4 + I_5 \\ B_4 &= I_5 \\ B_5 &= I_6 + I_7 \\ B_6 &= I_7 \end{aligned} \tag{3}$$

This means that the relation between the load currents of the busbars and the branch currents can be written as follows:

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{pmatrix}$$

In general, equation (3) can be written as follows:

[B] = [BIBC] [I].

where BIBC represent the matrix of currents injected into the busbars which are a matrix containing only the values 0 and 1.

2.3. Active and reactive power losses:

The determination of the power flow is essential to locate the overloaded electric lines and to calculate the value of the power losses.

The active power losses at a branch (i) are written as follows:

$$P_{loss,i} = R_i \frac{(P_i^2 + Q_i^2)}{|V|_i^2} \tag{4}$$

We can determine the total power losses of the network by summing the losses of all branches of the equation (4).

$$P_{T,loss} = \sum_{i=1}^{Nbr} P_{loss,i} \tag{5}$$

The reactive power losses at a branch (i) may be written as follows:

$$Q_{loss,i} = X_i \frac{(P_i^2 + Q_i^2)}{|V|_i^2} \tag{6}$$

We can also determine the total reactive power losses of the network by summing the losses of all branches of the equation (6)

$$Q_{T,loss} = \sum_{i=1}^{Nbr} Q_{loss,i} \tag{7}$$

2.4. Objective function:

The objective function is the measure of the system considered. The primary aim is improving the voltage profile and keep it between regulatory limits. But, this goal is transcribed mathematically as a constraint, because the voltage limits are strict

limits that the system cannot exceed and therefore they cannot be expressed as an objective function.

There are several possible objectives for solving the problem of choosing the optimal location of DGs in the distribution system. Among the most common goals are:

- Improvement of voltage profile: we seek to balance voltage profile while trying to minimize the sum of the relative deviations of the voltage in each node compared to the nominal tension.
- Minimization of online losses: Reducing Joules losses is becoming a priority for distribution network operators. This allows an increase in transit margins on power lines in addition to improving the voltage profile.

$$F_{objectif} = \min \sum_{i=1}^{Nbr} P_{loss,i} \quad (8)$$

So the objective function is associated with the nature of decision variables, constraints, and depends on the type of optimization adopted.

2.5. Technical constraints related to the problem

Voltage constraint: The value of the voltage profile of each busbar must be limited as follows:

$$V_{min} \leqslant V_i \leqslant V_{max}$$

Where Vmin = 0.95 p.u. and Vmax = 1.05 p.u. are the minimum and maximum of voltage amplitude values at each busbar, respectively. Active and reactive power balance constraints :

$$P_G = P_{loss} + P_D \quad \text{and} \quad Q_G = Q_{loss} + Q_D$$

where

Q_G and P_G are the reactive and active powers injected by the D_G while Q_D and P_D are the reactive and active load powers at the kth node.

Q_{loss} and P_{Loss} are reactive and active power losses in the system.

3. Test Network

The network studied is the MV network of the Kenitra city in Morroco it consists of a source substation with an installed power of 36MVA and a short-circuit power of 397.15 MVA, which supplies two distribution lines L1 and L2, its lengths are 9.4 km and 12 km respectively. This line has various characteristics (Blue, green and orange) which are shown in Table 1. This feeder aliments a load of more than 4.5 MW and 2.76 MVAR.

Table 1: Cables Parameters

Cables	Parameters		
	Impedance Ω/km	Capacitance nF/km	Maximum current A
Green	0.239+j0.11	300	300
orange	0.372+j0.13	220	225
blue	0.145+j0.1	370	400

The characteristics of this distribution network in Figure 4 are given as follows:

- Number of busbars = 40;
- Number of branches = 39;
- Slack bus N ° = 1;
- Nominal voltage 20 kV

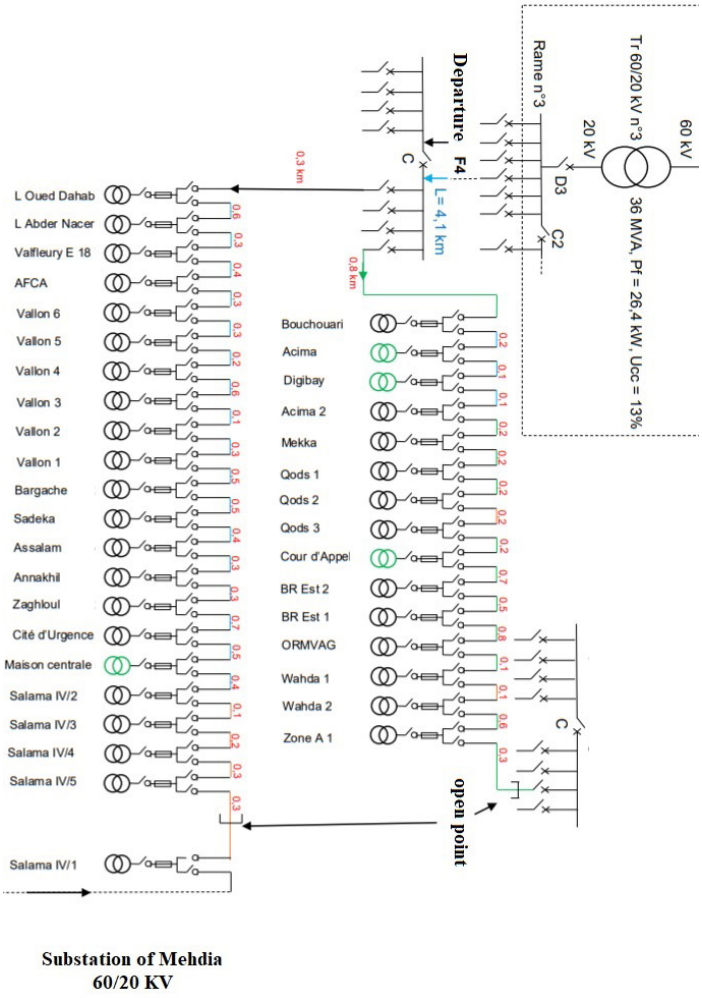


Figure 4: Studied distribution network

4. Simulation and discussion

For simulation purposes, the network studied was implemented in ETAP environment. The objective of this part is to choose the best position of DG for our distribution networks, so here we are interested in two constraints, the power losses and the voltage profile.

In this paper, we considered two cases:
Case 1 we fixed the position of the NOP and we remove the DG to have the minimum of power losses and we ameliorate the voltage profile.
Case 2 we fixed the DG in the optimal position and we change the position of NOP to minimize the power losses.

4.1. Simulation of case 1

We study the network for the first time by removing the DG place. The attitude of the distribution network after changing the position of DG is illustrated in Figures 5 & 6.

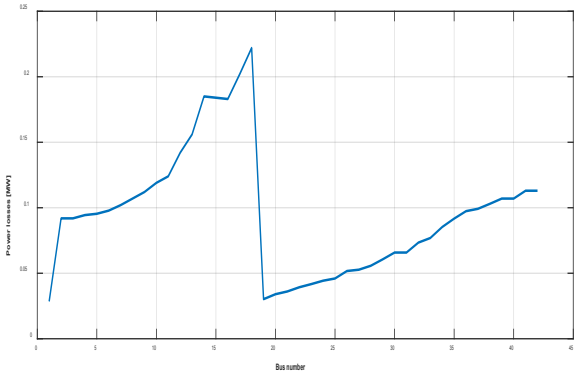


Figure 5: Power losses after the integration of DG in different buses

As presented in Fig. 5 we can see that the power losses change after removing DGs for different positions in our distribution system. However, we can remarque also that the power losses are minimum after the integration of DG in busbar number 18.

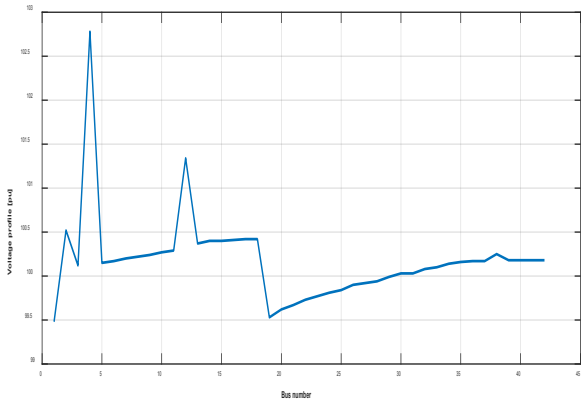


Figure 6: The voltage profile after the integration of DG in different buses

For the voltage profile, we took the average voltage value for the buses because for each DG location we have 41 voltage values.

As we can see in fig.6 the voltage profile for all buses of our distribution system is improved after the integration of DG in different positions, so with the integration of DGs, we can avoid the voltage drop. for the best place of DGs, we can remarque that for bus 18 the voltage profile is near to 1 pu.

4.2. Discussion of case 1

We can note in Fig. 5 that the integration of DG in our system can greatly reduce the total power losses of 0.15 MW, this reduction of power losses are important in bus bar 18; so the best place to integrate our DG is in this busbar, however, we can also observe in fig.6 the improvement of the voltage profile after using dg in our distribution network, we can remarque that for bus 18 the voltage profile is near to 1 pu, we can express this by the integration of DG in this place.

4.3. Simulation of case 2

We study now the performance of our network with DG fix in busbar 20 and we will move the position of NOP. The results are given in Fig.7

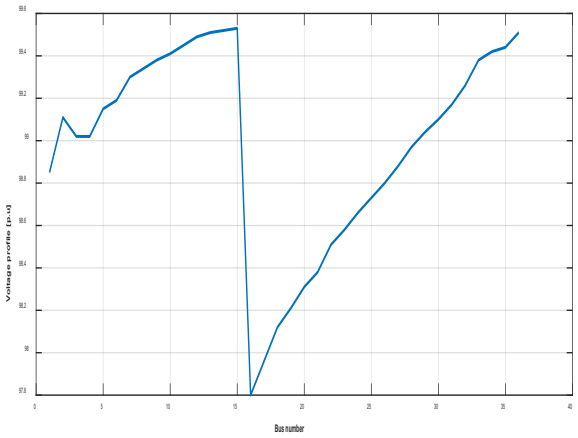


Figure 7: Voltage profile after removing the position of NOP

As presented in Fig. 7 the voltage profile of our distribution network changes proportionally after removing NOP in the different positions, however, we show that the voltage profile is great when the position of the NOP is between bus 16-17 and 36-37.

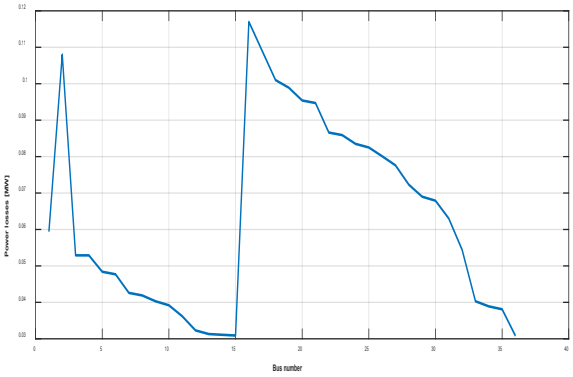


Figure 8: Power losses after removing the position of NOP

For Fig. 8 we can observe that removing NOP between buses influence also the power losses, but we can remarque that the power losses are minimum when the NOP position is is between bus 16-17 and 36-37.

4.4. Discussion of case 2

We can remarque in Fig.7 that the removing NOP greatly reduces the power losses of 0.03 MW, this reduction of power losses are important when it's between bus 16-17 and 36-37, so the best for NOP is between this busbar.

We can also remarque in fig. 8 that the voltage profile is maximum when NOP is between the same buses.

We can explain that by balancing the loads between the feeders of our system.

4.5. Synthesis

To compare the two cases, we will draw the power losses of the two cases in the same graph.

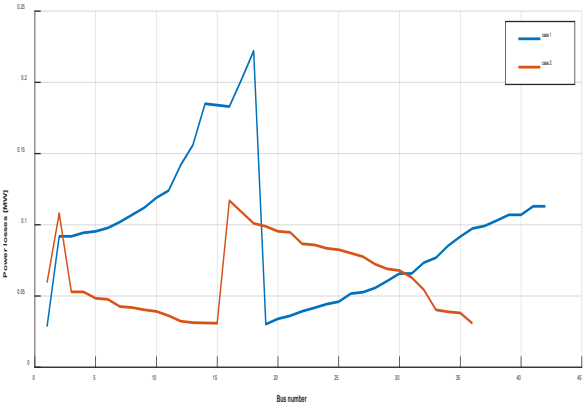


Figure 9: Power losses in case 1 & 2

As presented in Fig.9 the power losses are lower in case 2 compared to case 1.

We can recapitulate that by integrating NOP to balance the loads between feeder we can minimize power losses in the distribution network more than integrating DG in a different position.

5. Conclusion & Perspective

This paper has proposed two approaches to reduce power losses in the distribution network, so we considered two constraints, reduce power losses and improving the voltage profile. To evaluate the effectiveness of this approach we worked on two cases:

Case 1 we fixed the position of the NOP and we remove the DG to have the minimum of power losses.

Case 2 we fixed the DG in the optimal position from the first case and we change the position of NOP. The principal objective was to decrease the power losses in the distribution network. To study the influence on grid losses we considered a Medium Voltage distribution network which is the MV network of the Kenitra city in Morroco it consists of a source substation with an installed power of 36MVA and a short-circuit power of 397.15 MVA was implemented using Etap and Matlab/Simulink. The results presented illustrate that we can decrease power losses in distribution systems by displacing the position of NOP more than integrating DGs in our system.

For Future work, it will be interesting to integrate metaheuristics algorithms like Butterfly optimization algorithm BOA [21] and particle swarm optimization algorithm PSO [22] to reduce power losses and choosing the best place for DGs and NOP.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] L. Thurner, A. Scheidler, A. Probst, M. Braun, "Analysing the Degree of Meshing in Medium Voltage Target Grids-An Automated Technical and Economical Impact Assessment," ArXiv Preprint ArXiv:1802.01492, 2018, doi:arXiv:1802.01492.

[2] T. Schnelle, M. Schmidt, P. Schegner, "Power converters in distribution grids-new alternatives for grid planning and operation," IEEE: 1–6, 2015, doi:arXiv:1802.01492.

[3] N. Acharya, P. Mahat, N. Mithulananthan, "An analytical approach for DG allocation in primary distribution network," International Journal of Electrical Power & Energy Systems, **28**(10), 669–678, 2006, doi:10.1016/j.ijepes.2006.02.013.

[4] A. Singh, S. Parida, "Combined optimal placement of solar; wind and fuel cell based DGs using AHP," Linköping University Electronic Press: 3113–3120, 2011, doi:10.3384/ecp110573113.

[5] M. Kashem, G. Ledwich, "Multiple distributed generators for distribution feeder voltage support," IEEE Transactions on Energy Conversion, **20**(3), 676–684, 2005, doi: 10.1109/TEC.2004.832090.

[6] S.-H. Lee, J.-W. Park, "Selection of optimal location and size of multiple distributed generations by using Kalman filter algorithm," IEEE Transactions on Power Systems, **24**(3), 1393–1400, 2009 doi: 10.1109/TPWRS.2009.2016540.

[7] N. Soukaina, E. Hassane, L. Tijani, "Hosting capacity estimation of underground distribution feeder in Urbain Areas," IEEE: 1–5, 2019, doi:10.1109/WITS.2019.8723822.

[8] M. Gomez-Gonzalez, A. López, F. Jurado, "Optimization of distributed generation systems using a new discrete PSO and OPF," Electric Power Systems Research, **84**(1), 174–180, 2012doi:org/10.1016/j.epsr.2011.11.016.

[9] D.K. Khatod, V. Pant, J. Sharma, "Evolutionary programming based optimal placement of renewable distributed generators," IEEE Transactions on Power Systems, **28**(2), 683–695, 2012 doi: 10.1109/TPWRS.2012.2211044.

[10] D. Singh, D. Singh, K. Verma, "Multiobjective optimization for DG planning with load models," IEEE Transactions on Power Systems, **24**(1), 427–436, 2009, doi: 10.1109/TPWRS.2008.2009483.

[11] M. Sedighi, A. Igderi, A. Parastar, "Sitting and sizing of distributed generation in distribution network to improve of several parameters by PSO algorithm," IEEE: 1083–1087, 2010, doi: 10.1109/IPECON.2010.5696977.

[12] M. Ettehadi, H. Ghasemi, S. Vaez-Zadeh, "Voltage stability-based DG placement in distribution networks," Ieee Transactions on Power Delivery, **28**(1), 171–178, 2012,doi:10.1109/TPWRD.2012.2214241.

[13] F. Ugranlı, E. Karatepe, "Multiple-distributed generation planning under load uncertainty and different penetration levels," International Journal of Electrical Power & Energy Systems, **46**, 132–144, 2013, doi:org/10.1016/j.ijepes.2012.10.043.

[14] M. Aman, G. Jasmon, H. Mokhlis, A. Bakar, "Optimal placement and sizing of a DG based on a new power stability index and line losses," International Journal of Electrical Power & Energy Systems, **43**(1), 1296–1304, 2012, doi:org/10.1016/j.ijepes.2012.05.053.

[15] S.K. Injeti, N.P. Kumar, "A novel approach to identify optimal access point and capacity of multiple DGs in a small, medium and large scale radial distribution systems," International Journal of Electrical Power & Energy Systems, **45**(1), 142–151, 2013, doi:org/10.1016/j.ijepes.2012.08.043.

[16] S. Ghosh, S.P. Ghoshal, S. Ghosh, "Optimal sizing and placement of distributed generation in a network system," International Journal of Electrical Power & Energy Systems, **32**(8), 849–856, 2010, doi:org/10.1016/j.ijepes.2010.01.029.

[17] M. Dehghani, Z. Montazeri, O. Malik, "Optimal sizing and placement of capacitor banks and distributed generation in distribution systems using spring search algorithm," International Journal of Emerging Electric Power Systems, 21(1), 2020, doi:https://doi.org/10.1515/ijeeps-2019-0217.

[18] M. Ismail, E. Hassane, E.M. Hassan, L. Tijani, "Power Losses Minimization in Distribution System Using Soft Open Point," IEEE: 1–5, 2020, doi: 10.1109/IRASET48871.2020.9092002.

[19] S. Ganguly, N. Sahoo, D. Das, "Multi-objective particle swarm optimization based on fuzzy-Pareto-dominance for possibilistic planning of electrical distribution systems incorporating distributed generation," Fuzzy Sets and Systems, **213**, 47–73, 2013, doi:org/10.1016/j.fss.2012.07.005.

[20] M. Aman, G. Jasmon, A. Bakar, H. Mokhlis, "A new approach for optimum DG placement and sizing based on voltage stability maximization and minimization of power losses," Energy Conversion and Management, 70, 202–210, 2013, doi:org/10.1016/j.enconman.2013.02.015.

[21] S. Arora, S. Singh, "Butterfly optimization algorithm: a novel approach for global optimization," Soft Computing, **23**(3), 715–734, 2019, doi:10.1007/s00500-018-3102-4.

[22] R. Pegado, Z. Ñaupari, Y. Molina, C. Castillo, "Radial distribution network reconfiguration for power losses reduction based on improved selective BPSO," Electric Power Systems Research, **169**, 206–213, 2019, doi:org/10.1016/j.epsr.2018.12.030.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 80-93 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Method of Modelling Prices for R&D Products in the Case of their Transfer from Engineering Universities to the Business

Oleksandra Mrykhina[1], Lidiya Lisovska[2,*], Ihor Novakivskyj[2], Terebukh Andrii[3], Valentyna Zhukovska[4]

[1]*Lviv Polytechnic National University, Department of Business Economics and Investment, 12, Banderu, Lviv, 79013, Ukraine*

[2]*Lviv Polytechnic National University, Management of Organizations Department, 12, Banderu, Lviv, 79013, Ukraine*

[3]*Lviv Polytechnic National University, Tourism Department, 12, Banderu, Lviv, 79013, Ukraine*

[4]*Kyiv National University of Trade & Economics, Department of Management, 54/1, Prospect Peremogy, Kyiv, 03057, Ukraine*

ARTICLE INFO

Article history:
Received: 02 July, 2020
Accepted: 21 August, 2020
Online: 09 September, 2020

Keywords:
Pricing
R&D product
Transfer
Commercialization

ABSTRACT

Global changes caused by the IV Industrial revolution and globalization processes resulted in a redistribution of roles of participants in innovative infrastructures of countries. Universities are leading both in terms of generating R&D products and in terms of developing business activities. Now there is a problem of insufficient methodological support of technological universities for pricing R&D products developed and prepared for transfer to the business environment. Existing methods and models do not meet the needs of the market, which is growing rapidly. At the same time, the market is characterized by a high degree of volatility. The purpose of the article is to develop a method for modelling prices for R&D products from universities to the business environment, which takes into account: the specifics of the R&D product, modern market features for this R&D product; the nature of the transfer and commercialization of this R&D product. The article identifies the factors that determine the processes of transfer, commercialization and market launch of R&D products, which affect the pricing of R&D products. Groups of characteristics that characterize systematize these factors: 1) consumer value of R&D product; 2) market susceptibility of R&D product; 3) transfer and commercialization processes of R&D product. Justified a number of factor attributes within the formed groups and assigned them the values of linguistic terms for adjusting the price of R&D product using fuzzy set theory algorithms. The method takes into account elements of cost, revenue and comparative estimation approaches. The method makes it possible to adjust prices for R&D products, taking into account heterogeneous features in the composition of R&D products and compare them with market analogues of R&D products. This contributes to achieving a higher level of pricing accuracy for R&D products when they are transferred from the university to the business environment. The resulting prices are compared with market prices for competitive analogues, which makes it possible to determine the scenario of transfer and commercialization of R&D product; justify the strategy of market development of R&D product; increase the level of manoeuvrability of pricing management for R&D product. The model was tested on a number of R&D products developed at Lviv Polytechnic National University (Ukraine). Application of the proposed method is advisable in the short-and medium-term forecasting period.

1. Introduction

1.1. Framework of theme relevance

This work is a continuation of the development of issues raised by the team of authors at the 14th International Scientific and

* Lidiya Lisovska, Address: Apt. 3, 136, Gen.Chuprunku str., Lviv, 79057, Ukraine, Contact No: +380679260337, Email: lida_lissovcka@ukr.net

Technical Conference on Computer Sciences and Information Technologies [1].

Global changes caused by the IV Industrial revolution, the popularization of the open innovation paradigm and the concept of Society 5.0, the virtualization of many areas of human activity, as well as other events and phenomena, caused the redistribution of roles of almost all participants in the innovation infrastructure of regions of countries. Leading positions are increasingly held by technological universities, both in terms of technology development opportunities and in terms of entrepreneurial development. Universities have become hotbeds for generating, transferring, commercializing, and even providing market support for technologies. Now they are the main providers of interaction between business, government and society.

This requires universities to constantly study the market needs for R&D products in the context of prospective requests due to scientific and technological progress and timely and reasonable response to such requests. Thus, today the processes of transferring R&D products from universities to the business environment are taking place against the background of market uncertainty. In contrast to business structures that are more mobile in the market competition, universities should act on the basis of careful strategic planning of R&D products transfer opportunities, evaluating the level of efficiency of technology transfer in general, being still at the beginning of the "R&D – market" chain.

The commercialization processes of University R&D products are both goals and tools. Goals, since commercialization means the relevance and expediency of further scientific research, and tools – since commercialization makes it possible to make a profit for their implementation. This requires providing universities with the necessary resources and methodological tools. In particular, pricing for R&D products prepared for transfer from universities to the business environment requires priority attention in this process.

As a result of unsubstantiated management decisions in pricing their R&D products, universities can have significant negative results both in the long term (loss of competitive positions, loss of business reputation in research schools) and in current activities (loss of revenue, revaluation of their own positions, loss of qualified personnel), which can lead to a weakening of interaction in the system of participants in the innovation infrastructure.

Until now, the world economic science has not paid much attention to the specifics of pricing for university R&D products as part of their transfer processes. However, the growing integration role of universities in the interaction of participants in the innovation infrastructure and the rapid development of technological universities in the world have led to the need to review this issue.

1.2. Statistical Background

The importance and relevance of research on pricing issues for R&D products prepared for transfer from universities to the business environment, confirm the pace of dynamic growth in the world of R&D products that are transferred by universities to the market.

According to statistics, the overall effect of transferring R&D products from universities to businesses to the global economy is estimated at Supporting regional economic growth and new job creation — up to $1.7 trillion in gross industrial output and 5.9 million jobs since 1996 [2].

According to this year's study, the global trend in R&D spending continues to grow worldwide, reflecting a surge to $240 billion in the Information & Technology sector. As in previous years, the growth in global R&D investments is being driven by spending in Asian countries, in particular China, which exceeds $500 billion in spending accounting for a 22 percent global share in investments [3].

Association of University Technology Managers (AUTM) in their research note that the contribution of universities to GDP as a whole appears to be growing faster than U.S. GDP as a whole, and faster than the manufacturing industries used in prior reports [4]. However, when considering the research-intensive industries, the modelled AUTM contribution to GDP appears to be growing, in relative terms, about as quickly as these research-intensive industries. In numbers it is the $723 billion contribution to gross output, $374 billion contribution to GDP, and providing support for 2.676 million jobs over the 22-year period, is based on an assumption of a 5% earned royalty rate on licensees' product sales.

The World University Rankings in their methodology for evaluating knowledge-intensive universities, they use achievement levels in all their main tasks: teaching, research, knowledge transfer, and worldview [5]. In the research task (takes 30%) on the rating scale, the profit from commercialization of R&D products is estimated at 9%. Despite the different size of universities and the scope of scientific activities, this indicator is adjusted to take into account the number of teachers.

Solving pricing problems when commercializing R&D products is particularly relevant for universities that have technology transfer centres or other similar divisions. A study of the world's leading university technology transfer centres (Oxford, MIT, Stanford, Cambridge, Ruprecht-Karl University of Heidelberg, the National University of Singapore, the University of Toronto, etc.) has shown that even if there is methodological support for their activities, they cannot always effectively apply it. The pace of development and market needs for R&D products is growing significantly faster than the corresponding methodological support is being developed for them, in particular in the field of pricing.

The world's leading science-intensive universities publish annual reports on commercialization to improve their reputation in business circles R&D products. Universities create specialized technology transfer units that implement programs to promote their own R&D products. In particular, start up companies based on discoveries and inventions by University researchers. The most popular form of knowledge transfer in the world, which is developing, is Creative works (no patented technologies). They may be trademarked or copyrighted and are often ready to be licensed when the inventor discloses the invention. In FY2019 [6], creative works earned $6.3 million in revenue from 109 licenses across 55 technologies and leveraged online automated licensing for over 1,000 additional transactions. There were 27 technologies made available via online app platforms with more than 47,000 app downloads. Creative works come from colleges and centres across all UMN campuses.

Paying attention to pricing for R&D products is important when universities enter into contract research based on its contractual obligations. Such transactions, providing business advice is an important source for the development of universities.

According to the annual reports of universities, sometimes such an item of income can reach 10% of the total cost of maintaining the university [7].

1.3. Formulation of the problem

Each engineering university is unique in terms of access to resources, the scale and focus of their research efforts, and the level of expertise in technology licensing, patenting, and sharing. However, most of them have similar problems related to taking into account these and other features in pricing for university-generated R&D products, in particular in a changing market and globalizing processes.

Existing approaches to pricing for R&D products cannot be applied in modern market conditions due to a number of factors, namely:

- accelerated pace and unpredictability of changes in environmental factors;
- increasing the number of participants in the innovation infrastructure, which is increasingly attracted to the region's communities;
- acceleration of changes in market conditions due to digitalization;
- spreading the use of open models of the innovation process;
- diversification of forms and methods of transmission R&D products (Express licenses available online, Fast-track opportunity licenses for early-stage technologies, Negotiated, exclusive, and nonexclusive license agreements, Sponsored research agreements (MN-IP), University start up companies (Venture Centre and Discovery Launchpad));
- limited financial and time resources for full development of R&D products by the developer (often before commercialization, R&D products can be offered at different levels of readiness);
- increasing the level of technology intellectualization;
- maturation of scientific technologies at early stages outside of scientific organizations, using the resources of commercial partners willing and able to continue the applied research and experimental development necessary for the market launch of the product;
- the emergence of unpredictable effects from the commercialization and implementation of R&D products in various sectors of the economy (convergence, spillover, crowd effect, diffusion, multiplicative effect, etc.);
- increasing dependence of the developed functional characteristics of R&D products, taking into account the specifics of the conditions for their future implementation, and so on.

The decrease in the level of predictability and objectivity of results from the use of existing pricing methods for R&D products, their low adaptability to the dynamics of the innovation environment, the growth of the impact of economic consequences from this, and other factors led to the search for new methodological tools for effective pricing of R&D products prepared for commercialization and transfer from universities to the business environment.

2. Theoretical Background

The processes of transfer and commercialization of R&D products is a complex multi-attribute mechanism, the effectiveness of which is largely determined by a reasonable choice of approaches, methods and pricing models for R&D products.

R&D product can result in different products (products, technology, organizational and management decision, etc.) and be at different Technology Readiness Levels (TRL's). Since R&D product is a commodity in the innovation market, General methodological approaches of the market economy are used for basic pricing [8].

Pricing refers to the field of strategic management decisions [9], so specialists in the system of strategic management mainly consider the approaches to its implementation. There are study three types of technology transfer strategies: income-generation strategy, service-to-faculty strategy, and local development strategy [10]. The highlighted strategies differ in the priorities for implementing the transfer strategy.

R&D product transfer, as a process of technology transfer, has its own pricing features in terms of technology transfer transactions between countries [11, 12] .

It is necessary to highlight the influence of such a factor on pricing as a form of calculation the main payment types in use by University Technology Transfer Units (TTUs) [13]. The more frequent types of payment are the running royalties, but other payment types are frequently included in the technology transfer agreements, such as the minimums, the milestone payments, and the patent costs reimbursement and maintenance.

Pricing strategy based on the subscription model could be an effective strategy for any organization to increase its revenue and maintain customer satisfaction at the same time [14]. The strategy enables the organizations to get more revenue from their subscribers over a longer period. It also reduces the upfront cost paid by users, making it more affordable, which would help increase, the number of subscribers.

The complexity of managing the transfer process and, consequently, technology pricing is complicated by the growing importance of the support package, which is «diversified and performance of three key business-community services – provision of consultancy, provision of continuous professional development courses (CPD), and leasing of facilities and equipment» [15].

The complexity of pricing for technology transfers also lies in the fact that the transaction process must be based on balance the needs of the university, researchers, licensing firms and financiers with government entities and the public that support – and ultimately benefit from – the products and services created, along with the economic vitality and job creation derived from the commercialization process.

Pricing is influenced by many factors, including [16]: (1) the characteristics of the property or service, (2) functional analysis, (3) contractual terms, (4) economic circumstances, and (5) business strategies.

In general, all the traditional approaches to R&D product pricing known in the world can be divided into three areas: cost-based, profitable and comparative, represented by many methods and methods of their application.

2.1. Cost approach

For situations with the pricing of R&D products developed in universities, with the purpose of their further transfer and

commercialisation are mostly used methods are the cost approach (for purposes of the base assessment, the definition of "bottom border" prices R&D product, development of cost estimates for studies, preparation of applications for grants and the like).

Scientists studies in modern research the valuation methods used by universities, the concerns of usage, and the difficulties in implementing the valuation methods. The author concluded that the universities typically use cost approach, market approach, income approach and auction to value the academic technologies [17]. Among which, cost approach is the most widely used one. The difficulties suffered by Taiwanese universities include lacking staffs specialized in valuation, expensive valuation service by external consultants, and the restrictions by government regulations. Based on the interview findings and previous studies, this study further designs a technology valuation framework for Taiwan universities and applies the framework to valuing the vaccines of duck viral hepatitis owned by a Taiwan university. At last, this study offers suggestion for valuing academic technologies by integrating the results from interviews and the experiences in empirical applications.

Scientists [18] use a cost-based approach in their research. When forming the price, they recommend operating the Setting standard cost. To reduce the cost of R&D product is offered the optimization method of standard cost setting.

Cost-based pricing methods for R&D product technology developed within the University have their drawbacks and limitations. Focusing on the amount of expenses incurred can completely separate the developer from the real economic situation on the market. Accordingly, the formation of estimates for development will be carried out according to the priorities of the University's development. Since research and development can take many years, often with significant interruptions, it is difficult to justify discount rates to bring costs to today's conditions, especially for countries with unstable economies.

2.2. Income approach

The revenue-based approach to pricing is widely used in the modern business environment. Scientists explore revenue approaches through econometric multifactor asset pricing models, which identifying the best factor-pricing model(s) is conspicuously lacking in investment research applications [19, 20].

In scientific works the authors use a generic reduced-form model economy with moderate risk premium nonlinearity to examine the size of the resulting misspecification-induced pricing errors [21].

Also some authors generalized model, what is presented to jointly characterize the optimal pricing and inventory policies to maximize the retailer's total expected profit [22].

The revenue approach was considered on the example of pricing for Massive Open Online Courses (MOOCs) [23]. Authors adopt a game-theoretic framework to model the interaction and strategic choices of a MOOC platform, learners, and universities. Based on the certificate prices and revenue sharing ratios chosen by the platform for courses with various certificate-purchasing rates, universities consider the competition intensity and decide their course quality levels, to attract learners.

The complexity of applying the revenue approach methods for university developments is related to the subjectivity of the justification of cash flows, which are often determined by the place of application of the development. Paying attention to the revenue approach is important in connection with the development of value-based pricing in recent years.

Some scientists pay special attention to conceptualize key stakeholders' perceptions of their experiences, opportunities, and barriers to implementing value-based pricing [24, 25].

2.3. Comparative approach

Among the methods of this direction, the most common are competitive methods, the essence of which is in comparison with its analogue to determine the competitive advantages and bottlenecks of R&D product. In particular, this method is used as a price formation tool Bert Rand price game Based on game theory, this paper studies the strategy of purchasing technology and upgrading from a technology supplier in a duopoly market, and analyses the pricing of products under different technology procurement strategies [26]. The research shows that, under the competition of product price, the game results of the enterprise's purchase of technology will be divided into symmetrical Nash equilibrium or asymmetric Nash equilibrium. Group of authors developed the complex of methodical support of the implementation and commercialization of domestic innovative devices [27].

The competitive pricing model has been developed for vintage capital model that combines a competitive market structure with an exogenous rapid rate of innovation [28].

There are many methods of estimating cost but as it is clear that we cannot consider any single technique to be the best one as each of the techniques have their own advantages and disadvantages. Efforts should be made to use a combination of the estimation techniques to arrive at a better cost and quality estimate [29]. This will allow you to justify the best pricing option for R&D products for the purpose of transferring it.

The developed approaches to determining the cost and price of R&D products by NASA and the Software Engineering Institute (federally funded research and development centre sponsored by the U.S. Department of Defense). Both teams of authors use the concept of technological readiness in their methods and promote the idea of combining different approaches.

NASA's pricing methodology is aimed at using yet-to-be-completed developments, for estimating the cost and schedule of low TRL technology research and development projects. The method involves using a system of 20 unique technology parameters that are subsequently reduced in number and suitable for use in characterizing these technologies. Further, a discussion of data acquisition effort and criteria established for data quality are provided. The authors offer recommendations on the gaps identified, description of a spreadsheet-based estimating tool initiated. The main cost analysis tool was parametric analysis.

TCASE generates anticipated ranges of cost and schedule duration for a technology development project by drawing analogies to historical and current project. Data for historical and current projects is stored in an accompanying database.

The Software Engineering Institute uses TRL Calculator (one for hardware and one for software) developed by Mr. Nolte at AFRL, it is the negotiation of the answers that is labour intensive. Thus, a good consensus building and conflict-resolution process is also needed. Therefore, the accuracy of calculating the cost

(price) of R&D products is determined by the quantity and quality of negotiations between participants in the technology transfer process. Therefore, an important issue is the choice of a rational range of interviewees. The use of this technique is limited to system components.

2.4. Theoretical generalization

Consequently, most existing methods and pricing models for R&D products are inflexible. They are difficult to adapt to the dynamic characteristics of R&D products. At the same time, appraisers often do not take into account a significant number of changing factors of the internal and external environment.

For a long time, the development of R&D products within universities leads to an increase in the duration of the chain of innovation "generation – commercialization – market diffusion", so the use of exclusively expensive methods is impossible.

Since the sources of generating R&D products are not the parameters of market demand, but the scientific achievements of the university, the use of comparative approaches can also be used in a limited format.

In turn, profitable approaches are also difficult to use in their pure form due to the growing level of convergence of markets and technologies, the emergence of new forms of business, the emergence of unexpected effects of use, and so on.

Taking into account the above, the use of quantitative methods alone will not give an adequate result of pricing, since there is a risk of not taking into account many important situational factors of a volatile market environment. The use of qualitative methods carries the risk of obtaining such an indicator of the price of R&D product that is not subject to adequate adjustment (the price, although competitive in the market, but it is difficult to adjust it under the conditions of market changes).

Currently, there are no methodological developments that would provide a flexible pricing mechanism for R&D products for their transfer from universities to the business environment in a changing market. The lack of such developments causes problems for universities:

- failure to provide an adequate contract price for R&D products during their transfer from universities to the business environment;
- it is impossible to take into account a number of important indicators of an investment project for R&D product (discount parameters, profitability indicators, cash flows, etc.);
- inefficiency of pricing strategies for R&D product (overestimating the price of R&D product in some cases, in others – underestimation, which leads to loss of profit);
- lying of R&D products projects "on the shelves", etc.

It is important to develop a method that would make it possible to adjust the price of an R&D product, depending on changes in market conditions, and overcome the disadvantages of existing methods. At the same time, this method should create the ground for changing the target setting of pricing tasks for R&D products, which will contribute to the formation of effective pricing strategies.

3. Research design

3.1. Methodology approach

The aim of this work is to develop a method for modelling prices for R&D products for their transfer from universities to the business environment, taking into account the conditions of market variability.

The development of a method for modelling the price of R&D products, taking into account the conditions of market variability, is based on a combination of cost, comparative and profitable approaches to pricing. The market is characterized by a variety of features that affect the pricing of R&D products. In practice, appraisers often neglect individual parameters of R&D products; artificially simplify the interdisciplinarity of indicators and their relationship, which reduces the effectiveness of pricing for R&D products. This complicates the processes of transfer and commercialization of R&D products, reduces the competitive stability of the product in the market.

A scientific search has shown that prices for R&D products that are being prepared for transfer from universities to the business environment can be adjusted to take into account the variability of market conditions, in particular by applying the appropriate correction factor. This coefficient is essentially an aggregate indicator that includes a number of relevant factors that influence the pricing of R&D products.

Therefore, for further research and implementation of the goal, the following hypotheses were formed.

Hypothesis 1. The methodological approach to modelling prices for R&D products when they are transferred from universities to the business environment, which takes into account changing market conditions, should be based on: 1) based on the amount of expenses incurred by developers, since the sources of their formation can be justified; 2) taking into account the correction coefficient developed to clarify the price of R&D product.

This coefficient shows the influence of a combination of factors (external and internal relative to the R&D product environment) on pricing for this R&D product.

To do this, you should develop a correction factor that takes into account expert assessments of the set of factors that affect this R&D product. These are the factors that can indicate a possible change (increase or decrease) in the value of the analysed R&D product, and which can be transferred to the pricing plane.

This hypothesis will be tested by a comparative analysis of the average market prices for a number of R&D products and prices obtained by applying a correction coefficient for the proposed author's method.

Hypothesis 2. The list of features used to model the price of an R&D product prepared for transfer from a university to a business environment is universal and exhaustive.

The signs that determine the change in the price adjustment coefficient must be analysed in the process of practical research of R&D products. This hypothesis will be tested by comparing the features used for price modelling between different R&D products.

Hypothesis 3. The methodological approach to calculating the correction factor for improving the pricing of R&D products when they are transferred from universities to the business environment is universal and effective for strategic pricing purposes.

When forming a correction coefficient, the problem arises not so much in the economic as in the mathematical context: indicators that reflect the specific features of R&D product and its transfer processes, market uncertainty, and so on are quite heterogeneous. These indicators are characterized by a complex level of mutual correlation, and their influence on the integral indicator (correction coefficient) is not always obvious.

In order to justify the method of adjusting prices for R&D products when they are transferred from universities to the business environment, which takes into account different types of indicators and their dependencies of different levels of complexity, it is advisable to use approaches within the theory of fuzzy sets.

It is proposed to develop an R&D product price adjustment coefficient based on the algorithms of fuzzy set theory, which will effectively respond to changes in market conditions (in particular, by adding or subtracting the corresponding scenarios in the knowledge base, adjusting features, etc.). One of the promising approaches of this theory for modelling and forecasting economic processes is the fuzzy logic approach, which gained popularity in economic applications in the second half of the last century.

The application of the fuzzy logic approach to develop a method for adjusting prices for R&D products during their transfer from universities to the business environment is justified by the following arguments.

- *The ability to take into account the interaction of factors of influence on the R&D product that are interrelated with different levels of complexity*. Factors that influence the price change for an R&D product can be both the specific characteristics of the R&D product itself, and the characteristics of the market to which this R&D product is displayed (in particular, the market susceptibility of this R&D product), as well as the processes of its transfer. These factors are not in a linear relationship, but are closely related, so considering them in the system using fuzzy set theory algorithms will help refine the price of an R&D product.
- *Adaptive capacity of expert assessments.* Methods based on the theory of fuzzy sets have a high adaptive capacity according to expert estimates and are quite adequate in practical use.
- *Combination of qualitative and quantitative assessments.* The development of the R&D product price correction coefficient within the framework of fuzzy set theory operates with both qualitative and quantitative forms of its implementation, which makes it possible to evaluate the features, external and internal features, properties of the R&D product market, and so on. The method also allows you to take into account estimates that pass from one state to another.
- *Feedback.* The R&D product price adjustment method is based on a number of indicators that can change during the calculation of the adjustment coefficient. For example, if some indicators are outdated or other indicators need to be introduced, they can be replaced, along with appropriate modification of scenarios in the knowledge base. This is based on the principles of feedback-from the market entity (consumer, enterprise, etc.) to the university.

By adjusting the price of R&D products based on fuzzy set theory, we understand the process by which, based on available data, you can set significant parameters that are not directly measured, but determine the nature of pricing for R&D products and make it possible to analyse changes in its state.

The criterion for the effectiveness of price adjustment for R&D products in the framework of fuzzy set theory is expressed by maximizing the degree of suitability of the obtained estimates for a specific market situation.

In economics, the most common models used for this type of problem are Mamdani, Sugeno, Tzukamoto, Larsen, and others. To adjust the price of R&D products based on fuzzy sets, we selected the Mamdani model, which differs from other models in that its rules in sequences contain fuzzy values (membership functions).

The methodology of estimation processes based on the theory of fuzzy sets is based on a number of theoretical assumptions, in particular, stated by Zadeh L. (1976).

First, it is assumed that the object of research can enter a state that belongs to a finite set of states S (the action of external and internal processes changes, but the lack of its measurement leads to a restriction of the set S). On the set S, there are two disjoint subsets $S_1 \cup S_2$, where S_1 is a subset of states that are consistent with the characteristics of the desired state by the level of development of features; S_2 is a subset of states that are not consistent with the characteristics of the desired (normative or reference) state by the level of development of features. The subset $S_1 = \{s_i\}, i = \overline{1,n}$ contains state types that allow for functioning and development. The subset $S_2 = \{s_j\}, i = \overline{1,m}$ contains state types that respond to changes in parameters or structural relationships that contribute to deviations from the desired result.

Second, the solution to the issue of inventory and evaluation of the real state of the object is based on the analysis of the set S or a subset of S_1 and S_2.

Third, the assessment is based on the appraisers' positions and methods of conducting it.

Fourth, detecting deviations from the characteristics of the desired state means that the object is in the process of transitioning from the S_k state to the S_l state, but the operating conditions may not be violated if S_k and S_l belong to a subset of S_1 states.

3.2. Data analysis

Based on the research, the factors that determine the processes of transfer, commercialization and market launch of R&D products, which affect the pricing of R&D products, are identified and systematized by groups of features:

Group A ($f(x_1)$) – characteristics that characterize the consumer value of an R&D product;

Group B ($f(x_2)$) – features that characterize the market susceptibility of the R&D product;

Group C ($f(x_3)$) – features that characterize the processes of transfer and commercialization of R&D product.

Group A combines comparative indicators that allow us to assess the value of R&D products for consumers. Consumer value characterizes the utility of an R&D product to meet the consumer's needs. The value that the developer creates in the form of an R&D product determines the consumer value of the R&D product in the form of a finished product and the consumer value of the product that is obtained using the R&D product. This group of attributes is not personalized and is directly related to the

characteristics of the R&D product. The level of this group of attributes allows us to assess the potential consumer value of R&D product in meeting the consumer's needs.

In particular, the indicators include three levels of R&D product readiness for commercialization (technological, patent, and analytical). These three levels of readiness are related to the assessment of the completeness (feasibility) of the R&D product and determine the level of consumer value of the R&D product as a commodity.

NASA development can be used to determine the level of technological readiness [30]. This organization has developed and successfully uses a methodology for assessing the levels of technological readiness (TRL).

The level of patent readiness includes an assessment of the availability (absence) of documents for intellectual property objects or the process of obtaining them. The level of patent readiness is evaluated as the level of protection of the R&D product and directly affects the consumer value of the R&D product.

The level of analytical readiness (ARL) can be assessed by the degree of development of analytical documents by R&D product developers. The lowest level of analytical readiness is associated with the justification of the R&D product concept. The highest level of ARL is considered a business plan for introducing innovations to the market. The development of analytical documents is based on marketing information and allows you to simplify the process of using R&D product.

Signs of the level of novelty and level of improvement of R&D product characterize the competitive advantages of R&D product as an object of innovation. The level of novelty of R&D product is related to the number of technological parameters that are introduced by the developer for the first time and distinguish R&D product from analogues. If R&D is product improving or modifying an innovation to address customer needs, R&D product has a certain level of improvement, such as increasing the level of existing technological options, R&D product.

The level of quality and competitiveness of a product that is obtained on the basis of R&D product, related to the evaluation of indicators of competitive advantages of the product (products, work, service), which is expected to be obtained as a result of the application of R&D product.

The level of possible unauthorized reproduction of an R&D product is related to the level of complexity of the R&D product. A low level of complexity may expose the consumer to the risk of losing the value of the R&D product.

The level of environmental friendliness of R&D product provides an assessment of the impact of R&D product on the environment when using R&D product or a product made on the basis of R&D product (environmental cleanliness of the development application, reducing environmental pollution, energy saving, reducing raw material costs, etc.) development.

R&D product may not affect the ecology of society, or reduce (increase) its influence on the analogue, both positive and negative. Accordingly, the level of environmental friendliness of the R&D product can increase or decrease the consumer value of the R&D product.

Group B includes features that are correlated with the assessment of the conditions for the introduction of R&D product in a particular market (segment). These indicators allow us to assess the level of attractiveness of market conditions for R & D products (direct action marketing environment).

In particular, the market susceptibility of an R&D product is directly determined by the state and intensity of competition in the target market. A high level of competition, i. e. a large number of similar R&D products and R&D product substitutes, impairs market sensitivity and reduces the level of consumer value of the R&D product [31, 32].

The size of the sales market and the rate of change in its value directly affect the market signs of the introduction of R&D product. These attributes are related to the quantitative parameters of the sales market. Their high level allows us to assess the market sensitivity of the technology as high.

The update level of the market for R&D product provides an opportunity to assess the level of intensity of innovation in a particular economic activity and is associated with the estimation of time expected use R&D to product replacement or improvement.

The level of readiness of consumers to use the product that is obtained based on R&D product is associated with the stage of the life cycle of the need and technology and affects the marketing costs for the consumer.

Group C allows you to evaluate the features of transferring R&D product to a specific consumer and the specifics of implementing R&D product in its activities.

Therefore, the second and third groups of attributes and their estimates are adaptive to the conditions of a specific R&D product transfer transaction.

The third group of features involves comparing and determining the level of attractiveness of the conditions for transferring rights to R&D product to the consumer and the conditions for introducing R&D product to the consumer.

The transfer terms are evaluated by the components of the transaction (the level of attractiveness of the financial terms of the R&D product transfer-providing discounts, instalment payments, rentals, etc.), as well as by the competitive advantages that the developer provides when implementing the R&D product (the level of completeness of technological support by technology developers).

Becoming increasingly important the signs associated with the direct introduction of R&D product from the consumer to evaluate the need for and scope of additional time, financial and other resources to conduct additional tests (the level of difficulty implementation R&D product, the level of necessary material resource support integration of R&D product, the level of initial investment costs that should be involved to implement the R&D product).

This group includes indicators of commercial (investment) attractiveness of R&D product (the level of payback, the level of indicators of economic efficiency of investment costs (commercialization)).

The hierarchical order of factors of influence (f (x)) and their features (p) on the price of R&D products is shown in Fig. 1.

The elements of the hierarchy of factors of influence and attributes in Fig. 1 are interpreted as follows: *S* – relative indicator of price change of R&D product (top of the hierarchy*), %; x1, x2, x3* – generalizing factors influencing the price of R&D product (thermal peaks); $p_{11} \dots p_{1n}$, $p_{21} \dots p_{2n}$, $p_{31} \dots p_{3n}$ –signs of factors influencing the price of the R&D product.

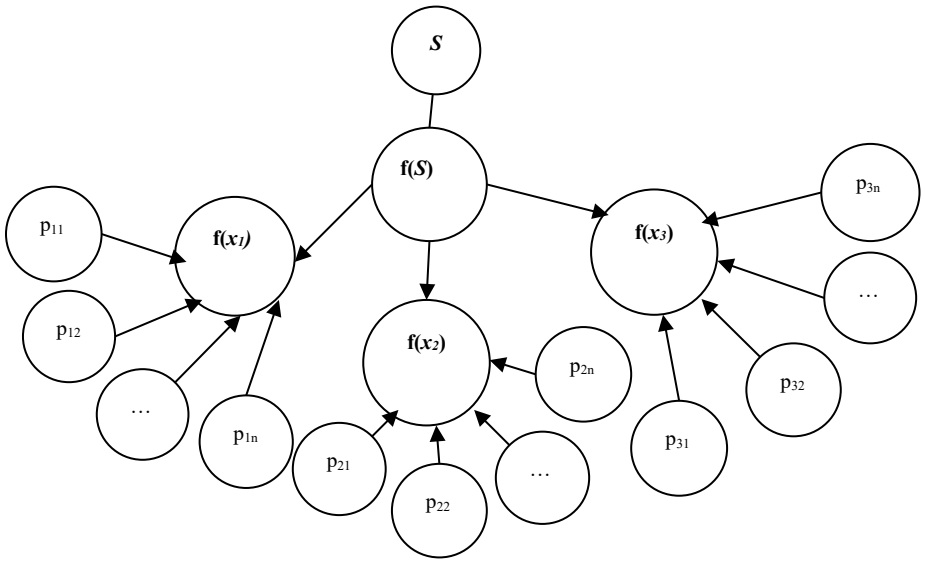


Figure 1: Hierarchical order of the influence of factors and attributes they contain on the price of R&D products

Reductions of f_S, f_{x1}, f_{x2}, f_{x3} are performed based on a logical output with fuzzy knowledge bases. A fuzzy subset of a set S is defined as a set of ordered pairs $A = \{x, \mu_A(x); x \in S\}$, where $\mu_A(x)$ is the characteristic function that takes values from some ordered set M = [0, 1] – many accessories $\mu_x\ (x) > 0, \forall x \in S, \mu_x\ (x) > 0, \forall x \nexists S, sup_{x \in S}[\mu_x\ (x)] = 1$.

In this case, the function $\mu_x\ (x)$ indicates the degree of belonging of element x to a subset of A and is a tool for converting linguistic variables to a mathematical language for further application of the fuzzy logic method.

The description of the R&D product price adjustment process based on the application of fuzzy set theory can be divided into stages: 1) fuzzification; 2) development of fuzzy rules (generalization of antecedents and consequents; inference using logical connections); 3) defazzification.

4. Empirical results

The analysis made it possible to identify the most popular features of factors within groups A, B, and C at this time and assign them the value of linguistic terms for adjusting the price of R&D product using the Mamdani model (table 1).

Table 1: Signs and values of linguistic terms of factors influencing the price of R&D product

Factors	Signs	Meaning and explanation of linguistic terms of the factor
1	2	3
Group A ($f(x_1)$) – features of the consumer value of an R&D product	The technological readiness level of R&D product	L – low indicator level [0; 1.25; 2.5]; T – tolerable indicator level [2.5; 3.75; 5]; A – applicable indicator level [5; 6.25; 7.5]; H – high indicator level [7.5; 8.75; 10].
	The patent readiness level of R&D product	
	The analytical readiness level of R&D product	
	The level of innovation of R&D product	
	The improvement of R&D product	
	The level of possible unauthorized reproduction of R&D product	
	The level of probability of unexpected effects from the introduction and use of R&D product	
	The level of quality and competitiveness of the product that is obtained on the basis of R&D product	
	The environmental level of R&D product	
Group B ($f(x_2)$) – features of the market susceptibility of the R&D product	The level of direct competition	L – low indicator level [0; 1.25; 2.5]; T – tolerable indicator level [2.5; 3.75; 5]; A – applicable indicator level [5; 6.25; 7.5]; H – high indicator level [7.5; 8.75; 10].
	The level of competition among substitutes	
	The size of the market for R&D product	
	The growth rate of the R&D product sales market	
	The update level of the market for R&D product	
	The level of readiness of consumers to use the product that is obtained on the basis of R&D product	
	The level of time expected to use an R&D product before replacement or improvement	
Group C ($f(x_3)$) – features of the	The level of initial investment costs that need to be raised for the implementation of R&D product	L – low indicator level [0; 1.25; 2.5];

processes of transfer and commercialization of R&D product	The cost level for ongoing R&D product maintenance	T – tolerable indicator level [2.5; 3.75; 5]; A – applicable indicator level [5; 6.25; 7.5]; H – high indicator level [7.5; 8.75; 10].
	The difficulty level of the integration of R&D product	
	The level of time spent on implementation and revision	
	The level of required regulatory support for R&D product implementation (obtaining permits, etc.)	
	The level of indicators of economic efficiency of investment costs (commercialization)	
	The level of attractiveness of financial terms of R&D product transfer	
	The level of recoupment	
	The level of expenses for marketing support of commercialization	
	The level of necessary material resource support for R&D product implementation	
	The level of access to the necessary resources to use R&D product	
	The level of need to establish various forms of interaction to use (or the level of sufficient capacity for self-implementation) R&D product	
	The level of completeness of technological support by technology developers	

The model for adjusting product R&D prices when they are transferred from universities to the business environment can be expressed as a function:

$S = f$ (the value of consumption of R&D products, and the market susceptibility of R&D products, production and transfer processes and commercialization of scientific and technical products), which is a system of functions in the following order:

$f(x_1) = f$ (*the technological readiness level of R&D product; the patent readiness level of R&D product; the analytical readiness level of R&D product; the level of innovation of R&D product; the improvement of R&D product; the level of possible unauthorized reproduction of R&D product; the level of probability of unexpected effects from the introduction and use of R&D product; the level of quality and competitiveness of the product that is obtained on the basis of R&D product; the environmental level of R&D product);*

$f(x_2) = f$ (*the level of direct competition; the level of competition among substitutes; the size of the market for R&D product; the growth rate of the R&D product sales market; the update level of the market for R&D product; the level of readiness of consumers to use the product that is obtained on the basis of R&D product; the level of time expected to use an R&D product before replacement or improvement);*

$f(x_3) = f$ (*the level of initial investment costs that need to be raised for the implementation of R&D product; the cost level for ongoing R&D product maintenance; the difficulty level of the integration of R&D product; the level of time spent on implementation and revision; the level of required regulatory support for R&D product implementation (obtaining permits, etc.); the level of indicators of economic efficiency of investment costs (commercialization); the level of attractiveness of financial terms of R&D product transfer; the level of recoupment; the level of expenses for marketing support of commercialization; the level of necessary material resource support for R&D product implementation; the level of access to the necessary resources to use R&D product; the level of need to establish various forms of interaction to use (or the level of sufficient capacity for self-implementation) R&D product; the level of completeness of technological support by technology developers).*

Using the *Mamdani* model in the Fuzzy Logic Toolbox component of the MATLAB software package, a model was developed for adjusting prices for R&D products when they are transferred from universities to the business environment.

Based on the generated set of factor terms (table 1), a knowledge base has been developed – possible scenarios for the ratio of R&D product consumer value factors (Group A), market susceptibility R&D product (Group B), R&D product transfer and commercialization processes (Group C). 33 rules have been established and justified. Note that in some scenarios, the weight of the rule is applied, which is measured in the range [0 ... 1] and indicates the level of significance of a particular scenario when adjusting R&D product prices. A fragment of this rule base is shown in figure 2.

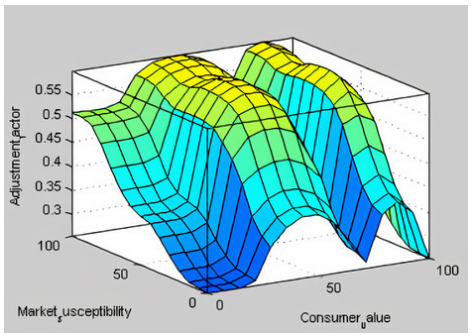
Developed model of R&D product price adjustment from universities in the business environment based on the theory of fuzzy sets enables:

- to enter grades for groups of signs, A, B, C (previously prepared by experts) and to obtain the correction factor prices R&D product;
- the resulting coefficient takes into account the current market volatility conditions and can serve as a basis for developing a strategy for R&D product market development.

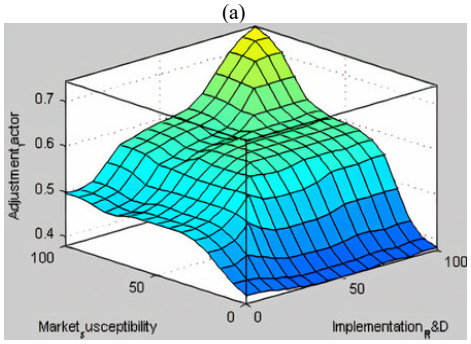
A fuzzy model for adjusting the price of R&D product when transferring from universities to the business environment is shown in figure 3.

1. If (Consumer_value is Low) or (Market_susceptibility is Low) or (Implementation_R&D is Low) then (Adjustment_factor is Low) (1)
2. If (Consumer_value is Low) and (Market_susceptibility is Average) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
3. If (Consumer_value is Low) and (Market_susceptibility is High) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
4. If (Consumer_value is Low) and (Market_susceptibility is Average) and (Implementation_R&D is Average) then (Adjustment_factor is Average) (1)
5. If (Consumer_value is Low) and (Market_susceptibility is Average) and (Implementation_R&D is High) then (Adjustment_factor is Tolerable) (1)
6. If (Consumer_value is Low) and (Market_susceptibility is High) and (Implementation_R&D is High) then (Adjustment_factor is Tolerable) (1)
7. If (Consumer_value is Low) and (Market_susceptibility is High) and (Implementation_R&D is Average) then (Adjustment_factor is Tolerable) (1)
8. If (Consumer_value is Average) or (Market_susceptibility is Average) or (Implementation_R&D is Average) then (Adjustment_factor is Average) (1)
9. If (Consumer_value is Average) and (Market_susceptibility is Average) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
10. If (Consumer_value is Average) and (Market_susceptibility is Average) and (Implementation_R&D is High) then (Adjustment_factor is Average) (1)
11. If (Consumer_value is Average) and (Market_susceptibility is High) and (Implementation_R&D is Average) then (Adjustment_factor is Average) (1)
12. If (Consumer_value is Average) and (Market_susceptibility is High) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
13. If (Consumer_value is Average) and (Market_susceptibility is Low) and (Implementation_R&D is Average) then (Adjustment_factor is Average) (1)
14. If (Consumer_value is Tolerable) and (Market_susceptibility is Low) and (Implementation_R&D is Low) then (Adjustment_factor is Low) (1)
15. If (Consumer_value is Tolerable) and (Market_susceptibility is Average) and (Implementation_R&D is Average) then (Adjustment_factor is Average) (1)
16. If (Consumer_value is Tolerable) and (Market_susceptibility is High) and (Implementation_R&D is High) then (Adjustment_factor is Tolerable) (1)
17. If (Consumer_value is Tolerable) and (Market_susceptibility is Low) and (Implementation_R&D is High) then (Adjustment_factor is Average) (1)
18. If (Consumer_value is Tolerable) and (Market_susceptibility is Average) and (Implementation_R&D is Average) then (Adjustment_factor is Average) (1)
19. If (Consumer_value is Tolerable) or (Market_susceptibility is High) or (Implementation_R&D is High) then (Adjustment_factor is High) (1)
20. If (Consumer_value is Tolerable) and (Market_susceptibility is Average) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
21. If (Consumer_value is Tolerable) and (Market_susceptibility is High) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
22. If (Consumer_value is High) or (Market_susceptibility is High) or (Implementation_R&D is High) then (Adjustment_factor is High) (1)
23. If (Consumer_value is High) and (Market_susceptibility is Average) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
24. If (Consumer_value is High) and (Market_susceptibility is Average) and (Implementation_R&D is High) then (Adjustment_factor is Tolerable) (1)
25. If (Consumer_value is High) and (Market_susceptibility is High) and (Implementation_R&D is Low) then (Adjustment_factor is Average) (1)
26. If (Consumer_value is High) and (Market_susceptibility is High) and (Implementation_R&D is Average) then (Adjustment_factor is Tolerable) (1)

Figure 2: Fragment of the database of rules for adjusting prices for R&D products when they are transferred from universities to the business environment



(a)



(b)

Fig. 3. Fuzzy model of R&D product price adjustment transfer from universities to the business environment: (a) influence customer value R&D product and market receptivity to corrective cost ratio R&D product; (b) influence of market receptivity, R&D product and the nature of its transfer and commercialization at the correct price ratio of R&D product

Using the described methodology, we will adjust the prices of a number of R&D products in the sequence (table 2):

1) *Expert assessment.* Each feature of R&D product is characterized by the level of its manifestation, which is proposed to be assessed by experts, in points, in the range 1...3 where 1 is little or no effect symptom, 2 – moderate effect of the symptom, 3 – severe impact signs (col. *3*), and then summarize the points and assign them a percentage (for example, on the first group: 9 points – the probability of a minimum 27 – maximum probable). Based on this, a generalized estimate is formed for a group of factor attributes, in %. Indicators for the corresponding R&D product are evaluated by experts based on questionnaires. The reliability of the evaluation results is confirmed by a sample of at least 15 experts;

2) *identification of the obtained factor value*, % (col. *4*);

3) *entering the obtained values into the developed model for determining the correction coefficient*. The obtained factor value with the average values of this indicator in the Fuzzy Logic Toolbox component of the MATLAB software package (based on a pre-formed knowledge base) are compared.

Therefore, the proposed method takes into account the ratio of elements of income, cost and comparative estimation approaches.

The method makes it possible to conduct economic forecasting in conditions of uncertainty; in particular, it is especially important at the stage of conceptual decision-making, planning the transfer and commercialization of R&D products from the university to the business environment, and so on. The proposed method has been tested on a number of R&D products developed at Lviv Polytechnic National University (Ukraine) and prepared for transfer to the business environment.

Table 2: Factors, signs of R&D product and their estimates

Factors	Signs	The level of feature manifestation, points	Meaning and explanation of linguistic terms of the factor
1	2	3	4
Group A ($f(x_1)$)	The technological readiness level of R&D product	1…3	L – low level of sign – 0 …25%; T – tolerable level of sign – 26 …50%; A – applicable level of sign – 51 … 75%; H – high level of sign – 76 …100%
	The patent readiness level of R&D product	1…3	
	The analytical readiness level of R&D product	1…3	
	The level of innovation of R&D product	1…3	
	The improvement of R&D product	1…3	
	The level of possible unauthorized reproduction of R&D product	1…3	
	The level of probability of unexpected effects from the introduction and use of R&D product	1…3	
	The level of quality and competitiveness of the product that is obtained on the basis of R&D product	1…3	
	The environmental level of R&D product	1…3	
	Factor estimate, %	0…100%	
Group B ($f(x_2)$)	The level of direct competition	1…3	L – low level of sign – 0 …25%; T – tolerable level of sign – 26 …50%; A – applicable level of sign – 51 … 75%; H – high level of sign – 76 …100%
	The level of competition among substitutes	1…3	
	The size of the market for R&D product	1…3	
	The growth rate of the R&D product sales market	1…3	
	The update level of the market for R&D product	1…3	
	The level of readiness of consumers to use the product that is obtained on the basis of R&D product	1…3	
	The level of time expected to use an R&D product before replacement or improvement	1…3	

	Factor estimate, %	0…100%	
Group C ($f(x_3)$)	The level of initial investment costs that need to be raised for the implementation of R&D product	1…3	L – low level of sign – 0 …25%; T – tolerable level of sign – 26 …50%; A – applicable level of sign – 51 … 75%; H – high level of sign – 76 …100%
	The cost level for ongoing R&D product maintenance	1…3	
	The difficulty level of the integration of R&D product	1…3	
	The level of time spent on implementation and revision	1…3	
	The level of required regulatory support for R&D product implementation (obtaining permits, etc.)	1…3	
	The level of indicators of economic efficiency of investment costs (commercialization)	1…3	
	The level of attractiveness of financial terms of R&D product transfer	1…3	
	The level of recoupment	1…3	
	The level of expenses for marketing support of commercialization	1…3	
	The level of necessary material resource support for R&D product implementation	1…3	
	The level of access to the necessary resources to use R&D product	1…3	
	The level of need to establish various forms of interaction to use (or the level of sufficient capacity for self-implementation) R&D product	1…3	
	The level of completeness of technological support by technology developers	1…3	
	Factor estimate, %	0…100%	

5. Discussion

Taking into account peculiarities of market situation, groups of attributes can be supplemented and / or modified.

The proposed method is tested on a number of scientific and technological complexes, developed at Lviv Polytechnic National University and prepared for transfer to business environment. The obtained results are shown in Table 3.

Table 3: Results of testing the R&D product price adjustment method prepared for transfer from the university to the business environment

№	R&D products	Factor / factor estimate, %			Correction factor, %	Current price, ths. USD	Adjusted price, ths. USD	Price change, + / -; ths.USD	Market price, analogue price, ths. USD
		Group A	Group B	Group C					
1	2	3	4	5	6	7	8	9	10
1	Technology of optically-stimulated luminescent dosimetry	13% (L)	44% (T)	87% (H)	-0,0876	23 020,80	22 999,63	-21,16	25 600
2	Ensuring Technological Strength of Welded Joints with Armored Steel of ARMSTAL 500-Type	11% (L)	34% (T)	61% (A)	0,0698	120,890	129,328	+8,438	142,300
3	Technology of surface friction strengthening of working surfaces of machine parts	27% (T)	44% (T)	71% (A)	0,0321	156,90	161,936	+5,036	165,89
4	Mobile robotic platforms: MRP-05 "Borsuk" and MRP-07 "Kubik"	47% (T)	73% (H)	82% (H)	-0,023	345,90	334,94	-10,96	340,90
5	Continuation of the	41% (T)	44%	73%	0,067	110,78	118,202	+7,42	129,97

	life of machine parts by the method of surfacing under a layer of flux		(T)	(H)					
6	Autonomous system for detection of smoke and carbon leakage monoxide	51% (A)	78% (A)	77% (H)	-0,036	0,084	0,081	-0,003	0,205
7	Predicting and providing a set value of the initial resistance of heterogeneous and homogeneous materials of small thicknesses and cross sections with a point capacitor contact micro-welding	39% (T)	42% (T)	72% (A)	0,0912	0,370	0,403	+0,033	0,518
8	Equipment for contact point capacitor micro-welding	34% (T)	67% (A)	45% (T)	-0,027	0,074	0,072	+0,002	0,096
9	Technology of surface friction displacement of working surfaces of machine parts	43% (T)	32% (H)	68% (A)	0,33	2,10	2,73	+0,63	22,00
10	Pocket City	90% (H)	45% (T)	81% (A)	0,002	3,20	3,2064	+0,0064	28,00
11	Bulk food mixer	50% (T)	48% (T)	63% (A)	0, 0731	3,10	3,32	+0,22	17,80

From the obtained results it can be seen, for example, in the cases of R&D products from the positions 2,3, 5, 7, 9, 10 and 7 (according to table. 3) it is quite possible to set the price at a higher level (in particular, at the level of corrected estimates).

It is predicted that sales of such products will not decrease, but will remain at the same level. After all, the adjusted estimate is based on a number of indicators that reflect the market perception of the product, the level of which in these cases significantly exceeds the level previously included in the cost of R&D product.

In cases 1, 4, 6 and 8, the price of the R&D product should be slightly reduced, which will help to increase its price competitiveness in the market, and, consequently, increase sales volumes.

The values obtained as a result of applying the model are characterized by an indistinct number with a certain range of values, which makes it possible to operate not with probabilistic estimates, but with project data. This helps to achieve a higher level of price accuracy for R&D product when it is transferred from the university to the business environment.

6. Conclusions

Hypothesis 1 comes true. A methodological approach to pricing for R&D products during their transfer from universities to the business environment is developed based on taking into account changing market conditions, which in particular were given in table 2 and on the basis of certain expenses for the R&D products. The result can become the initial price of R&D products in the process of discussing the transaction for its transfer.

A practical study of R&D products for signs that determine the impact on their prices has shown that *hypothesis 2* does not come true. Since each specific situation with R&D product has its own pricing features, due to many factors. For example, those features from factor groups A, B, and C that are relevant for their consideration when modelling prices for R&D products in the field of industrial electronics will not be relevant for R&D products in the field of it. In addition, the same evaluation criteria for R&D products in different industries may have different approaches to the interpretation of their essence. Therefore, the proposed list of features is not exhaustive, and may vary depending on the types of R&D products, their branches of application, features of transfer from the university to the business environment, and so on.

Hypothesis 3 is confirmed, since the methodological approach to calculating the correction factor for improving the pricing of R&D products when they are transferred from universities to the business environment is universal. This approach can be applied regardless of the stage of readiness of R&D products, and depending on the possibility of its transfer (in particular, in situations where the buyer is ready to purchase these R&D products at the initial stage of readiness). The methodological approach is effective for strategic pricing purposes.

The criterion for the effectiveness of modelling prices for

R&D products in the framework of fuzzy set theory is expressed by maximizing the degree of suitability of the obtained estimates for a specific market situation.

The advantages of the proposed methodological approach to pricing for R&D products are as follows:

- justification of the price based on the principles of the author's method gives a more reliable result (takes into account changing market conditions, implementation conditions for a specific buyer, is characterized by flexibility in accounting indicators, etc., the level of novelty);
- the method is aimed at accounting not only for static expenses (which sometimes even become irrelevant because they are spread over a long period of time), but also for dynamic market conditions of R&D products, conditions of the consumer's operating environment, and so on;
- the developer (university) has a lot of flexibility in terms of profit management (in particular, according to the results of the study, in most cases, the margin profit increases);
- the level of trust of business partners is growing at the university, as the procedures for modelling prices for R&D products prepared for transfer from universities to the business environment become significantly more transparent;
- the author's approach can be used not only for the price, but also for other pricing parameters for R&D products;
- the metrics used in the method can be modified depending on the type of R&D product (technological, product, etc.) and market conditions;
- Disadvantages:
- the use of the proposed method is advisable only in the short- and medium-term forecasting period;
- the method is essentially based on the competence of experts. Experts should have a proper level of knowledge, rely on the results of marketing research, and have operational data on the given characteristics.

The approach developed by the authors develops the principles of marketing pricing, providing opportunities to clarify the mechanism of influence of factors on the formation of the value of R&D products, and therefore their prices.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We thank the developers of R&D products and scientists of Lviv Polytechnic National University (Ukraine), who shared the pearls of their knowledge and experience, which made it possible to conduct research and justify the results.

References

[1] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, L. Lisovska, "Consumer aspects in assessing the suitability of technologies for the transfer", in 2019 Proceedings IEEE /14th International Scientific and Technical Conference on Computer Sciences and Information Technologies, 142-147, 2019, doi:10.1109/STC-CSIT.2019.8929879.

[2] About Tech Transfer [online] Available at: https://autm.net/about-tech-transfer / Pages/ https://www.autm.net/AUTMMain/media/Advocacy/Documents/Points_to_Consider.pdf [Accessed 21.03.2020]

[3] Global Funding Forecast Predicts Growth of R&D Spending Worldwide [online] Available at: https://www.rdworldonline.com/global-funding-forecast-predicts-growth-of-rd-spending-worldwide/ [Accessed 11.02.2020]

[4] The Economic Contribution of University/Nonprofit Inventions in the United States: 1996 – 2017. [online] Available at: https://autm.net/AUTM/media/About-Tech-Transfer/Documents/Economic_Contribution_Report_BIO_AUTM_JUN2019_web.pdf/ [Accessed 12.02.2020].

[5] Young University Rankings 2019 [online] Available at: https://www.timeshighereducation.com/world-university-rankings/2019/young-university-rankings#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats/ [Accessed 12.04.2020].

[6] Technology Commercialization [online] Available at: https://research.umn.edu/units/techcomm/about-us/overview/ [Accessed 19.04.2020].

[7] Deakin University Annual Report 2018 [online] Available at: https://www.deakin.edu.au/__data/assets/pdf_file/0007/1906819/300519-Deakin_AR_2018_pages_web.pdf/ [Accessed 17.04.2020].

[8] D. Hague, Pricing in business (Vol. 23), Routledge, 2018.

[9] F. Brescia, G. Colombo, P. Landoni, "Organizational structures of Knowledge Transfer Offices: an analysis of the world's top-ranked universities", J. Technol. Transf,. **41**(1),132–151, 2016, doi:10.1007/s10961-014-9384-5.

[10] P. Giuri, F. Munari, A. Scandura, L. Toschi, "The strategic orientation of universities in knowledge transfer activities", Technological Forecasting and Social Change, **138**, 261-278, 2019, doi:10.1016/j.techfore.2018.09.030.

[11] A. M. Rugman, L.Eden, (Eds.), Multinationals and transfer pricing. Routledge, 2017.

[12] D. Mescall, K. J. Klassen, "How Does Transfer Pricing Risk Affect Premiums in Cross-Border Mergers and Acquisitions? ", Contemporary Accounting Research, **35**(2), 830-865, 2018, doi:10.1111/1911-3846.12397.

[13] A. Rocha, R. Lima, M. Amorim, F. Romero, "Payment types included on technology licensing agreements and earnings distribution among Portuguese universities", Tékhne, **15**(2), 100-107, 2017, doi:10.1016/j.tekhne.2017.11.001

[14] A. Singh, Technology software pricing in cloud computing era, Ph. D Thesis, Massachusetts Institute of Technology, 2017.

[15] R. Zhou, P. Tang, "The role of university Knowledge Transfer Offices: Not just commercialize research outputs!", Technovation, 102100, 2019, doi:10.1016/j.technovation.2019.102100.

[16] C. P. Rossing, M.Cools, C. Rohde, "International transfer pricing in multinational enterprises", Journal of Accounting Education, **39**, 55-67, 2017, doi:10.1016/j.jaccedu.2017.02.002.

[17] M. Y. Wang, "The valuation methods and applications for academic technologies in Taiwan", in 2016 Portland International Conference on Management of Engineering and Technology (PICMET), IEEE, 1320-1327, 2016. doi:10.1109/PICMET.2016.7806739.

[18] Q. Meng, J. Lou, J. Zhu, X. Bai, "Standard cost setting and application of improved ant colony optimization algorithm", Systems Engineering-Theory & Practice, (**7**), 9, 2016.

[19] F. Barillas, J. Shanken, "Comparing asset pricing models", The Journal of Finance, **73**(2), 715-754, 2018, doi:10.1111/jofi.12607.

[20] Y. Jin, P. Lu, Q. Qi, Z. G. Tang, T. Xiao, "Tight approximation ratio of anonymous pricing", in 2019 Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, 674-685, 2019, doi:10.1145/3313276.3316331.

[21] M. W. Brandt, D. A. Chapman, "Linear approximations and tests of conditional pricing models", Review of Finance, **22**(2), 455-489, 2018, doi:10.1093/rof/rfy003.

[22] Q.Wang, N.Zhao, J.Wu, Q.Zhu, "Optimal Pricing and Inventory Policies with Reference Price Effect and Loss-Averse Customers", Omega,102174,2019, doi:10.1016/j.omega.2019.102174.

[23] L. C. Kung, W. C. Lee, "Pricing and Diversification of Massive Online Open Course Platforms", in 2017 PACIS , Twenty First Pacific Asia Conference on Information Systems Langkawi, 2017, 202. [online] Available at: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1077&context=pacis2017 [Accessed 23.11.2019].

[24] A. Mejía, R. Gilardino, F. B.Kristensen, L. P. Garrison, I. H. T. A. L. America, "Value-Based Pricing in Latin America: How Far Away Are We?", Value in health regional issues, **17**, 219-223, 2018, doi:10.1016/j.vhri.2018.09.007.

[25] P. M. Danzon, "Affordability challenges to value-based pricing: mass diseases, orphan diseases, and cures", Value in Health, **21**(3), 252-257. 2018, doi:10.1016/j.jval.2017.12.018.

[26] X. Qian, W. Liu, J. Yang, "Game theory analysis of technology adoption timing and pricing decision in supply chain system under asymmetric nash

equilibrium", Journal of Intelligent & Fuzzy Systems, **35**(3), 3101-3111, 2018, doi:10.3233/JIFS-169664.

[27] T. J. Kudryavtseva, E. A. Ivanova, E. A. Kozlova, A. E. Skhvediani, "Pricing and assessment of competitiveness of innovative medical devices in the context of commercialization strategy", Academy of Strategic Management Journal, **16**, 110, 2017.

[28] A. Copeland, A. H. Shapiro, "Price setting and rapid technology adoption: The case of the PC industry", Review of Economics and Statistics, **98**(3), 601-616, 2016, doi:10.1162/REST_a_00539.

[29] S. Shekhar, U. Kumar, "Review of various software cost estimation techniques", International Journal of Computer Applications, **141**(11), 31-34, 2016, doi:10.5120/ijca2016909867.

[30] S. Hirshorn, S. Jefferies, "Final Report of the NASA Technology Readiness Assessment (TRA)", Study Team, 2016. [online] Available at: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20170005794.pdf[Accessed 27.12.2019]

[31] O. I. Karyy, H. V. Podvalna, "Relationship marketing of automobile transportation companies: The need of establishing mutual understanding with a client", Actual problems of economics, **10**(184), 149-158, 2016.

[32] G. T. Piatnytska, V. M. Zhukovska, "Domestic trade development: Current transformation and priopities in socialization", Actual problems of economics, **11**(173), 106-119, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 81-93 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Maximum Power-Point Tracking and Stall Control with Eddy Current Brake System on Small-Scaled Wind Turbines and its Application on Agricultural Harvesting

Anupa Koswatta[1], Faramarz Alsharif[2], Yasushi Shiroma[1], Shiro Tamaki[1,*], Junji Tamura[2]

[1]*Graduate school of Engineering, University of the Ryukyus, 903-0213, Japan*

[2]*School of Earth, Energy and Environmental Engineering, Kitami Institute of Technology, 090-8501, Japan*

ARTICLE INFO

Article history:
Received: 21 May, 2020
Accepted: 21 June, 2020
Online: 12 July, 2020

Keywords:
Small-scaled Wind Turbine
MPPT control
Eddy current brake
Over-rotation
Strong wind

ABSTRACT

This research aims to enhance the generated power of the small-scaled wind turbine using the eddy current brake system and Maximum Power Point Tracking (MPPT) control method. We analyzed the behavior of the generated power and power factor, with and without the MPPT control which implemented by eddy current brake system. Also, the feasibility of the system investigated using different wind conditions such as strong and calm wind conditions. The load data has different voltage respond to the system since its conditions depend on the day/night loads pattern, weather conditions, soil moisture. Moreover, the analogical experiment for small-scaled wind turbine blade destruction is analyzed to determine the maximum penetration value of mechanical power in order to retrieve an optimal angular velocity which resulting in provides a possible maximum power to loads. At the same time, emergency break is operated when angular velocity reaches to critical speed to avoid destruction. In the simulation, we collected the real load data from a mango farm in Okinawa prefecture in Japan. The results were analyzed through simulations for the different wind conditions. In the end of simulation, we could verify that either Maximum Power Point and emergency control are activated correspondingly.

1. Introduction

As a renewable energy generating equipment, small-scaled wind turbines have rising demand to fulfill the modern energy requirements. The small-scaled wind turbines have several merits such as less area needed for setting up the entire system, low cost maintenance and easy to assemble which are advantages for household use [1]. Comparing to the conventional large-scale wind turbines, small-scaled wind turbines produces small amount of power while large scale wind turbines produces maximum 8MW power. Therefore small-scaled wind turbines are suitable for household and small-scaled green houses and farms.

To supply a stable electricity output, small-scaled wind turbines uses several control methods such as pitch control [2-3], yaw control, brake control [4-6]. Considering brake control, it is mainly focus on brake pad-based system for the safety of the wind turbine [7] while contactless brake systems related research is ongoing currently. In [8] we proposed to realize a contactless angular velocity in emergency case by eddy current method, however, the optimal control was not concerned. In this research, we are focusing to enhance the power through maximum power point tracking (MPPT) using eddy current brake system and eddy current brake system for emergency rotational control.

Since the wind speed stability is always unpredictable, it is a difficult task to generate the maximum power continuously and keep the generated power in a stable state. In this regard MPPT control can be considered as an important controlling method to generate the maximum power. Many types of MPPT control methods are used for generating maximum power such as using fuzzy logic controllers [9], adaptive control [10], etc. In this research, we use the eddy current brake as the main controller tool for the MPPT control. Here, MPPT operation is to convey the maximum possible power from turbine to generator. Therefore, this operation contains power enhancement characteristics. Meanwhile, emergency break is implemented as well to avoid any destruction or malfunction of entire system since small-scaled wind turbine's angular velocity rise up drastically during the high wind penetration.

*Corresponding Author: Shiro Tamaki, shiro@ie.u-ryukyu.ac.jp

In this study, the mango green house is a mango cultivating farm where the location is in Okinawa, Japan. The mango plants are required to be grown in a controlled environment in order to make best condition for cultivating. Also, the conditions are monitored using Internet of Things (IoT) technology. Therefore, the farmer can analyze the real time mango plants conditions such as its soil moisture, farm temperature, CO_2 conditions, light conditions and humidity.

In this paper, first we explain about the structure of the proposed system including proposed brake system, control diagram, simulation block diagram. Next, we explain the mathematical modeling of the system. Afterwards, we describe the simulation and results. Finally, we describe the conclusion of our research. This research is an extended work of [11].

2. Blade destruction analysis

This section explains the analogical experiment of the blade destruction. We performed this experiment to analyze the blade destruction for different penetration values. In other words, we did this experiment to decide the safety margins of the small-scaled wind turbine. In Figure 1, the experimented broken small-scaled wind turbine blade is displayed. In the following equation (1) F_{max} means the maximum allowable force on the wind turbine blade and m means the mass of the turbine's blades. Also, ω_{max} represents the blade destruction angular velocity. Table 1 shows the specification of the blade. The material of the blade is made using Fiber Reinforced Plastic (FRP).



Figure 1. Broken wind turbine blade

$$F_{max} = mr\omega^2_{max} \quad (1)$$

Table 1. Specification of the blade

Parameter	Value	Unit
F_{max}	4.5	kN
m	2.57	kg
r	2.5	m
ω_{max}	26.45	rad/s

In Figure 2, the A point stands for the maximum force safety margin. If the penetration exceeds the point A then the turbine's blades start to brake. When the penetration value reaches nearly 28kN (point D), the turbine's blades were destroyed. Therefore, to protect the turbine, we put c as the primary safety margin which has penetration value of 2.5kN. B means the secondary safety margin that 4kN is penetrated. For secondary safety margin, the angular velocity is 25rad/s and angular velocity for primary safety margin is 20rad/s. Hence, for the simulation purpose, we set the primary safety margin angular velocity ω_{lim} as follows:

$$\omega_{lim} = 20\text{rad/s}.$$

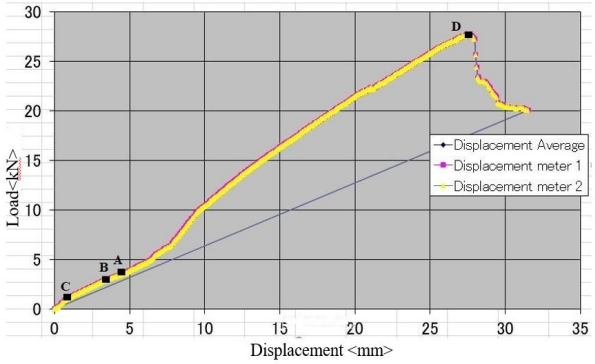


Figure 2. Blade destruction chart

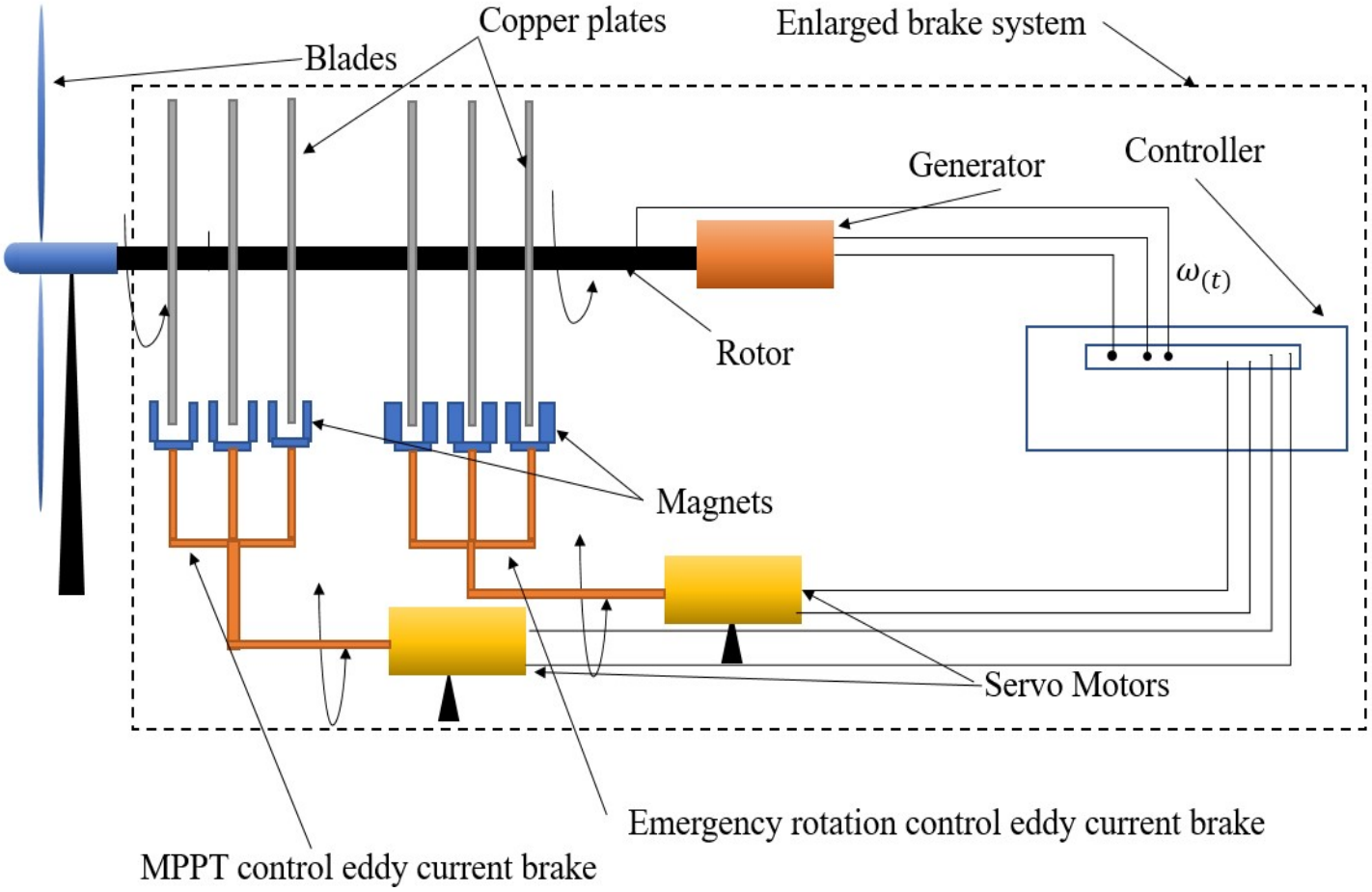
3. Structure of the system

It is necessary to consider the electrical and mechanical aspect of the proposed system structure. For the mechanical section, we consider the brake system controlling structure. Here, the emergency rotation control eddy current brake system and MPPT control eddy current brake works independently which is illustrated in Figure 3. They always control the reference value of the angular velocity ω_{ref} and the angular velocity output ω. When the controller receives the angular velocity of the turbine, controller will send its signal to two servo motors which are operates as emergency rotation control and MPPT control unit to place the magnets nearby the copper plates to control the rotation.

When the wind turbine rotates, the signal controller will receive the angular velocity measurement and it will send the signals to the emergency rotation control eddy current brake and MPPT control eddy current brake. According to the current rotation situation, above mentioned two brake systems will control the $f_{emergency(\omega)}$ which corresponds to emergency rotation control eddy current brake and $f_{MPPT(\omega)}$. Here, $f_{MPPT(\omega)}$ corresponds to MPPT control eddy current brake. This process is illustrated in Figure 4.

Figure 5, shows the entire block diagram of the small-scaled wind turbine system for realization in hardware. Here, *1/s* symbol represent the integrator symbol. For this system, the generator is considered as DC generator. Figure 6, displays the equivalent circuit of DC generator including the protection diode, load and battery.

Regarding the electrical aspect of the proposed system, important conditions are explained as follows. When the battery charge becomes low, the additional current from the generator will charge the battery. In the other hand, if the generator unable to supply demanded current, then battery will supply the required current to the load. If the battery is full and the required power is supplied to the load, then the turbine rotation is reduced using the eddy current brake system.



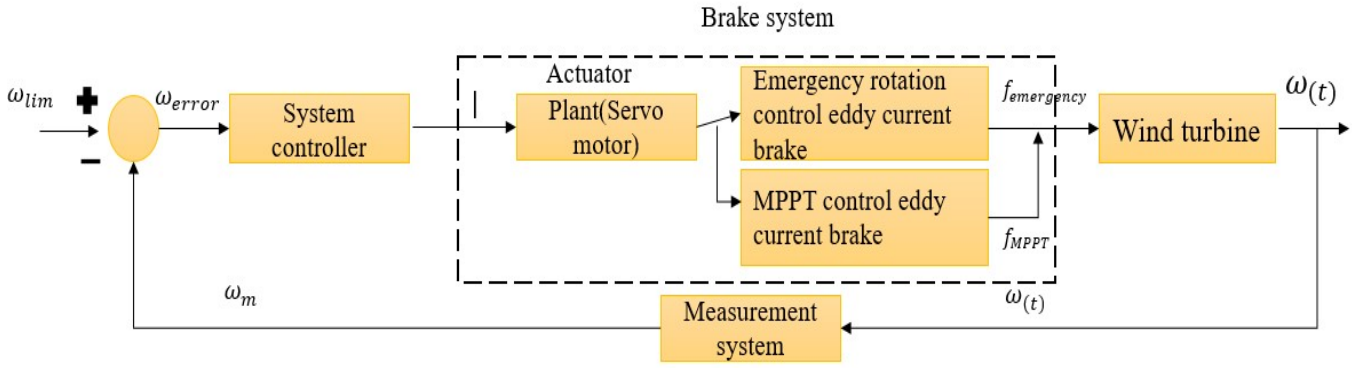
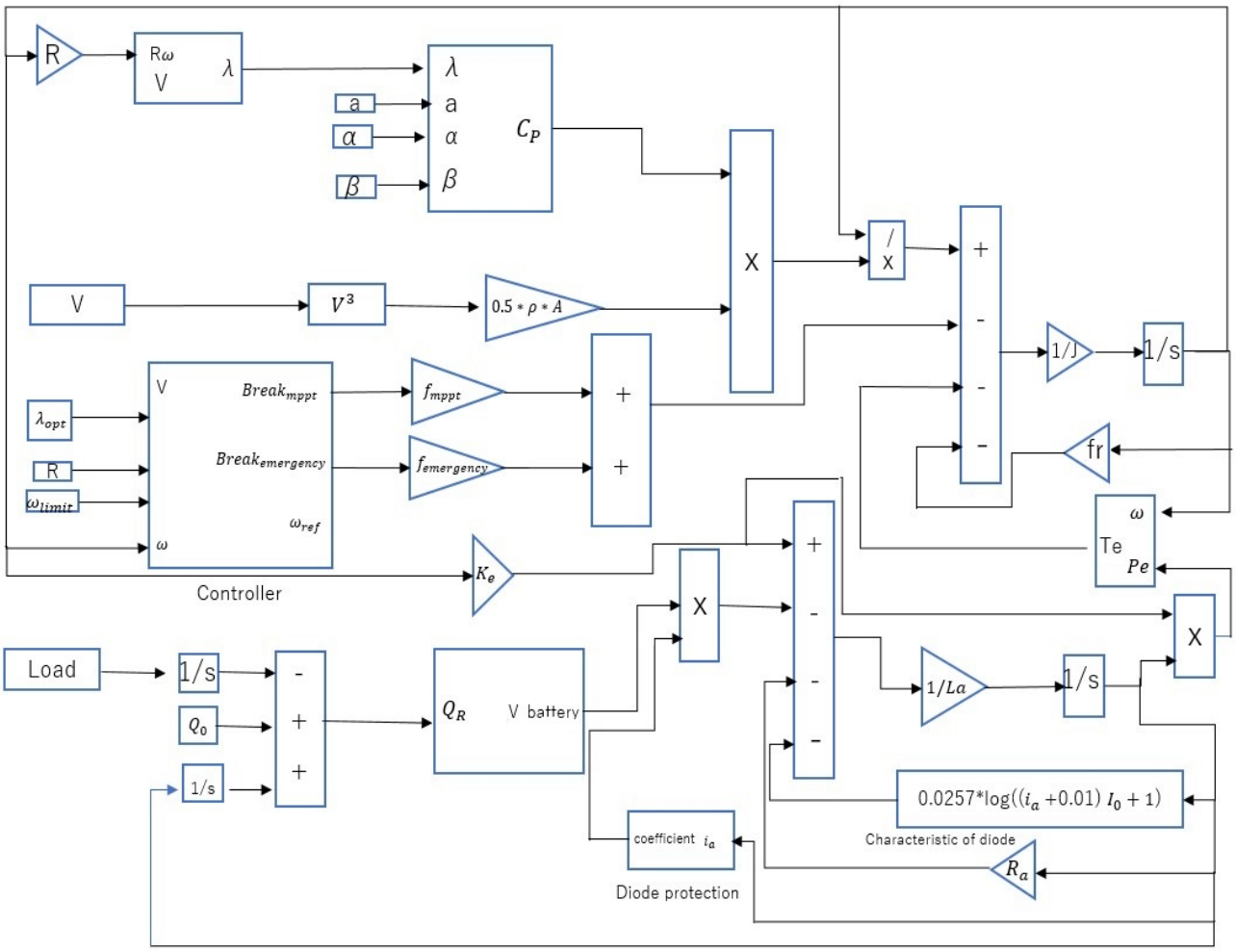


Figure 4. Control diagram



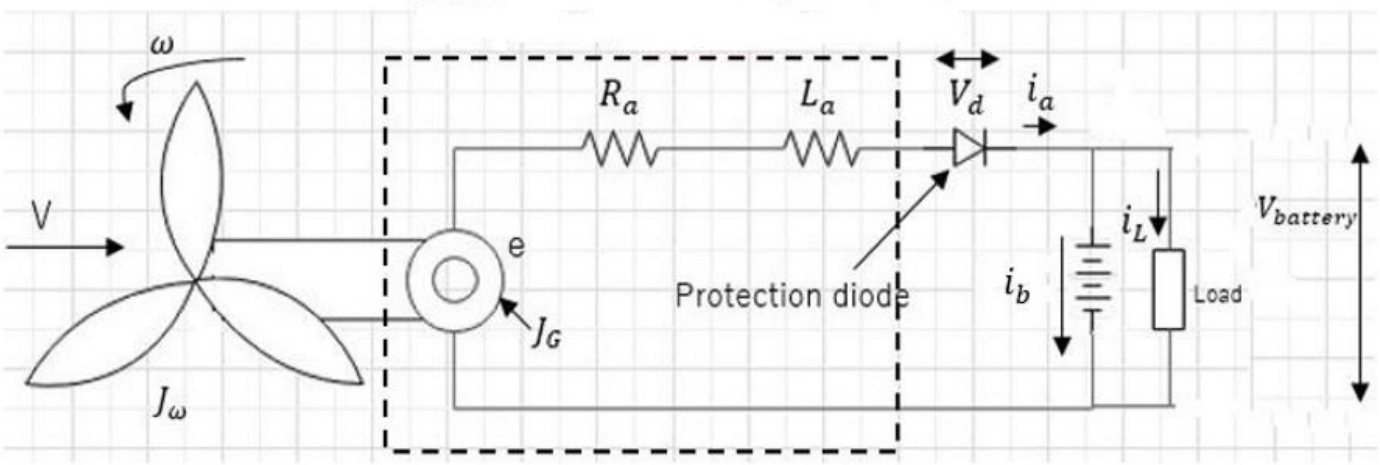


Figure 6. DC generator equivalent circuit diagram

Electromagnetic brake systems can be seen in modern world applications such as Maglev train brake system, gym instruments, and elevators, etc. Unlike the non-contactless braking operation, eddy current brake system can be considered as an efficient brake system [12].

Here, let us describe the eddy break system specifically. When the rotor shaft rotates in ω angular velocity, the magnetic flux between the two magnets are changing due to the interaction with the copper plate. Due to this magnetic flux change, eddy current i_{eddy} will induce on the copper plate according to the faraday law of induction [13]. Due to the induced current, there will be

opposing force or in the rotational case the torque generate against the rotation. This torque will act as the eddy current brake torque. In Figure 7, q_m means the magnitude of magnetic charge on magnetic poles and q_c means magnitude of induced magnetic charge on copper plate.

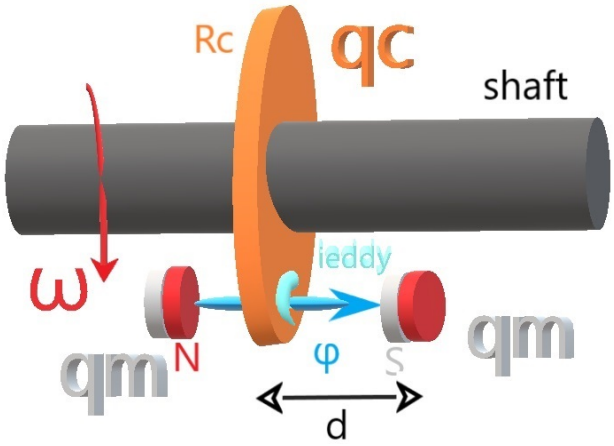


Figure 7. Eddy current brake

$$\varepsilon = - \frac{d\varphi}{dt} \tag{2}$$

where, ε stands for electromotive force and φ for magnetic flux.

Here, R_C is the resistance of the copper plate and I_{eddy} is the induced eddy current on the copper plate. The relation with the R_C and I_{eddy} is shown in (3), which is based on the ohm's law,

$$\varepsilon = R_c I_{eddy} \tag{3}$$

From above equations, we can write down the eddy current as follows.

$$I_{eddy} = - \frac{1}{Rc} \frac{d\varphi}{dt} \tag{4}$$

Due to occurrence of above current, the break system is activated. Next section describes the dynamics of whole system.

4. Mathematical modeling of the system

This section introduces the mathematical modeling of the entire system. Therefore, it is important to consider the dynamics representation of mechanical and electrical part of the proposed system. The parameters were used for the mechanical section is displayed in table 2 and the electrical section is displayed in table 3 along with combined electrical and mechanical equations.

Equation (5) displays the summation of the inertia moments of the wind turbine and blades

$$J = J_\omega + J_G \tag{5}$$

The conventional mechanical and electrical dynamics of the system displayed in (6)

$$J \frac{d\omega(t)}{dt} = \frac{1}{2} C_T \rho A R V_{(t)}^2 - f_r \omega_{(t)} - T_e \tag{6}$$

Table 2: Mechanical Parameters

Symbol	Name
J_ω	Inertia moment of the blade
J_G	Inertia moment of the generator
ω	Angular Velocity
ρ	Density of atmosphere
A	Area swept by blade
R	Length of blade
$V_{(t)}$	Speed of wind
f_r	Axis friction factor
μ	Permeability
q_m	Magnitude of the magnetic field of the magnet
q_c	Magnitude of the magnetic field of the copper plate
d	Distance between copper plate and magnet
a_{adj}	Decision functional adjustment
ω_{ref}	Reference value of angular velocity
ω_{lim}	Allowed maximum angular velocity
ω_{stall}	Stall controlled angular velocity
I_{eddy}	Induced eddy current
C_P	Power coefficient
C_t	Torque coefficient
k	Boltzmann's constant
f_{MPPT}	Induced eddy current brake torque for MPPT control
f_{em}	Induced eddy current brake torque for emergency brake

This formula stands for controlling the brake plate rotation in a contact manner. However, in [8] we have proposed for contactless break system that is shown as follows.

$$J \frac{d\omega(t)}{dt} = \frac{1}{2} C_T \rho A R V_{(t)}^2 - f_r \omega_{(t)} - f_m f(\omega_{(t)}) - T_e \tag{6-1}$$

where, *fm* is the eddy current brake torque in following equation and *f(.)* is the decision function.

$$f_m = \mu \frac{2 q_m q_c}{4 \pi d^2} \tag{7}$$

Here, μ stands for permeability of the space. q_m and q_c stands for magnitude of the magnetic field of the magnet and copper plate respectively. This formula is based on gilbert model magnetic force [14].

Next, we are going to model the electrical part of the system. In the following table the parameters are described.

Generated power is displayed in (8).

$$P_e = \varepsilon \, i_a \tag{8}$$

Electrical torque of the generator displayed in (9)

$$T_e = \frac{P_e}{\omega} \tag{9}$$

where, e_a stands for induced voltage and i_a for armature current. Equation (10) displays the battery electrical charge dynamics

$$Q_R = Q_0 + \int i_a \, dt - \int i_L \, dt \tag{10}$$

As it is shown in above equation the remaining amount of battery capacity depends on the accumulation of consuming current and armature current that provided from turbine and battery

initial capacity. Adversely, the current exchange among battery, turbine and loads can be retrieved by derivative of equation (10) which is

$$i_R = i_a - i_L \tag{11}$$

Table 3: Electrical Parameters

Symbol	Name
i_a	Output current of generator (armature current)
i_b	Charging current of the battery
i_L	Loads current
ε	Induced emf
V_L	Loads voltage
V_{Battey}	Battery Voltage
Q_0	Initial condition of the battery capacity
I_0	Saturated current
I_d	Flowing current
V_d	Voltage drop of diode
I_s	Saturated current
q	Electron charge
T	Temperature
K	Boltzmann constant

Therefore, whenever equation (11) becomes zero, the consuming and generating currents are balanced or if i_R is greater than zero it is both charging mode to battery and providing currents to loads, simultaneously. Adversely, if i_R is less than zero then the battery only provides the power to loads that is discharging mode.

From equivalent circuit of DC generator in Figure 6, the electrical dynamics of the system is obtained as follows.

$$\varepsilon - R_a i_a - L \frac{di_a}{dt} - V_d = V_L \tag{12}$$

The purpose of setting diode in the output terminal of dc generator is to avoid being motor mode for dc generator when there is no wind penetration since battery is connected to generator's terminal. Thus, diode characteristic should be considered as well. The drop voltage of diode is shown in following equation.

$$V_d = \frac{KT}{q} ln(\frac{I_d}{Is} + 1) \tag{13}$$

Where, I_d corresponds to armature currents as following expression (14).

$$I_d = I_s(e^{\frac{qV_d}{KT}} - 1) \tag{14}$$

Next, the induced electromotive force (emf) of the generator is displayed in (15).

$$\varepsilon = K_e \omega \tag{15}$$

Here, K_e stands for Induced emf constant.

Up to here, the electrical model were described. From next equation the mechanical parts including the break terms are expressed.

The following equation shows the mechanical torque.

$$T_m = \frac{P_m}{\omega} \tag{16}$$

Below equation is based on the conventional break method, however, *Cp* is variable with respect to tip speed ratio.

$$J\frac{d\omega(t)}{dt} = \frac{C_{P(\lambda)}\rho A R V_{(t)}^2}{2\omega} - f_r\omega_{(t)} - f_m f(\omega_{(t)}) - T_e \tag{17}$$

Then we can reform the above equation for proposed method by adding the eddy current brake for emergency brake and MPPT control that is shown as (18).

$$J\frac{d\omega(t)}{dt} = \frac{C_{P(\lambda)}\rho A V_{(t)}^2}{2\omega} - f_r\omega_{(t)} - T_e - (f_{mp}f_{MPPT(\omega)} + f_{em}f_{emergency(\omega)}) \tag{18}$$

In above equation, f_{MPPT} and $f_{emergency}$ has different values and f_{em} value is larger than f_{MPPT} since $f_{emergency}$ requires more braking force to control the over-rotation of the turbine than the force use to realize MPPT.

Equation 19 task is to work as a switch to keep the output power in a maximum state which means keep the angular velocity ω in the following condition: ($\omega < \omega_{ref}$) that is in the MPPT operation status. However, if ω exceed ω_{ref} value ($\omega > \omega_{ref}$), then the output power is not in a Maximum power state. Nevertheless, still the system is generating power since ω value isn't reached to ω_{lim} yet which is the safety margin of the wind turbine. If this situation going to happen then the emergency brake will start to reduce the ω value to keep the state as $\omega < \omega_{lim}$ condition. Therefore, equation 20 [15] task is to work as a switch to determine ($\omega < \omega_{lim}$) condition.

$$f_{MPPT(\omega)} = \frac{1}{2} + \frac{1}{\pi} tan^{-1}(\alpha_{adj}(\omega - \omega_{ref})) \tag{19}$$

$$f_{emergency(\omega)} = \frac{1}{2} + \frac{1}{\pi} tan^{-1}(\alpha_{adj}(\omega - \omega_{lim})) \tag{20}$$

The reference value of the angular velocity is shown in (21). Here the λ_{opt} means the optimum value of tip speed ratio.

$$\omega_{ref} = V\frac{\lambda_{opt}}{R} \tag{21}$$

ω_{ref}, value is a function of λ_{opt} and V. Since this is a small wind turbine, λ_{opt} value require small value to obtain C_P^{opt} value. Due to the compatibility of small-scaled turbine, we designed the mathematical expression of the C_P as mentioned below in (22) and C_P-λ relationship shown in Figure 8.

$$C_P = -a(\lambda - \alpha)(\lambda - \beta) \tag{22}$$

Where, the condition of coefficients is $a > 0$ and $\beta > \alpha$.

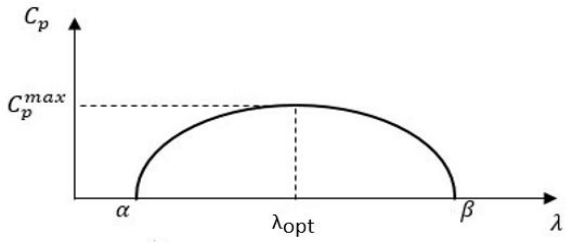


Figure 8. C_P-λ curve

Optimum value of wind turbine power factor is shown in (23).

$$C_P^{max} = \frac{\alpha + \beta}{2} \tag{23}$$

As it is shown in Figure 7 the output terminal of generator is connected to a battery and battery is connected to the loads in parallel. Therefore, in order to verify the battery voltage, we have designed a battery capacity charging in which the battery voltage can be obtained by the Equation (24) represents the characteristic of the accumulated consumed current.

$$V_{Battery} = \frac{V_{max} - V_{min}}{0.5 - \frac{1}{\pi}\tan^{-1}(20\times(-5))}\left(\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(1(Q_R - 5))\right) \quad (24)$$

Notice that the load voltage and battery voltage is identical since as it has been mentioned they are connected in parallel.

5. Simulation and results

In this section, we conduct the simulation in two steps. First, we simulated the entire system for different average wind velocities from 7m/s, 14m/s, 21m/s and 28m/s, respectively. Then plotted the stall control charts for above velocities. We did this step to analyze the wind turbines stall control behavior for different wind velocities. Afterwards, we plotted the Mechanical power (Pm) and power coefficient of the system for different conditions. For the second step, we simulated the entire system for calm and storm days wind conditions for one day and analyzed the behavior of outputs of the system.

5.1. Simulation conditions

This section explains the conditions of the simulation and system we used in order to conduct the simulation. The load data for the system is applied using real mango greenhouse system in Okinawa [16]. Load pattern is displayed in Figure 9. This load consists of Light emitting diode (LED) lights, Compact Fluorescent Light (CFL) lights, Electrical Fans. According to the Figure 9 the highest load value is 12.8A. Here, the sample time for the simulation is 0.01s. MATLAB/Simulink software platform is used for the simulation. For the simplicity, simulation time is set to be 100s from Figure 10 to Figure 21. From Figure 22 to Figure 33, simulation time is one day long. Values of the parameters for simulation are displayed specifically in Table 4.

Table 4: Parameters of simulation

Symbol	Value	Unit
ω	Variable	$rads^{-1}$
ρ	1.225	kg.m^3
A	19.625	m^2
R	2.5	m
$V_{(t)}$	Variable	ms^{-1}
f_r	0.1	kg.m^2s^{-1}
f_{mppt}	2200	kg. ms^{-2}
$f_{emergency}$	5800	kg. ms^{-2}
μ	$4\pi10^{-7}$	T. m. A^{-1}
q_m	62.5*10^9	A.m
q_c	0.1*10^{-4}	A.m
d	0.05	m
a_{adj}	10	
ω_{lim}	20	$rads^{-1}$
ω_{ref}	Variable	$rads^{-1}$
J_ω	1.1	kg.m^2
J_g	0.1	kg.m^2
L_a	2.87	mH
R_a	4.2	Ω
q	1.602*10^{-19}	C
k	1.380*10^{-23}	J/K
I_0	1*10^{-22}	A
K_e	2.45	V.S
Q_0	60	A.h
α	1	
b	3.5	

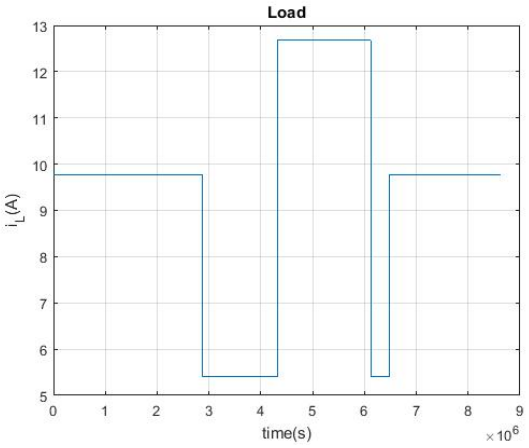


Figure 9. Load

5.2. Results

From figure 10 to figure 13, they display the stall control chart for 7m/s, 14m/s, 21m/s, 28m/s, respectively.

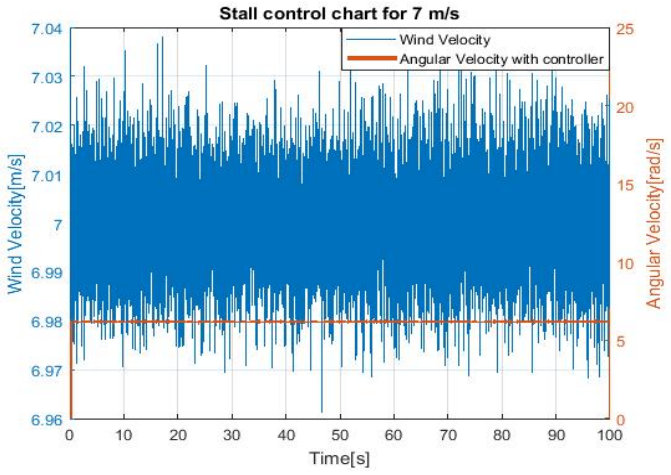


Figure 10. Stall control chart for 7 m/s

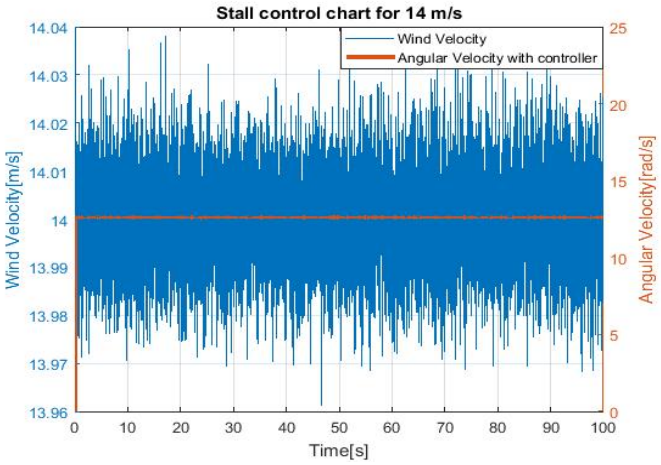


Figure 11. Stall control chart for 14 m/s

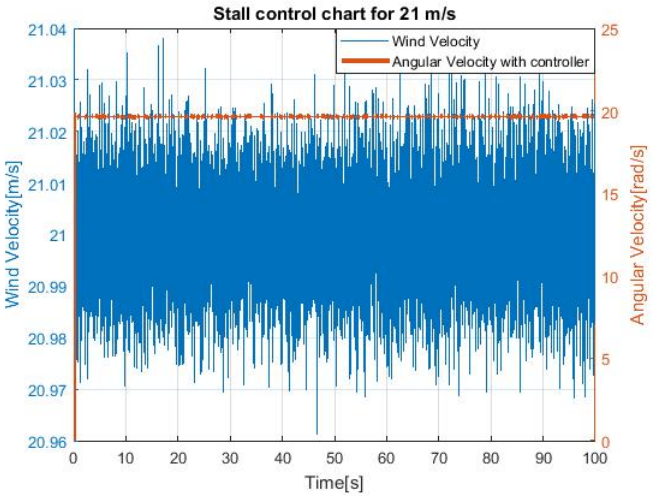


Figure 12. Stall control chart for 21 m/s

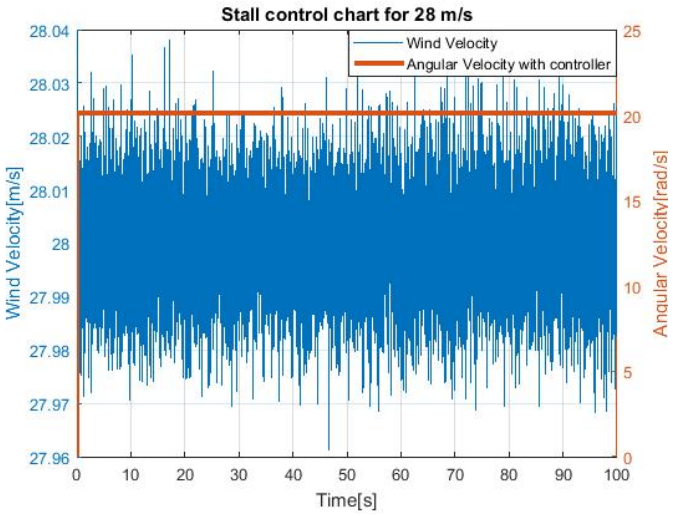


Figure 13. Stall control chart for 28 m/s

According to above stall control figures, the wind turbine angular velocity emergency control is valid up to 28m/s velocity. Therefore, the eddy current brake system can control the over-rotation of the wind turbine up to 28m/s.

Table 5 shows the ω_{ref} which is optimal angular velocity to realize the maximum power operation and ω_{stall} is the actual output of angular velocity.

Table 5: ω_{ref} and ω_{stall} comparison

Input velocity	ω_{ref}	ω_{stall}
7	6.299	6.1849
14	12.5999	12.5994
21	18.8999	19.6377
28	25.199	20.094

As it is obvious in the above table, the output angular velocity for input wind is lower than the optimal angular velocity since in high wind penetration emergency break is activated. Specifically, comparing to the ω_{stall}, ω_{ref} value is becoming lower when wind velocity increase. That means when the input wind velocity is 28m/s the ω_{ref} value is 25.199 the MPPT control is stopped because $\omega > \omega_{ref}$. However wind turbine is still operating using the emergency eddy current brake since ω_{stall} is nearly the value of ω_{lim}. Next, from Figure 14 to Figure 17, The instantaneous angular velocity deviation $\Delta\omega = \omega_{stall} - \omega_{ref}$ show their behavior for 7m/s, 14m/s, 21m/s, 28m/s average wind velocities, respectively.

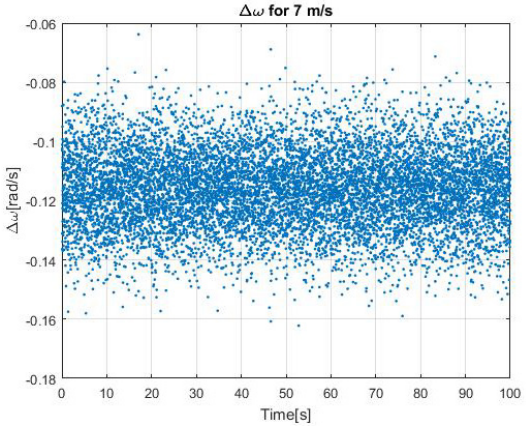


Figure 14. $\Delta\omega$ for 7 m/s

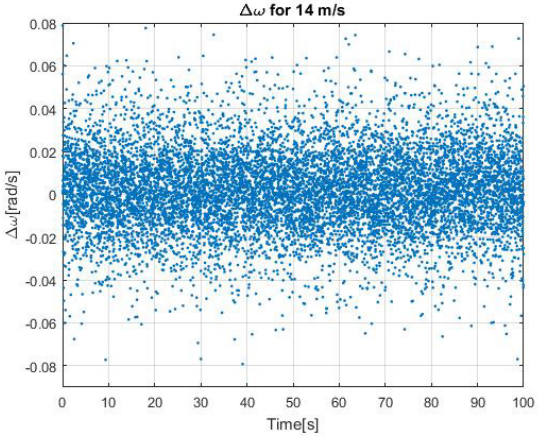


Figure 15. $\Delta\omega$ for 14 m/s

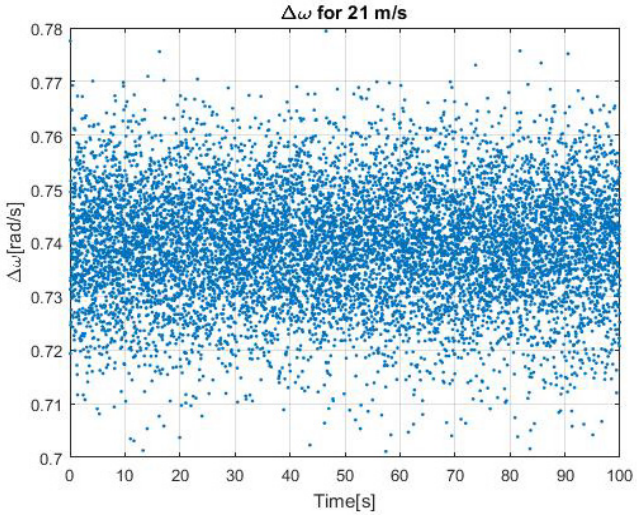


Figure 16. $\Delta\omega$ for 21 m/s

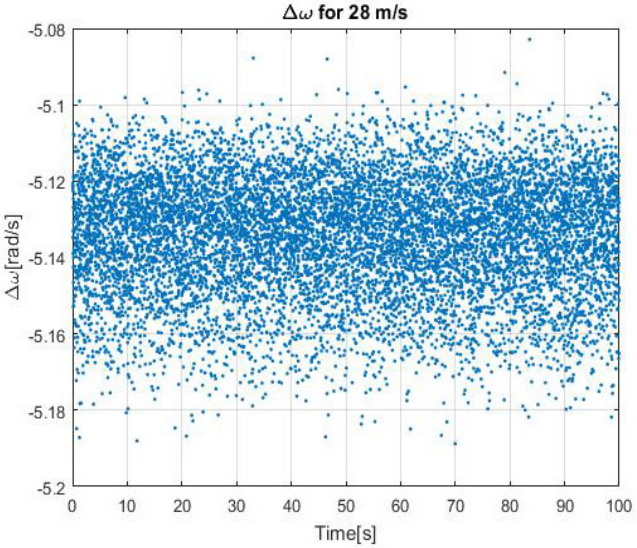


Figure 17. $\Delta\omega$ for 28 m/s

The mechanical power comparison with and without MPPT and emergency brake and Cp behavior comparison with and without the MPPT and emergency brake are shown from Figure 18 to Figure 21.

According to the Figure 18, when the average wind velocity is 7m/s the mechanical power with the MPPT and emergency brake has higher value comparing to without the MPPT and emergency brake. As well, for Cp, when the MPPT and emergency brake is triggered it contains higher value. But, when Cp is without the control of MPPT and emergency brake, the value is low.

In Figure 19, Figure 20 and Figure 21 which are correspond to average wind velocity of 14m/s, 21m/s and 28m/s, they have the same comparison results as mentioned in above for 7m/s. Therefore, it is obvious that when the system is under the controlled state for MPPT and over-rotation, the mechanical power of the wind turbine and the power factor (Cp) have their maximum values.

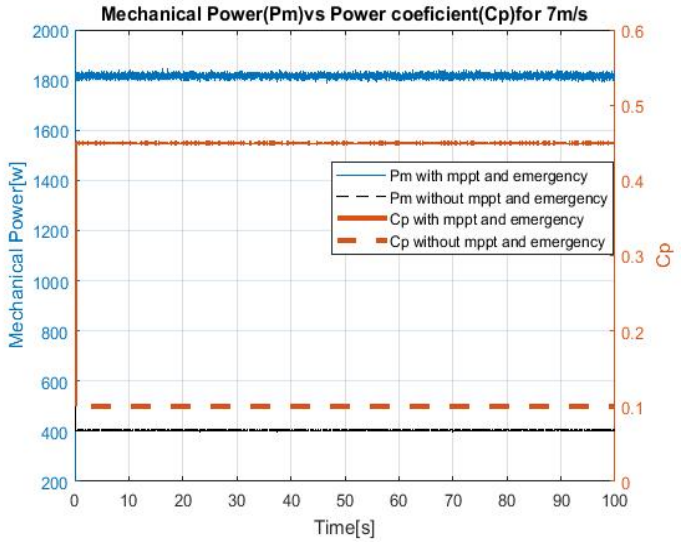


Figure 18. Pm vs Cp for 7m/s

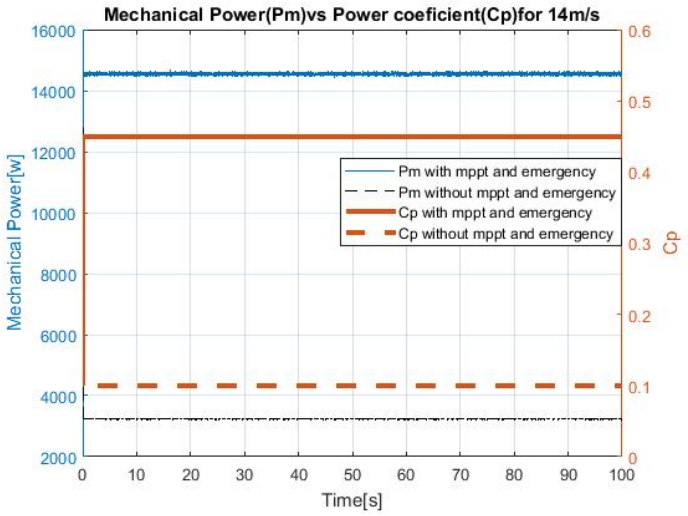


Figure 19. Pm vs Cp for 14m/s

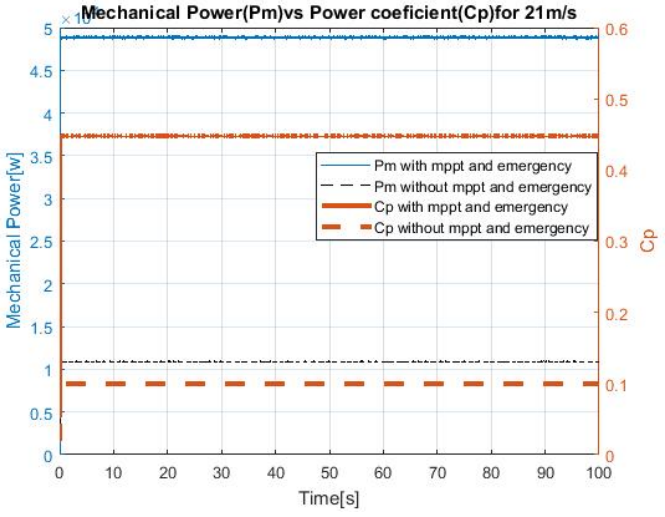


Figure 20. Pm vs Cp for 21m/s

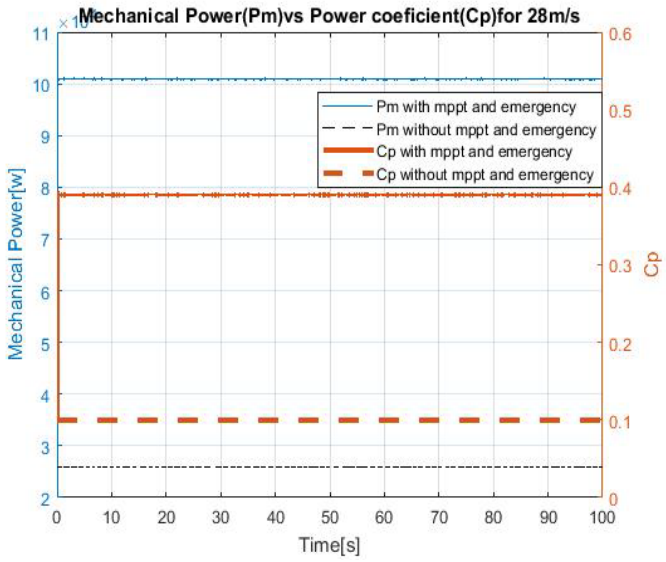


Figure 21. Pm vs Cp for 28m/s

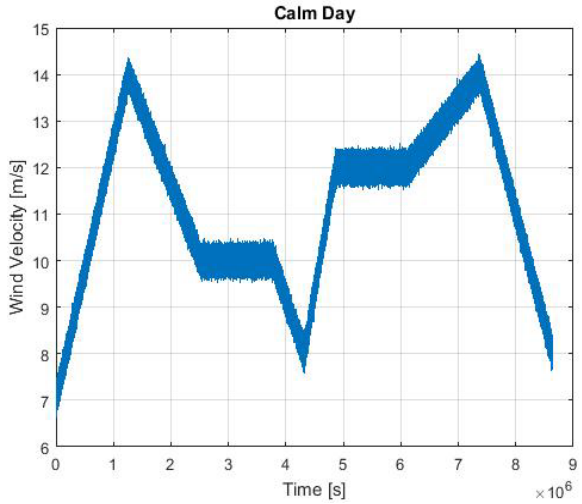


Figure 22. Calm day wind pattern

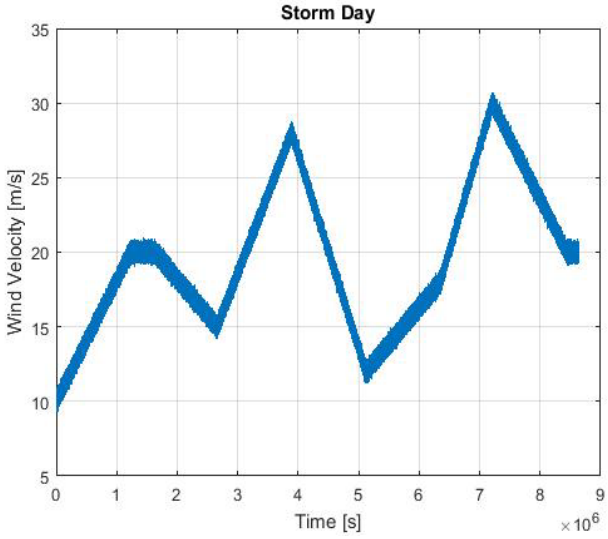


Figure 23. Storm day wind pattern

We have artificially created two wind patterns for the simulation process. First, we have done the simulation for calm day and then for the storm day condition. For calm day and storm day wind patterns are shown in Figure 22 and Figure 23. For the first phase of calm day the velocity gradually increases from 7m/s to 14m/s. Afterward, it decreases gradually from 14m/s to 10m/s. Then the 10m/s wind velocity maintain stable for few hours. Afterwards the velocity will decrease until 8m/s. Then, it will increase from 8m/s to 12m/s. Again, it will increase from 12m/s to 14m/s gradually. Finally, the velocity will decrease from 14m/s to 8m/s. For calm day the highest wind speed is approximately 14.3m/s and storm day the highest wind speed is approximately 30m/s. Lowest wind speed for calm day and storm day are 7m/s and 10m/s.

Next, Figure 24 and Figure 25 shows the angular velocity deviation $\Delta\omega$ values which changes with respect to the wind condition. There are some peaks occurred according to the Figure 22. These peaks happen due to the emergency braking for sudden wind condition changes. Here, the $\Delta\omega$ is nearly zero means the wind turbine produces its maximum power.

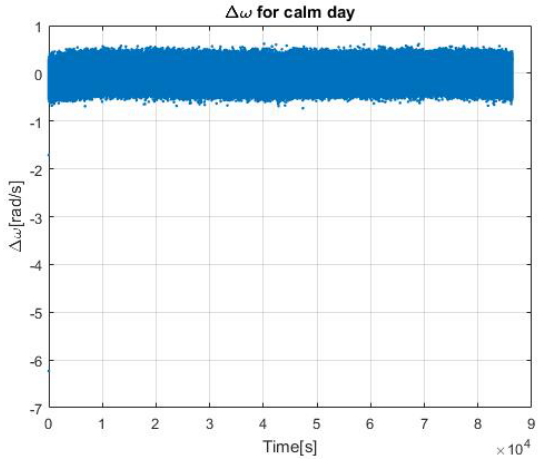


Figure 24. $\Delta\omega$ for calm day

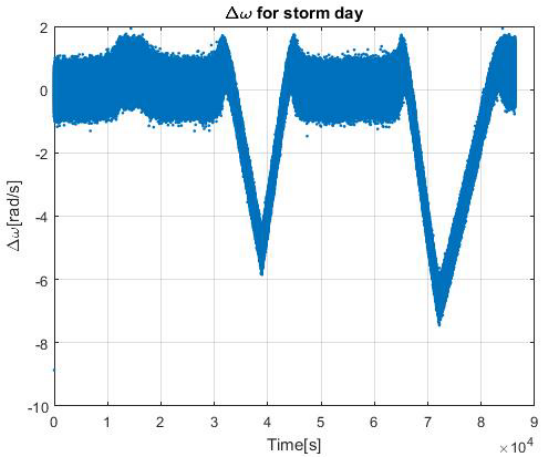


Figure 25. $\Delta\omega$ for storm day

In calm day, the angular velocity does not achieve to critical value which is 20 rad/s. Nevertheless, according to angular velocity deviation in Figure 24, it is attempting to reach the angular velocity that operates at maximum power point. Same to Storm day, angular velocity tries to operate at its maximum power point.

However, as it can be seen Figure 25, the deviation of angular velocity has negative two peaks for some determined period. Correspondingly, it is obvious in Figure 27 that during the peaks of angular velocity deviations the output of angular velocity is suppress by 20 rad/s. Thus, the emergency brake is also activated in storm situation.

The angular velocity always kept under $\omega_{lim} = 20$ for both calm day and storm day stall control patterns which are shown in Figure 26 and Figure 27.

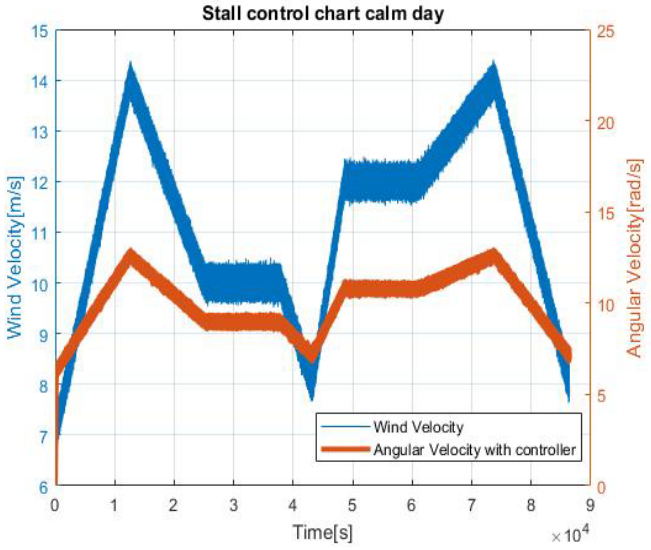


Figure 26. Stall control chart for calm day

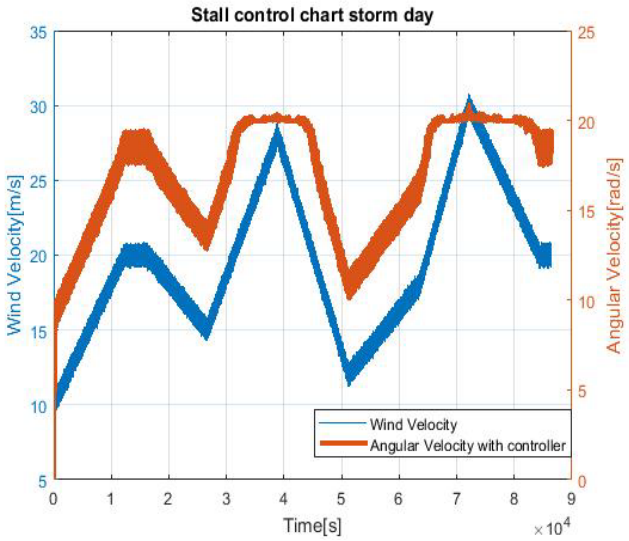


Figure 27. Stall control chart for storm day

Figure 28 and Figure 29 displays the Mechanical power (Pm), Electrical power (Pe), Electrical torque (Te) for calm day and Pe, Pm, Te, Tm for storm day. According to those figures the Pe, Pm, Te, Tm have higher values comparing to the calm day for emergency brake and MPPT brake activated state. Therefore, the small wind turbine is capable of generating high power even in a storm day.

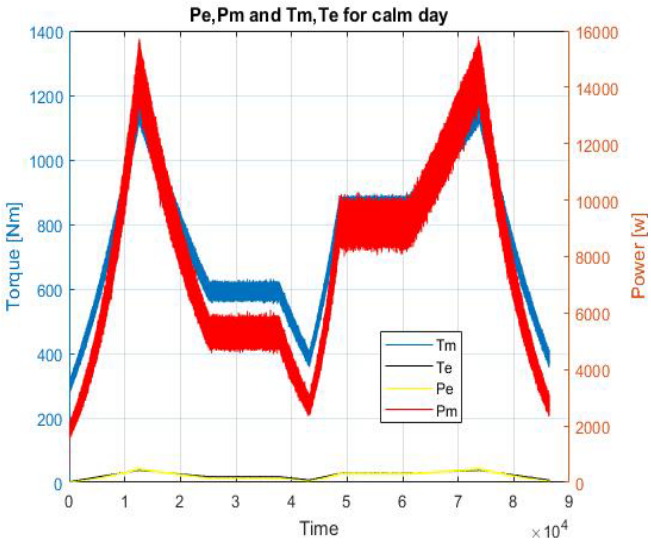


Figure 28. Pe,Pm, Tm and Te for calm day

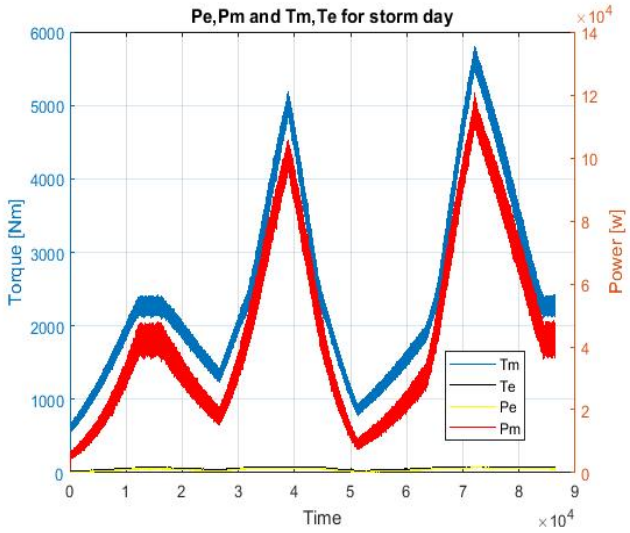


Figure 29. Pe,Pm, Tm and Te for storm day

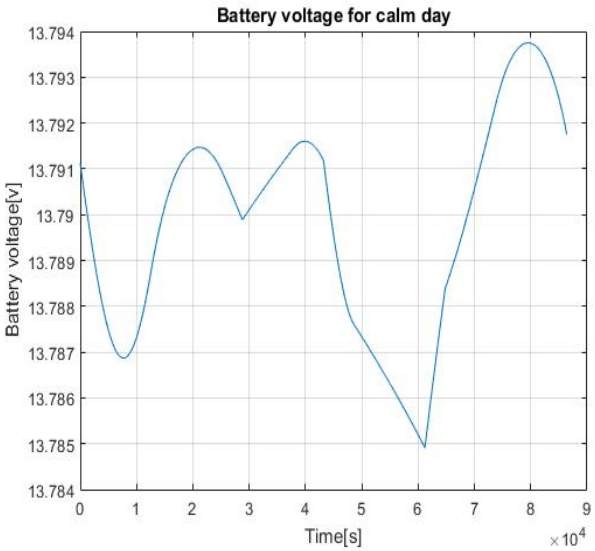


Figure 30. Battery voltage for calm day

According to Figure 30 and Figure 31, the battery voltage for storm day has good performance comparing to the battery voltage for calm day. Thus, for both cases we can say it has sufficient amount of voltage.

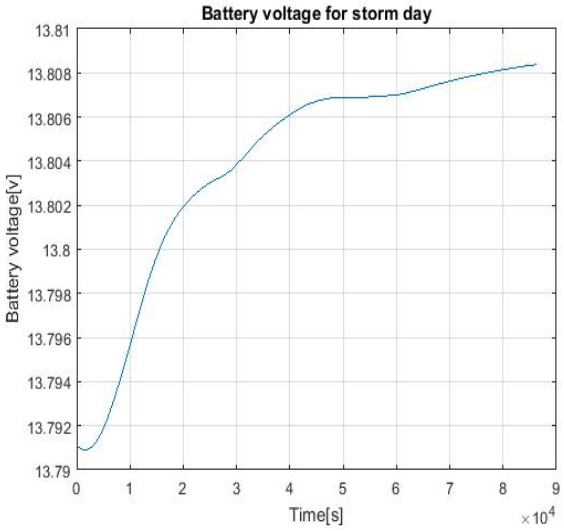


Figure 31. Battery voltage for storm day

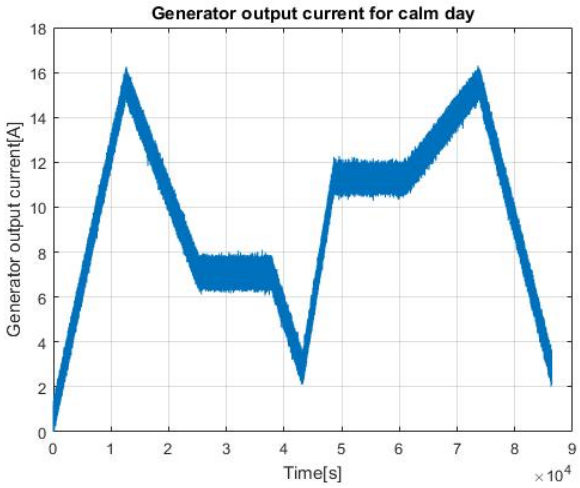


Figure 32. Generator output current for calm day

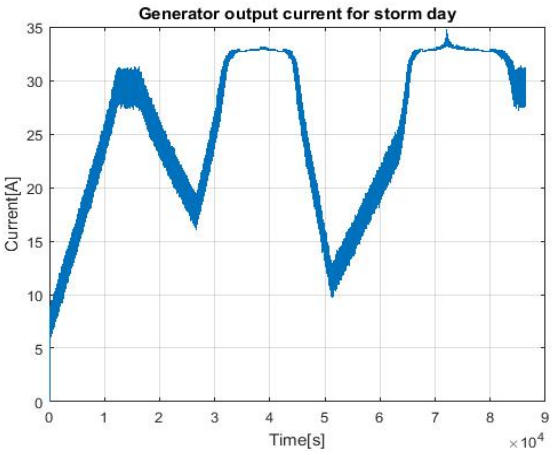


Figure 33. Generator output current for storm day

Figure 32 and Figure 33 shows the generator output current for calm day and storm day. For storm day, the generator output current has significant increase comparing to the clam day current in the MPPT and emergency brake-controlled condition. Nevertheless, even in calm day sufficient capacity of battery is maintained which is sustainable for providing the power to green-house environmental loads. Thus, we can conclude that efficiency of power extraction is improved with cooperative of the emergency operation.

6. Conclusion

We have conducted the simulation for different wind condition and analyzed the behavior of the system when the system is in control of MPPT and Eddy current brake conditions. Moreover, the analogical experiment for the small-scaled wind turbines to decide the maximum penetration value and maximum angular velocity the wind turbine blades can withhold was performed. Therefore, we can conclude that the behavior of the system is in controlled state by using above mentioned control methods. We believe the eddy current brake system can be implemented in a small-scaled wind turbine system for the purpose of MPPT and control the over-rotation by strong wind. As future works, we will add pitch control for establish the system more robust. Also, the real system will be implemented.

Conflict of Interest

The authors declare no conflict of interest.

Reference

[1] J.F Manwell, J.G. Mcgowan, A.L. Rogers, Wind Energy Explained, John Wiley and Sons, 2009.

[2] F. Asharif, S. Tamaki, T. Nagado, T. Nagata M.R. Asharif, "Analysis of non-linear adaptive friction and pitch angle control of small-scaled wind turbine system," Future Generation Information Technology FGIT 2011, Control and Automation CA, Springer-Verlag, Lecture Note: Communication in Computer and Information Science, 26-35, December 2011. https://doi.org/10.1007/978-3-642-26010-0_4

[3] F. Alsharif, J. Tamura, S. Tamaki, K. Anupa, M. Futami, "Dynamics modification of passive pitch control for small-scaled wind turbine" Grand Renewable Energy 2018, Yokohama, Japan, June 2018. https://doi.org/10.24752/gre.1.0_147

[4] F. Asharif, S. Tamaki, H. Teppei, T. Nagado, T. Nagata, "Feasibility confirmation of angular velocity stall control for small-scaled wind turbine system by phase plane method" IEEK Transactions on Smart Processing and Computing, **2**(4), 240-247, 2013.

[5] F. Asharif, S. Tamaki, T. Nagado, M. R. Alsharif, "Comparison and evaluation of restrain control in wind turbine with various shock absorber considering the time-delay" International Journal of Control and Automation, Vol. **5**(3), 111-131, September 2012.

[6] F. Asharif, M. Futami, S. Tamaki, T. Nagado, K. Asato, "Stability and performance analysis of passive and active stall control of small-scaled wind turbine system by phase plane method" Proceedings of IEEE, International Conference on Intelligent Information and BioMedical Sciences (ICIIBMS 2015), .27-33, OIST, Okinawa, Japan, November 2015. https://doi.org/10.1109/ICIIBMS.2015.7439483

[7] M.Entezami, S.Hillmansen, P.Weston, M.Ph.Papaelias "Fault detection and diagnosis within a wind turbine mechanical braking system using condition monitoring" Renewable energy, 47, 175-182, November 2012. https://doi.org/10.1016/j.renene.2012.04.031

[8] A Koswatta, F Alsharif, Y Shiroma, M Khosravy, S Tamaki, J Tamura, "Alternative braking method for small-scaled wind turbines connected DC green house with analogical experiment on blade destruction" Advances in Science, Technology and Engineering Systems Journal, 5(2), 500-511 (2020). https://doi.org/10.25046/aj050264

[9] A. Z. Mohamed, M. N. Eskander, and F. A. Ghali, "Fuzzy logic controlbased maximum power tracking of a wind energy system," Renew. Energy, 23(2), pp. 235–245, Jun. 2001. https://doi.org/10.1016/S0960-1481(00)00099-9

[10] H. Zhao, Q. Wu, C. N. Rasmussen, M. Blanke, "Adaptive speed control of a small wind energy conversion system for maximum power point tracking ", IEEE Trans. Energy Convers., 29(3), 576-584, Sep. 2014. https://doi.org/10.1109/TEC.2014.2312978

[11] K. Anupa, F. Asharif, S. Yasushi, K. Ken, S. Tamaki, J. Tamura, "Power analysis of eddy current brake system applied on small-scaled wind turbine for dc greenhouse" International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Shanghai, China, 2019. https://doi.org 10.1109/ICIIBMS46890.2019.8991519

[12] K.J. Lee, K. Park, "Optimal robust control of a contactless brake system using an eddy current" Mechatronics, **9**(6), 615–631, 1999. https://doi.org/10.1016/S0957-4158(99)00008-2

[13] J. D. Kraus, K. R. Carver, Electromagnetics, 2nd ed. McGraw Hill, 1973.

[14] D. Gubbins, E. Herrerobervera, ENCYCLOPEDIA of GEOMAGNETISM AND PALEOMAGNETISM, Springer, 2007.

[15] F. Asharif, S. Tamaki, T. Hirata, T. Nagado, T. Nagata, "Design of electromagnetic and centrifugal force pitch angle stall controller on application to small-scaled wind turbine system" International Symposium on Stochastic Systems Theory and Its Applications, 2013. https://doi.org/10.5687/sss.2014.231

[16] Y. Shiroma, H. Afuso, R. Suwa, A. Kinjo, Y. Tonooka, T. Kaga, Itaru. Nagayama, S. Tamaki, G. Maharjan, "Development of higher yield and high-quality mango production system based on Internet of Things" IEEJ Transactions on Industry Applications, **139**(2), 166-173, Feb. 2019. https://doi.org/10.1541/ieejias.139.166

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 94-102 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Transient Analysis of a Line-Start Synchronous Reluctance Motor with Symmetrical Distributed Brass Rotor Bars

Mbika Muteba*

Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg, 2193, Republic of South Africa

ARTICLE INFO

Article history:
Received: 14 July, 2020
Accepted: 31 August, 2020
Online: 09 September, 2020

Keywords:
Brass Rotor Bars
Dynamic Response
Line-Start AC Motor
Starting Transients
Synchronous Reluctance Motor

ABSTRACT

In this paper, the performance evaluation of a line-start three-phase Synchronous Reluctance Motor (SynRM) with symmetrical distributed brass rotor bars is presented. The machine, which has been designed from a conventional three-phase induction motor (IM) NEMA frame stator is proposed as an alternative to a squirrel cage induction motor (SCIM). The 2D Finite Element Analysis (FEA) under ac magnetic transient solution was used to study some performance parameters of interest during starting transients. The experimental measurements were carried out in order to validate the numerical computation, to analyze the starting transients, and to explore the dynamic responses due to load variations. The FEA and experimental results of the synchronous reluctance motor with brass rotor bars (SynRM-BRBs) are compared to the results of a conventional three-phase SCIM of the same NEMA frame stator. The results evidenced that the reluctance torque developed by the SynRM-BRBs has a compounding effect on the accelerating torque, reaching its steady-state operational condition faster than the SCIM. The dynamic response of the SynRM-BRBs is faster in contrast to the SCIM during load variations. Furthermore, it was noted through measured results that the proposed line-start three-phase SynRM had a reduced dynamic no-load, and load current as opposed to the SCIM, thus positioning itself as a good candidate to replace the SCIM in applications that require a line-start ac motor with good starting transients and fast dynamic responses.

1. Introduction

In recent years, the increase in consciousness towards future sustainable growth, as required by international energy regulations, electric motor technologies for households, irrigation systems and industrial applications are endlessly under exploration for additional enhancements as far as the performance and prime costs are concerned [1]. Thus, the need to broaden ac machines' technologies with good starting and dynamic abilities has surfaced [2]. The SCIMs are the most used singly fed line-start ac motors in households, irrigations systems and industrial applications. Despite their robustness and ease of usability, they suffer low efficiency, low power factor and longer time dynamic responses. The possibility to replace squirrel cage IMs with synchronous machines has emerged with several work focusing on the line-start permanent magnet synchronous motors (PMSynMs) that sought to bring solution to problems associated with the squirrel cage IMs low efficiency and low power factor. Therefore, the line-start PMSynMs are high-class alternatives to SCIMs, as they are well within the proposed IE4 efficiency levels, despite having excessive cost per kW [3]. Aside from the high cost of PM materials and high risk of demagnetization, the manufacturing process turns to be expensive, especially when permanent magnets with high hardness are necessitated [4], [5]. Consequently, line-start PMSynMs cannot easily substitute the very entrenched SCIMs [4].

During the last few decades, various scholars have extended their interests beyond the line-start PMSynMs by investigating the line-start SynRM that incorporates the constructional attributes of three-phase SCIM and SynRM. The line-start SynRM can develop an induction torque at starting and a reluctance torque during acceleration, and at synchronous speed. The rotor copper loss is zero during synchronous operation because there is no current induced in rotor bars. To solidify the possibility to replace the SCIMs with line-start SynRMs, several good work have

*Mbika Muteba, 5151 John Orr, Doornfontein, +27115996089, mmuteba@uj.ac.za

sought to compare the performance between the three-phase SCIM and three-phase SynRM with cage bars on the rotor, for various applications [3-4, 6-7]. The Rotor design and dynamic equations that govern the performance of line-start SynRM with round copper bars are investigated in [8]. The dynamic model and transient analysis of line-start SynRM with copper bars in flux barriers are reported in [9] and [10]. Although the copper cage bars have lower rotor resistance and offer a finer synchronization ability than the aluminium, they contribute to the increase in rotor moment of inertia due to their high net weight.

However, modern line-start SynRMs possess a die-casted aluminium cage inside the rotor air barriers [4], [6], [7]. In principle, manufacturing costs of a squirrel cage IM are the same compared to SynRM with die-casted aluminium cage, but at full-load synchronous speed, the efficiency of the line-start SynRM with die-casted aluminium is higher, for the same frame size [7]. The weight of an aluminium conductor is less than the weight of a copper conductor, subjecting the rotor to less stress from centrifugal forces and reduced starting inertia, slighter vibration while running, and it is easily movable than an analogous copper rotor. Moreover, the low yield strength, low Young's Modulus of elasticity and low melting point of the aluminium conductor serve as a good motivation to investigate other conducting materials to be used on the rotors of line-start SynRMs.

Therefore, this paper proposes a line-start SynRM with brass rotor bars. The line-start capabilities are acquired through round brass bars placed inside the air barriers on *direct*-axis and in the rotor core on the *quadrature*-axis as shown in Figure 1 (a). The characteristics of brass used in the proposed machine are well presented in [11]. Despite having a high electric resistivity compared to aluminium and copper, the brass presents a high yield strength, high Young's Modulus of elasticity and high melting point. In the proposed motor, the brass bars are surrounded by cast epoxy resin. The latter permits to keep the bars steady and provide good electric insulation between the brass bars and the steel laminations. The bars are bolted, soldered and brazed to a brass end-ring at both end-points of stack lamination to form a cage.

The organization of this paper is in this manner: section 2 elaborates on ratings and specifications of the proposed SynRM-BRBs, and section 3 presents the model of the SynRM-BRBs. Section 4 deals with the ramifications of the number of rotor brass bars on synchronous parasitic torques during starting, while section 5 focuses on the performance evaluation of the proposed SynRM-BRBs by means of FEA. Section 6 provides details related to experimental validation, while section 7 summarizes the key findings of this paper and further elaborations regarding possible future work.

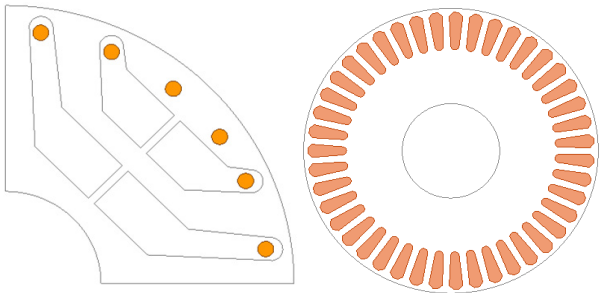
2. Motor Specifications and Ratings

In this paper, the starting transients and dynamic responses of the SynRM-BRBs is compared to a squirrel cage IM that has identical stator frame. Table 1 shows the ratings and specifications of the SynRM-BRBs and SCIM, while Figure1 depicts the rotor cross sections of both machines. The proposed SynRM has 24 symmetrical distributed brass rotor bars, thus only a pole is illustrated in Figure 1 (a). At the same time, the three-phase SCIM has 43 rotor copper bars as depicted in Figure 1 (b). From a design point of view, the number of stator slots and rotor brass bars should not equate [12]. The selection of the number of brass rotor bars depends on the number of air barriers per pole, and on the adequate combination between the number of stator slots and rotor bars for an un-skewed rotor.

Figure 2 depicts the brass bars and end rings for the SynRM, while Figures 3 (a) and (b) picture the prototype rotors for the SynRM with brass bars and squirrel cage induction motor respectively.

Table 1: Ratings and Specifications

Description	Values
Stator slots	36
Poles pairs	2
Full-load line voltage, V	380
Base frequency, Hz	50
Full-load current, A	12
Full-load power, kW	5.5
Series conductors per phase	144
Stator external radius, mm	105
Stator inner radius, mm	73.33
Rotor inner radius, mm	72.8
Shaft radius, mm	24
Airgap Length, mm	0.35
Stack length , mm	160



(a) (b)

Figure 1: Rotors 'cross section, (a) SynRM with brass bars, (b) squirrel cage induction motor

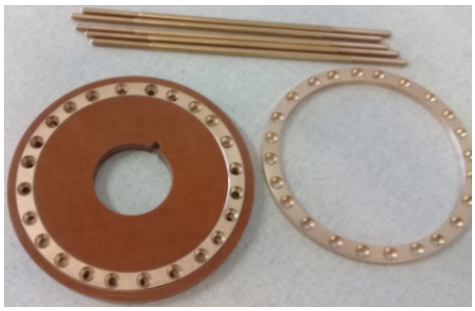


Figure 2: Photograph of brass bars and end rings

The adequate number of stator slots and rotor bars combinations for a 4-pole, 36-slot machine with a skewed rotor should be 36/25, 36/27, 36/28, 36/29, 36/30 and 36/43 [12]. In this paper, the combination of stator slots and rotor bars is 36/24 for the SynRM and 36/43 for SCIM. The rotor of the SCIM is skewed by a slot pitch. Alternatively, a non-skewed rotor design was

opted for the SynRM-BRBs, thus making the combination of 36/24 to be adequate.



(a)



(b)

Figure 3: Photographs of prototype rotors (a) SynRM-BRBs (b) SCIM

3. Model of the SynRM with Brass Rotor Bars

The brass rotor winding is modelled as two identical windings, one on the *direct*-axis and the other on the *quadrature*-axis. Figure.4 illustrates the *d*-and *q*-axis equivalent circuital models of the SynRM-BRBs. The voltage equations that describe the electrical characteristic of the SynRM-BRBs, in rotating arbitrary reference frame are in (1)-(4).

$$V_{sd}^{r} = R_s\, i_{sd}^{r} - \omega\lambda_{sq}^{r} + \rho\lambda_{sd}^{r} \tag{1}$$

$$V_{sq}^{r} = R_s\, i_{sq}^{r} + \omega\lambda_{sd}^{r} + \rho\lambda_{sq}^{r} \tag{2}$$

$$0 = R_{rd}\, i_{rd} + \rho\lambda_{rd} - \left(\omega - \omega_r\right)\lambda_q \tag{3}$$

$$0 = R_{rq}\, i_{rq} + \rho\lambda_{rq} + \left(\omega - \omega_r\right)\lambda_d \tag{4}$$

where $\rho = (d/dt)$, the subscripts *s* and *r* indicate the variables associated with the stator and rotor respectively, while the superscript *r* refers to the rotating reference frame, λ_{sd} and λ_{sq} are the *d*-and *q*-axis stator flux-linkages, λ_{rd} and λ_{rq} are the *d*- and *q*-axis rotor flux-linkages, ω and ω_r are the synchronous and rotor speeds, in electric radians respectively. The two values of rotor resistance for *d*-and *q*-axis in (3) and (4) contemplate the non-isotropic character of the SynRM rotor. The flux-linkages equations are given in (5)-(10).

$$\lambda_{sd}^{r} = L_{\sigma s}\, i_{sd}^{r} + L_{md}\, i_{md}^{r} \tag{5}$$

$$\lambda_{sq}^{r} = L_{\sigma s}\, i_{sq}^{r} + L_{mq}\, i_{mq}^{r} \tag{6}$$

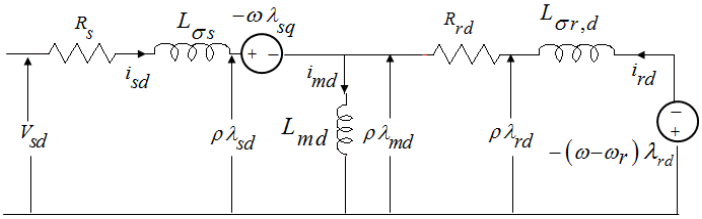
$$\lambda_{md}^{r} = L_{md}\ \ i_{md}^{r} \tag{7}$$

$$\lambda_{mq}^{r} = L_{mq}\ \ i_{mq}^{r} \tag{8}$$

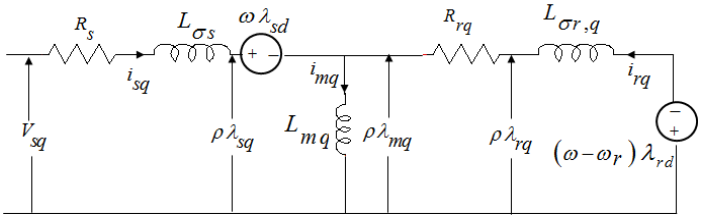
$$\lambda_{rd} = L_{\sigma r,d}\ \ i_{rd} \tag{9}$$

$$\lambda_{rq} = L_{\sigma r,q}\ \ i_{rq} \tag{10}$$

here $L_{\sigma s}$ and $L_{\sigma s}$ are the stator and rotor leakage inductances, L_{md} and L_{mq} are *direct*-and *quadrature*-axis magnetizing inductances.



(a)



(b)

Figure 4: Equivalent circuit of the SynRM-BRBs, (a) *direct*-axis circuital model, (b) *quadrature*-axis circuital model

The torque developed by the three-phase SynRM with brass rotor bars is expressed in (11).

$$T_{em} = \frac{3}{4}\, p\left[\left(L_{ds} - L_{qs}\right) i_{ds}^{r} i_{qs}^{r} + \left(L_{mds}\, i_{dr} i_{qs}^{r} - L_{mqs}\, i_{qr} i_{ds}^{r}\right)\right] \tag{11}$$

Ignoring the friction coefficient, the dynamic equation of the SynRM with cage bars that governs the torque balance at the shaft is generally expressed as

$$T_{em} = \frac{2J}{p}\frac{d\omega_r}{dt} + T_L \tag{12}$$

here *p* is the number of poles, *J* is the total rotor moment of inertia and T_L is the load torque. The electromagnetic torque in (11) has two integrant expressions. The first expression is the reluctance torque, which is the torque at rated synchronous speed, and this torque pulsates at two times the slip frequency during holding phase of the motor [4]. The second expression is the asynchronous torque, which is the torque developed at standstill, also known as

pullout torque. This torque exists only when the SynRM is out of synchronism. A detailed derivation of the asynchronous and reluctance torques is delineated in [7].

4. Space harmonics and Parasitic Torques

The field harmonic caused by the rotor bar current, stator slotting and stator winding phase-spread can be obtained using (13), (14) and (15) respectively [11-12].

$$n_r = x\frac{R_b}{p_1} + 1 \tag{13}$$

$$v_s = x\left(\frac{Q_s}{p_1}\right) + 1 \tag{14}$$

$$u_s = 2x\left(\frac{\pi}{\sigma}\right) + 1 \tag{15}$$

where x is any positive or negative number, R_b is the number of rotor bars, p_1 is the number of fundamental pairs of poles, Q_s is the total number of stator slots and σ is the phase-belt angle. The analysis pertaining to the existence of field harmonics owing to stator slotting, stator winding phase-belt and number of rotor bars in the SynRM-BRBs and SCIM is well documented in [11]. Furthermore, the field harmonic components emanated from the rotor gave rise to synchronous parasitic torques when linking with the field harmonics arisen from the stator with the same order [12-13]. From [11], it was noted that the synchronous parasitic torques occur in both the SynRM-BRBs and the squirrel cage IM. The effect of parasitic torques is more pronounced in the SynRM-BRBs than the squirrel cage IM [11]. To circumvent the synchronous parasitic torque caused by the interactivity between the stator field harmonics and rotor slot harmonics, (13) should not correlate (14) [11-12]. The effects of synchronous parasitic torques on starting transients of both machines are discussed in the succeeding section through FEA.

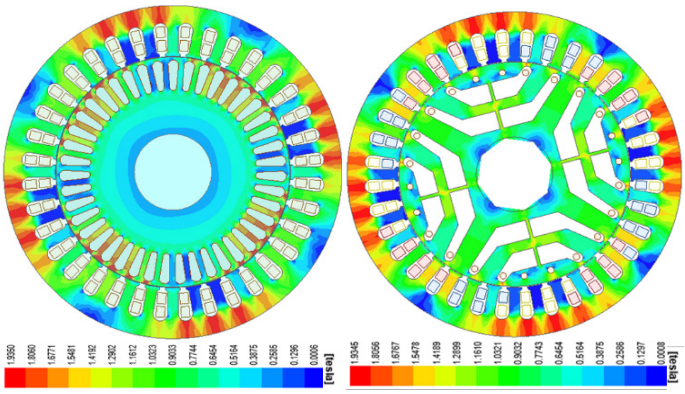
5. Finite Element Analysis

5.1. Machines' flux density

The FEA has been carried out at nameplate voltage. The distributed three-phase double layer stator windings are supplied by three-phase sinusoidal voltage. The skin effect and core loss are not disregarded in the FEA. The skew of the SCIM rotor is accounted for in the FEA.

Figure 5 shows the flux density distribution plots; Figure 6 depicts the airgap flux density profiles, and their FFTs are given in Figure 7. It is clear from the FEA results in Figure 5 that there are localized saturations of stator back iron in both the conventional squirrel cage IM and SynRM with brass rotor bars. Localized saturations are also observed between some rotor teeth, and on some magnetic wedges of conventional squirrel cage IM. High flux density of about ± 1.2 tesla is noticed on the magnetic radial ribs of the SynRM-BRBs. In the latter, some magnetic tangential bridges of the upper air barriers exhibit a flux density of about ± 1.4 tesla, while the flux density in other magnetic tangential bridges levels between 1.2 tesla and 1.3 tesla.

The effects of stator slot openings are noticeable in the airgap flux density waveforms illustrated in Figure 6. The FFT results in Figure 7 (a) evidenced that the SynRM-BRBs has achieved a fundamental airgap flux density of 0.61 tesla. Elseways, the achieved fundamental airgap flux density of SCIM is 0.57 tesla. The 17th space harmonic are dominant in both machines. These harmonics are mainly due to stator slotting, and are reliant on the stator slots and pole pairs' numbers. The 17th airgap space harmonics do not contribute to any asynchronous parasitic torque that may occur during starting because they do not produce rotor bar currents in both machines, whose magneto-motive force (MMF) harmonic has the same order. The 3rd airgap space harmonic is high in the SynRM-BRBs compared to the SCIM. The net airgap flux density is the result of the MMF harmonics and the airgap conductance harmonics. The latter is due to slot opening, leakage, skew, and saturation harmonics. The 3rd space harmonics are mainly caused by main flux path saturation in the stator back iron as noticed in Figure 5, and they are also due to airgap conductance harmonics caused by leakage slot magnetic saturation.



(a) (b)

Figure 5: Flux density distribution, (a) Squirrel cage induction motor, (b) SynRM with brass bars

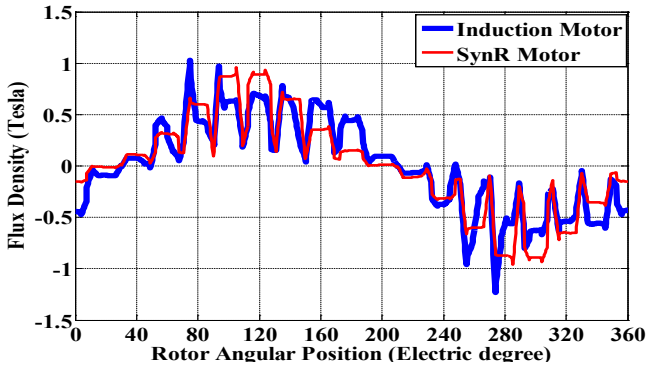
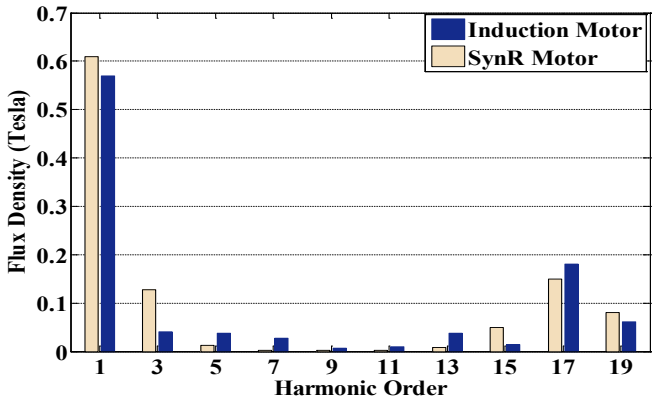


Figure 6: Airgap flux density profile

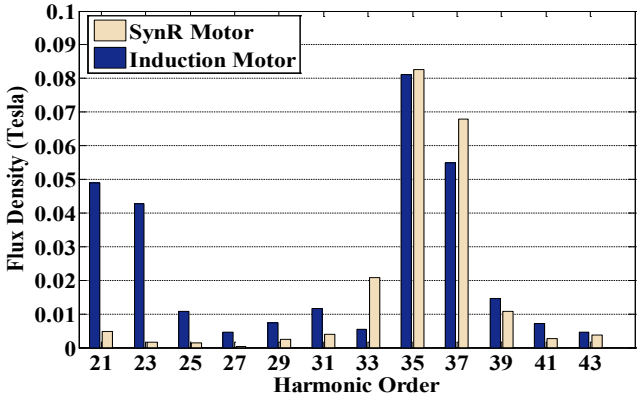
Although both conventional SCIM and SynRM with brass rotor bars show almost the same level of localized flux saturation in the stator back iron, the rotor of SynRM with brass bars shows a well-distributed main flux path saturation between air barriers, thus increasing the airgap 3rd flux density harmonic components.

These airgap 3rd flux density harmonics have the advantage to positively contribute to the average airgap flux density and the machine torque density [14-16].

From figure 7 (b), it is clear that the 35th and 37th airgap flux density harmonics are dominant as far as high order number is concerned. They are mainly due to stator slot and phase-belt magneto-motive force harmonics. In the SynRM-BRBs the 35th and 37th airgap flux density harmonics produced in the brass bars currents with the 35th and 37th MMF harmonics, thus producing synchronous parasitic torques. On the other hand, the 21st and 23rd airgap flux density harmonics in the squirrel cage IM produced rotor currents with the 21st and 23rd MMF harmonics. The same can be stated for the 41st and 43rd airgap flux density harmonics. The latter produced rotor currents with 41st and 43rd MMF harmonics in the SCIM. The interaction of these space harmonics which are originated from different sources (stator and rotor), and having the same order number would contribute to the occurrence of asynchronous or synchronous parasitic torques [12-13].



(a)



(b)

Figure 7: Airgap flux density harmonic components, (a) low order harmonics (b) high order harmonics

5.2. Motors' performance parameters

The rated performance parameters in Table 2 were directly obtained from FEA at the nameplate values. The magnetizing reactance, which depends on the fundamental stator winding number of turns, winding factor, effective stack length, effective inverse airgap and saturation factor, is higher for the SynRM-BRBs compared to the SCIM. This justifies the SynRM with brass bars high value of the airgap fundamental flux density as observed in the previous subsection. The effective inverse airgap is proportional to the carter factor. It should be noted that the carter factor of the SynRM with brass bars does not account for rotor slot opening, which is not the case for the SCIM.

Table 2: Performance Parameters

Description	Value	
	IM	SynRM
Stator resistance , Ω	0.559	0.559
Stator leakage reactance, Ω	0.976	1.030
Stator slot leakage reactance	0.602	0.599
Stator end leakage reactance	0.212	0.212
Stator differential leakage reactance, Ω	0.162	0.217
Magnetizing reactance, Ω	31.21	41.95
Rotor resistance, Ω	1.002	2.377
Rotor leakage reactance, Ω	4.605	5.832
Rotor end leakage reactance, Ω	0.071	0.062
Rotor slot leakage reactance, Ω	4.069	6.659
Rotor differential leakage reactance, Ω	0.250	1.077
Rotor skewing leakage reactance, Ω	0.111	0

From Table 2, it is noticed that the value of the SynRM-BRBs rotor resistance is more than twice the rotor resistance of the SCIM. In addition to a high value of the brass resistivity compared to copper resistivity, the brass bars have a low section area compared to the squirrel cage bars. The skin effects, which are associated with the flux and current density distribution in brass bars or copper cage bars, influence both the rotor resistances and rotor slot leakage reactances. The skin effects are more significant in the SynRM with brass bars because 16 of the 24 brass rotor bars are surrounded by great amount of cast epoxy resin having small air bubbles. The skewing of the SCIM rotor has introduced the skewing magneto-motive force responsible for additional leakage component known as rotor skewing leakage reactance, thus slightly reducing the magnetizing reactance, while increasing the rotor leakage inductance. Though the SynRM with brass bars rotor differential leakage reactance is high, there is a huge difference as far as the stator differential leakage reactances are concerned. It should be noted that the stator differential leakage reactances are attenuated by the reaction of the rotor bars [12]. The two machines use the same stator NEMA frame and they have the same series conductors per phase, therefore given rise to almost the same values of stator resistance, slot leakage reactance and end connection leakage reactance.

5.3. Transient currents and mass rotor torques

Excluding the shaft, the inertia of the rotor is 0.08087 kg.m^2 for the SynRM with brass bars and 0.122 kg.m^2 for the squirrel cage IM. In the latter, the total weight of bars and end rings is found to be 6.418 kg, while it is found to be 3.279 kg for the SynRM-BRBs. The low rotor inertia in the SynRM with brass bars contributes to a fast response as far as the dynamic associated with the mass rotor torque is concerned, as evidenced in Figure 9. Furthermore, the transient currents shown in Figure 8 evidence that the SynRM with brass bars draws low starting currents compared to the SCIM.

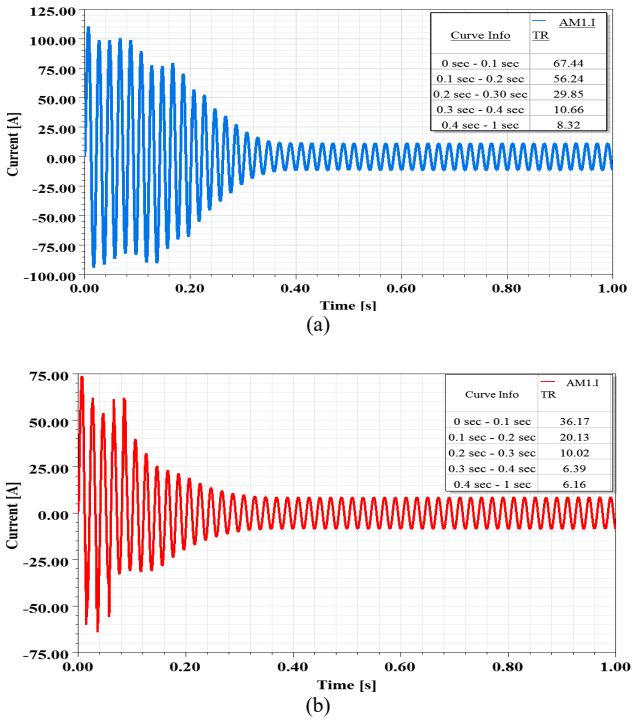


Figure 8: Transient current, (a) Squirrel cage IM, (b) SynRM with brass bars

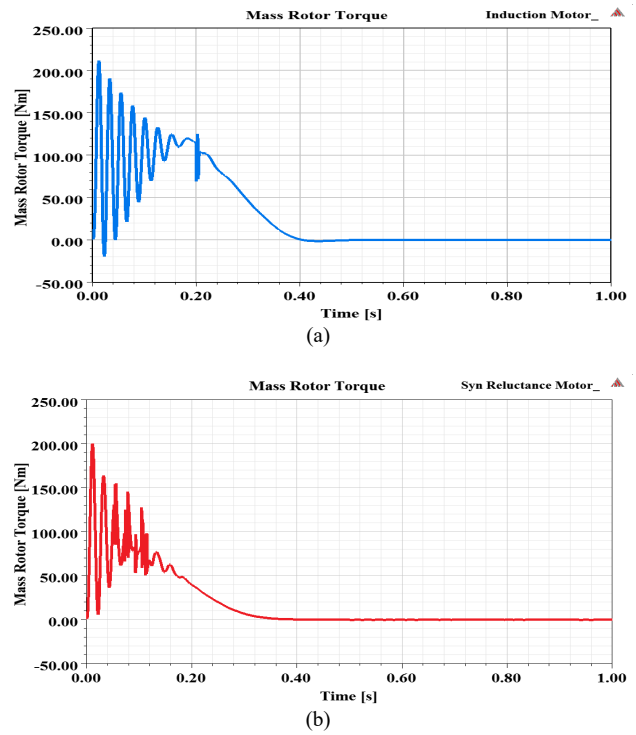


Figure 9: Transient mass rotor torque, (a) Squirrel cage IM, (b) SynRM with brass bars

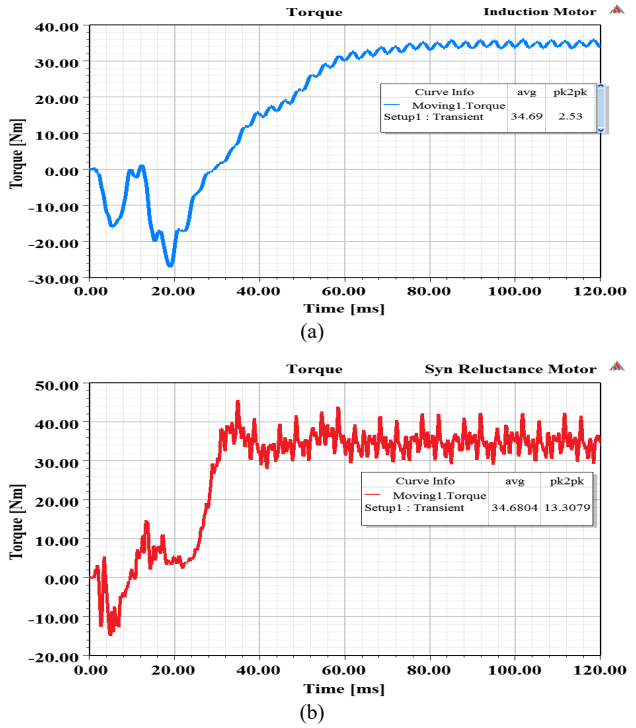


Figure 10: Starting torque, (a) Squirrel cage IM, (b) SynRM with brass bars

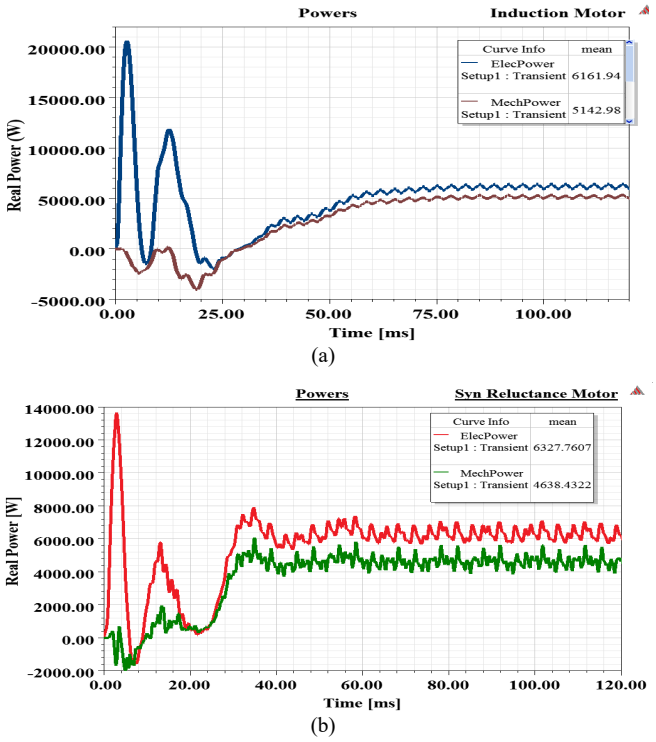


Figure 11. Transient input and output powers, (a) Squirrel cage IM, (b) SynRM with brass bars

5.4. Transient torques and Powers

The total winding energy losses during motor starting and acceleration depend on the stator current, rotor current, starting stator resistance and starting rotor resistance. The starting torque characteristics in Figure 10 indicate that the SynRM-BRBs has a shorter transient response time compared to the squirrel cage IM.

The SynRM with brass bars acceleration begins as early as 6 mSec, crawls for about ± 8 mSec before accelerating up to synchronous operation. The interactivity of the 1st rotor slot field harmonic (11th order) and the 2nd stator winding phase-spread (11th order) field harmonic has given rise to synchronous parasitic torque which are partly responsible for the crawling effect in SynRM with brass bars as discussed earlier on. However, the

squirrel cage IM has shown to have good acceleration without the crawling effect. Both motors have almost the same steady-state average torque, but the SynRM with brass bars exhibits high torque ripple contents compared to the squirrel cage IM. The interactivity of field harmonics of the electrical loading and the rotor anisotropy is responsible for the high ripple contents in the SynRM with brass bars. In addition to skewing the rotor, the optimization of rotor flux barriers microscopic design variables may assist with the mitigation of some torque harmonics that contribute to excessive torque ripple in the SynRM with brass rotor bars [17-23].

Operating at 80% of full-load, the input electric and output mechanical powers transient behaviors of the squirrel cage IM and the SynRM with brass bars are given in Figure 11 (a) and (b) respectively. The FEA results evidenced that the squirrel cage IM absorbs a high real power during starting as opposed to the SynRM-BRBs. This is because the squirrel cage IM locked rotor current is also high.

However, the steady-state squirrel cage IM absorbed real power is less as opposed to the power absorbed by the SynRM with brass bars. The mechanical out powers follow the torque pattern as observed in Figure 10. For the same torque density, the squirrel cage IM delivers more output power in comparison to the SynRM-BRBs. The low output mechanical power of the SynRM-BRBs is caused by high transient rotor winding losses and high critical slip. The latter is reached faster in the SynRM-BRBs, and it is mainly dependent on the rotor resistance and total rotor leakage reactance. Operating at 80 % of full-load, the efficiency of the SynRM-BRBs is found to be 73.3 %, while that of the squirrel cage IM is 79.58 % for the same operational condition.

6. Experimental Results

6.1. Motors' starting transients

The SynRM with brass rotor bars analysed in this paper starts directly online with applied rated line voltage of 380 V. Neglecting the stator resistance, the maximum pullout torque that the motor is proficient to develop at a designated voltage V is [4]

$$\frac{\partial T_s}{\partial \theta} = 0 \rightarrow T_s = \frac{3}{4} p \frac{V_s^2}{2} \frac{\left(L_d - L_q\right)}{\left(X_d X_q\right)} \tag{16}$$

where $L_d = (L_{\sigma s} + L_{md})$, $L_q = (L_{\sigma s} + L_{mq})$, θ is the load angle, X_d and X_q are the d-and q-axis synchronous reactances respectively. The d-and q-axis magnetizing inductances measured values are 34.3 mH and 10.77 mH respectively. Furthermore, the mesured value of stator leakage inductance is 4.2 mH. Figure 12 shows a photo of the experimental setup.

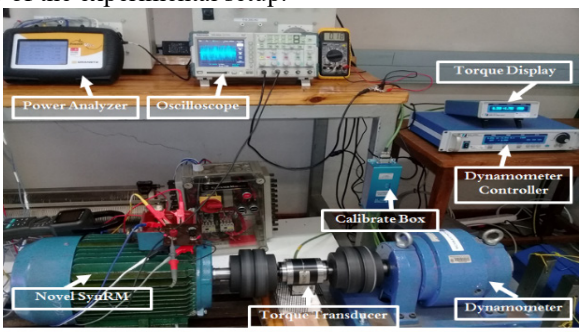
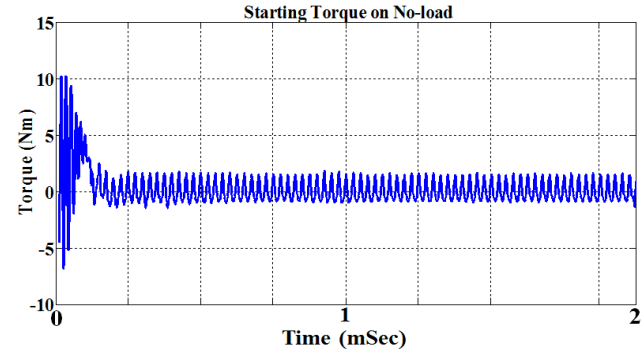
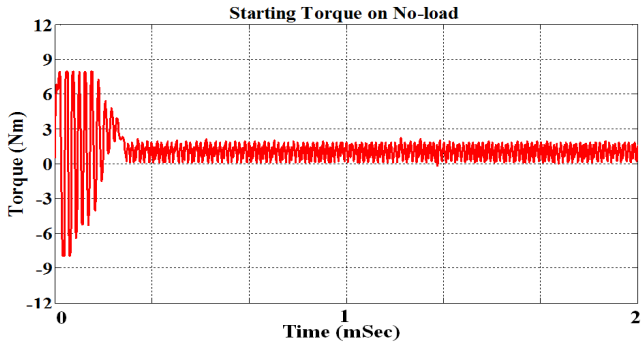


Figure 12: Experimental setup rig photo

Figures 13 (a) and (b) express the no-load starting transient torques of the squirrel cage IM and SynRM with brass bars respectively, while Figures 14 (a) and (b) depict the no-load starting transient currents. At starting, the SynRM developed an asynchronous torque owing to great value of induced current in the brass rotor bars.



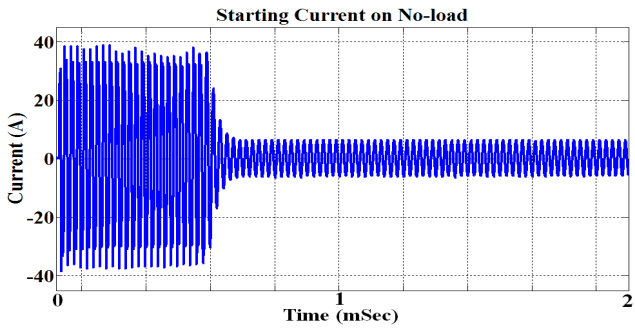
(a)



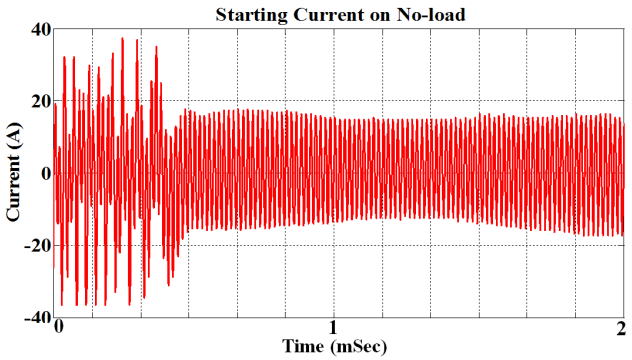
(b)

Figure 13: Starting transient torque at no-load (a) squirrel cage IM (b) SynRM with brass bars

Furthermore, the starting currents on no-load of both motors are more than three times the rated currents. The squirrel cage IM motor starting current dropped fast once it begins to accelerate. The crawling effects during acceleration mode are significant in SynRM with brass bars as it was the case in the FEA results illustrated in the previous section. The crawling effects are caused by synchronous parasitic torques, and they may be mitigated by exploring new designs of the SynRM with brass rotor bars, with adequate combination between the number of brass rotor bars, flux-barriers per pole and that of stator slots for an un-skewed rotor [12-13].



(a)



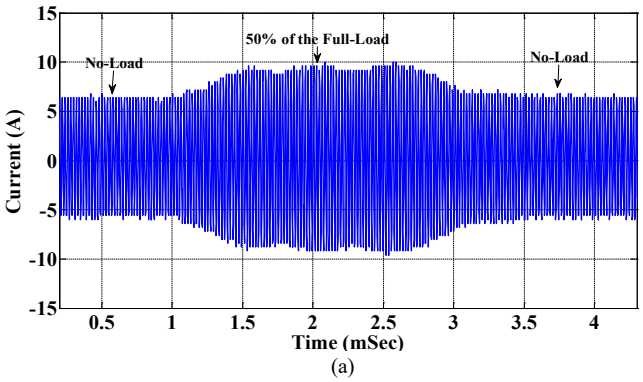
(b)

Figure 14: Starting transient current at no-load (a) squirrel cage IM, (b) starting current of SynRM

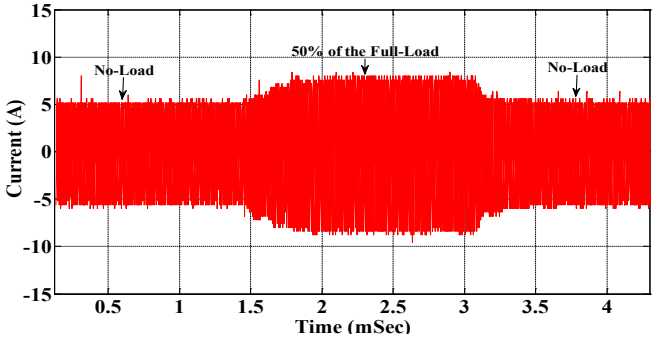
6.2. Motors' dynamic responses

This subsection deals with transients associated with the dynamic responses when the motor mechanical load is abruptly varied. The dynamic responses for currents of the SCIM and SynRM with brass rotor bars are shown in Figures 15 and 16 respectively. The results in Figure 15 denote that the no-load peak-to-peak current is low in the SynRM-BRBs compared to the squirrel cage IM. In the latter, there is a steady increase of currents for about 0.5 mSec and 1 mSec to peak-to-peak values of ± 9 A and ± 12 A, when the load is varied from no-load to 50% and 60% of the full-load respectively. There is a fast decrease of peak-to-peak current for about 0.35 mSec and 0.6 mSec, when the load is varied to no-load from 50% and 60% of the full-load respectively.

However, the results in Figure 16 indicate that the SynRM-BRBs exhibits a fast dynamic response compared to the squirrel cage IM, with a steady increase in currents for about 0.27 mSec and 0.35 mSec, to a peak-to-peak values of ± 8 A and ± 10 A, when the loads of 50% and 60% of the full-load are respectively applied. The decrease of peak-to-peak current lasts for about 0.1 mSec and 0.2 mSec when the loads of 50% and 60% of the full-load are respectively removed. Simply put, the SynRM-BRBs' dynamic response is 0.23 mSec and 0.65 mSec faster than the squirrel cage IM when there is a load variation from no-load to 50% and 60% of the full-load respectively. Furthermore, the SynRM-BRBs' dynamic response is 0.25 mSec and 0.4 mSec faster than the squirrel cage IM when the loads of 50% and 60% of full-load are respectively removed.

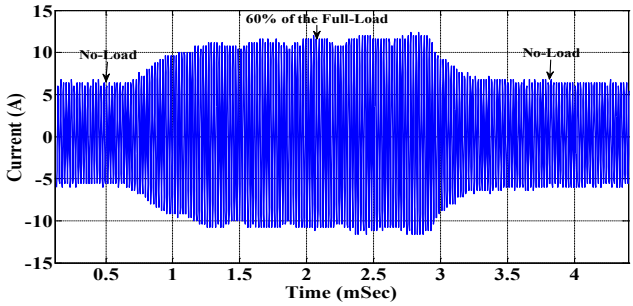


(a)

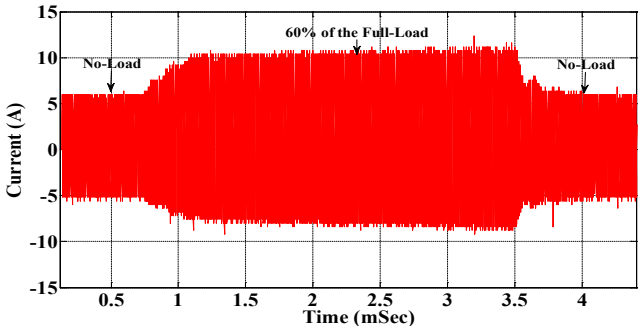


(b)

Figure 15: Dynamic responses, no-load to 50% of the full-load current (a) squirrel cage IM, (b) SynRM with brass bars



(a)



(b)

Figure 16: Dynamic responses, no-load to 60% of the full-load (a) squirrel cage IM, (b) SynRM with brass bars

7. Conclusion

This paper presented the analysis of a SynRM-BRBs as a strong candidate to replace the squirrel cage IM in applications that require a line-start ac motor with good starting transients and fast dynamic responses. The FEA results evidenced that the low rotor inertia of the SynRM-BRBs contributed to fast transient responses. Furthermore, the reluctance torque has been observed to have a compounding effect in both FEA and experimental results, on the accelerating torque of the proposed SynRM-BRBs.

The proposed line-start motor has proved to have fast dynamic responses, reduced steady state no-load, and load current compared to the squirrel cage IM. The main contribution of the proposed design is in the use of brass bars and end-rings on the rotor, having a high yield strength, high Young's Modulus of elasticity and high melting point in contrast to copper and aluminium cages used in modern line-start SynRMs.

For future work, new designs of the SynRM with brass rotor bars, with adequate combination between the number of brass rotor bars, air barriers per pole and that of stator slots for a skewed and an un-skewed rotor may intensively be explored in order to alleviate the crawling effect due to synchronous parasitic torques. The optimization of rotor flux barriers microscopic design variables may also assist with the mitigation of some torque harmonics that contribute to high torque ripple contents in SynRM with brass rotor bars during synchronous operation.

References

[1] A.T. de Almeida, F.J.T.E. Ferreira and G. Baoming, Beyond Induction Motors Technology Trends to Move up Efficiency, IEEE Tran. Ind. Appl. **50** (3), 2103-2114, 2014, https://doi.org/10.1109/TIA.2013.2288425

[2] A. Castagnini, T. Känsäkangas, J. Kolehmainen, and P. Savio Termini, Analysis of the starting transient of a synchronous reluctance motor for direct-on-line applications, Coeur d'Alene, ID, USA, May 10-13, 2015, https://doi.org/10.1109/IEMDC.2015.7409047

[3] A.H. Isfahani and S. Vaez-Zadeh, Line start permanent magnet synchronous motors: Challenges and opportunities, Energy, **34**(11), 1755-1763, 2009, https/doi.org/10.1016/j.energy.2009.04.022

[4] M. Gamba, E. Armando, G. Pellegrino, A. Vagati, B. Janjic and J. Schaab, Line-start synchronous reluctance motors: Design guidelines and testing via active inertia emulation, in IEEE Energy Conversion Congress and Exposition, Montreal, QC, Canada, Sept 20-24, 2015, ,https://doi.org/10.1109/ECCE.2015.7310340

[5] D. Mingardi and N. Bianchi, Line-start PM-assisted synchronous motor design, optimization, and tests, IEEE Tran. Ind. Elec, **64** (12), 9739-9747, 2017, https://doi.org/10.1109/TIE.2017.2711557

[6] A. Kersten, Y. Liu and D. Pehrman, Rotor Design of a Line-Start Synchronous Reluctance Machine with Respect to Induction Machine for Industrial Applications, in 13th IEEE International Conference on Electrical Machines, Alexandroupoli, Greece, Sept 3-6, 2018, https://doi.org/10.1109/ICELMACH.2018.8507053

[7] M. Gamba, G. Pellegrino, A. Vagati, and F. Villata, Design of a line-start synchronous reluctance motor," in IEEE International Electric Machines Drives Conference, Chicago, IL, USA, May 12-15, 2013, https://doi.org/10.1109/IEMDC.2013.6556163

[8] A. Kersten, Y. Liu, D. Pehrman, and T. Thiringer, Rotor design of line-start synchronous reluctance machine with round bars, IEEE Tran. Ind. Appl, **55** (4), 3685-3696, 2019, https://doi.org/ 10.1109/TIA.2019.2914010

[9] V. Aguba, M. Muteba, D.V. Nicolae, Dynamic modelling and transient analysis of synchronous reluctance motor with cage bars in the rotor structure, in International Symposium on Power Electronics, Electrical Drives, Automation and Motion, Amalfi, Italy, June 20-22, 2018, https://doi.org/10.1109/SPEEDAM.2018.8445292

[10] V. Aguba., M. Muteba, D.V. Nicolae, Transient analysis of a start-up synchronous reluctance motor with symmetrical distributed rotor cage bars, in 2017 IEEE AFRICON, Cape Town, South Africa, Sept 28-20, 2017, https://doi.org/10.1109/AFRCON.2017.8095668

[11] M. Muteba, Comparison of dynamic behaviors between a synchronous reluctance motor with brass rotor bars and a squirrel cage induction motor, in IEEE PES/IAS PowerAfrica, Abuja, Nigeria, Aug 20-23, 2019, https://doi.org/10.1109/PowerAfrica.2019.8928746

[12] I. Boldea, S. A. Nasar, The induction machine handbook , CRC Press, New York, 2002.

[13] L. Wang, X. Bao, C. Di, Y. Zhou and Q. Lu, Analysis of synchronous parasitic torque in dual skew cage rotor induction motors with equivalent slot number, IET Elect. Power Appl., 2017, Vol. **11**(8), 1357-1365, 2017, https://doi.org/10.1049/iet-epa.2016.0565

[14] K. Wang, Z. Q. Zhu, G. Ombach and W. Chlebosz, Average torque improvement of interior permanent magnet machine using Third Harmonic in Rotor Shape, IEEE Tran. Ind. Elec, **61** (9), 5047-5057, 2014, https://doi.org/10.1109/TIE.2013.2286085

[15] R. O. C. Lyra and T.A. Lipo, Torque Density Improvement in a Six-Phase Induction Motor With Third Harmonic Current Injection, IEEE Tran. Ind. Appl, **48** (5), 1351-1360, 2002, https://doi.org/10.1109/IAS.2001.955773

[16] H. A. Toliyat, S. P. Waikar, and T.A. Lipo, Analysis and Simulation of Five-Phase Synchronous Reluctance Machines Including Third Harmonic of Airgap MMF, IEEE Tran. Ind. Appl, **34** (2), 332-339, 1998, https://doi.org/10.1109/28.663476

[17] X. L. Bomela and J. Kamper, Effect of stator chording and rotor skewing on performance of reluctance synchronous machines. IEEE Tran. Ind. Appl., **38**(1), 91-100, 2002, https://doi.org/10.1109/28.980362

[18] N. Bianchi, S. Bolognami, D. Bond, D and M. D. Pre, Rotor flux-barrier design for torque ripple reduction in Synchronous Reluctance and PM-assisted Synchronous Reluctance Motors. IEEE Trans. Ind. Appl, **45**(3), pp. 921-928, 2009, https://doi.org/10.1109/TIA.2009.2018960

[19] V. Bilyi, D. Gerling and D. Bilyi, Flux barrier design method for torque ripple reduction in synchronous reluctance machines. in IEEE Transportation Electrification Conference and Expo, Asia- Pacific, Busan, Korea, June 1-4, 2016, https://doi.org/10.1109/ITEC-AP.2016.7512918

[20] M. N. F. Ibrahim, P. Sergeant and E. E. M. Rashad, Simple Design Approach for Low Torque Ripple and High Output Torque Synchronous Reluctance Motors, Energies, **9** (11), 1-14, 2016, https://doi.org/10.3390/en9110942

[21] N. Bianchi, M. Degano and E. Fornasiero, Sensitivity Analysis of Torque Ripple Reduction of Synchronous Reluctance and Interior PM Motors, IEEE Trans. Ind. Appl, **51**(1), 2014, 187-195, https://doi.org/10.1109/TIA.2014.2327143

[22] N. Bianchi, S. Bolognani, D. Bon, and M. Dai Pre, ''Torque harmonic compensation in a synchronous reluctance motor,'' IEEE Trans. Energy. Con., **23**(2), 466–473, 2008, https://doi.org/ 10.1109/TEC.2007.914357

[23] M. Muteba, B. Twala and D.V. Nicolae, Torque ripple minimization in synchronous reluctance motor using a sinusoidal rotor lamination shape, in the IEEE 22nd International Conference on Electrical Machines, Lausanne, Switzerland, Sept 4-7, 2016, https://doi.org/ 10.1109/ICELMACH.2016.7732588

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 96-107 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Using TOST in Teaching Operating Systems and Concurrent Programming Concepts

Tzanko Golemanov*, Emilia Golemanova

Department of Computer Systems and Technologies, University of Ruse, Ruse, 7020, Bulgaria

ARTICLE INFO

Article history:
Received: 30 July, 2020
Accepted: 15 October, 2020
Online: 08 November, 2020

Keywords:
Operating Systems
Concurrent Programming
Teaching Tools

ABSTRACT

The paper is aimed as a concise and relatively self-contained description of the educational environment TOST, used in teaching and learning Operating Systems basics such as Processes, Multiprogramming, Timesharing, Scheduling strategies, and Memory management. TOST also aids education in some important IT concepts such as Deadlock, Mutual exclusion, and Concurrent processes synchronization. The presented integrated environment allows the students to develop and run programs in two simple programming languages, and at the same time, the data in the main system tables can be monitored. The paper consists of a description of TOST system, the features of the built-in programming languages, and demonstrations of teaching the basic Operating Systems principles. In addition, some of the well-known concurrent processing problems are solved to illustrate TOST usage in parallel programming teaching.

1. Introduction

This paper is an extension of work originally presented in 29th Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE) [1].

Operating Systems (OS) is one of the most fundamental courses in Computer Engineering, Computer Science, and Information Systems curricula. At the same time, students are faced with a lot of problems in learning OS basics, due to not fully understanding the wide range of techniques, strategies, and architectures involved in OS modules. The complexity and depth of concepts in this area requires a careful and detailed explanation from the educator in order to reach a better apprehending. Developing concurrent programs in an operating system environment (Unix, Linux, etc.) asks for students to already have prerequisite skills in both the C-programming and the operating system. Usually, such level of experience is difficult to be obtained at an undergraduate degree, thus a preferable approach would be to use a convenient and effective teaching tool that helps the students in their uneasy endeavor to apprehend and master the OS concepts.

We have assigned the widely spread tools for studying OS into the following two groups:

- Instructional Operating Systems
- Visual OS simulators

The systems from the first group (MINIX [2], Nachos [3], Xinu [4], Pintos [5], GeekOS [6]) run on real hardware and are too complicated to install and use. The main disadvantage of the instructional OS is the requirement students to be familiar with system-level programming in C, C++, Java or assembly language, and to be able to study large program codes (more than 4K program lines of Geek-OS and 15K of MINIX). Most of these systems do not have graphical visualizations, which makes them not very suitable for teaching purposes. In such cases, students spend a lot of time learning the systems themselves and how to use them properly.

In contrast, the visual tools from the second group (SOsim [7], SchedulerSim [8], CPU Sim [9]) are easier to use and convenient, however, they present rather restricted and pre-specified abilities (e.g. simulation of just one OS module or subsystem). For example, the TROJAN [10] simulator helps in teaching multiprocessor organization, the cache utilization, and the network traffic issues, while SIME [11] visualizations focus on memory management, but without shared memory concepts. Additionally, some of these systems (Proc OS [12], Alg OS [13]) do not connect theoretical concepts to a running code. Others present simulations of only one example problem (e.g. the Producer/Consumer problem [14]).

The authors propose a newly-developed by them OS teaching system named ***TOST*** *(Teaching in Operating Systems Tool)*. It has

*Corresponding Author: Tzanko Golemanov, Ruse University, Ruse, Bulgaria, tgolemanov@uni-ruse.bg

been created in response to the requirement of the educational tool to be straightforward to use, yet powerful enough, allowing the students to not only experiment with OS settings, architectures, and strategies but also to create and run any concurrent programs they want. This imposes the need for TOST to be developed as an integrated environment consisting of a code editor, a high-level language compiler, a virtual processor emulator, an interface for setting and modifying OS parameters, and a visualization of the content of the main OS tables.

The structure of the paper is as follows: Section 2 is an introduction to TOST giving an overview of the TOST interface, the grammars of the built-in programming languages, the manners to develop and execute concurrent processes, as well as the ways to set and modify system options, and to monitor the system tables' data. Section 3 is devoted to teaching Operating System basics, such as Process Management, Process Scheduling, and Memory Management. Section 4 describes the use of TOST in teaching Concurrent Programming concepts, specifically Mutual Exclusion (Dekker's algorithm, Peterson's algorithm, Test_And_Set instruction, and Semaphores), Interprocess Synchronization (Sleeping Barber problem), and Deadlock (Dining Philosophers problem). In Section 5, a student evaluation is presented. We conclude in Section 6 and discuss some further improvements to TOST.

2. The TOST System

The main objective of an educational tool in OS is to give students an inside view of the operating system. At the same time, in our opinion, the tool should meet the following basic requirements:

- to operate on an accessible hardware and software platform
- to support multiprogramming and concurrent processes
- to be with a well-known interface
- to support well-known programming languages with concurrency features
- to allow students the experimentation with the basic OS parameters and options.

TOST has been developed with exactly this idea in mind. It allows students to develop and execute concurrent processes, while at the same time they can change the basic settings of the OS (CPU scheduling, Quantum size, Memory management, Virtual memory strategies, etc.) and to monitor the main system tables (Processes, Blocked, Ready, Semaphores, etc.).

The proposed integrated environment includes:

- a multi-tasking, and a multi-window operating system
- editors and compilers for two simple programming languages
- a virtual processor emulator.

The last TOST version is developed with Embarcadero® Delphi within Embarcadero RAD Studio 10.3.3 Rio. It has no installation procedure, the executable file of the tool is very small in size (less than one MB), and it operates on any MS Windows environment. TOST is easy to learn and use, and allows students to become familiar with the capabilities and the limitations of the built-in programming languages within a short time (about an hour).

The first TOST environment is introduced in [15], while the focus of [1] is on using TOST in teaching mutual exclusion, synchronization, and deadlock. This paper expands the application area of TOST, describing its usage for teaching and learning OS concepts.

2.1. Programming in TOST

A general view of the current TOST main menu is depicted in Figure 1.

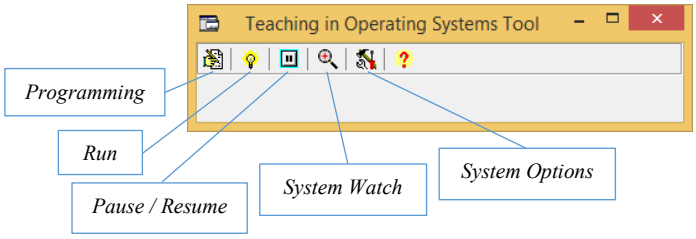


Figure 1: TOST interface

Using TOST, students can develop and execute concurrent programs in two simple programming languages - PASCAL-style and C-style. The grammars of these languages, consist of roughly twenty grammar productions only, which are presented in Figure 2 and Figure 3 in Extended BNF (ISO/IEC 14977).

```
program    = "void" "main" "(" ")" body
body       = [ { type identifier { "," identifier } } ] block
type       = "int" | "char" | "bool" | "shared" | "semaphore"
block      = "{" statement { ";" statement } "}"
statement  = block | assignment | if | while | read | write | lock | unlock |
             init | wait | signal
assignment = identifier "=" expression
expression = expr [ ( "<" | ">" | "==" ) expr ]
expr       = term { ( "+" | "-" | "or" ) term }
term       = factor { ( "*" | "/" | "and" ) factor }
factor     = [ "not" ] ( identifier | constant | random | "(" expression
             ")" )
constant   = number | """" char """" | "#" ascii_code | "true" | "false"
random     = "random" "(" expression ")"
if         = "if" "(" expression ")" statement
while      = "while" "(" expression ")" statement
read       = "cin" ">>" identifier
write      = "cout" "<<" expression { "<<" expression }
lock       = "lock" "(" identifier ")"
unlock     = "unlock" "(" identifier ")"
init       = "init" "(" identifier "," expression ")"
wait       = "wait" "(" identifier ")"
signal     = "signal" "(" identifier ")"
```

Figure 2: C-style language grammar productions

```
program    = [ "program" identifier ";" ] body "."
body       = [ "var" { identifier { "," identifier } ":" type ";" } ] block
type       = "integer" | "char" | "boolean" | "shared" | "semaphore"
block      = "begin" statement { ";" statement } "end"
statement  = block | assignment | if | while | read | write | lock | unlock |
             init | wait | signal
assignment = identifier ":=" expression
expression = expr { ( "<" | ">" | "=" ) expr }
expr       = term { ( "+" | "-" | "or" ) term }
term       = factor { ( "*" | "/" | "and" ) factor }
```

```
factor    = [ "not" ] ( identifier | constant | random | "(" expression
            ")" )
constant  = number | "'" char "'" | "#" ascii_code | "true" | "false"
random    = "random" "(" expression ")"
if        = "if" expression "then" statement
while     = "while" expression "do" statement
read      = "read" "(" identifier ")"
write     = "write" "(" expression { "," expression } ")"
lock      = "lock" "(" identifier ")"
unlock    = "unlock" "(" identifier ")"
init      = "init" "(" identifier "," expression ")"
wait      = "wait" "(" identifier ")"
signal    = "signal" "(" identifier ")"
```

Figure 3: PASCAL-style language grammar productions

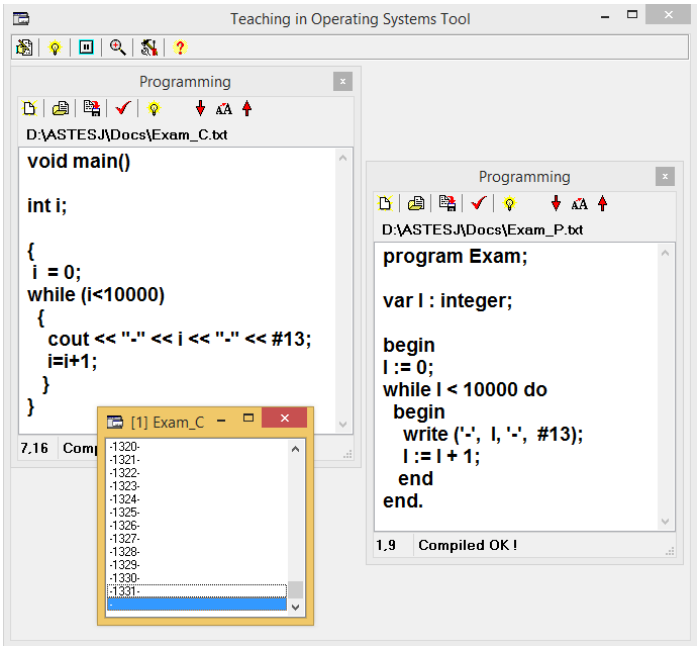


Figure 4: Process editing and executing in TOST

In the text bellow we will use C-style only for the programs' definition.

The programming languages in TOST are exclusively for learning purposes and while they are quite simple, there are some features that need to be addressed:

2.1.1. Types of variables

Only a few standard simple data types *integer*, *char* and *boolean* are available for local variables, nevertheless, some data types for concurrent programming purposes are also provided. Variables from the ***shared*** type can be declared in a TOST program. These variables become common for all concurrent processes running in the system. A shared variable is unified, i.e. it can accept any integer, char or boolean values, as well as any data can be extracted from it. As a result, passing through such variable can also be used in order to convert data from one type to another.

Another type of data used in parallel programming is the ***semaphore***. In TOST we implement the original Dijkstra's semaphores concept [16]. A semaphore variable ***S*** consists of two properties - ***S.value*** and ***S.queue***. The property S.value is a non-negative integer, while S.queue is a FIFO list of identifiers of the blocked processes (process ID).

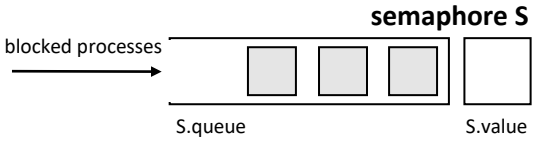


Figure 5: Semaphore's structure

2.1.2. Statements

Although TOST languages maintain only basic statements like assignment, conditional, loop and input/output, some statements for concurrent programming purposes are also available.

The semaphore operations ***init***, ***wait***, and ***signal***, which are defined below, are functions from the OS kernel:

init (S, value)

```
S.value = value
```

wait(S)

```
if (S.value > 0)
  decrement S.value;
else
  block this process and add it's ID in S.queue
```

signal (S)

```
if (S.queue is empty)
     increment S.value;
else
     unblock the first process from S.queue
```

Figure 6: Semaphore operations

Students can investigate the advantages and disadvantages of using the special assembly language instruction ***Test_and_Set*** [17] through the statements ***Lock*** (CS) and ***Unlock*** (CS), where CS (***C****ritical* ***S****ection*) is a boolean shared variable.

Function ***random***(*expression*) which returns a random integer value from *0* to *expression-1* can also be used in some synchronization primitives' implementations.

2.2. Processes Execution Management in TOST

After the successful compilation, an object code file is generated (with COD extension) and it is ready to be run in the environment. The compiled processes can be started from the ***Run*** at the main menu (as well as from the editing windows).

Figure 7 presents the general system view where in the ***Run*** dialog box students can select an object file to be executed, and set the initial parameters of the running process.

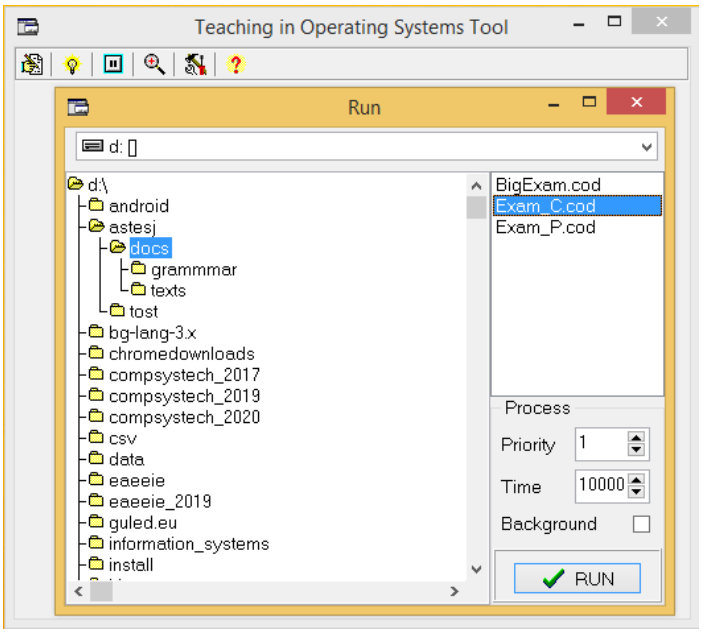


Figure 7: Process running

Priority spin-edit box sets the initial process priority while the total process execution time (used in some scheduling strategies) can be entered in ***Time***. By checking ***Background*** the students can specify the process visual mode (execution in a separate window or behind the scenes). Pressing the ***RUN*** button multiple times will cause the start of several concurrent processes.

If detailed monitoring of running processes is required, a temporary "freeze" can be performed via ***Pause / Resume*** at the main menu.

2.3. TOST System Options

Some basic OS parameters can be set and modified while several processes are executed in a concurrent mode.

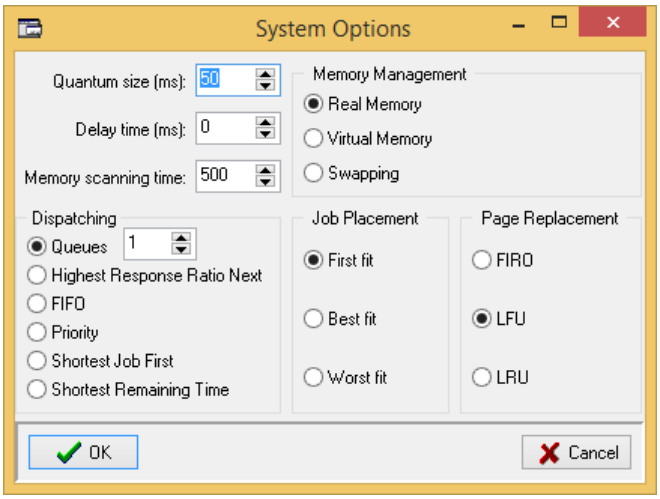


Figure 8: System Options dialog box

In System Options dialog box (Figure 8) students can modify:

- ***Quantum Size***: system time sharing quantum size in ms
- ***Delay time***: processes speed control
- ***Memory scanning time***: time interval of RAM scanning
- ***Dispatching***: current scheduling strategy
- ***Memory Management***: current memory management strategy – *Real Memory*, *Virtual Paging* or *Swapping*
- ***Job Placement***: jobs placement strategy in real memory mode
- ***Page Replacement***: page replacement strategy in virtual memory mode

These modifications of system parameters allow real-time monitoring of how each of them affects the processes executed.

2.4. TOST System Watch

For achieving an OS inside view, while several processes are executed in parallel mode, students can monitor the information dynamics in selected system tables (Figure 9): ***Processes***, ***Ready***, ***Blocked***, ***Memory Allocation*** map and ***Semaphores*** info. The information from each of the system tables is presented in a separate window, as it is shown in Figure 10.

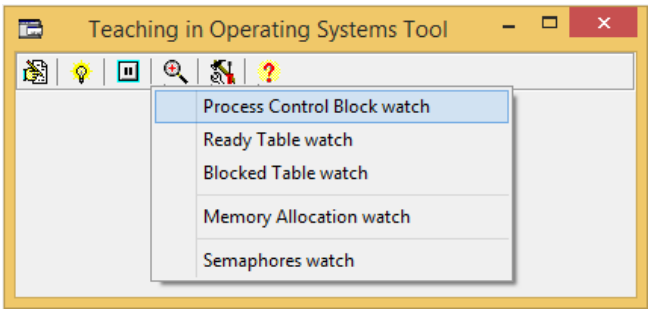


Figure 9: System tables Watch

2.4.1. Processes

The ***Processes*** table (Figure 10) contains the current ***Process Control Block*** of each of the existing concurrent processes.

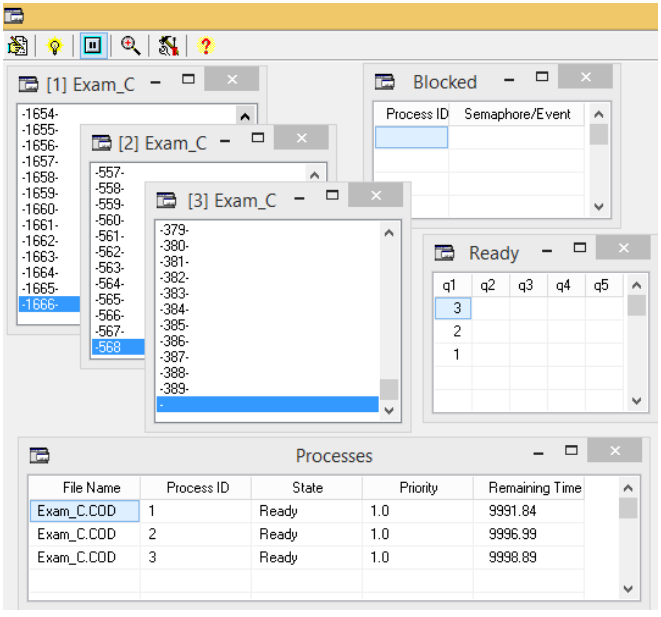


Figure 10: Processes execution and system tables monitoring

The information is presented in five fields:

- ***File Name***: the name of the compiled object file
- ***Process ID***: an integer unique identifier of the process
- ***State***: current process status – *Ready*, *Running* or *Blocked*
- ***Priority***: a real value determining the process priority
- ***Remaining Time***: the time remaining until the process completion.

In this table, students can manipulate the priority or the remaining time of the selected process.

2.4.2. Ready table

The ***Ready*** table contains IDs of all processes with status "Ready". It is possible (according the settings in System Options) to have up to 5 priority queues. Thus, students can experiment with scheduling strategies like ***Round Robin*** (one queue only) or ***Multilevel Feedback Queue*** [18] for the separation of the processes with different behavior (with more I/O operations or with more computations).

2.4.3. Blocked table

The identifiers of blocked processes and the names of semaphores/events, related to their blocking are presented in the ***Blocked*** table. In fact, these are FIFO-queues with blocked processes of all semaphores used. The information in this table is especially helpful when students have to monitor semaphores usage in processes ***synchronization***, ***mutual exclusion*** and ***deadlock*** prevention cases.

2.4.4. Memory Allocation map

Important information about memory allocation and referencing is given by ***Memory Allocation Watch***. When this submenu item is selected, the system periodically scans the TOST RAM and displays results in a separate window. This feature allows students to monitor a map of the memory blocks, where the occupied blocks are marked with the process' ID. The detailed referencing of the pages in the virtual memory is also displayed.

2.4.5. Semaphores information

Semaphores Watch submenu item can be selected for monitoring all semaphore variables' current values in a separate window. This helps students to master the mechanism of using both binary and counting semaphores.

3. Teaching Operating System Basics

3.1. Teaching Process Management

A process [2] is basically a program in execution. During its existence, the process goes through several different states, and the basic three are ***Ready***, ***Running***, and ***Blocked***.

Once students run several processes, they can monitor (as shown in Figure 10) their life cycles in the *Processes* (ID, the current State and the state changing), *Ready* (ID list of all ready processes), and *Blocked* (ID list of all blocked processes, and the event each process is waiting for) tables.

3.2. Teaching Process Scheduling

The act of choosing which of all the ready processes should be moved to the running state is not a trivial task [19]. It is known as Process Scheduling and is performed by OS Dispatcher.

There are several goals of the process scheduling system:

- to maximize the fairness according to the processes priorities
- to keep the CPU busy at all time
- to maximize throughput and to deliver minimum response time for all processes
- to minimize resource starvation
- to support a lot of interactive users
- to minimize system overheads (OS CPU time, OS RAM).

In fact, very often these goals contradict with each other, so the Dispatcher should implement an appropriate compromise [20]. Through experimentation, students are expected to gain understanding in the scheduling strategies and to be able to choose the most appropriate one, depending on the user's needs and objectives. In TOST students can change (at any time) the current scheduling strategy using ***Dispatching*** radio-buttons in ***System Options*** (see Figure 8).

Scheduling disciplines of the two general categories are supported in TOST:

Non-Preemptive Scheduling: once the OS assigns the CPU to a process, the process does not release the CPU until it has finished:

- ***FIFO:*** Simplest scheduling algorithm with only one queue (see queue q1 in *Ready*). New processes are appended in q1 where context switches only occur upon running process' termination. Then the top placed process in q1 becomes *Running*. Students can monitor:
  - scheduling overheads are minimal
  - throughput can be low
  - no starvation
  - waiting time and response time can be high
  - no prioritization occurs
  - interactive users are not allowed
- ***Priority:*** Whenever a scheduling event occurs, the *Processes* table will be searched for the process with the maximum priority number (the most important one), which will be the next one to be scheduled for execution. Students can monitor:
  - scheduling overheads are low
  - throughput can be low
  - starvation can occur in a busy system with many high-priority processes in the Ready table

- waiting time and response time can be high
- interactive users are not allowed

- ***Shortest Job First (SJF)***: Dispatcher selects the process with the least estimated processing time to be executed. The estimated processing time in TOST can be set while process is initialized (Figure 7). Students can monitor how the algorithm increases throughput, and how the starvation can become a problem, in a system with many short processes in the *Ready* table.
- ***Highest Response Ratio Next (HRRN):*** is a dynamic priority discipline proposed by Brinch Hansen to mitigate the problem of the large processes starvation. When CPU becomes free, Dispatcher calculates the priorities of all ready processes according to:

  Priority = 1 + waiting time / processing time

  and selects the high priority process to be executed. Students can monitor the priority calculations (in *Processes* table) which gives preference to short processes, but with each subsequent recalculation the priority of the large ones grows and they can be selected as well.

Preemptive Scheduling: Dispatcher can interrupt a running process any time in middle of the execution and the process is returned back in the *Ready* table:

- ***Shortest Remaining Time (SRT):*** is a modification of SJF with a preemption. When a new shortest process arrives Dispatcher interrupts the current process and assigns the CPU to the new one. Students can monitor that this discipline achieves maximum throughput in most cases, and the starvation is possible for large processes where a lot of small processes arrive.
- ***Queues:*** By specifying the number of queues (Figure 8), students can set two strategies:
  Queues (1): Round-robin (RR): is a FIFO modification with the process preempting where *Ready* table is a queue. Dispatcher assigns a small time interval (Quantum) per process, then interrupts and appends it at the end of the *Ready* queue. When students set *Delay time* in *System Settings* they can monitor processes preempting in the *Ready* table in detail. When students run processes with different priorities, they can monitor how the more important processes are with proportionally larger quantum and get a better service, and the starvation can never occur.
  Queues (2-5): Multilevel feedback queues: In this case, processes are grouped according to their behavior - interactive processes vs batch processes, and processes with more I/O vs. computational processes. In order to obtain a balance in system's resources utilization Dispatcher has to provide better service for interactive and I/O processes. Students can monitor in detail how processes pass into the multilevel set of priority queues (from 1 to 5) in *Ready* table.

3.3. Teaching Memory Management

In general, when discussing the topic of memory management [2], [19] the focus is on the virtual memory organization. However, before reaching this point, students need to understand in detail what the issues are in using the real memory organization. Only then they can comprehend why this kind of organization has been replaced everywhere by the virtual one.

3.3.1. Real Memory

Real memory organization is a classical (one of the oldest) memory allocation model according to which the process is placed in a consecutive memory blocks [21]. Each process takes up exactly as much memory as it needs, and due to the fact that the memory blocks have sequential addresses, process loading, executing and the memory releasing are extremely fast. Then *"Where is the problem?"* comes to be the next instructor's question towards the students.

Students are instructed to develop and try to run several large processes. Later, they run five small processes and with using the TOST feature ***Memory Allocation Watch,*** monitor the memory map with real memory organization. Then students kill the first, second and fourth process, so these memory blocks are freed. Figure 11 shows how the memory blocks occupied by a process are marked with its identifier.

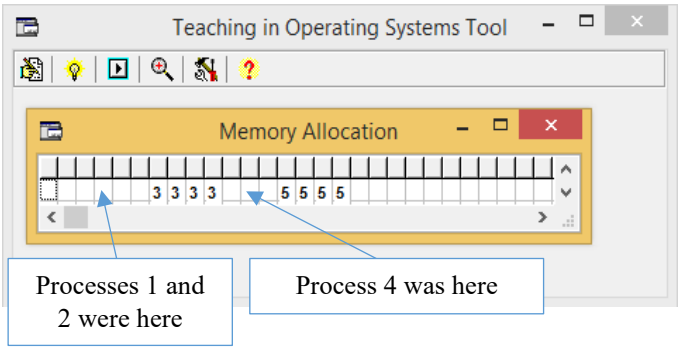


Figure 11: Memory allocation map

Following these instructions, students quickly realize the deficiencies of the real memory:

- A process cannot be started if it is even one byte larger than the available physical memory.
- There is a low CPU utilization, due to the low degree of multiprogramming (number of ready processes loaded in physical memory).
- The occurrence of memory fragmentation: In high dynamics of started and finished processes, the free memory area is broken into small parts, in which none of the ready processes can be allocated. Thus, a lot of memory blocks can remain unusable.

In order to minimize the problem of memory fragmentation, the placement strategies ***First-fit***, ***Best-fit*** and ***Worst-fit*** can be applied (Figure 8). To assess their usefulness students can run a new process and monitor its placement in the memory.

3.3.2. Virtual Memory

Although the problem of memory fragmentation can be solved, other serious problems of real memory remain: small maximum process size and low degree of multiprogramming.

Virtual memory organization allows each process to be divided into blocks (***pages***) while only a small part of the code and data (that are currently needed) are located in the physical memory. The physical memory is divided into blocks, which are called ***frames***. Figure 12 illustrates a moment in the execution of a large process where ***Virtual Memory*** as *Memory management* is set (Figure 8).

At that stage, students can monitor the occupation and references to the blocks (pages) in progress, with the memory scanned at a certain interval (0.5s for example). When the process starts, only a small part of its blocks are loaded - those that are needed at the beginning. Later, the OS loads additional blocks into the memory, those which are needed by the process but are not available. The blocks that the process refers to during the scan interval are colored red.

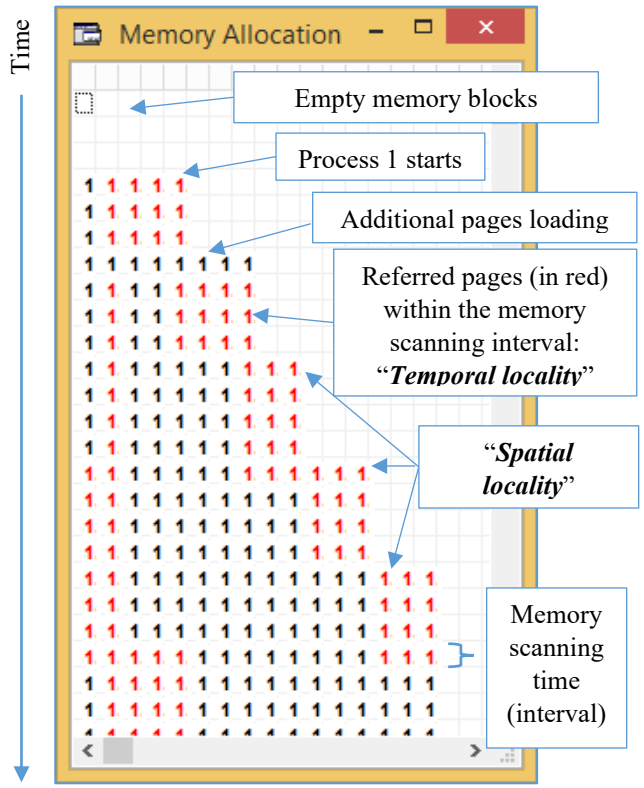


Figure 12: Memory allocation and pages referencing

Memory allocation map allows the instructor to clarify the ***principles of the locality*** [22] - ***Temporal locality*** and ***Spatial locality***.

Principles of the locality can be used for anticipatory paging (swap prefetch) to increase processes' execution speed. These principles are useful when the optimal replacement strategy has to be chosen.

In TOST students can set (Figure 8) and examine various replacement strategies such as *Least Recently Used* (***LRU***), *Least Frequently Used* (***LFU***) and ***FIFO***.

4. Teaching Concurrent Programming Concepts

Even if the students are familiar with the theoretical aspects of parallel programming, in order to gain intuition about the subject, it is vital to be provided with examples of incorrect operation of competitive processes. Therefore, our approach is to firstly develop concurrent programs without any protections and then, analyze the faulty actions of the processes. Consequently, modify the programs in order to achieve correct results. In addition, to strengthen this knowledge, we solve and program some of the well-known toy-problems in concurrent processing.

Representative examples of using TOST in teaching ***Mutual Exclusion***, ***Interprocess Synchronization*** and ***Deadlock*** are considered in the next sections.

4.1. Teaching Mutual Exclusion

The parts of the concurrent processes where they access a shared resource are called ***Critical Sections*** (CS) [23]. Critical Sections must be protected by synchronization primitives, assuring only one of the processes can be in the CS.

Figure 13 presents programs with no protection of the critical section. Students develop and execute an initialization process (init.txt), and two concurrent processes (p1.txt and p2.txt), incrementing the shared variable C 1000 times. After finishing the processes, the students can see the final result value of C = 1839, instead of the correct value of 2000.

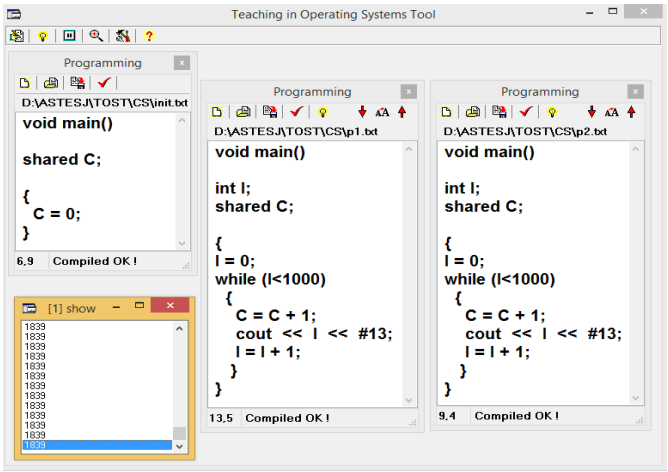


Figure 13: Concurrent processes with Critical Section

After examining the reasons for the wrong result, the students transform the programs gradually, seeking the correct solution for synchronization primitives. Thus, they examine program versions with one, two, and three shared variables, and at last, they come to the consideration of the ***Dekker's algorithm*** (Figure 14).

The ***Peterson's algorithm*** [24] is another well-known algorithm that students can try and compare with Dekker's algorithm. A simple solution of CS protection by the Peterson's algorithm is given in Figure 15.

In Figure 16 a simple example of CS protection, using primitives ***Lock*** and ***Unlock***, based on the hardware instruction ***Test_And_Set***, is presented.

The last example considered is the use of semaphores to protect the critical section (Figure 17).

```
void main()
shared C, ReqP1, ReqP2, Turn ;
{
 C = 0;
 ReqP1 = false;
 ReqP1 = false;
 Turn = 1;
}
```

```
void main()
int I;
shared C, ReqP1, ReqP2, Turn ;
{
I = 0;
while ( I < 1000 )
 {
  ReqP1 = true;
  while ( ReqP2 == true )
   If  (Turn  == 2)
    {
     ReqP1 = false;
       while (Turn  == 2) ;
     ReqP1 = true;
    };
        C = C + 1;
  ReqP1 = false;
  Turn  = 2;
  I = I + 1;
 }
}
```

(a) Initialization process (b) Process 1

Figure 14: Critical Section protection by Dekker's algorithm

```
void main()
shared C, ReqP1, ReqP2, Turn;
{
 C = 0;
 ReqP1 = false;
 ReqP1 = false;
}
```

```
void main()
int I;
shared C, ReqP1, ReqP2, Turn ;
{
I = 0;
while ( I < 1000 )
 {
  ReqP1 = true;
  Turn  = 2;
  while (ReqP2 and (Turn == 2)) ;
        C = C + 1;
  ReqP1 = false;
  I = I + 1;
 }
}
```

(a) Initialization process (b) Process 1

Figure 15: Critical Section protection by Peterson's algorithm

```
void main()
shared C, CS;
{
 C = 0;
 CS = false;
}
```

```
void main()
int I;
shared C, CS ;
{
I = 0;
while ( I < 1000 )
 {
  Lock (CS);
        C = C + 1;
  Unlock (CS);
  I = I + 1;
 }
}
```

(a) Initialization process (b) Process 1

Figure 16: Critical Section protection by Test_And_Set instruction

```
void main()
semaphore  CS;
shared  C;
{
 C = 0;
 init (CS, 1);
}
```

```
void main()
semaphore  CS;
shared  C;
int  I;
{
 I = 0;
 while (I<1000)
  {
   wait (CS);
     C = C + 1;
   signal (CS);
   I = I + 1;
  }
}
```

Figure 17: Critical Section protection by a semaphore

Figure 18 presents the contents of two main OS tables (Processes and Blocked) during the concurrent execution of three identical processes with the source code of Figure 17. Processes *P1_S* and *P2_S* are blocked, while the ready process *P2_3* is inside the critical section. Students also have the possibility of monitoring the semaphore variable CS in a separate window

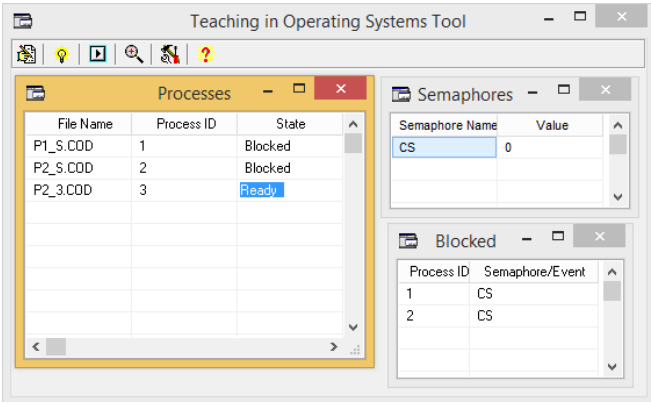


Figure 18: Monitoring TOST system tables

After finishing the processes, the value of the shared variable C is correct and equal to 3000.

4.2. Teaching Interprocess Synchronization

A good illustration of the Interprocess Synchronization is the. ***Sleeping Barber problem*** [25] depicted in Figure 19. As soon as the Barber or Client process reaches a certain point in the execution, it must stop and wait for an event to occur.



Figure 19: Sleeping Barber problem

The problem can be stated as follows: In a barbershop, there is a barber chair and an area for the waiting clients with five chairs. The algorithms of the concurrent processes of the Barber and the Client include the following steps:

Barber:

1) *The barber signals that he is ready to work (free).*
2) *The barber sits on the barber's chair and sleeps, waiting for a new client (indoor bell).*
3) *The barber gets up and waits for the client to sit on the barber's chair.*
4) *The barber begins the haircut and signals when he completes the work.*
5) *The barber waits until the client leaves the barbershop (outdoor bell).*
6) *Go to 1.*

```
void main()
semaphore Chairs, Barber,
          Client, Sit, Finish,
          Out;
{
   init ( Chairs, 5 );
   init ( Barber, 0 );
   init ( Client, 0 );
   init ( Sit, 0 );
   init ( Finish, 0 );
   init ( Out, 0 );
}
```

There are 5 chairs available
The barber is not available
No new client
No one is sitting in barber's chair
The barber has not finished work
The client has not come out

(a) Initialization

```
void main()
semaphore Chairs, Barber,
          Client, Sit, Finish,
          Out;
{
   while (true)
  {
    signal ( Barber );
    wait ( Client );
    wait ( Sit );
    signal (Finish );
    wait ( Out );
  }
}
```

(b) Barber

```
void main()
semaphore Chairs, Barber,
          Client, Sit, Finish,
          Out;
{
   wait ( Chairs );
   signal ( Client );
   wait ( Barber );
   signal ( Chairs );
   signal ( Sit );
   wait ( Finish );
   signal (Out );
}
```

(c) Client

Figure 20: Sleeping Barber solution

Client:

1) *The client checks if there is a free chair in the waiting room and eventually waits until a chair is available.*
2) *The client signals that there is a new client (input bell).*
3) *The client is waiting for the barber to be free.*
4) *The client frees the chair in the waiting room.*
5) *The client sits on the barber's chair and signals that he is ready for a haircut.*
6) *The client is waiting for the barber to complete the haircut.*
7) *The client gets out of the barber's chair and leaves the barbershop (output bell).*

The first thing students have to do is to identify processes synchronization events and implement one semaphore variable for each of the controlled events. Then they develop three programs - Initialization (for setting semaphore values), Barber, and Client (Figure 20).

The implementation of the synchronization of one barber's process and eight clients' processes are depicted in Figure 21.

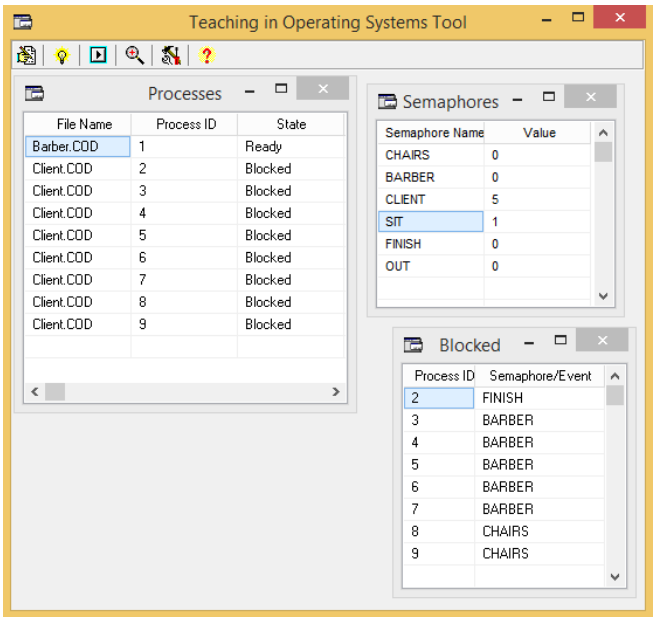


Figure 21: Sleeping Barber execution monitoring

According to the information in TOST system tables:

- the barber is busy (Semaphores: *Barber = 0*)
- a client is seated in the barber chair (Semaphores: *Sit = 1*) and waiting for the barber to complete the haircut (Blocked: *Finish*)
- five new clients are in the waiting room (Semaphores: *Clients = 5*) and all the chairs are occupied (Semaphores: *Chairs = 0*)
- all five new clients are waiting for the barber (Blocked: *Barber*)
- two clients are waiting for a chair outside of the barbershop (Blocked: *Chairs*).

4.3. Teaching Deadlock

The ***Dining Philosophers*** toy-example [26], depicted in Figure 22, is often used to analyze the synchronization and deadlock problems in the concurrent processes execution, as well as to demonstrate the approaches for solving them.

The students solve this problem by developing five programs (for the philosophers) while the correct access (mutual exclusion) to the shared objects (forks) is implemented by semaphores. The example programs for the first two philosophers are presented in Figure 23.

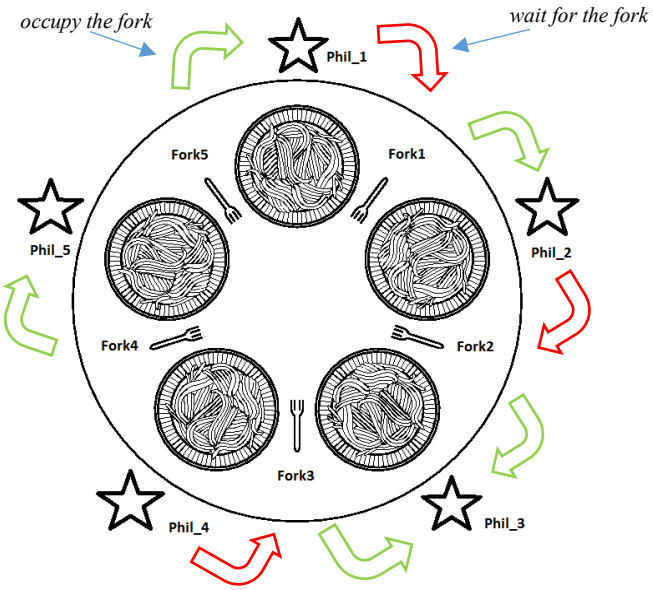


Figure 22: Dining Philosophers problem

```
void main()
semaphore FORK5, FORK1 ;
{
while ( true )
 {
    cout << "T" << #13;
   wait ( FORK5 );
   wait ( FORK1 );
    cout << "E" << #13;
   signal ( FORK1 );
   signal ( FORK5 );
 }
}
```

```
void main()
semaphore FORK1, FORK2 ;
{
while ( true )
 {
    cout << "T" << #13;
   wait ( FORK1 );
   wait ( FORK2 );
    cout << "E" << #13;
   signal ( FORK1 );
   signal ( FORK2 );
 }
}
```

Figure 23: Phil_1 and Phil_2 program samples

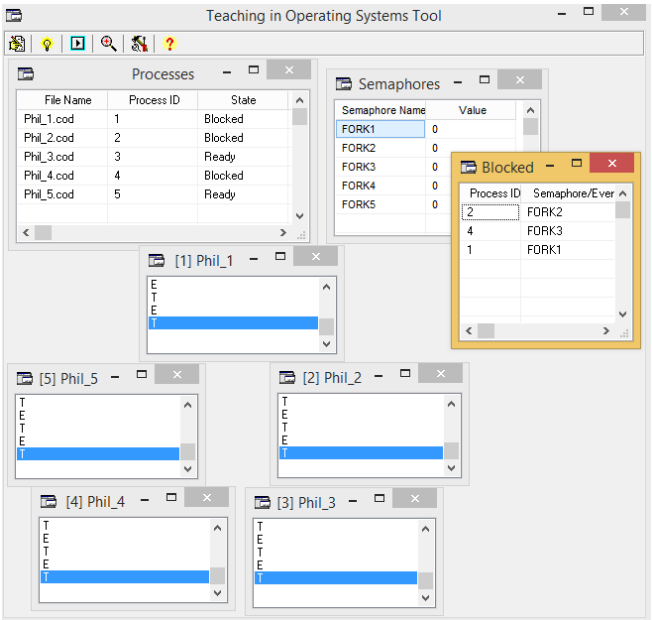


Figure 24: Dining Philosophers execution

TOST allows students to monitor (Figure 24) how philosophers are waiting for the forks (in *Processes* and *Blocked*) and how they take and put the forks on the table (in *Semaphores*).

Another aim of the Dining Philosophers problem is to demonstrate the prevention of ***Deadlock***.

Students can monitor how in the case of all philosophers taking a fork with their right hand, while expecting the release of a second fork, a *Deadlock* occurs (Figure 25 and Figure 26).

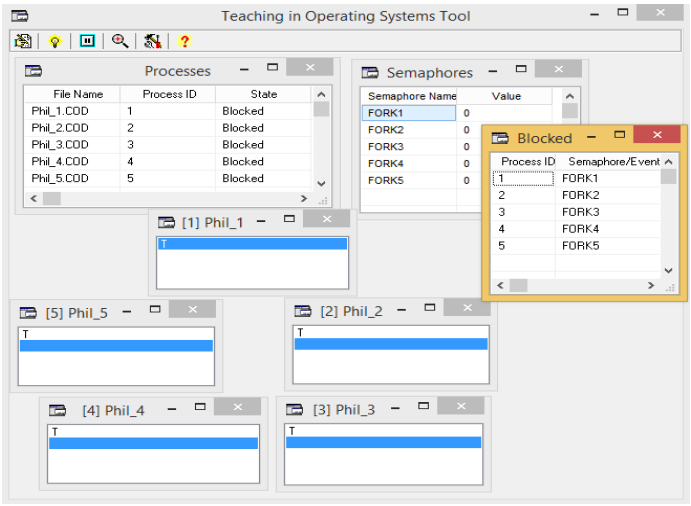


Figure 25: Dining Philosophers Deadlock example

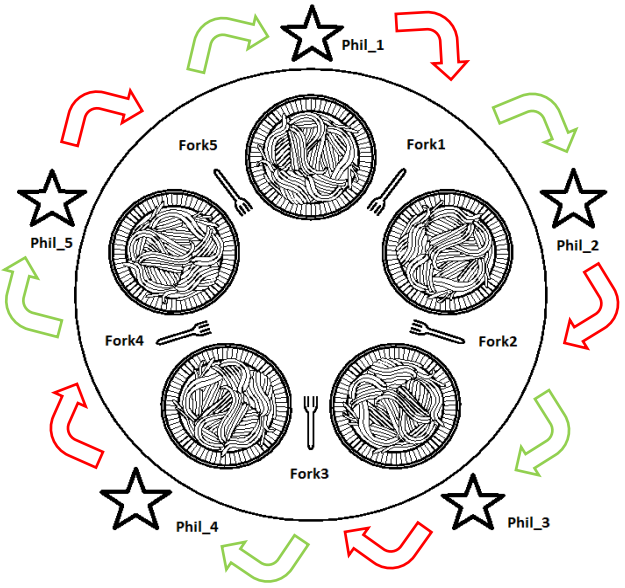


Figure 26: Dining Philosophers Deadlock example (Circular Wait)

Students have to find a solution and save philosophers from starvation dead. One idea is to break the fourth necessary condition for Deadlock - ***Circular Wait*** [27]. Students apply this approach by reprogramming Phil_1 (Figure 27) and introducing the general requirement for all processes to request resources in **ascending order only**.

```
void main()
semaphore FORK5, FORK1 ;
{
while ( true )
  {
      cout << "T" << #13;
     wait ( FORK5 );
     wait ( FORK1 );
       cout << "E" << #13;
     signal ( FORK1 );
     signal ( FORK5 );
   }
}
```

```
void main()
semaphore FORK5, FORK1 ;
{
while ( true )
  {
      cout << "T" << #13;
     wait ( FORK1 );
     wait ( FORK5 );
       cout << "E" << #13;
     signal ( FORK1 );
     signal ( FORK5 );
   }
}
```

Figure 27: Phil_1 program – **before** and **after** changes

5. Students' Opinion

Trying to improve the quality and usefulness of the integrated environment TOST, moreover to get better results from its usage, we also rely on end-users' opinions. For this reason, in a span of two consecutive years, we conduct qualitative surveys with students at the end of the course. The survey used a 5-point Likert scale and is based on a relatively small sample of 131 bachelor students. The summary data for user acceptance of the presented tool are in Table 1.

Table 1: Opinion Survey Results

Survey questions	**A**	**B**	**C**	**D**	**E**
I like TOST environment	2	13	30	55	31
TOST is easy to learn	0	3	15	40	73
I think that programming in TOST is complex	75	29	11	12	4
TOST is visually very attractive	7	22	26	61	15
Overall, I am satisfied with the TOST approach to studying Operating systems	5	11	15	41	59
Overall, I am satisfied with the performance of the TOST environment	9	9	12	21	80
The installation process of TOST is quick and easy	0	0	3	5	123
TOST often freeze, crash, or does not behave as expected	106	9	11	4	1
Overall, I am satisfied with the experimenting with TOST example solutions in understanding concepts related to Mutual Exclusion, Synchronization, Deadlock, and Operating system in general	4	2	3	111	11
I think the interface of the TOST environment is user-friendly	8	5	39	57	22
TOST is successful and effective in stimulating Operating system algorithms and strategies	0	0	0	95	36

A: *strongly disagree*
B: *disagree*
C: *undecided*
D: *agree*
E: *strongly agree*

Based on the data presented, it can be assumed that the responses are positive. According to the majority, the TOST environment is easy to learn and use, user-friendly, useful, and supportive for their understanding of the fundamental concepts of operating systems and the problems concerning concurrent programming.

6. Conclusions and Further Research

Learning the concept and basic principles of Operating systems poses a great challenge to Computer Systems and Technologies students. In this paper, the authors present an integrated environment named TOST, specially designed to be used in Operating systems and Concurrent programming courses at Ruse University. A description of the TOST environment and a demonstration of how it can be very effectively used in teaching Operating systems principles are discussed, as well as some of the TOST features that address the learning of Mutual Exclusion, Interprocess Synchronization, and Deadlock.

The TOST includes a time-sharing operating system with compilers for two simple and easy to use programming languages. There are shared variables and semaphores variables for concurrent algorithms development. The ability to monitor the real-time information from basic OS tables allow students to gain a better apprehending of the algorithms, concepts, and theories behind the design, construction, and insights of operating systems.

The presented environment might be extended in the following directions:

- developing a built-in file system with a set of file allocation methods and management strategies;
- developing a spooling system for mediation between a process and slow peripheral devices;
- implementing complex data types and subroutines for the programming languages integrated.

We have been using TOST for several years in the Department of Computer Systems and Technologies at Ruse University and the student's feedback and results are positive. The lack of any installation procedure, the small size, and the 'copyleft' license allow TOST to be obtained and used without limitation and with full inclusivity.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This work/paper is supported by project 20-FEEA-01 "Methods and tools for multimedia content analysis and automated document and big data processing", funded by the Research Fund of the "Angel Kanchev" University of Ruse.

References

[1] T. Golemanov, E. Golemanova, "Using TOST in Teaching Mutual Exclusion, Synchronization, and Deadlock," in 29th Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE) 2019, Ruse, Bulgaria, 2019, doi:10.1109/EAEEIE46886.2019.9000441.
[2] A.S. Tanenbaum, A.S. Woodhull, Operating Systems Design and Implementation (Second ed.), 2011.
[3] W. Christopher, and T.A. S. Procter, "The Nachos instructional operating

system," in 1993 Winter USENIX Conference, San Diego, California, 1993, doi:10.5555/1267303.1267307.
[4] D. Comer, Operating System Design: The XINU Approach, Prentice-Hall, 1984.
[5] B. Pfaff, G. Romano A. and Back, "The Pintos Instructional Operating System Kernel," in SIGCSE 2009, Chattanooga, TN, 2009, doi:10.1145/1539024.1509023.
[6] D. Hovemeyer, J.K. Hollingsworth, B. Bhattacharjee, "Running on the bare metal with GeekOS," ACM SIGCSE Bulletin, **36**(1), 2004, doi:10.1145/1028174.
[7] L.P. Maia, F.B. Machado, P.J.A. C., "A Constructivist Framework for Operating Systems Education: a Pedagogic Proposal Using the SOsim," in ITiCSE'05, Monte de Caparica, Portugal, 2005, doi:10.1145/1067445.1067505.
[8] T.W. Chan, "A Software Tool in Java for Teaching CPU Scheduling," Journal of Computing Sciences in Colleges, **19**(4), 257–263, 2004, doi:10.5555/1050231.1050269.
[9] D. Skrien, "CPU Sim 3.1: A tool for simulating computer architectures for computer organization classes," Journal on Educational Resources in Computing (JERIC), **1**(4), 2001, doi:10.1145/514144.514731.
[10] D. Park, R.H. Saavedra, "Trojan: A high-performance simulator for shared memory architectures," in Proceedings of the 29th Annual Simulation Symposium. IEEE, 1996, doi:10.5555/829528.831210.
[11] A.R. Lopes, D.A. de Souza, J.R.. de Carvalho, W.O. de Silva, S.V.L. P., "SIME: Memory simulator for the teaching of operating systems," in Computers in Education (SIIE), 2012 International Symposium on. IEEE, 2012.
[12] P. Gayet, B. Bradu, "Procos: A real-time process simulator coupled to the control system," in Proceedings of ICALEPCS, Kobe, Japan: 794–796, 2009.
[13] A. Garmpis, "Design and development of a web-based interactive software tool for teaching operating systems," Journal of Information Technology Education, **10**(10), 1–17, 2011, doi:10.28945/1357.
[14] B. Mustafa, "Visualizing the modern operating system: simulation experiments supporting enhanced learning," in Proceedings of the 2011 Conference on Information Technology Education. ACM, 209–214, 2011, doi:10.1145/2047594.2047650.
[15] T. Golemanov, E. Golemanova, "A Teaching in Operating Systems Tool," in CompSysTech'06, 2006.
[16] E.W. Dijkstra, "Cooperating sequential processes (EWD-123)," Center for American History, University of Texas at Austin, 1965, doi:10.5555/1102034.
[17] M. Herlihy, "Wait-free synchronization," ACM Trans. Program. Lang. Syst. **13**(1), 1991, doi:10.1145/114005.102808.
[18] A.C. Arpaci-Dusseau, Remzi H.; Arpaci-Dusseau, Operating Systems: Three Easy Pieces, Arpaci-Dusseau Books, 2014.
[19] A. Silberschatz, P.B. Galvin, G. Gagne, Operating System Concepts. 9., John Wiley & Sons, Inc., 2013.
[20] P. Krzyzanowski, Process Scheduling: Who gets to run next?, 2015.
[21] D. Samanta, Classic Data Structures, Prentice Hall India Pvt, 2004.
[22] S. William., Computer organization and architecture : designing for performance (8th ed.), Prentice Hall, Upper Saddle River, NJ, 2010.
[23] M. Raynal, Concurrent Programming: Algorithms, Principles, and Foundations, Springer Science & Business Media, 2012.
[24] G.L. Peterson, "Myths About the Mutual Exclusion Problem," Information Processing Letters, **12(3)**, 115–116, 1981, doi:10.1016/0020-0190(81)90106-X.
[25] H.M. Deitel, P.J. Deitel, D.R. Choffnes, Operating systems, Pearson/Prentice Hall, 2004.
[26] C.A.R. Hoare, Communicating Sequential Processes, Prentice Hall International, 2004.
[27] J. Coffman, Edward G., M.J. Elphick, A. Shoshani, "System Deadlocks," ACM Computing Surveys, 1971, doi:10.1145/356586.356588.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 100-107 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Personality Measurement Design for Ontology Based Platform using Social Media Text

Andry Alamsyah*, Sri Widiyanesti, Rizqy Dwi Putra, Puspita Kencana Sari

School of Economics and Business, Telkom University, Bandung 40257, Indonesia

ARTICLE INFO

Article history:
Received: 03 March, 2020
Accepted: 22 April, 2020
Online: 03 May, 2020

Keywords:
Big Five Personality
Ontology Model
Personality Measurement
Personality Platform
Social Media
Radix Tree

ABSTRACT

Human behavior quantification is an essential part of psychological science. One of the cases is measuring human personality. Social media provide rich text, which can be beneficial as a data source to get valuable insight. Previous researches show that social media offered favorable circumstances for psychological researchers by tracking, analyzing, and predicting human character. In this research, we propose a personality measurement design to help to assess human character through linguistic usage from human digital traces. We construct our model by classifying social media text to the pre-determined personality facet from Big Five personality traits, mapping the knowledge to the ontology model, and implementing the model as a platform dictionary. Our model is based on the Indonesian language, which to the best of our knowledge is the first in the subject area. The platform is running effectively by using a well-established sorting algorithm, called the radix tree. Our objective is to support psychological science in adapting to a new technological era.

1. Introduction

The presence of advanced technologies, such as social media platforms and mobile devices, are shifting the way on how people communicate. Social media users have the freedom to upload daily routine information, exchanging messages, or even basic conversations [1]. The content created by the users is formed into digital traces. The personality of users revealed through their writing or textual content [2]. Each personality has its own charm that deserves attention over its complex arrangement. There are thoughts, hearts, and feelings that can change over time [3]. Hence, measuring human personality is a hard problem regarding the dynamic feeling's alteration.

Personality measurement has become the most extensive research in the field of Psychology [4]. The legacy methodology of personality assessment performed by interview and written examination [5]. Those methods are integrated method, where an interview is conducted to validate the test result. The characteristic of legacy methodology requires fulfillment instruments, such as the physical tools and psychologists. In order to adapt to the new technological era, some alternative methodologies are developed for bringing a faster process and result. One of them is themeasurement through its natural environment or using digital traces on social media [6]. In this research, we utilize digital traces to put forward an automation personality measurement for minimizing fulfill instruments in the legacy methodology. Thus, it would significantly reduce the cost and reduce the time process of getting the results. This approach needs to be developed, considering that automation is the most demanded characteristic of the Industry 4.0 era.

According to Madden et al. [7], digital traces, which consist of user information, e.g., personal information, shared texts, pictures, and videos, are a proof dataset that cannot be ignored, which expressed online human activity. These footprints are offered valuable opportunities for psychology research in understanding human characteristics [8]. Previous researches have assessed human personality through social media, such as Bhardwaj et al. [9], who assess personality through Facebook and LinkedIn and Park et al. [6] who applying the regression model to predict human character based on social media language.

Most of the research generally uses a machine learning approach, in contrast to our study, which utilizing the ontology approach. Machine learning provides us some leverage, such as the speed of analysis regarding large-scale data [10]. It is also able to predict personality on various forms of data like text, speech, and image [11]. However, the machine learning approach has some weaknesses in processing each meaning and intention of words due to language uniqueness [12]. Ontology afford us a better understanding of contextual knowledge [13]. There is an opportunity to use ontology as a basis for measuring human personality through the words on linguistic usage. Thus, in terms of measuring human nature through social media textual data, the

*Andry Alamsyah, Email : andrya@telkomuniversity.ac.id

ontology model provides a more accurate result than a machine learning approach as long as the collection word type in the ontology model's corpus is prosperous. The following approach scheme is shown in Fig. 1.

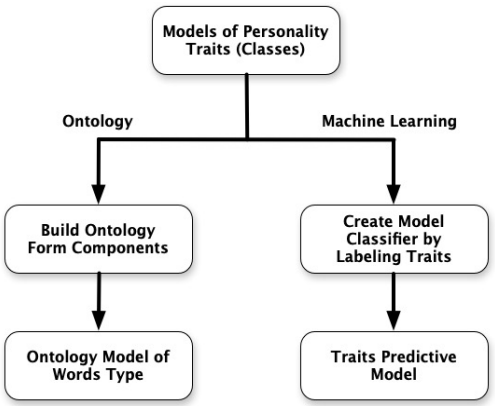


Figure 1: Personality Traits Model Scheme between Ontology and Machine Learning Approach

Fig. 1 shows the difference between ontology and machine learning model schemes. On the ontology model scheme, the personality trait model established by mapping words that reflect personality as a model's forming component. Those collections of words then verified by experts and utilized as a dictionary for measuring personality. This model is a representation of an expert's knowledge and adequate to map more than one personality trait in the complex sentence. Meanwhile, the personality trait model on the machine learning model scheme is a depiction of a machine's algorithm according to the labeling process. The machine interprets the label on the training dataset as a rule for predicting personality. Therefore, the approach of the machine learning model scheme is entirely different than ontology. The ontology model scheme is only mapping the words with reference to an expert's knowledge, while machine learning is predicting based on the learning result of the machine.

To the best of our knowledge, research of personality measurement has put forward in several personality traits taxonomies based on specified purposes, such as for assessing job placement and natural human emotion. In this research, we want to use the most general taxonomy of human's essential persona. The Big Five Personality theory gets a consensus of a general taxonomy regarding human personality traits [14]. It also able to represent and simplify the diverse characteristics of a human's personality [15]. The Big Five Personality Traits is built by examining several unique human attributes in their linguistic usage. This theory also offers prized terms called Revised Neuroticism-Extraversion-Openness Personality Inventory (NEO-PI-R) metric, which facilitate the exposure psychological characteristic of broad trait [16]. In this research, we use the Big Five Personality Traits Theory with the NEO-PI-R metric as a domain knowledge of our ontology model. The personality divided into five domains and further divided to thirty facet scales. Our research also applies the lexical hypothesis as a basis for analyzing textual content in social media. Based on research of Raad and Mlacic [17], the lexical hypothesis is a process of understanding the meaning of textual data. The lexical hypothesis is qualified to map the personality traits through words in a language [18].

In terms of model development, the needs of an open-source platform are essential. The advantage of creating the platform are model crowdsourcing, public corpus enrichment, correction, and verification. Hence, the platform is significantly enhancing the model's value over time. Nowadays, a platform with an ontology-based model for measuring human personality is still rare. In favor of getting a better model for measuring personality, we implement the model into a platform. This research aim is to show the way on how we develop a design in mapping human character by utilizing the ontology model. The model constructed by a collection of words that refer to personality in Bahasa. Our study mapped 2,331 instances in different facets and traits — those instances used as the corpus in our platform.

2. Literature Review

This chapter provides theoretical foundations related to the personality measurement proposed model based on the ontology approach. Literature review is sorted according to the data flow from social media data collection, personality definition, the personality traits, the ontology as the knowledge representation literature, and at last, is the radix tree algorithm used to parse and sort social media texts.

2.1. User-Generated Content

According to Naab and Sehl [19], user-generated content (UGC) has three criteria, there are 1). UGC is characterized by a rate of personal contribution, 2). UGC must be disclosed, and 3). UGC is built outside the sphere of occupation and professional routines. Besides, Wyrwoll stated that user-generated content is a content that is published online via various platforms by its user [20]. The users are not only the person but also the organizations. The substance classified as a UGC has similar characteristics, such as publicly accessible to other users, need a creative effort to create, and not a result of expertise routines and practices. UGC also defined in many forms, like blogs, posts, chat, podcasts, images, videos, tweets, and many other ways. In this modern era, UGC is a substantial information source for discovering knowledge from human digital activities [21]. Existing studies reveal that there is a high correlation between personality and personal inclination [22]. UGC indirectly shows beneficial information, like the user's demand, lifestyle, and personality. Hence, we utilize UGC in social media to measure human personality.

2.2. Personality

Personality is defined as a set of a person's characteristics, including acting, thinking, and feeling. Personality also correlates with emotions, values, attitudes, and talents [23, 24]. These attributes are establishing a unique persona of an individual and differentiate one person from another [25]. According to Stemmler, a person's personality closely related to language usage in speaking or writing [2]. Language is the most prevalent and reliable tool for people to convey their internal thoughts and emotions in giving comprehension to others [26, 27]. Hence, the linguistic usage of a person is an essential subject in the field of psychology and communication.

Previous research has examined human personality through linguistic usage. Howlader et al. predict Facebook user's personality through status and linguistic features by applying

regression models [28]. Boyd and Pennebaker propose a complementary model that provides big data solutions to measure human personality based on words people use [29]. Pietro et al. and Bogolyubova et al. [30, 31] who discovered the dark triad personality of social media users their online communication. Flekova and Gurevych [32] predict the personality of fictional characters in the novel using lexical-semantic features. Another research is Wei et al. [33], who predict human personality via information in digital traces and conversation logs.

2.3. Big Five personality Traits

Big Five Personality Traits is a model that identifies five characteristics of personality. It is also recognized as the OCEAN model, which stands for Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN) [34, 35]. This model is related to the lexical hypothesis and stated that language in daily interactions is a reflection of the most personality characteristics [36]. Thus, the Big Five Personality Traits predict and describe essential personality differences. In favor of getting a clearer explanation, Rossberger in [37] describes five traits below:

Table 1: Big Five Personality Traits

Big Five Personality Traits	Definition
Openness	or usually called openness to experience: extent to which individuals exhibit intellectual curiosity, self-awareness, and individualism/ nonconformance.
Conscientiousness	extent to which individuals value planning, acquire the tenacity quality, and achievement oriented.
Extraversion	extent to which individuals involved with the external world, encounter enthusiasm and other positive emotions.
Agreeableness	extent to which individuals' value mutual effort and social harmony, modesty, dignity, and trustworthiness.
Neuroticism	extent to which individuals deal with negative feelings and their propensity to emotionally overreact.

According to Costa and McCrae's study [16], NEO-PI-R was advanced in samples of middle-age and older adults. The NEO-PI-R included scales to measure six conceptually derived facets in each OCEAN. The scales show generous internal consistency, temporal cohesion, convergent and discriminant validity against partners and peer ratings. Table. 2 shows the six facets which explaining each of the factors in personality traits.

2.4. Ontology and Knowledge Representation

Ontology is a collection of concepts which able to model terms of vocabulary into a domain knowledge [38]. From the perspective of computational science, ontology is explained as a concept to model the system structure. For example, the relevant entities and relationships that exist from observations are useful for specific purposes [39].

Table 2: NEO-PI-R Metric

Personality Traits	Facet	Description
Openness	Aesthetics	Artistic interests; believe in the value of art/do not love poetry
	Fantasy	Imagination; have an expressive imagination/seldom daydream
	Actions	Adventurousness; prefer variety /dislike changes
	Ideas	Intellect; admire convoluted problems/avert discussion
	Feelings	Emotionality; encounter emotions intensely/rarely get emotional
	Values	Liberalism; tend to elect for liberals/believe in one true divinity
Conscientiousness	Competence	Self-efficacy; accomplish tasks successfully/misunderstand the situation
	Order	Orderliness; prefer to order/abandon a mess
	Dutifulness	Follow the rules/break the rules
	Achievement-Striving	Work hard/just do the task
	Self-Discipline	Get task done right away/waste the time
	Deliberation	Cautiousness; avoid error carefully/charge into things directly
Extraversion	Warmth	Friendliness; make friends effortlessly/hard to make friends
	Gregariousness	Love large parties/prefer to be alone
	Assertiveness	Take action/wait for others to get the way
	Activity-Level	Regularly busy/love to take it easy
	Excitement-seeking	Love excitement/dislike loud music
	Positive Emotions	Cheerfulness; radiate joy/am seldom amused
Agreeableness	Trust	Trust others/distrust people
	Compliance	Morality; never deceive/use false praise
	Altruism	Make people feel pleasant/despise on others

Personality Traits	Facet	Description
	Straightforward-ness	Cooperation; easy to fascinate/have a sharp tongue
	Modesty	Dislike being center of attention/think highly of myself
	Tender-mindedness	Sympathy; sympathize with the destitute/believe in people
Neuroticism	Anxiety	Worry about things/relaxed most of the time
	Angry-Hostility	Get enraged easily/rarely get annoyed
	Depression	Often feel miserable/feel satisfying
	Self-Consciousness	Easily frightened/not being intimidated easily
	Impulsiveness	Immoderation; often be affected by other/easily resist temptations
	Vulnerability	Panic easily/remain calm under pressure

Ontology associated with discovering and modeling reality under particular perspectives [40]. It focused on the structure and nature of an object [41]. Ontology also pointed out to a representational knowledge which indicates the type of class or entity associated with the relationship of the subtype [42].

Ontology generally has fundamental form components, i.e., class, instance, and relation [43]. Class is referring to a set of multiple instances, like words and phrases. An instance is a scope that considered in the ontology domain knowledge. A relation is defined as a relation among classes or instances [44]. This method is flexible, easy to modify, understood by humans and machines, and able to integrate with machine learning. Ontology has at least four evaluation methods [45]: 1). The Golden Standard; 2). The Application-based; 3). The Human Assessment; 4). The Data-Driven. Our study needs validation and evaluation for each instance in the model before implementing it on the platform.

Recent research has shown that ontology is able to represent knowledge for measuring personality. Some of the studies are that applying linguistic feature analysis and ontology model to measure human nature through social media data [46]. Another research computes human personality derivation in the modern physiognomy domain [47]. [48] also shows that ontology is a representative way to measure contextual knowledge, by building a complex domain of music and see how it relates to the domain of personality. A few research examples were clarifying how ontology might contribute to the analysis, comprehension, and research about human behavior and psychological research. Most of these existed researches are mainly focused in the English language, in contrast to our research which concerns in Bahasa.

2.5. Radix Tree

The dictionary which consists of strings data type is very-time consuming. A sorting algorithm is needed in order to get a more efficient process. Radix Tree is one of several algorithms that able to sort data in a database. Radix tree also beneficial for constructing associative arrays that expressed through the keyword in the form of strings. According to Mauro, radix tree is worked by labeling edges with a sequence of strings rather than characters, and constricting chains of nodes into a single element. Hence, radix tree is running more efficiently compared to a regular tree [49]. The instance of radix tree displayed in Fig. 2.

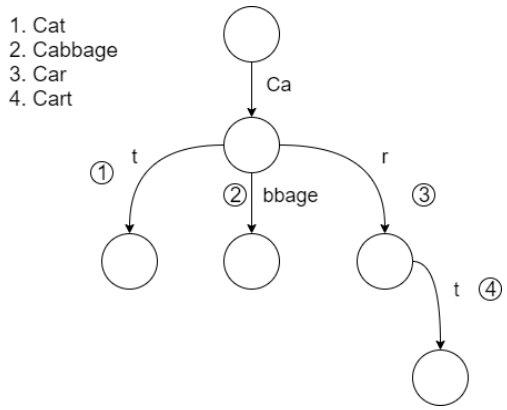


Figure 2: The Instance of Radix Tree Algorithm

3. Methodology

Our idea is to collect social media posts of people who considered influential in society. Their posts most likely to be responded by the public, which will generate conversations, and often set standard for informal language in Bahasa. In the model construction, we map words and phrases to a certain personality, which will validate by experts. There are several steps to construct the proposed model. Those steps are data collection, data preparation, model construction, model validation, platform construction, and conclusion. The research workflow is shown in Fig. 3. In contemplation of earning comprehensive understanding, we also illustrate the conception of our research methodology in Fig. 4.

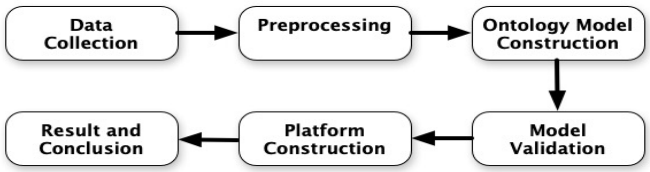


Figure 3: The Proposed Model Workflow

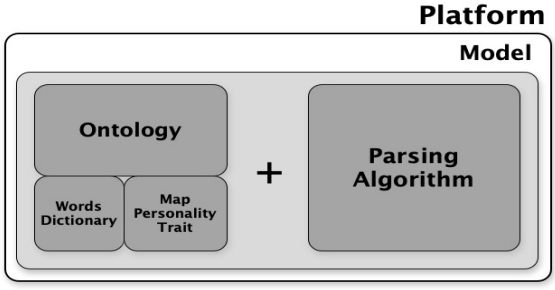


Figure 4: The Conception of Research Methodology

3.1. Data Collection

Our research utilizes real-world conversations on *Twitter* social media. A recent study shows that *Twitter* provides a valuable chance to study human behavior in a natural environment [1]. In the data retrieval process, we use three samples i.e., the famous user in *Twitter* social media with specific criteria. Our samples criteria are:

1. Verified accounts or having tweets with more than 1000 tweets or 500,000 followers.
2. Shows the latest activity with different tweets.
3. Shows many interactions with other accounts.
4. Not a protected account.

Based on those criteria, we have collected 13,047 tweets from three selected users. We comprehend that despite having the same principles in data collection, this collection of famous people who have different character tendencies, and this is the reason why we conduct this research.

3.2. Preprocessing

The data preprocessing is required to clear irrelevant data, such as URLs, symbols, and other terms, which is not beneficial for this research [50]. This process objective is to get substantial information over the data [51]. According to Khadim, preprocessing is a process of improving data quality while reducing barriers that will occur in the classification process [52]. The preprocessing is divided into several steps shown in Table. 3.

Table 3: Preprocessing Steps

Steps	Name of Preprocessing Steps	Definition
1	Case Folding	Transforming all capital letters to lowercase letters.
2	URLs Removal	Eliminating web page links
3	Symbol and Number Removal	Delete the symbol and number in the document
4	Tokenizing	Splitting the sentences into a token
5	Phrases Lookup	Check if there are any phrases in the token
6	Synonym Recognition	Determine the synonym of a word and replace it based on a dictionary
7.	Word Generalization	Replacing a word into a more general word in order to reduce data redundancy

3.3. Ontology Model Construction

This model is established by classifying words in each tweet into thirty classes in the NEO-PI-R metrics and afterward generalize it into Big Five Personality traits. Those collections of words are defined as a corpus or dictionary of our model. In this study, we state the personality class in Big Five Personality Theory as a class, facet in NEO-PI-R metric as a sub-class, and each composed word into an instance. For better understanding, we show our ontology domain model hierarchy in Fig. 5.

The hierarchy of our ontology model is settled in the bottom-up paradigm since this model starting from specific to general or instances to class [53]. There are no properties needed in personality measurement ontology model. As an instance, from Fig. 5., the words *terluka*, *melelahkan*, *penderitaan*, which means hurting, exhausting, and suffering are classified into vulnerability in Neuroticism class. These words arranged to vulnerability facets because it indicates the feelings of flimsy due to the risk of harm from some experiences.

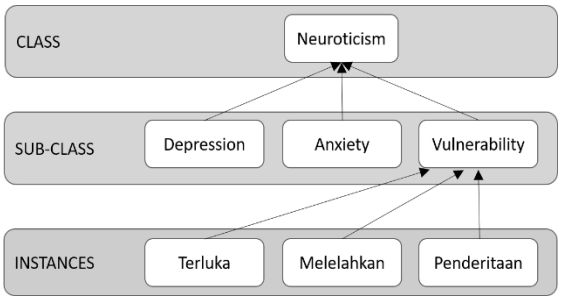


Figure 5: The Hierarchy of Personality Measurement Ontology Model

3.4. Model Validation

The validation process is essential to prevent and minimize any bias from the previous classification process. The human assessment approach is an assessment method with the help of domain experts [45]. For this reason, we assign psychologists to validate our classification result before deployment. In this process, the expert ensures that the classified word has entered the correct personality trait according to the Big Five theory and NEO-PI-R metric.

We conduct a human assessment because a human can represent language in terms of circumstances. The words in Bahasa ordinarily have various meanings depending on contextual purposes. For example, the words *bisa* can be represented to “able” or “can” or “poison from the snake.” Hence, we need the psychologist to check and validate the ambiguity of the words that frequently appeared in people's linguistic usage, especially in social media.

3.5. Platform Construction

Our platform is designed as an open-source platform to encourage the public or crowds to get involved in model development. People can openly provide some enrichment, correction, and verification of the model. Web-based platform is suitable for our research purposes due to its characteristics compared to other forms such as a mobile app or desktop app. The features of this platform are to measure personality based on the input. The common format of the collection of short text sentences and large-scale text is in comma-separated value (csv); thus, we use this .csv format file as one of our input files. At last, we visualize the result in a radar chart as the output. The platform’s workflow displayed in Fig. 6.

In this research, we utilize Personality Measurement Ontology (PMO) Platform which has been built in our prior research [51]. The platform is running by employing a radix tree algorithm as an effort to sort and parse the sentence, then find the semantic similarities between the input text and the corpus. The pseudocode presented in Fig. 7. The trait calculation is conducted

by finding similar keywords between processed textual data and the existing corpus in the ontology model. If matched, the data then considered representing one or more personality traits depending on the number of keywords that paired.

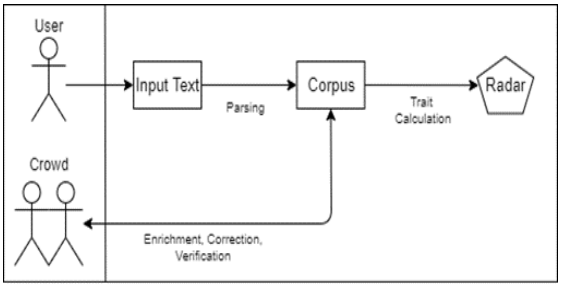


Figure 6: The Platform Workflow

The algorithm of our platform consists of several stages, i.e., convert the model to radix tree, saving database, input processing (sentence and csv format), and trait calculation. The following explanation of each stage are:

1. Algorithm 1: convert the model to radix tree. The established model is sorted using the radix tree algorithm. This algorithm composed of a) csv input file, which a text file contains of collections of keywords, b) a radix tree algorithm named 'tree'; c) an algorithm that is occupied by the data in csv called 'trie'.
2. Algorithm 2: saving database (keywords, traits, and facets). Generate a row for each keyword, facet, and traits into a 'database'.
3. Algorithm 3: input processing (sentence). Parsing sentence by utilizing 'trie'. The result of this process is called 'set_of_words,' which contains detected keywords.
4. Algorithm 4: input processing (csv). Parsing csv data into the database. The result is 'set_of_word' which loaded by detected keywords.
5. Algorithm 5: trait calculation. Calculate ('set_of_word') to get the desired form in the personality measurement results.

```
algorithm 1    convert model to radix tree
input          csv (keywords)
output         radix_tree (keywords)
1              csv = input (csv file)
2              tree = new radix_tree
3              for each row in csv do:
                   tree <- row.keyword
4              trie <- dump(tree)
5              output (trie)
```

```
algorithm 2    saving database (keyword, trait, facet)
input          csv = [keyword, trait, facet]
output         database(keyword, trait, facet)
1              csv = input (csv file)
2              data = [keyword, trait, facet]
3              for each row in csv do:
                   data.keyword = row.keyword
                   data.trait = row.trait
                   data.facet = row.facet
               database <- import(data)
               output(database)
```

```
algorithm 3    input processing(sentence)
input          sentence
output         result(set_of_words)
               array = input(sentence)
               trie <- load(database)
               for each word in array do:
                   if word is in trie
                       set_of_words <- word
               output(set_of_words)
```

```
algorithm 4    input processing(csv)
input          csv(set_of_sentence)
output         result(set_of_words)
               csv = input(csv file)
               load radix_tree
               trie <- load (database)
               for each row in csv do:
                   for each word in row do:
                       if word is in trie:
                           set_of_words <- word
               output(set_of_words)
```

```
algorithm 5    trait calculation
input          set_of_words
output         result(trait composition)
               array = input (set_of_words)
               for each word in array do:
                   trait = trait(where facet =
                   = word.facet)
                   trait =+ 1
               output(trait)
```

Figure 7: The Pseudocode of PMO

To facilitate user's convenience, the platform provides a personality measurement process in two ways: 1) by entering textual data such as opinions, statements, or conversations; 2) by uploading the csv file, which consists of numerous textual data. The interface of our platform is shown in Fig. 8.



Figure 8: The PMO Platform Interface

4. Result and Analysis

In order to examine our platform, we measure the personality of the three samples through the platform. We calculate a set of sample tweets in the form of a csv file. The result visualized into a radar chart for getting a better comparison of each personality trait. The measurement results displayed in Table 4.

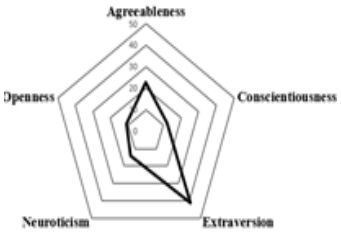
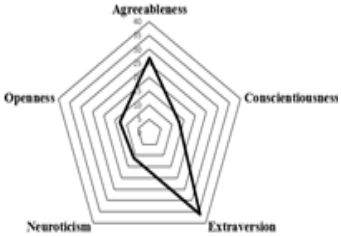
From our measurement result, each sample tends to have one trait as a dominant personality trait. It is a good result because human always has at least one dominant persona over all of the existed character. It also proves that our platform is able to measure all of the traits in the Big Five Personality theory.

As shown in Table 4., our samples have a similar result but not entirely identical. The result also indicates that our samples have a high score in Extraversion personality traits. It means that our samples often show positive feelings, friendliness, and activity-level on their social media activity. For example, *@shitlicious*, who regularly displays his activities, both significant and trivial matters in his daily routine.

The second-highest score of our sample's measurement result is Agreeableness personality trait. This character is representing individuals who have value cooperation and social harmony with another person. *@bepe20* is repeatedly showing his compliance and altruism. As a famous professional football player in Indonesia, *@bepe20* has a tremendous number of fans. He

generally answers the message of his supporters on social media. He also does not hesitate to praise and welcome others who even he does not know. Thus, this persona makes *@bepe20* also has a high score in Agreeableness besides the Extraversion personality trait.

Table 4: The Result of PMO Platform

Account	Result
@jojosuherman	
@bepe20	
@shitlicious	

Even though the platform has proven capacity in measuring personality through linguistic usage, it still has some limitations. Our platform cannot measure complex phrases with different contextual meanings. For example, the sentence *keren gila*, which means he or she expresses a fantastic feeling, is measured into Extraversion and Neuroticism personality traits. The words *keren*, which means impressive, is reflecting Extraversion and *gila*, which means crazy or stupid is representing Neuroticism. That instance shows our platform's limitation, which only running by mapping word by word. Hence, our platform still needs enrichment and development for measuring sentences with phrases that consist of a different word with different personalities.

5. Conclusion

In this research, we explore human's linguistic usage in social media to build an ontology model for measuring human character. We have successfully implemented the ontology model to a web-based platform for measuring personality automatically. In this study, the ontology form component defined by assigning personality as a class, facet as a sub-class, and words that are referring personality as instances. The platform is generally running well but cannot handle the complex phrases.

Our model helps us to measure human personality in social media. This research significantly contributed to a psychology study, especially in Indonesia. Our model is adequate to support a psychologist to speed-up the personality measurement process. It also can be improved for reading complex phrases by generating another algorithm. Thus, our suggestion for future research is to employ another parsing algorithm such as n-gram for receiving better results. We also suggest enriching the platform's corpus by engaging the words that are reflecting personality from different social media. Another recommendation is adding weight to each classified word. Since we only measure the frequency of words, and not considering the context of the tweets, adding weight to the indexed words is required to detect context on a sentence.

References

[1] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," 9th WEBKDD and 1st SNA-KDD Workshop, San Jose, California, USA, 2007.

[2] G. Stemmler, and J. Wacker, "Personality, Emotions and Individual Differences in Physiological Responses," Journal of Biological Psychology, vol. 84, July 2010, pp. 541-551.

[3] S. I. Wahjono, "Perilaku Organisasi," Graha Ilmu, 2010.

[4] L. Li, A. Li, B. Hao, Z. Guan, and T. Zhu, "Predicting Active Users' Personality Based on Micro-Blogging Behaviors," PLoS ONE 9(1): e84997, January 2014.

[5] J. L. Farr and N. T. Tippins, "Handbook of Employee Selection 2nd Edition," Taylor & Francis Group, 2017.

[6] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. P. Seligman, "Automatic Personality Assessment Through Social Media Language," Journal of Personality and Social Psychology, November 2014.

[7] M. Madden, S, S. Fox, A. Smith, and J. Vitak, "Digital Footprints", Internet: https://www.pewresearch.org/internet/2007/12/16/digital-footprints/, 2007 [October 18, 2019].

[8] W. Bleidorn, C. J. Hopwood, A. G. C. Wright, "Using Big Data to Advance Personality Theory," Current Opinion in Behavioral Sciences, vol. 18, December 2017, pp. 79-82.

[9] S. Bhardwaj, P. K. Atrey, M. K. Saini, A. E. Saddik, "Personality Assessment using Multiple Online Social Networks," Multimedia Tools and Applications, vol: 75November 2016, pp. 13237-13269.

[10] B. Y. Pratama and R. Sarno, "Personality Classification Based on Twitter Text Using Naive Bayes, KNN Aand SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), November 2015.

[11] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. F. Moens, and M. D. Cock, "Computational Personality Recognition in Social Media," User Modeling and User-Adapted Interaction, vol:26, June 2016, pp. 109-142.

[12] B. Liu, L. Yao, and D. Han, "Harnessing Ontology and Machine Learning for RSO Classification," Springer, 2016.

[13] N.F. Noy, and D.L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems Laboratory, Technical Report KSL-01-05, 2001.

[14] O. P. John, R. Robins, and L. Pervin, "Handbook of Personality Theory and Research," New York: The Guilford Press, 2008.

[15] P. T. Costa and R. R. McCrae, "Normal Personality Assessment in Clinical Practice: The NEO Personality Inventory," Psychological Assessment, 4(1): 5-13, March 1992.

[16] P.T. Costa Jr and R.R. McCrae, "Domains And Facets: Hierarchical Personality Assessments using The Revised Neo Personality Inventory," Journal of Personality Assessment, vol. 64, February 1995, pp. 21-50.

[17] B. D. Raad and B. Mlacic, "The Lexical Foundation of The Big Five Factor Model.," in Oxford handbook of the Five Factor Model, US: Oxford University Press, 2017, pp. 191-216.

[18] B. De Raad, "The Big Five Personality Factors: The Psycholexical Approach to Personality," Hogrefe & Huber Publishers, Göttingen, 2000.

[19] T. K. Naab and A. Sehl, "Studies of User-Generated Content: A Systematic Review," Journalism 18(10), Sagepub, October 2016.

[20] C. Wyrwoll, "Social Media Fundamentals, Models, and Ranking of User-Generated Content," Wiesbaden: Springer Fachmedien Wiesbaden, 2014

[21] M. F. Moens, J. Li, and T. S. Chua, "Mining User Generated Content," Boca Raton: CRC Press, Taylor dan Francis Group, 2014.
[22] R. Gao, B. Hao, S. Bai, L. Li, A. Li, and T. Zhu, "Improving User Profile with Personality Traits Predicted from Social Media Content," Proceedings of the 7th ACM conference on Recommender systems, October 2013.
[23] H. Woo and H. J. Ahn, "Big Five Personality and Different Meanings of Happiness of Consumers," Economics and Sociology, vol 8(3), November 2015, pp.145-154.
[24] L. R. Goldberg, "An Alternative "Description of Personality": The Big-Five Factor Structure," Journal of Personality and Social Psychology, 59(6) December 1990, pp. 1216–1229.
[25] B. B. Lahey, "Psychology: An Introduction (Ed.11)", New York: McGraw-Hill, May 2011.
[26] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,", PLoS ONE 8(9): e73791, September 2013.
[27] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," Journal of Language and Social Psychology, vol 29(1), March 2010, pp. 24-54.
[28] P. Howlader, K. K. Pal, A. Cuzzocrea, and S. M. Kumar, "Predicting Facebook-Users' Personality Based on Status and Linguistic Features via Flexible Regression Analysis Techniques," the 33rd Annual ACM Symposium, April 2018.
[29] R. L. Boyd and J. W. Pennebaker, "Language-based personality: a new approach to personality in a digital world,' Current Opinion in Behavioral Sciences, vol:18, December 2017, pp. 63-68.
[30] D. P. Pietro, J. Carpenter, S. Giorgi, L. H. Ungar, "Studying the Dark Triad of Personality through Twitter Behavior," the 25th ACM International, October 2016.
[31] O. Bogolyubova, P. Panicheva, R. Tikhonov, V. Ivanov, and Y. Ledovaya, "Dark Personalities on Facebook: Harmful Online Behaviors and Language," Computers in Human Behavior, vol: 78, January 2018, pp. 151-159.
[32] L. Flekova and I. Gurevych, "Personality Profiling of Fictional Characters using Sense-Level Links between Lexical Resources," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, January 2015.
[33] H. Wei, F. Zhang, N. J. Yuan, C. Chao, H. Fu, X. Xie, Y. Rui, and W. Y. Ma, "Beyond the Words: Predicting User Personality from Heterogeneous Information," the Tenth ACM International Conference, February 2017.
[34] R. A. Power and M. Pluess, "Heritability Estimates of The Big Five Personality Traits based on Common Genetic Variants," Translational Psychiatry 5(7): e604, July 2015.
[35] M. Ziegler, K. T. Horstmann, and J. Ziegler, "Personality in Situations: Going Beyond the OCEAN and Introducing the Situation Five," Psychological Assessment 31(4), March 2019.
[36] N. Ramdhani, "Adaptasi Bahasa dan Budaya dari Skala Kepribadian Big Five," Jurnal Psikologi, vol: 39, December 2012, pp. 189-207.
[37] I. Ali, "Personality Traits, Individual, Innovativeness and Satisfaction with Life," Journal of Innovative & Knowledge, February 2018.
[38] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A Probabilistic Taxonomy for Text Understanding," Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, May 2012.
[39] S. Staab and R. Studer, "Handbook on Ontologies 2nd Edition," Springer Publishing Company, 2009.
[40] G. A. Silver, J. A. Miller, M. Hybinette, G. Baramidze, and W. S. York, "DeMO: An Ontology for Discrete-event Modeling and Simulation," Simulation, vol. 87(9), September 2011, pp. 747–773.
[41] V. Devedzic, "Understanding Ontological Engineering," Communications of the ACM - Supporting Community and Building Social Capital, vol: 45, April 2002, pp. 136-144.
[42] R. Arp, B. Smith, and A. D. Spear, "Building Ontologies with Basic Formal Ontology". The MIT Press, 2015.
[43] C. Roussey, F. Pinet, M. A. Kang, O. Corcho, "An Introduction to Ontologies and Ontology Engineering," Ontologies in Urban Development Projects, Springer, pp. 9-38.
[44] A. Alamsyah, M. R. D. Putra and D. D. Fadhilah, "Ontology Modelling Approach for Personality Measurement based on Social Media Activity," in International Conference on Information and Communication Technology, Bandung, 2018.
[45] K. Sekandar, A Quality Measure for Automatic Ontology Evaluation and Improvement, Utrecht: Utrecht University, 2018.
[46] D. Sewwandi, K. Perera, S. Sandaruwan, O. Lakchani, A. Nugaliyadde, and S. Thelijjagoda, "Linguistic Features based Personality Recognition Using Social Media Data," 6th NCTM, Malabe, Sri Lanka, January 2017.
[47] A. K. Awaja, "Human Personality Derivation Using Ontology Based Modern Physiognomy," PhD and MSc Theses- Faculty of Information Technology, Islam University of Gaza, 2017.
[48] C. M. Marques, J. V. Zuben, and I. Z. Guilherme, "FTMOntology: An Ontology to Fill the Semantic Gap between Music, Mood, Personality, and Human Physiology," Model-Driven Approach to XML Schema Evolution, October 2011, pp. 27-28.
[49] C. Blochwitz, J. Wolff, J. M. Joseph, S. Werner, D. Heinrich, S. Groppe, and T. Pionteck, "Hardware-Accelerated Radix-Tree Based String Sorting for Big Data Applications," Architecture of Computing Systems - ARCS 2017: 30th International Conference, March 2017, pp. 47-58.
[50] M.D.N. Arusada, N.A.S. Putri, and A. Alamsyah, "Training Data Optimization Strategy for Multiclass Text Classification," 5th International Conference on Information and Communication Technology, 2017.
[51] A. Alamsyah, M. F. Rachman, C. S. Hudaya, and R. R. Putra, "A Progress on the Personality Measurement Model using Ontology based on Social Media Text," 2019 International Conference on Information Management and Technology (ICIMTech), August 2019.
[52] A.I. Khadim, Y. N. Cheah, and N.H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, 2014.
[53] A. Herdiani, N. Selviandro, and M.F.A. Azka, "Pemanfaatan Ontologi dengan Paradigma Pembangunan Combined Hierarchy Dalam Pengukuran Indeks Kebahagiaan Masyarakat Kota Bandung," e-Proceeding of Engineering, vol.3, no.2, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 108-114 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Experimental Studies of the Silicon Photomultiplier Readout Electronics Based on the Array Chip MH2XA030

Oleg Dvornikov[1], Vladimir Tchekhovski[2], Yaroslav Galkin[3], Alexei Kunz[3], Nikolay Prokopenko[4,5,*]

[1]*Department No. 1, JSC "Minsk Research Instrument-Making Institute", MNIPI JSC, Kolasa St., 73, Minsk, 220113, Belarus*

[2]*Laboratory of electronics methods and tools of experiments, Institute for Nuclear Problems of Belarusian State University, Bobruyskaya St, 11, Minsk, 220030, Belarus*

[3]*Micro- and Nanoelectronics Department, Belarusian State University of Informatics and Radioelectronics, Brovka St., 6, Minsk, 220013, Belarus*

[4]*Information System and Radio Engineering Department, Don State Technical University, Gagarin Sq., 1, Rostov-on-Don, 344000, Russia*

[5]*Institute for Design Problems in Microelectronics of Russian Academy of Sciences, IPPM RAS, Sovetskaya St., 3, Zelenograd, 124681, Russia*

ARTICLE INFO

Article history:
Received: 26 March, 2020
Accepted: 22 April, 2020
Online: 03 May, 2020

Keywords:
Silicon Photomultiplier
Readout Electronics
Array Chip
Charge-Sensitive Amplifier
Comparator

ABSTRACT

The experimental findings of the main units of readout electronics of silicon photomultipliers (SiPMs) based on array chip (AC) MH2XA030: a charge-sensitive amplifier (CSA) with an adjustable conversion factor and a base line restorer (BLR) circuit and two types of voltage comparators are considered. The electrical circuits of the units, the measurement results of static and dynamic parameters are given.

1. Introduction

Silicon photomultipliers are successfully used in a number of fields of science and technology for recording various types of electromagnetic radiation [1-10], because they have smaller dimensions, supply voltage and noise [11-13] in comparison with vacuum PM tubes. To process SiPM signals are usually used specialized analog integrated circuits (ICs), optimized according to the parameters of the signal source – internal capacitance, amplitude and duration of the current pulse.

SiPM readout electronics are relatively not required and their development costs too much. This explains the importance of creating specialized ICs on the ACs. So, we developed the main units of SiPM readout electronics at AC MH2XA030:

- the CSA with an adjustable conversion factor and a BLR circuit called ADPreampl3 [11];
- analog interface with a large dynamic range IBUF [14];
- different types of voltage comparators [15].

The purpose of this article is to consider the experimental findings of amplifier ADPreampl3 and comparators ADComp1, ADComp3 made on AC MH2XA030.

2. Circuitry Features of the Analog Units under Study

A detailed description of the operation of amplifier ADPreampl3 and comparators ADComp1, ADComp3 is given in [11,15]. When they were implemented at the AC, minimal corrections were made, caused by a slight change in the resistances of the resistors and the formation of diodes at the p-n junctions of the field effect transistor.

*Corresponding Author: Nikolay Prokopenko, Don State Technical University, Russia | Email: prokopenko@sssu.ru

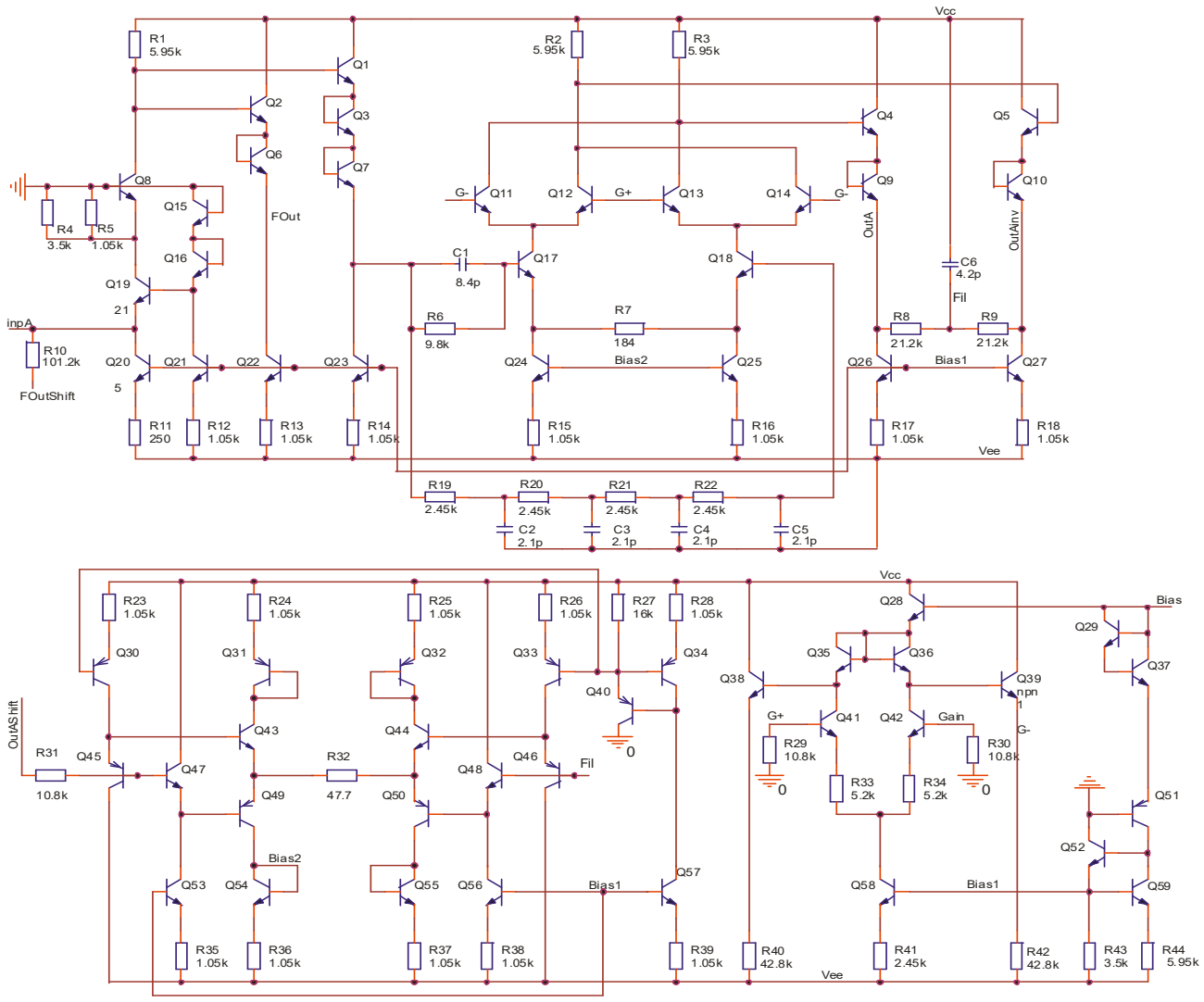


Figure 1: The electrical circuit of amplifier ADPreampl3

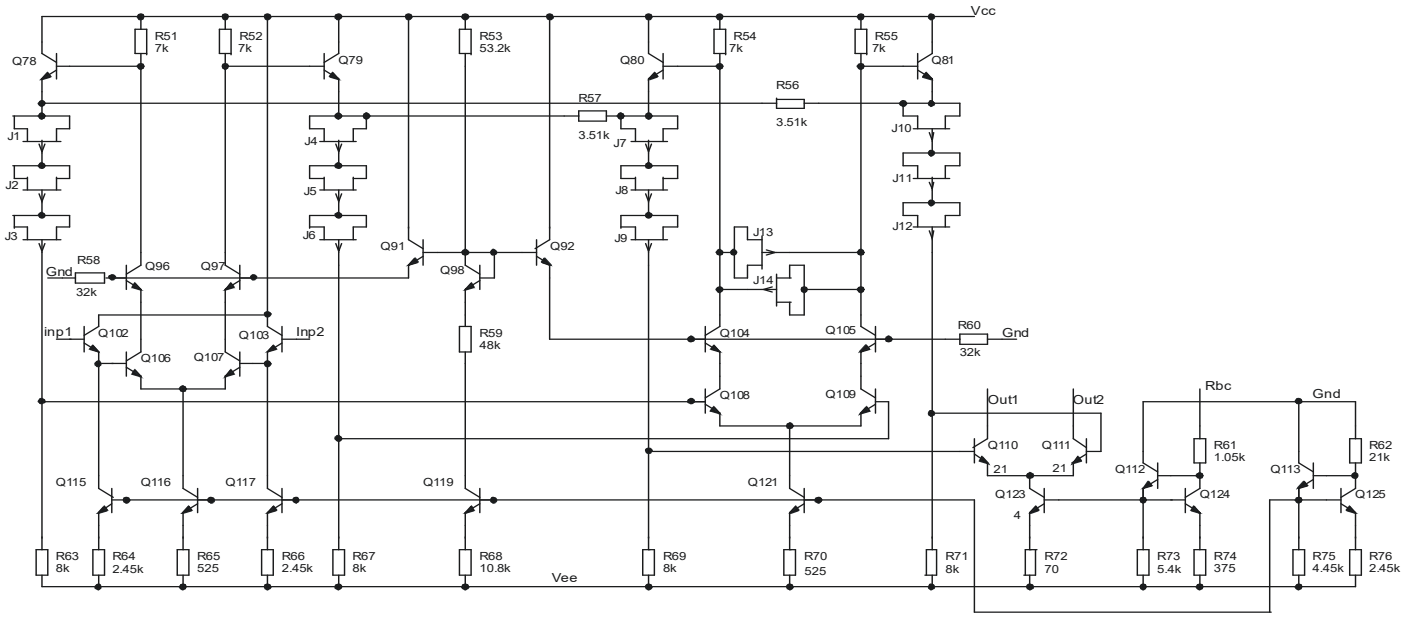


Figure 2: The electrical circuit of comparator ADComp1

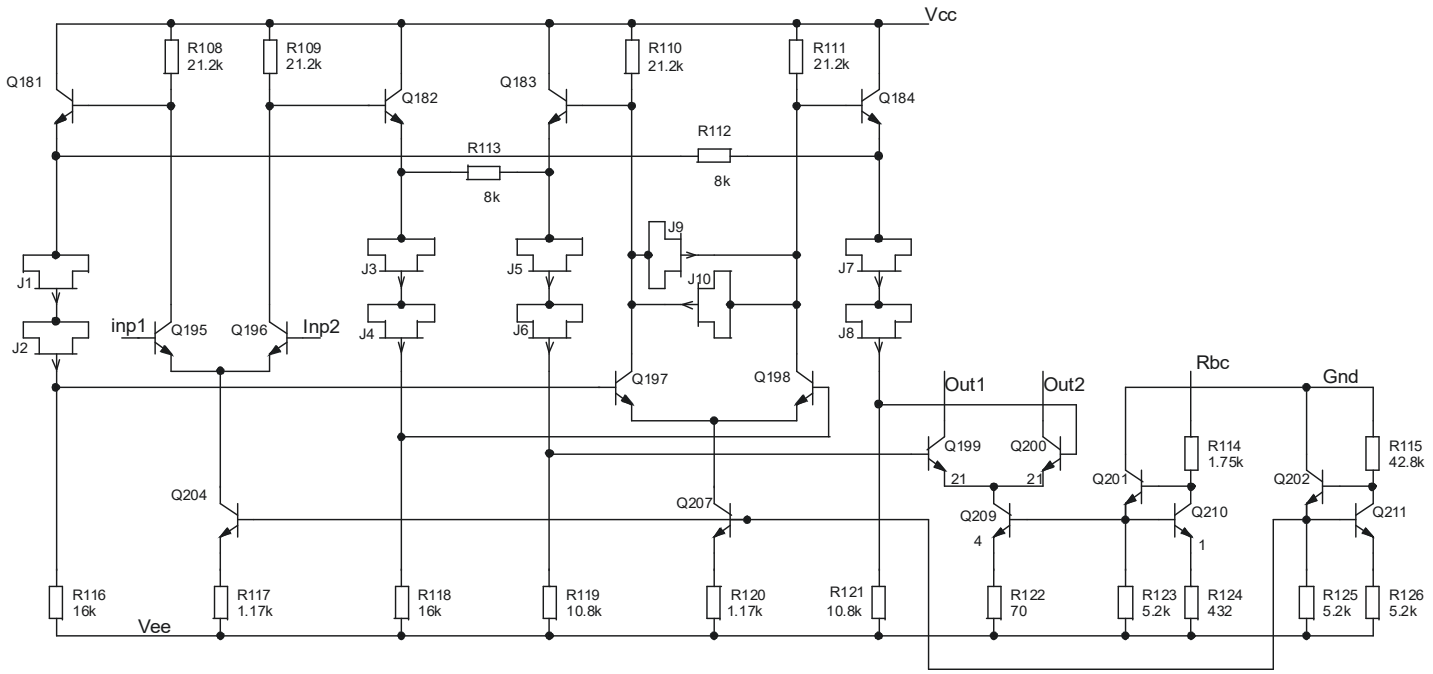


Figure 3: The electrical circuit of comparatorADComp3

Electrical circuits of ADPreampl3, ADComp1, ADComp3 are shown in Figures 1-3. Note that all nodes with the same name, for example, V_{CC}, V_{EE}, G-, G+, Bias1, Bias2, Fil in Figure 1 are interconnected, and, if necessary, the scaling factors are given in the diagrams, showing the number of the AC transistors connected in parallel. So, the emitter junction area of transistor Q19 in Figure 1 corresponds to the area of the emitter junctions of the 21st parallel-connected n-p-n-transistor of the AC.

The values of the resistances R7 = 184 Ohms, R74 = 375 Ohms, R124 = 432 Ohms are shown in Figure 1, Figure 2 and Figure 3 with great accuracy, but these are not necessary conditions. In the circuit, these resistor values are obtained with parallel and serial connection of the 1.05 kOhms and 2.45 kOhms resistors available on the AC.

It is worth noting that there is no need for special use of the p-JFET. These transistors are formed on the AC and are used only in diode connections.

Assignment of the nodes in the circuits of Figures 1-3 are the following.

For ADPreampl3: V_{CC}, V_{EE} – positive and negative supply voltage; InpA – amplifier input; FOut – output of the fast signal circuit (SC), OutA, OutAinv – direct and inverse output of the slow SC, FOutShift – voltage that sets the base line (dc output voltage in the absence of an input signal) by the output of the fast SC, OutAShift – voltage that sets the base line by the outputs of the slow SC, Gain – voltage that sets the value of the charge-voltage conversion factor K_{QV} by the outputs of the slow SC, Bias – node for connecting a current source that sets the operation mode of the amplifier.

For ADComp1, ADComp3: V_{CC}, V_{EE} - positive and negative supply voltage; Gnd - bus of null potential ("ground"); Inp1, Inp2 – inputs; Out1, Out2 – outputs; Rbc – node for connecting a voltage source that sets the maximum output current of the comparator, usually this node is connected to null potential.

Comparator ADComp3 is a simplified version of ADComp1 and is designed to reduce the AC area occupied by the comparator circuit. So, in total 8 units of ADComp1 or 16 units of ADComp3 can be manufactured at AC MH2XA030. On the manufactured experimental samples, comparator ADComp3 was made in two-channel design.

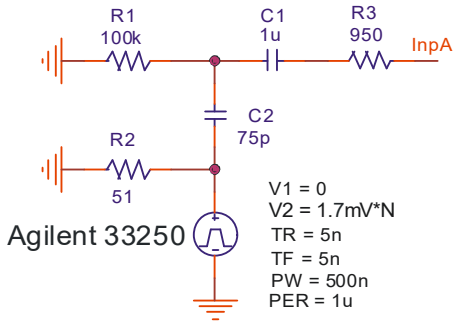


Figure 4: The simplified electric equivalent circuit of SiPM Photonique with 516 microcells, used in measurements

3. Measurement Results

Experimental samples of AC MH2XA030 with analog units were manufactured at JSC "Integral" (http://www.integral.by/). Measurements were performed using a set of equipment, including signal generators Agilent 33250 and Tektronix AFG3152C, a set of fixed attenuators, digital oscilloscope GDS 73354, dual-channel voltage supply source Agilent 3646A, data acquisition system 34970A with a set of modules, and digital multimeter Agilent 34410A. The signal came at the input of ADPreampl3 from the electric equivalent circuit SiPM Photonique with 516 microcells [11] (Figure 4), where TR, TF, PW, PER, V1, V2 are parameters of the rectangular voltage source used, namely, the rise and fall times, pulse duration, period, initial and final voltage values, correspondingly. By setting different voltages of V2, it is possible to continuously change the input charge of the amplifier in

accordance with the relation $Q_{INP}=C2 \cdot V2$, or when the voltage of V2 is a multiple of 1.7 mV, which corresponds to one switched on cell, set the value of input charge N of the switched-on microcells of SiPM Photonique. The pulse duration of the generator is chosen equal to 500 ns in order to register the output pulses of ADPreampl3 of different polarity after the output signal reaches the base line.

In the comparator measurements, the same common-mode signal V_{CM} came simultaneously to both inputs, and a rectangular voltage pulse symmetrical with respect to the common-mode signal with a rise and fall front duration of 3 ns and a peak-to-peak value equal to V_{INP} was additionally received at one of the inputs. Both outputs of the comparators were connected through resistors with a resistance of 50.5 Ohms to the bus of null potential.

In total 20 units of amplifiers ADPreampl3, 10 units of dual-channel comparators ADComp3 and single-channel comparators ADComp1 were measured. Figures 5-14 and the table show the main measurement results, and the figures show the dependencies closest to the average.

Measurements made it possible to establish the following.

1) The spread of the base line by output FOut ranged from minus 24 mV to 276 mV with an average value of 85.6 mV. In this case, a voltage change in the node FOutShift from -3 V to 3 V is sufficient to establish a base line value of FOut output close to zero.

2) When the BLR circuit was switched off (the node OutAShift was not connected to any voltage), the spread of the base line at the output OutA was from 300 mV to 800 mV. When connecting the node OutAShift to the bus of null potential, the average base line value for the output OutA was 3.72 mV, and for the output OutAinv it was minus 2.42 mV.

3) The BLR circuit enables to smoothly change the base line at the outputs OutA, OutAinv in the range of ± 0.9 V.

4) When developing analog units, it was assumed that the output signal of amplifier ADPreampl3 would come to the input of the comparator with a small input capacitance. In connection with the above, the output stages of the amplifier are simple emitter followers on n-p-n-transistors and the voltage pulse shape at the outputs FOut, OutA, OutAinv depends on the input capacitance of the oscilloscope, which leads to different output pulses in the positive and negative half-waves. If necessary, the influence of the load capacitance on the shape of the output pulse can be reduced by using an emitter follower on complementary bipolar transistors, as in operational amplifiers previously developed on AC MH2XA030 [16]. Such an output stage increases the current consumption of the amplifier and therefore its use is carried out only in justified cases. To exclude the influence of the capacitance of the oscilloscope on the shape of the output pulses, a non-inverting voltage follower on chip AD8132 connected to the pins FOut, OutA, OutAinv was used in the measurements.

As can be seen from Figure 5, Figure 6, the inclusion of the follower provides almost the same shapes of the output pulses of both polarities.

5) At the maximum gain required to register the signal of 2 microcells of SiPM Photonique switched-on, the dynamic range of ADPreampl3 exceeds 20 dB, however, in this case, the conversion factor K_{QV} depends on the input charge Q_{INP} (Figure 7, 8).

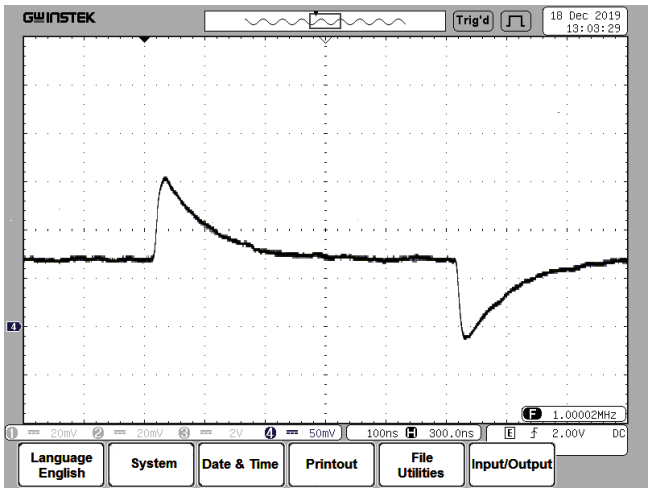
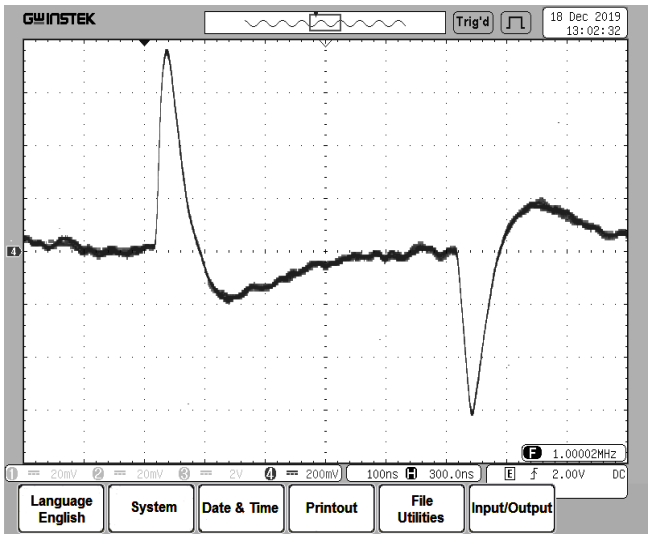
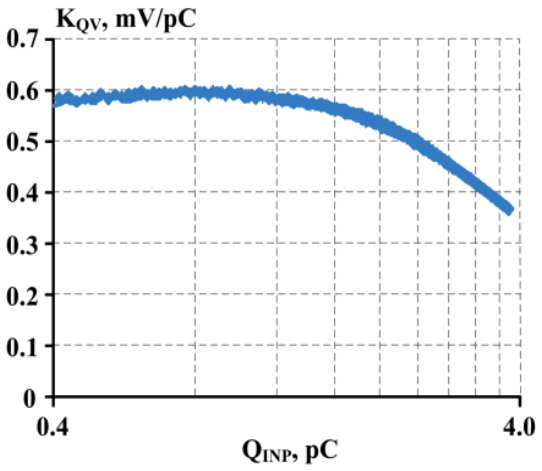


Figure 5: The voltage pulses at the output FOut for 10 switched-on microcells of SiPM Photonique when the voltage follower is used





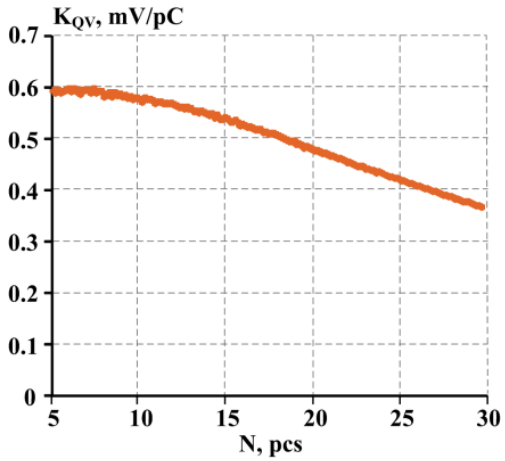


Figure 8: The dependence of the maximum value of the conversion factor K_{QV} at the pin OutA on the value of N switched-on microcells of SiPM Photonique

In the case when it is necessary to register the input charge from more than 30 switched-on microcells of SiPM Photonique, it is recommended to reduce the output pulse by reducing the voltage at the Gain pin (Figure 9) or use the signal from the pin FOut (Figure 10). To significantly increase the dynamic range, it is advisable to use new circuitry solutions, for example, considered in [14].

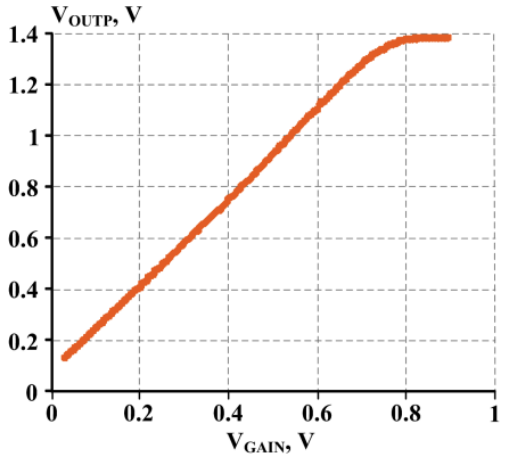
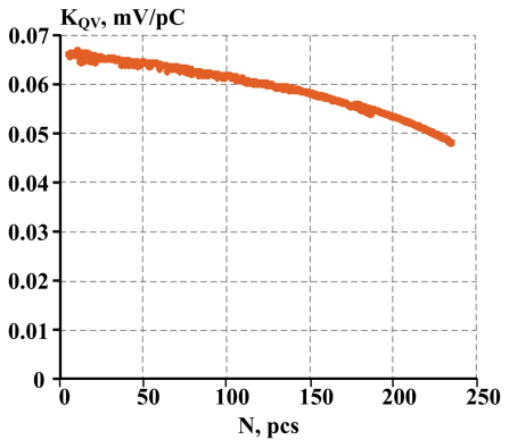


Figure 9: The dependence of the value of the pulse at the output OutA (V_{OUTP}) on the voltage at the pin of the Gain (V_{GAIN})



6) As follows from the table, the measured values of the static parameters of the comparators are close to the simulation results.

Parameter Name	Parameter value			
	ADComp1		ADComp3	
	measurements	simulation	measurements	simulation
Input current I_{INP}, μA	0.7-0.8	0.58	0.5-0.7	0.5
Maximum output current I_{OUT}, mA	3.2-3.4	4.03	2.8-3.0	4.0
Sensitivity, mV	2.0	2.0	2.0	2.0

The output current of the comparators is less according to the measurement results than in the simulation, as shown in Table 1. The dynamic parameters of the comparators characterize the dependencies shown in Figure 11 - Figure 14. It is possible to connect the Rbc node to an external voltage source by adjusting the output current. However, this requires an additional highly stable voltage source. We plan to reduce the resistance of the current-setting resistors with the closest adjustment of the interconnections of the elements of the AC. In this case, the output current of the comparators will increase even when the Rbc node is connected to zero voltage.

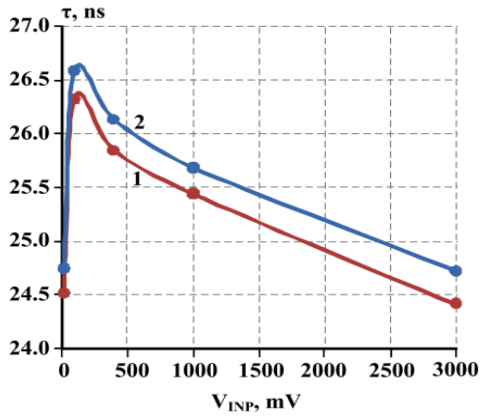


Figure 11: The dependence of the average signal propagation delay in two-channel comparator ADComp3 on the amplitude of the input signal for the first (1) and second (2) channels

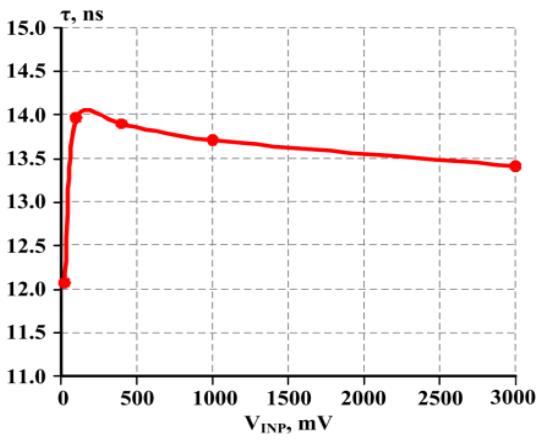


Figure 12: The dependence of the average signal propagation delay in comparator ADComp1 on the amplitude of the input signal

7) The measurements revealed a lower value of the maximum output current of the comparators and a significant excess of the measured values of the peak time of ADPreampl3 and the signal propagation delay of the comparators over the values obtained during the simulation.

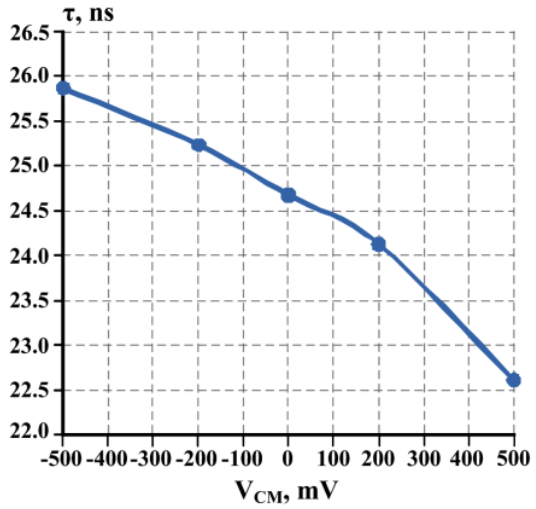


Figure 13: The dependence of the average signal propagation delay in two-channel comparator ADComp3 on the value of the common-mode signal at V_{INP} = 20 mV

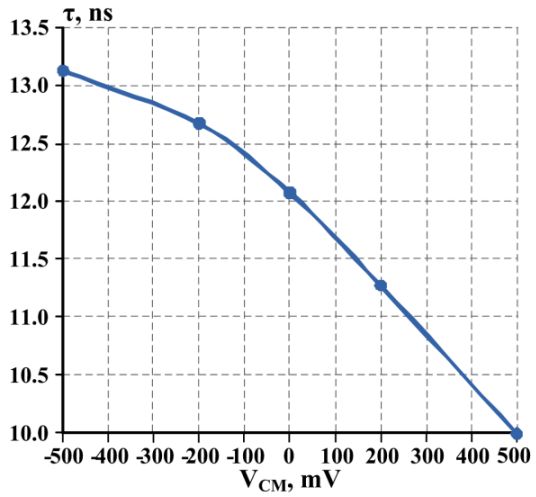


Figure 14: The dependence of the average signal propagation delay in comparator ADComp1 on the value of the common mode signal at V_{INP} = 20 mV

From our point of view, the insufficient value of the maximum output current of the comparators is due to the difference in the current-setting resistors of the output stages (R72 in Figure 2, R122 in Figure 3) from the required value and can be eliminated when adjusting the interconnects of the AC elements.

The Spice parameters CJE, CJC, CJS of the AC MH2XA030 transistors used in the simulation were obtained from the identified parameters for transistors with a large layout area by decreasing in proportion to the area of the emitter junction. For the transistors of AC MH2XA030 made according to the minimum design rules, such scaling of the Spice parameters turned out to be incorrect and identification of the Spice parameters CJE, CJC, CJS by the results of measurements of experimental samples is required.

4. Conclusion

It was experimentally established that the created base line restorer circuit reduces the spread of the base line of amplifier ADPreampl3 by almost 100 times and enables to smoothly change the base line at the outputs OutA, OutAinv in the range of ± 0.9 V.

At maximum gain, the dynamic range of ADPreampl3 exceeds 20 dB, however, in this case, the conversion factor depends on the value of the input charge. To register large input charges, it is recommended to reduce the output pulse by reducing the voltage at the Gain pin or process the signal from the pin FOut.

The static and dynamic parameters of created comparators ADComp1, ADComp3 satisfy the requirements for readout electronics of silicon photomultipliers.

The measurements revealed a mismatch between the results of simulation and measurements of the maximum output current of the comparators, the peak time of the amplifier and the signal propagation delays of the comparators, therefore, in the near future, it is planned to adjust the interconnects of the AC elements to increase the maximum output current and refine the Spice parameters of transistors according to the measurement results of experimental samples.

The developed units can be used to create multichannel signal processing circuits of SiPM based on AC MH2XA030.

Conflict of Interest

The authors declare that there is no conflict of interests regarding publication of this paper.

Acknowledgment

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

References

[1] G. Paternoster, L. Ferrario, F. Acerbi, A. G. Gola and P. Bellutti, “Silicon Photomultipliers Technology at Fondazione Bruno Kessler and 3D Integration Perspectives,” ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC), Cracow, Poland, 2019, pp. 50-53. DOI: 10.1109/ESSDERC.2019.8901738

[2] W. Jiang, Y. Chalich, M. J. Deen, “Sensors for Positron Emission Tomography Applications,” J. Sensors 2019, 19, 5019; DOI: 10.3390/s19225019.

[3] R.R. Raylman, A. Stolin, S. Majewski, and J. Proffitt, “A large area, silicon photomultiplier-based PET detector module,” Nucl. Instrum. and Methods A, vol. 735, pp. 602-609, 2014. DOI: 10.1016/j.nima.2013.10.008

[4] A.L. Goertzen, *et al.*, “Design and performance of a resistor multiplexing readout circuit for a SiPM detector,” IEEE Trans. Nucl. Sci., vol. 60, no 3, part 1, pp. 1541-1549, 2013. DOI: 10.1109/TNS.2013.2251661

[5] S. Dey, E. Myers, T.K. Lewellen, R.S. Miyaoka, and J.C. Rudell, “A row-column summing readout architecture for SiPM based PET imaging systems,” IEEE Nucl. Sci. Symp. and Med. Imag. Conf. (NSS/MIC), pp. 1-5, Oct. 27 2013-Nov. 2 2013. DOI: 10.1109/NSSMIC.2013.6829062

[6] S. Gundacker, *et al.*, “A systematic study to optimize SiPM photodetectors for highest time resolution in PET,” IEEE Trans. on Nucl. Sci., vol. 59, no 5, part 1, pp. 1798-1804, 2012. DOI: 10.1109/TNS.2012.2202918

[7] A. Del Guerra, *et al.*, “Silicon Photomultipliers (SiPM) as novel photodetectors for PET,” Nucl. Instrum. and Methods A, vol. 648, pp. S232 – S235, 2011. DOI: 10.1016/j.nima.2010.11.128

[8] Xiaoli Li, C. Lockhart, T.K. Lewellen, and R.S. Miyaoka, “Study of PET detector performance with varying SiPM parameters and readout schemes,” IEEE Trans. on Nucl. Sci., vol. 58, no 3, part 1, pp. 590-596, 2011. DOI: 10.1109/TNS.2011.2119378

[9] B. Seitz, A. G. Stewart, K. O'Neill, L. Wall, and C. Jackson, "Performance Evaluation of Novel SiPM for Medical Imaging Applications," J. Photonics, 2014, no. 1 (43), Pp.104-113. DOI: 10.1109/NSSMIC.2013.6829685

[10] H. Sabet, et al., "High-performance and cost-effective detector using microcolumnar CsI:Tl and SiPM," IEEE Trans. on Nucl. Sci., vol. 59, no 5, part 1. pp. 1841-1849, 2012. DOI: 10.1109/TNS.2012.2202248

[11] O.V. Dvornikov, V. A. Tchekhovski, N. N. Prokopenko, Ya. D. Galkin, A. V. Kunts, A. V. Bugakova, "Implementation of Reading Electronics of Silicone Photomultiplier Tubes on the Array Chip MH2XA030," Visnyk NTUUKPI Seriia–Radiotekhnika Radioapara to buduvannia, 2019, Iss.78, pp.60–66. DOI: 10.20535/RADAP.2019.78.60-66.

[12] O. V. Dvornikov, V. A. Tchekhovski, V. L. Diatlov, "Equipments to single photon registration. Part 1. Features and possibilities of multi-channel photodetectors with intrinsic amplification," Devices and methods of measurements : Scientific and Engineering J., 2012, no. 2(5), Pp. 5–13. (In Russian)

[13] O. V. Dvornikov, V. A. Tchekhovski, V. L. Diatlov, "Equipments to single photon registration. Part 2. Silicon photomultiplier signal preliminary processing. (Review) ," Devices and methods of measurements : Scientific and Engineering J., 2013, no. 1 (6), Pp. 5–13. (In Russian)

[14] O. V. Dvornikov, N. N. Prokopenko, V. A. Tchekhovski, Y. D. Galkin, A. E. Titov and A. V. Bugakova, "Silicon Photomultipliers' Analog Interface with Wide Dynamic Range," Proceedings of 17th IEEE East-West Design & Test Symposium (EWDTS-2019), September 13-16, 2019, Batumi, Georgia, pp. 270-273. DOI: 10.1109/EWDTS.2019.8884430.

[15] O.V. Dvornikov, N. N. Prokopenko, V. A. Tchekhovski , Ya. D. Galkin, A. V. Kunz, A. E. Titov, "Radiation-Hardened Voltage Comparators for the Synthesis of Microcircuits on the Array Chip MH2XA030," Proceedings of IEEE International Conference on Computation, Automation and Knowledge Management (ICCAKM - 2020), 9-11 January 2020, Dubai, Pp. 1-4.

[16] O. V. Dvornikov, V. L. Dziatlau, V. A. Tchekhovski, N. N. Prokopenko and A. V. Bugakova, "BiJFet Array Chip MH2XA030 — a Design Tool for Radiation-Hardened and Cryogenic Analog Integrated Circuits," 2018 IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), St. Petersburg, 2018, pp. 13-17. DOI: 10.1109/EExPolytech.2018.8564415

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 3, 547-556 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Efficient Discretization Approaches for Machine Learning Techniques to Improve Disease Classification on Gut Microbiome Composition Data

Hai Thanh Nguyen[*,1], Nhi Yen Kim Phan[1], Huong Hoang Luong[2] , Trung Phuoc Le[2], Nghi Cong Tran[3]

[1]*College of Information and Communication Technology, Can Tho University, Can Tho city, 900100, Vietnam.*

[2]*Department of Information Technology, FPT University, Can Tho city, 900000, Vietnam.*

[3]*National Central University, Taoyuan, 320317, Taiwan, R.O.C.*

ARTICLE INFO

Article history:
Received: 10 May, 2020
Accepted: 18 June, 2020
Online: 28 June, 2020

Keywords:
Personalized medicine
Bacterial composition
Classic machine learning
Discretization
Species abundance
Read counts
Disease prediction
Deep learning

ABSTRACT

The human gut environment can contain hundreds to thousands bacterial species which are proven that they are associated with various diseases. Although Machine learning has been supporting and developing metagenomic researches to obtain great achievements in personalized medicine approaches to improve human health, we still face overfitting issues in Bioinformatics tasks related to metagenomic data classification where the performance in the training phase is rather high while we get low performance in testing. In this study, we present discretization methods on metagenomic data which include Microbial Compositions to obtain better results in disease prediction tasks. Data types used in the experiments consist of species abundance and read counts on various taxonomic ranks such as Genus, Family, Order, etc. The proposed data discretization approaches for metagenomic data in this work are unsupervised binning approaches including binning with equal width bins, considering the frequency of values and data distribution. The prediction results with the proposed methods on eight datasets with more than 2000 samples related to different diseases such as liver cirrhosis, colorectal cancer, Inflammatory bowel disease, obesity, type 2 diabetes and HIV reveal potential improvements on classification performances of classic machine learning as well as deep learning algorithms. These binning approaches are expected to be promising pre-processing techniques on various data domains to improve the performance of prediction tasks in metagenomics.

1 Introduction

This paper is an extension of work originally presented in The 11th IEEE International Conference on Knowledge and Systems Engineering (KSE) 2019 in Da Nang, Vietnam [1].

Recent years, the field of health care has been receiving great attention from the world. Many services and high-tech applied equipment are manufactured for medical use. Medical results require high accuracy and meet a wide range of diseases, so it is necessary to deploy diagnostic methods and treatments with new technologies. Also, dangerous diseases such as Liver Cirrhosis, Colorectal, HIV, etc. have an increasing rate due to lifestyle, way of life, diet. Liver Cirrhosis is a disease of concern due to the increasing trend each year. The main cause of cirrhosis is the living environment, using a lot of alcohol, toxic chemicals. The level and duration of consumption is an important determinant of the development of liver pathology. In 2015, cirrhosis was the 12th leading cause of death in the United States, with a total of 42,443 deaths 2,494 compared to 2014 [2]. Colorectal cancer is the third most common disease in the United States. The main object of the disease is elderly but the proportion of young people with the disease tends to increase. According to statistics in early 2020, an estimated 53,200 deaths (28,630 males and 24,570 females) are due to the disease, but there is an increased incidence in young people. Although the incidence decreases by 3.6% per year from 2007 to 2016 in adults 55 years and older, they increase 2% per year in adults under 55 years. This year, colorectal cancer is estimated to be the fourth most commonly diagnosed cancer in the United States for men and women aged 30 to 39 [3]. Heart failure is a common complication of Cardiometabolic diseases (CMD). When the pumping action of the heart weakens, the amount of blood pumped out

[*]Corresponding Author: Hai Thanh Nguyen, Email: nthai@cit.ctu.edu.vn, nthai.cit@ctu.edu.vn

is insufficient for the body to make it difficult for the person to feel short of breath, or chest pain, which is called heart failure. Genetic diversity has not been considered in the diagnosis and prognosis of the disease. We apply a single method of treatment to all patients with a similar diagnosis. The results showed that some patients' health did not improve, and the rest gradually recovered. This shows that many methods need to be applied for an effective treatment for each patient. Advances in data processing technology make us understand the importance of metagenomic to human health and explore the diversity of genetics. Deep learning has provided many algorithms to help scientists propose models, methods of diagnosis and treatment.

Modern techniques in healthcare are still developing at a great speed. One of them is Personalized Medicine which defines the impact goals that will work for a patient based on the patient's environmental factors, genes, etc. and is used on a group of patients. Today, scientists have studied numerous methods for Personalized Medicine and metagenomic is one of them. As we have known, metagenomics is a method of sequencing and analyzing the DNA of microorganisms collected from the environment without culturing them. We are looking at the human gut environment. Bacteria are often very diverse, they are classified into seven basic types: domain, kingdom, phylum, class, order, family, genus, and species. This diversity helps to provide more information about diseases to support more effective diagnosis and prognosis. The diseases under consideration are complex and we only have a limited number of observation data samples, so the prediction tasks are produced in inconsistent results with comparable diseases.

To test and propose models for healthcare services, Machine learning algorithms have been strongly researched and developed to solve metagenomics-related problems , mainly prediction genes [4, 5], Operational Taxonomic Unit clustering [6–9], comparative metagenomics [10–13], binning, taxonomic profiling and assignment [14]. All of the issues mentioned are given in [14].

The authors presented the basic principles of machine learning in [15] and a typical pipeline was introduced [16, 17]. This model is clustered or classified based on pre-processed results and feature extraction. A machine learning program consists of three basic components, such as data (or experience), task (formed by the output of the algorithm) and target (possibly in the form of measuring the efficiency of the output). MetAML was introduced in [17] by Edoardo et al. which is a computational framework that learns about the presence of specific species signs and the relative abundance of species, these two collectively called are quantitative microbiome configurations. Using machine learning to independently evaluate the accuracy of models on large metagenomic datasets, thereby analyzing and comparing the practical microbiome usage strategies that were recommended in [18]. With the task of classifying metagenomics, Ph-CNN [19] was introduced to the OTU hierarchical structure and it was also compared to other technologies in machine learning such as random forests. PopPhy-CNN [20] is introduced by D. Reiman et al. PopPhy-CNN is deep learning framework, using embedded information based on a phylogenetic tree to predict diseases from metagenomic data. PopPhy-CNN has superior performance compared to random forests, support vector machines, LASSO and the basic 1D-CNN model built with bacterial vectors. In addition to retrieving information microbiological classifications from trained CNN models, PopPhy-CNN also visualizes phylogenetic tree classifications. Several deep learning algorithms have been evaluated as a feasible approach to speed up DNA sequencing [21]. To identify viruses by deep learning with metagenomic data, J. Ren et al. [22] proposed using DeepVir Downloader (a reference-free and alignment-free machine learning method) to improve accuracy and support virus research. We will present 1D representations through packaging and expansion methods, and demonstrate the effectiveness of applying Multi-layer Perceptron (MLP), traditional artificial neural networks to perform predictive tasks.

In the study [23], the authors presented the MSC algorithm with the goal of classification to detect the circulating rate and estimate their relative level. MSC is a metagenomic sequence classification algorithm that has accuracy, memory and runtime and gives an approximate estimate of abundance over other algorithms. Jolanta Kawulok [24] presented research on the environmental classification of metagenomic data to build a microbiome fingerprint. Another study by Lo Chieh et al. [25] proposed the MetaNN model, this is a model of host phenotypes classification from metagenomic data using neural networks. The results show that MetaNN is superior to the exact classification team for synthetic and real metagenomic data compared to other models, contributing to the development of microbiome-related disease treatments.

In this research, we have investigated and implemented a variety of binning methods to improve predictive performance. We have proposed different binning methods including binning based on frequency, width, the proposed breaks and combination between scaler and binning. After applying binning approaches, the data are fetched into machine learning algorithms including both classical machine learning and deep learning. We present the results which are produced with more running times in [1] for deeper comparison and include p-values for finding significant results. Additionally, we include another dataset, Crohn disease, in the experiments, for a complete comparison with the state-of-the-art. The results of [1] run by only MultiLayer Perceptron, we extend to run with a variety of machine learning algorithms including classical algorithms such as Random Forest, Linear Regression and deep learning (Convolutional Neural Network (CNN)) technique. The number of bins also carried out to consider in this work for choosing the number of bins for metagenomic data binning approaches. The work is expected to provide robust pre-processing methods to enhance the performance of machine learning algorithms applying to metagenomic data. Results, datasets and scripts for the experiments are uploaded to the public GitHub repository. To sum up, the contributions of this work include:

- The study presents various binning methods. The width of all bins can be equal or the width of each bin is conducted from the frequency of values. We also consider binary bins to determine whether the feature exists in the considered sample or not. The proposed methods as shown results can improve the performance comparing the original data.

- A combination of scaler algorithms and binning methods is also introduced for the comparison. Some scalers such as logarithm calculations and quantile transformation reveal good performance on some datasets.

- Methods are evaluated by disease prediction tasks on a variety of diseases including liver cirrhosis, colorectal cancer, IBD, obesity, HIV, Type 2 diabetes. Considered data types include species abundance and counts at other taxonomic ranks such as genus, family, etc.
- Several machine learning techniques including both classic machine learning and deep learning are investigated with Classification tasks on metagenomic data.
- The proposed framework, namely Metagenomic-To-Bins (**Met2Bin**), including scripts, results and datasets is published at https://github.com/thnguyencit/met2bin.

In the next sections of the paper, we describe 8 metagenomic datasets used in the experiments including the total number of features, the total number of samples, both the number of disease and non-disease (Section 2). In Section 3, we introduce the metagenomic data binning with various approaches along with scaler algorithms. Section 4 describes the empirical results and Section 5 provides insightful remarks of the study.

2 Metagenomic data benchmarks

To evaluate the performance of classifiers, we run the prediction tasks on a variety of datasets (8 metagenomic datasets) including species abundance datasets and read counts related to different specific diseases, such as Liver cirrhosis (CIR), colorectal cancer (COL), Crohn's disease, Human Immunodeficiency Virus (HIV) infection, Inflammatory Bowel Disease (IBD), Obesity (OBE), Type 2 Diabetes (T2D and T2W). Each dataset consists of 4 main parameters: (1) the number of features, (2) the number of samples, (3) the number of samples affected by the disease, (4) the number of healthy samples.

CIR dataset comprises 542 features with 232 samples including 118 patients and 114 healthy individuals. COL dataset consists of 121 individuals with 48 patients. The number of patients affected by Crohn's disease is 663 out of 975 people were considered. For HIV dataset, the number of positive cases is 129 out of 155. IBD dataset includes 253 samples of which 164 are affected by the disease, OBE dataset consists of 174 non-obese and 170 obese individuals. T2D and WT2 datasets include 344 samples and 96 samples, respectively.

The HIV, Crohn's datasets contain 155 and 975 samples, respectively, which have values greater than 1, evaluated using the recommended method [26] with the number of reads for microbial taxa at the levels which are higher than species. Crohn's disease is a type of inflammatory bowel disease (IBD). This disease can affect any segment of the gastrointestinal tract from the mouth to the anus. The features in two these datasets can be genus counts, family counts, or order counts. We bring read counts datasets from the analysis in [27] to compare to our method.

For species abundance datasets, each sample, species abundance is a relative proportion and it is revealed as a real number that has the total abundance of all species summing to 1 (The details are shown in Table 1).

Let D be the set of considered datasets, D = $\{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8\}$, with d_1 = CIR, d_2 = COL, d_3 = Crohn, d_4 = HIV, d_5 = IBD, d_6 = OBE, d_7 = T2D, d_8 = WT2, d = 1..8

$F_i = \{f_1, f_2, ..., f_m\}$ includes m features corresponding to d_i

$S_i = \{s_1, s_2, ..., s_n\}$ includes n samples corresponding to d_i

$P_i = \{p_1, p_2, ..., p_k\}$ includes k patients who affected by diseases corresponds to d_i

$C_i = \{c_1, c_2, ..., c_k\}$ includes x controls / healthy individuals that correspond to d_i

$$Matrix(C) = \begin{pmatrix} d_1 & F_1 & S_1 & P_1 & C_1 \\ d_2 & F_2 & S_2 & P_2 & C_2 \\ d_3 & F_3 & S_3 & P_3 & C_3 \\ d_4 & F_4 & S_4 & P_4 & C_4 \\ d_5 & F_5 & S_5 & P_5 & C_5 \\ d_6 & F_6 & S_6 & P_6 & C_6 \\ d_7 & F_7 & S_7 & P_7 & C_7 \\ d_8 & F_8 & S_8 & P_8 & C_8 \end{pmatrix}$$

$$= \begin{pmatrix} CIR & 542 & 232 & 118 & 114 \\ COL & 503 & 121 & 48 & 73 \\ Crohn & 48 & 975 & 663 & 312 \\ HIV & 60 & 155 & 129 & 26 \\ IBD & 465 & 253 & 164 & 89 \\ OBE & 572 & 344 & 170 & 174 \\ T2D & 381 & 96 & 53 & 43 \\ WT2 & 443 & 110 & 25 & 85 \end{pmatrix}$$

The read counts of each feature in HIV and Crohn datasets can be greater than 1 while total species abundance of all features in one sample of other species abundance datasets is sum up to 1:

$$\sum_{i=1}^{k} f_i = 1$$

with:

- k is the number of features for a sample.
- f_i is the value of the i-th feature.

The next section, we will introduce pre-processing methods based on binning approaches on these metagenomic datasets.

3 Metagenomic data binning

Data binning or Data Discretization is a data processing method which transforms continuous value into discrete value. To discretize continuous values into "bins", we need to determine "breaks" where indicates which bin these values belong to. "Breaks" are real values which can be 0.1, 0.35, etc. that are considered as "boundaries" of bins. Let say, we have an array of values including 0.000012, 0.02, 0.56, 0.92. We would like to divide 10 bins which own an equal width for each bin on a considered value range from 0 to 1. The width of each bin or interval width, in this case, is $\frac{1-0}{10} = 0.1$. The value range of the first bin is from 0 to 0.1, the second bin is from 0.1 to 0.2, etc. In our study, we do not consider the values of 0 (zeros), so values which are greater than 0 and lower than 0.1 (such as values of 0.000012, 0.02) will belong to the first bin while the second bin contains values which are greater than or equal to 0.1 and lower than 0.2. For other values with the computation as above, 0.56 will belong to the 6th bin while the last bin (10th bin) contains 0.92. The breaks as the example mentioned consist of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. We also add 0 to the breaks to

Table 1: Information on eight considered datasets.

	CIR	COL	Crohn	HIV	OBE	T2D	WT2	IBD
#Features	542	503	48	60	465	572	381	443
#Samples	232	121	975	155	253	344	96	110
#Patients	118	48	663	129	164	170	53	25
#Controls	114	73	312	26	89	174	43	85

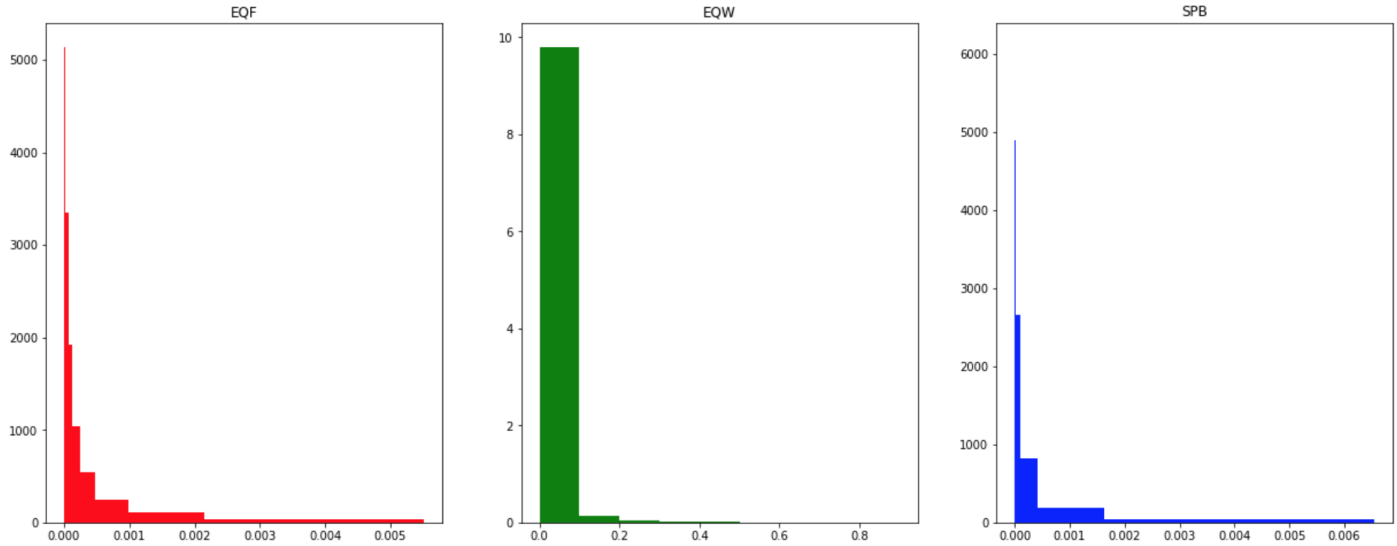


Figure 1: Data density with various binning methods on Liver Cirrhosis bacterial species abundance dataset using the same 10 bins. X-axis shows a value range of abundance.

compare whether the values are greater than 0 to distribute values to bins. Metagenomic data can exist outliers that cause many wrong many prediction results by learning machine algorithms. Binning approaches are expected to improve the performance by reducing the effects of minor observation errors and to get rid of noise in the data.

This section will present various binning approaches which can be binning with **equal width** or **equal frequency** of values or basing on **species abundance distribution** of several considered species abundance datasets or simply only considering whether the feature exists in the sample or not ($value > 0$), namely **Binary Binning**. Some methods combining between binning approach and transformation with scaler algorithms are also presented.

3.1 Equal Width binning

EQual **W**idth binning (**EQW**) divides and delivers continuous values to bins which have equal width. Each bin has the equal width which is computed by $\frac{Max\_value - Min\_value}{number of bin}$ in the range of [Min, Max] of the data. For instance, we would like to deliver original values to 5 equal width bins ($k = 5$) using a range of [Min=0,Max=1], then width of each bin is 0.1 (w = 0.1). The interval boundaries include $Min + w, Min + 2 \times w, ..., Min + (k-1) \times w$. The idea is simple but this method show improvements in prediction tasks.

3.2 *EQual Frequency binning (EQF)*

Equal frequency binning method cuts the data into n parts (bins) which each part contains approximately the same number of values. The breaks are identified using the training set so the performance in the testing phase will be poor if the training set cannot reflect exact general data distribution of the considered disease. Breaks depend totally on data distribution so the width of each bin can vary significantly.

3.3 Binning based on species abundance distribution

SPecies **B**ins (**SPB**) is extended from EQW combining species abundance distribution conducted from 6 species abundance datasets in [1]. Authors in [1] presented breaks including $0, 10^{-7}, 4 * 10^{-7}, 1.6 * 10^{-6}, 6.4 * 10^{-06}, 2.56 * 10^{-05}, 0.0001024, 0.0004096, 0.0016384, 0.0065536$ for Species Bins. The first break ranges from 0 to 10^{-7} which is the smallest value of species abundance known in six species abundance datasets of CIR, COL, IBD, OBE, T2D, and WT2 [1]. The width of each bin is equivalent to a 4-fold increase from the previous bin.

3.4 Equal Width binning combining scaler algorithms

Some transformation algorithms applying to original data can be useful for binning. Standardization method is a widely-used technique for numerous machine learning algorithms to resolve the problem of different data distributions. Quantile Transformation (**QTF**), MinMaxScaler (**MMS**), and logarithmic computations scalers are considered to convert data before binning.

Quantile Transformation is implemented to combine with EQW in these experiments. QTF is considered as a robust pre-processing

technique because it can reduce the effect of the outliers. Samples in test and validation sets which are smaller or larger than the fitted range then will be assigned to the bounds of the output distribution. Another algorithm illustrated in this study is MinMaxScaler, to make a comparison with QTF and logarithmic computations. MinMaxScaler converts each feature to a given range by (1) and (2) formulas:

$$X_{std} = \frac{X - X.min}{X.max - X.min} \tag{1}$$

$$X_{scaled} = X_{std} * (max - min) + min \tag{2}$$

Functions which perform the transformation as above are now available in scikit-learn library.

As described in [1], metagenomic data usually follow the zero-inflated distribution. Data scaled with the methods of transformation based logarithm calculation reveal more normally-distributed. In this study, we use logarithm computation base 4 and base 100 for comparison.

3.5 Binary Bin

Binary Bin (**B2**) which can be considered as the one-hot encoding method, also is brought to compare. B2 indicates whether a feature is present or absent in a sample. If values are greater than 0, Bin 1 contains them. Otherwise, they are delivered to Bin 0 (with all values=0).

3.6 Data distribution Visualization of binning methods and scalers

Figure 1 shows various binning approaches on CIR dataset. The breaks of EQF include $0, 4 * 10^{-07}, 2.47 * 10^{-05}, 6.2 * 10^{-05}, 1.27*10^{-04}, 2.466*10^{-04}, 4.788*10^{-04}, 9.783*10^{-04}, 2.1484*10^{-03}, 5.5149 * 10^{-03}$. These breaks are approximate to SPB. We note that some first bins (with EQF) own high density with the width of these are rather small ($4 * 10^{-07}$ for the first bin) while the width of the 9th bin is about $3.4 * 10^{-03}$. Similar results are exhibited for SPB. Width of each bin with EQW is equal, so we can see that the first bin contains most of the data.

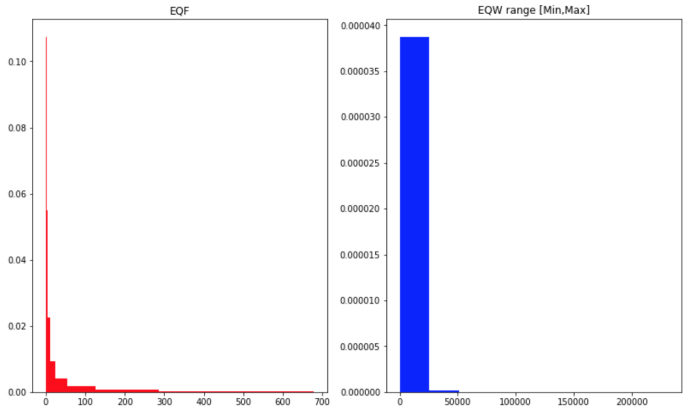


Figure 2: Data density on Crohn Read Counts dataset. X-axis shows a value range of counts with breaks using EQF and EQW on Min-Max range of training set.

For data type of counts, the values in features can be greater than 1, so SPB and EQW considering in a value range from 0 to 1 are not efficient for this type. EQW with a range between Min and Max values in the training set and EQF can work in this situation (Figure 2). Observed and conducted from Figure 2 for EQW method, we see that data distribution of metagenomic is the zero-inflated distribution, no matter what data is abundance or counts. However, a transformation with logarithm enables data to be more normally-distributed (Figure 3).

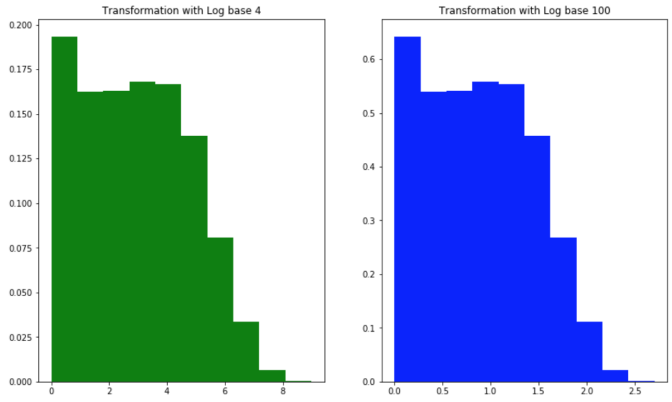


Figure 3: Data density on Crohn Read Counts dataset after transformation with logarithm base 4 and 100. The X-axis shows a transformed value range

The performance of the proposed binning methods will be evaluated in the next section (Section 4).

4 Experimental Results on Metagenomic data binning Approaches

To exhibit the efficiency of binning approaches on various machine learning algorithms, we present the results with a classic machine learning algorithm (Random Forest), Linear Regression and a famous deep learning technique that is Convolutional Neural Network on 1D data (CNN1D).

```
 OPERATION          DATA DIMENSIONS   WEIGHTS(N)   WEIGHTS(%)

     Input   #####    542    1
InputLayer     |   -------------------         0      0.0%
             #####    542    1
    Conv1D    \|/  -------------------       256      0.7%
      relu   #####    540   64
   Flatten   ||||| -------------------         0      0.0%
             #####       34560
     Dense   XXXXX -------------------     34561     99.3%
   sigmoid   #####           1
```

Figure 4: CNN1D architecture.

The algorithms of **CNN1D, Linear Regression** are both implemented using Adam optimization function, a learning rate of 0.001, and using an overall epoch of 500 along with a batch size of 16 and binary cross-entropy loss function. To reduce overfitting issue, we use "Early Stopping" with the number of epoch patience of 5. The learning will stop if the Loss is not improved after 5 consecutive epochs. As conducted from [1], we should use a shallow deep learning architecture instead of deeper architectures, so the proposed CNN1D architecture includes a convolutional layer consisting of 64

filters of the size of 3, following by a max-pooling of size 2, and an activation function of ReLU [28]. The details of CNN1D are visualized in Figure 4.

Random Forests algorithm is a robust learning algorithm and widely-used in numerous studies related to bioinformatics tasks. In this study, Random Forest algorithm is deployed with 500 trees, nodes are expanded until all leaves are pure or until all leaves contain less than *min_samples_split* = 2 where *min_samples_split* is the minimum number of samples required to split an internal node.

The performance of each classifier is measured by an average of Area Under the Curve (AUC) and an average Accuracy (ACC) on 10-stratified-fold-cross validation repeated 5 times. The same folds are used for all classifiers, i.e. training and test sets were identical for each classifier. Besides, results are visualized by Boxplot to exhibit graphically depicting groups of numerical data through their quartiles. Our results are compared to state-of-the-arts including ***MetAML*** [18] on 6 species abundance datasets and ***Selbal*** [27] on 2 read counts datasets. **MetAML** [18] is a framework for metagenomic data analysis running on species abundance with classic machine learning algorithms such as SVM and Random Forest. MetAML performed the best with Random Forest; hence in comparison with our methods, we also run the classification tasks using Random Forest with the same parameters with MetAML. **Selbal** uses balance score to find good features, then fetching the features into Linear Regression algorithms for the prediction tasks.

We need to specify the value range to divide bins for binning approaches. In this study, the considered value range can be either [0,1] or [Min,Max] to divide bins for data. [Min,Max] means we consider the range covered by the minimum value and the maximum value of all features in the training set to bin the data.

We present the experimental results as followings. First, we show the disease prediction performance of all considered binning approaches (Section 4.1). Next, we evaluate and compare the differences in performance when we change the number of bins. Then, promising methods are compared with the state-of-the-art including **MetAML** [18] and **Selbal** [27].

4.1 Evaluation on different data pre-processing methods for metagenomic data

We compare the performance of various pre-processing methods based on binning approaches with three different widely-used machine learning algorithms including Random Forests (Figure 5), Convolutional Neural Network (Figure 6) and Linear Regression (Figure 7).

Figure 5 shows the prediction performance of the considered binning methods performed by Random Forests on 8 considered datasets. We compare the efficiency of different binning approaches (B2, EQF, EQW, SPB) and various scalers (Logarithms, Min-Max Scaler, Quantile Transformation). Except for Crohn dataset, there are not too significant differences in the performances of Random Forests algorithm with different approaches. In the chart, NA ("Not Available") means the model running on the original data without using binning, so the value range for binning is also NA (for example, NA_log4_NA, NA_none_NA, etc.). For CIR and OBE datasets, EQF without using scaler achieves the best performance. The average AUC for predicting CIR, OBE datasets using EQF_none are 0.95582, 0.68238, respectively. For samples from COL dataset, the highest result is with QTF scaler. The datasets of Crohn's disease, HIV, IDB achieve the best results with EQW binning combined with QTF scaler on the value range of [Min,Max]. The best AUC for Crohn's disease dataset is 0.86976 and IBD obtains the best value at 0.88888 while HIV dataset obtains the best at 0.72438. The remaining 2 datasets including T2D and T2W using the binning method of EQW on the range of [Min,Max] without using scaler, reach the AUC best at 0.76286 and 0.80868, respectively. Crohn dataset shows worse results on the value range of [0,1]. It seems more appropriate because this dataset uses read count where the values are either equal 0 or greater than 1.

Figure 6 exhibits the results of CNN model on the considered datasets. As seen from the figure, binning approaches can outperform other methods. CIR dataset has the highest AUC value of 0.95986 while IBD dataset owns the best AUC value of 0.90894. Both two those results are evaluated with EQF without using any scalers. COL dataset obtains the best results with using the EQW combined Min-Max scaler on the value range of [Min,Max] of the training sets with AUC of 0.83732. Crohn's gets the best result using QTF scaler with AUC of 0.85698. The prediction results on HIV disease using EQF without scaler on the value range of [Min,Max] peak at the best AUC value of 0.72788. OBE dataset has the highest AUC when we use the SPB approach. Two datasets of T2D, T2W running with EQW binning on the range of [0, 1] exhibit the best AUCs of 0.75746, 0.80238, respectively.

We also present performances of Linear Regression algorithm in Figure 7. As exhibited, the results are rather similar to mentioned previous two algorithms. CIR dataset obtains the best AUC value of 0.95870 with using EQF binning on the range of [0,1] while we achieve the best AUC of 0.83990 on COL with EQW. Original data of Crohn dataset being run by QTF scaler reaches the highest AUC value of 0.86884. Some results on other datasets are similar to CNN's results.

From the shown experimental results, we notice that CNN, in general, achieves better results than using Random Forest and Linear Regression. Binning approaches appear to be more efficient to enhance significantly the performance with CNN and Linear Regression.

The methods of EQW on the value range [min, max] of training sets, EQF binning and scalers algorithms appear to be appropriate methods for the prediction tasks of HIV and Crohn's disease where the values of features can be greater than 1. Comparing to the performance of the original data (NA_none_NA), these methods can give significant improvements.

4.2 Number of bins for Metagenomic binning

A comparison among the numbers of bins for binning EQW approach is presented in Figure 8. Average AUCs on each number of bins applying to 8 considered datasets are calculated to compare. The numbers bins of 5, 10 give average AUCs (on 8 datasets) of 0.8022450, 0.8011900, respectively while using 100 and 255 bins reveal AUCs of 0.7621925, 0.7471600, respectively. The binary bin approach reaches the average AUC of 0.7958750. As observed, the numbers bins of 5 and 10 obtain significantly better results compared to 100 and 255 bins. As shown from the average performance on the all considered datasets, data discretization with 5 bins

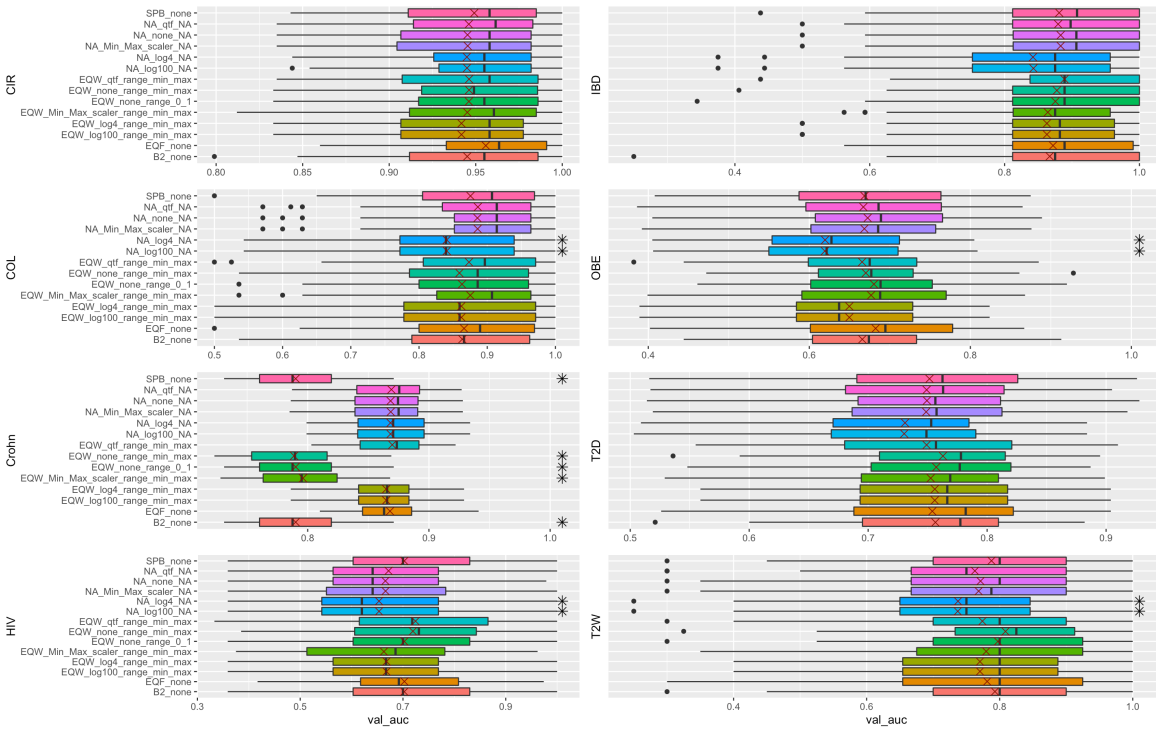


Figure 5: Different Methods Comparison using Random Forest. "*" reflects significant differences compared to the best result on each dataset. "X" reveals average performance on 10-fold cross-validation repeated 5 times. Methods names denote The binning method combining Scaler and data range for binning. For example, `EQW_Min_Max_scaler_range_min_max` denotes that we performed MinMax scaler and binning by EQW on the range of [Min,Max] of values in training set while `NA_none_NA` means no binning method or scaler is applied. Black Dots exhibit outliers in results.

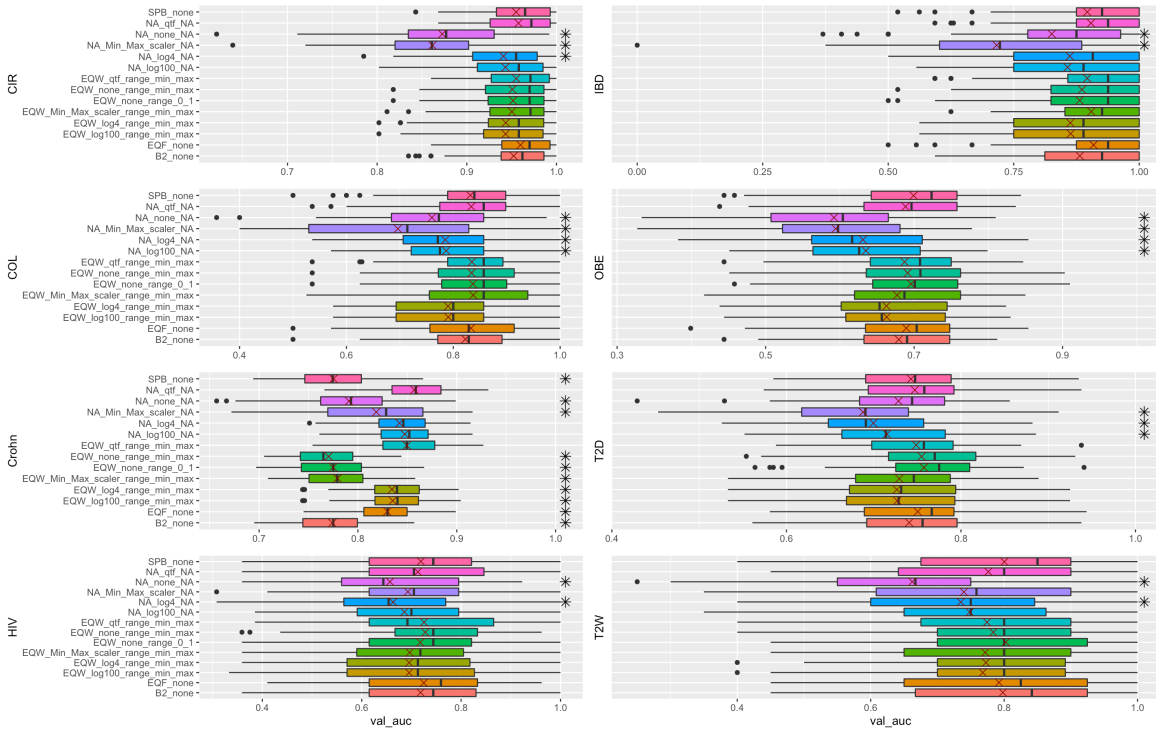


Figure 6: Different Methods Comparison using CNN. All symbols and stickers in the chart are the same as Figure 5.

achieves the best.

We see that CIR dataset obtains the best AUC value is 0.9522 while IBD dataset achieves the AUC value of 0.88162 with 2 bins. Additionally, The Crohn dataset reaches the best AUC of 0.77488 with 100 bins. Three datasets of COL, OBE and T2W reveal the best AUC values with 5 bins with AUCs of 0.83670, 0.69594, 0.80238,

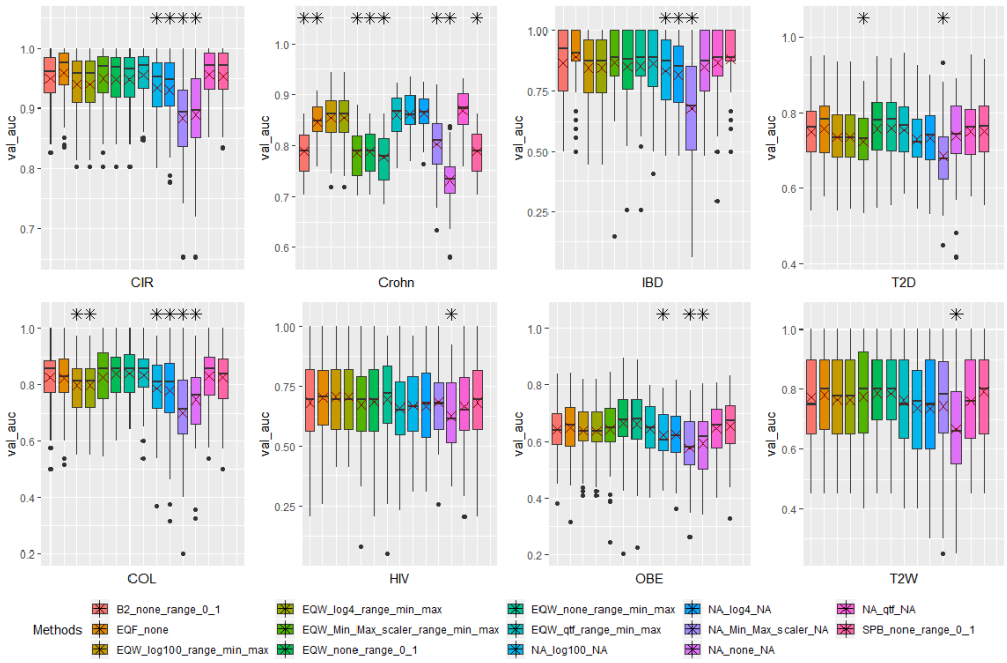


Figure 7: Different Methods Comparison using Linear Regression. All annotations in the chart are the same as Figure 5.

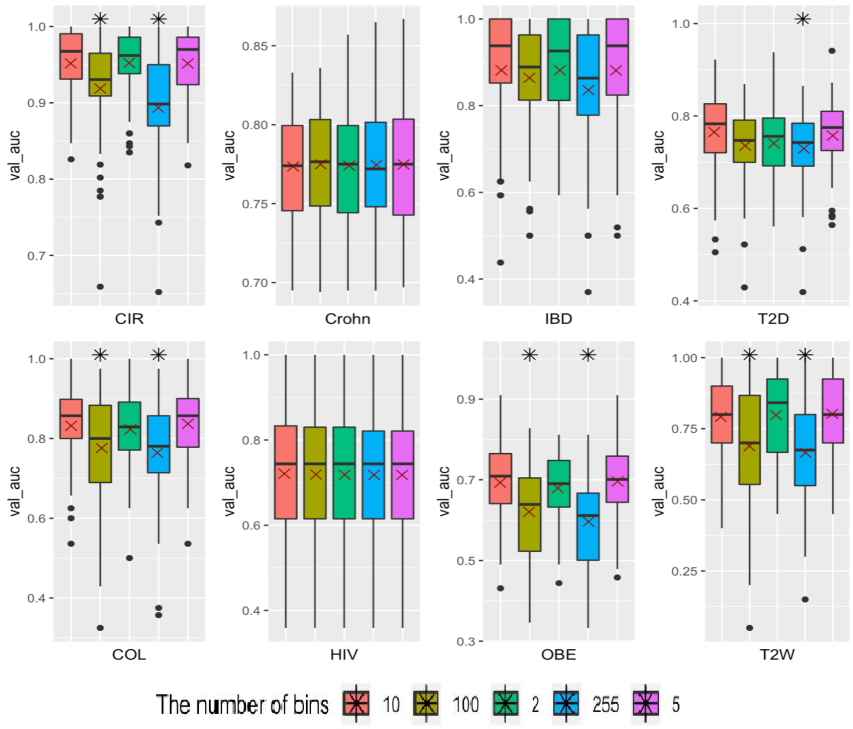


Figure 8: Performance of the various number of EQW bins using the CNN model. "*" reflects significant differences compared to the best result on each dataset. "X" reveals average performance on 10-fold cross-validation repeated 5 times.

respectively. Otherwise, prediction tasks on the diseases of HIV and T2D with 10 bins give the best results.

4.3 State-of-the-art comparison

To reflect the efficiency of binning approaches on metagenomic data, we compare the proposed binning approaches to some state-of-the-art including MetAML [18] and selbal [27].

In Table 2, we display the results using binning approaches of EQW, EQF, SPB with a total of 5 bins and comparing state-of-the-art in average ACC and average AUC on 10-fold cross-validation repeated 5 times. Three datasets including CIR, OBE and T2W all had better results than state-of-the-art. COL disease has only better

Table 2: Results Comparison of robust binning methods (the number of bins is 5 for EQW, EQF and SPB) and state-of-the-art (Selbal on HIV, Crohn's disease and MetAML on other datasets) in average ACC and average AUC on 10-fold cross-validation repeated 5 times. The results formatted in bold text are better compared to the state-of-the-art.

Datasets			CIR	COL	Crohn	HIV	IBD	OBE	T2D	T2W	AVG
	State-of-the-art	val_acc	0.877	0.805	NA	NA	0.809	0.644	0.664	0.703	0.750
EQF	CNN	val_acc	**0.906**	0.792	0.783	0.833	**0.829**	**0.680**	0.655	**0.722**	0.775
EQF	RF	val_acc	**0.887**	**0.808**	0.813	0.816	**0.810**	**0.657**	**0.675**	**0.720**	0.773
EQW	CNN	val_acc	**0.883**	0.785	0.739	0.827	**0.851**	**0.668**	**0.667**	**0.707**	0.766
EQW	RF	val_acc	**0.886**	0.791	0.745	0.809	0.807	**0.645**	**0.680**	**0.708**	0.759
SPB	CNN	val_acc	**0.903**	0.790	0.737	0.828	**0.830**	**0.672**	0.650	**0.727**	0.767
SPB	RF	val_acc	**0.883**	0.785	0.739	0.827	**0.851**	**0.668**	**0.667**	**0.707**	0.766
	State-of-the-art	val_auc	0.945	0.873	0.820	0.674	0.890	0.655	0.744	0.762	0.795
EQF	CNN	val_auc	**0.960**	0.833	**0.830**	**0.725**	**0.909**	**0.689**	**0.750**	**0.792**	0.811
EQF	RF	val_auc	**0.956**	0.866	**0.868**	**0.703**	0.872	**0.682**	**0.754**	**0.781**	0.810
EQW	CNN	val_auc	**0.952**	0.837	0.775	**0.718**	0.881	**0.696**	**0.757**	**0.802**	0.802
EQW	RF	val_auc	**0.946**	0.863	0.790	**0.703**	0.875	**0.680**	**0.757**	**0.797**	0.801
SPB	CNN	val_auc	**0.955**	0.832	0.775	**0.718**	**0.896**	**0.699**	0.742	**0.801**	0.802
SPB	RF	val_auc	**0.952**	0.837	0.775	**0.718**	0.881	**0.696**	**0.757**	**0.802**	0.802

ACC value when binning with EQF combined with Random Forest model (ACC value is 0.868). In Crohn's disease, when performing the binning method with EQF combined with 2 models CNN (ACC value is 0.830) and Random Forest (ACC value is 0.868) both give better results than state-of-the-art. AUC values for HIV disease are better than state-of-the-art in all models and methods, but there is no ACC result in this disease higher than state-of-the-art. IBD has 5 good results when done with ACC values but only 2 good results for AUC. T2D has most of the results better than state-of-the-art, only when doing SPB method with CNN model (both ACC and AUC values) and when binning with SPB, CNN model with ACC measurements are lower than state-of-the-art.

5 Conclusion

In this study, we presented **Met2Bin** with various binning approaches using Equal with binning, Equal frequency binning, Species bins and binary bins to reduce the effects of minor observation errors and to get rid of noise in the metagenomic data. Scaler with QTF and logarithm transformation also show potential improvements in the data type of reading counts. In most cases, binning approaches and scaler algorithms can improve performance for machine learning algorithms.

The binning and scaler approaches are examined on a vast of datasets including different diseases and various data types (species abundance and read counts at genus or family or order levels). We can see that the proposed method can work on any value ranges. This research only takes into account unsupervised binning methods. Considerations on the labels of samples should be carried out in further studies.

As revealed from the performance of disease prediction, we can predict Liver cirrhosis, IBD with high accuracy while Obesity, HIV and T2D diagnosis are still challenges. Further research should investigate to improve those diseases.

In general, CNN produces better results than classic machine learning. However, the considered CNN architecture in this work is rather small and modest but its performance exhibits promising results. Further investigations on the CNN architectures should be considered to improve the performance. We also do not consider the labels of samples when we build the breaks for data binning. In the future, the research should consider and investigate the supervised binning approaches to evaluate whether those can be efficient or not on metagenomic data.

The binning methods are potential methods so that we can use such bins for converting numeric data and showing them in 2D images. A bin which represents the magnitude of value can be shown in the image with a specific colour. Binning techniques enable us to visualize bio-markers in images as well as to leverage advancements in deep learning algorithms for images to do prediction tasks.

The results and other materials of this work can be downloaded from https://github.com/thnguyencit/met2bin.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Thanh Hai Nguyen, Jean-Daniel Zucker. Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches. The 2019 11th International Conference on Knowledge and Systems Engineering (KSE-IEEE), ISBN: 978-1-7281-3003-3, pp 381-385. 2019. https://doi.org/10.1109/KSE.2019.8919295

[2] Colorectal Cancer: Statistics Approved by the Cancer.Net Editorial Board. 2020. https://www.cancer.net/cancer-types/colorectal-cancer/statistics

[3] Young-Hee Yoon, Chiung M. Chen; Liver cirrhosis Mortality in the United States: National, State and Regional Trends, 2000-2015. 2018. https://pubs.niaaa.nih.gov/publications/surveillance111/Cirr15 .htm.

[4] C. Mathee et al. ;SURVEY AND SUM- MARY: Current methods of gene prediction, their strengths and weaknesses; p. 4103-4117.ISSN 0305-1048. 2002. https://doi.org/10.1093/nar/gkf543

[5] Z. Wang, Y. Chen and Y. Li;A Brief Review of Computational Gene Prediction Methods; p. 216-221. ISSN 1672-0229. 2004. https://doi.org/10.1016/S1672-0229(04)02028-5

[6] N.P. Nguyen et al.; A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity; ISSN 2055-5008. Nature. https://doi.org/10.1038/npjbiofilms.2016.4

[7] M. Arumugam et al.; Enterotypes of the human gut microbiome; 473, p. 174-180. ISSN 1476-4687. 2011 https://www.nature. com/articles/nature09944

[8] S. Park et al.; hc-OTU: A Fast and Accurate Method for Clustering Operational Taxonomic Units Based on Homopolymer Compaction; 15, p. 441-451. ISSN 1545-5963. 2016. https://doi.org/10.1109/TCBB.2016.2535326

[9] Wei ZG et al.: A Dynamic Multi-Seeds Method for Clustering 16S rRNA Sequences Into OTUs. PubMed. 2019. https://doi.org/10.3389/fmicb.2019.00428

[10] D. H. Huson, D. C. Richter, S. Mitra, A. F. Auch and S. C. Schuster; Methods for comparative metagenomics; 10, p. S12. ISSN 1471-2105. 2009. https://doi.org/10.1186/1471-2105-10-S1-S12

[11] L.-x. Chen, M. Hu, L.-n. Huang, Z.-s. Hua, J.-l. Kuang, S.-j. Li and W.-s. Shu; Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine drainage; 9, p. 1579-1592. ISSN 1751-7370. 2015. https://doi.org/10.1038/ismej.2014.245

[12] S. Nayfach et al.; Toward Accurate and Quantitative Comparative Metagenomics; 166, p. 1103-1116. ISSN 0092-8674. https://doi.org/10.1186/1471-2105-10-S1-S12

[13] S. M. Dabdoub et al.; Comparative metagenomics reveals taxonomically idiosyncratic yet functionally congruent communities in periodontitis; 6, p. 38993. ISSN 2045-2322. https://doi.org/10.1038/srep38993

[14] K. Sedlar et al.; Bioinformatics strategies for taxon- omy independent binning and visualization of sequences in shotgun metagenomics; 15, p. 48-55. ISSN 2001-0370. 2016. https://doi.org/10.1016/j.csbj.2016.11.005

[15] H. Soueidan et al.; Machine learning for metagenomics: methods and tools; Metagenomics 1. 2017. https://doi.org/10.1515/metgen-2016-0001

[16] G. Ditzler et al.; Forensic identification with environmental samples; IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012. https://doi.org/10.1109/ICASSP.2012.6288265

[17] G. Ditzler et al.; MultiLayer and Recursive Neural Networks for Metagenomic Classification; IEEE Transaction Nanobioscience 14, p. 608-616. 2015. https://doi.org/10.1109/TNB.2015.2461219

[18] E. Pasolli et al.; Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights; PLoS Comput. 2016. https://doi.org/10.1371/journal.pcbi.1004977

[19] D. Fioravanti, Y. Giarratano, V. Maggio, C. Agostinelli, M. Chierici, G. Jurman and C. Furlanello; Phylogenetic convolutional neural networks in metagenomics; 19, p. 49. 2018. ISSN 1471-2105. https://doi.org/10.1186/s12859-018-2033-5

[20] D. Reiman, A. A. Metwally and Y. Dai; PopPhy-CNN: Ation Neural Networks for Metage- nomic D Phylogenetic Tree Embedded Architecture for Convoluata. 2018. http://biorxiv.org/lookup/doi/10.1101/257931

[21] F. Celesti, A. Celesti, J. Wan and M. Villari; Why Deep Learning Is Changing the Way to Approach NGS Data Processing: a Review; ISSN 1937- 3333. 2018. https://doi.org/10.1109/RBME.2018.2825987

[22] J. Ren et al.; Identifying viruses from metagenomic data by deep learning. 2018. http://arxiv.org/abs/1806.07810

[23] Subrata Saha, J. Johnson, S. Pal, G. M. Weinstock, S. Rajasekaran, MSC: a metagenomic sequence classification algorithm, Bioinformatics; vol.35, 17,p. 2932-2940. 2019. https://doi.org/10.1093/bioinformatics/bty1071

[24] Jolanta Kawulok, M. Kawulok and S. Deorowicz; Environmental metagenome classification for constructing a microbiome fingerprint. Biol Direct 14. 2019. https://doi.org/10.1186/s13062-019-0251-z

[25] Lo Chieh, Marculescu, Radu; MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. CM University. Journal contribution. 2019. https://doi.org/10.1186/s12859-019-2833-2

[26] K. Sedlar et al.; Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics; 15, p. 48-55. ISSN 2001-0370. 2018. https://doi.org/10.1016/j.csbj.2016.11.005

[27] Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. Balances: a New Perspective for Microbiome Analysis. mSystems. 2018;3(4):e00053-18. Published 2018 Jul 17. https://doi.org/10.1128/mSystems.00053-18

[28] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU). 2018. https://arxiv.org/abs/1803.08375

ASTES

Advances in Science, Technology and Engineering Systems Journal 5, No. 4, 109-118 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Optimization of Dual Motion Mechanism with Double Grooved Cams for High-voltage Gas Circuit Breaker

Masanao Terada[1,2,*], Yuki Nakai[3], Hiroaki Hashimoto[1], Daisuke Ebisawa[3], Hajime Urai[1], Yasunobu Yokomizu[2]

[1]*Research & Development Group, Hitachi, Ltd., 319-1292, Japan*

[2]*Department of Electrical Engineering, Nagoya University, 464-8603, Japan*

[3]*Energy Business Unit, Hitachi, Ltd., 316-8501, Japan*

ARTICLE INFO

Article history:
Received: 02 April, 2020
Accepted: 19 June, 2020
Online: 12 July, 2020

Keywords:
Circuit Breaker
Capacitive Current Switching
Dual Motion
Grooved Cam

ABSTRACT

A novel design of a dual motion mechanism for a high-voltage gas circuit breaker is presented. The contact motion of the circuit breaker due to the operating mechanism increases capacitive current switching (CCS) performance. CCS is one of the interrupting duties of the circuit breaker, where high voltage is applied during the half cycle from contact separation. The dual motion mechanism drives two contacts in opposite directions from each other. Operating energy is reduced because the maximum displacement of the moving parts linked to the operating mechanism is shortened. To increase CCS performance at lower operating energies, the contact on the opposite side of the contact linked to the operating mechanism requires quick motion in the CCS period with a short displacement. The dual motion mechanism reported here is composed of two grooved cams that cross each other (double grooved cams). A pin positioned at the intersection point of the grooved cams rotates a lever linked to both contacts while changing the lever ratio that shortens the path length of the pin movement. To implement an optimized displacement curve with low operating energy and low mechanical stress while keeping the CCS performance high, a shape optimization method was developed that uses a direct search to minimize the local contact forces acting on the contact positions between the grooved cams and the pin. In order to maintain the stability of the pin in motion, a position holding part was designed by considering size of the gaps between the grooved cams and the pin. The measured displacement curve was in good agreement with the ideal one. In addition, a full-scale prototype was fabricated that successfully passed a 10,000-motion test.

1. Introduction

This paper is an extension of the work originally presented at the 2019 5th International Conference on Electric Power Equipment - Switching Technology (ICEPE-ST) [1]. The changes from the conference paper are as follows. First, in section 2, the role of a high-voltage gas circuit breaker is described in Figure 1 to clarify the relationship with the target motion mechanism. Second, in section 3, the design method of the displacement curve is described in Figure 4. The method aims at achieving reductions in the operating energy and mechanical load while maintaining high capacitive current switching (CCS) performance as is required for the high-voltage gas circuit breaker. An optimal solution is determined from the trade-off relationship between the operating energy and the mechanical load, as shown in Figure 8. Third, section 4 describes the method for optimizing the shape of the double grooved cams in detail by referring to the flowchart shown in Figure 9 and the position-holding parts for stabilizing the movements of the cams shown in Figure 14. In addition, the critical intersection angle of the double grooved cams is described in Figure 13. Furthermore, a comparison of optimization results for two different initial configurations and an analysis of the convergence process of the optimal calculation are given in Figure 15. Finally, section 5 describes the method used to verify the operation of the double grooved cams in Figure 17.

This paper contributes to design studies of general motion mechanisms by proposing the motion mechanism suitable for the

*Corresponding Author: Masanao Terada, Email: masanao.terada.br@hitachi.com

gas circuit breaker, optimizing the displacement curve that contributes to low operating energy, and providing the shape optimization method of the motion mechanism.

High-voltage circuit breakers require low operating energy for the sake of high reliability and low cost. Several manufacturers have proposed motion mechanisms to increase separation speed of contact with low operating energy [2-4]. Reference [2] describes a double-speed mechanism which moves a movable-side contact linked to the operating mechanism quickly with an auxiliary cam. Moreover, reference [3,4] describes dual motion systems that drive two contacts in opposite directions from each other. The dual motion mechanisms can shorten the moving distance of the movable-side (drive-side) contact, which speeds up the heavy drive-side moving parts. By driving a light contact on the opposite side (driven-side) parts, which is fixed in the conventional single motion system, the relative speed between the driven-side contact and drive-side contact becomes sufficiently high with lower operating energy.

CCS is one of the interrupting duties, which requires withstanding of a recovery voltage to be applied during a half-cycle from contact separation. In particular, CCS requires a high gas density inside the insulation nozzle that covers the driven-side contact and a low electric field between the contacts. To increase the gas density between contacts, the density drop inside the nozzle due to flow acceleration caused by the supersonic flow should be suppressed [5-11]. To decrease the electric field between the contacts, their separation in the period of the half-cycle from contact separation has to be increased; that is, the contacts must move at a fast speed. Reference [12] shows the optimal opening speeds for CCS depending on rated voltage. Reference [13] indicates that faster motions resulting in a reduced electric field and a longer nozzle throat length that avoids a density drop are effective ways to increase CCS performance. The first goal of this paper is to present the optimal motion properties for increasing CCS performance and the second one is to describe a motion mechanism that realizes these properties, presupposing a nozzle with a sufficiently narrow cross-sectional area between the inner nozzle wall and the driven-side contact so as not to affect the density drop.

To optimize the displacement curve for higher CCS performance, it is important to control the speed at every moment and the timing when the driven-side contact starts to move. Reference [14] shows that CCS performance is improved when the driven-side contact starts to move 3 ms before contact separation. Moreover, if the CCS performance with the optimum displacement curve is the same as that of the initial displacement curve, the operating energy with the optimum displacement curve is smaller than that with the initial displacement curve. To ensure reliability, it is also important to reduce mechanical stress. The higher the speeds of the contacts are, the higher the load acting on the dual motion mechanism becomes. This requires that the acceleration be limited at the beginning and the end of the motion. Therefore, it is desirable for a motion mechanism to be able to express any displacement curve. Accordingly, a new type of dual motion mechanism is proposed, which is composed of two arbitrarily shaped grooved cams crossing each other. A cam design method using an arbitrarily shaped spline function has been proposed [15]. Reference [16] applied this cam design method to a rotating cam of an operating mechanism. The idea for the motion mechanism involving crossing several grooved cams comes from other fields and has been embodied in refuelling devices [17].

This paper describes the design of a displacement curve on the driven-side of the high-voltage circuit breaker in a way that minimizes the operating energy and mechanical load acting on the mechanism while keeping CCS performance high. Besides, it describes an optimal design and a mounting structure for a dual motion mechanism with double grooved cams. Furthermore, the operation of the mechanism is verified with a small-sized test setup and a real-scaled prototype of the high-voltage circuit breaker.

2. Proposed Dual Motion Mechanism

Figure 1 shows the high-voltage gas circuit breaker with the dual motion mechanism. During normal operation in electric power transmission, the main drive-side and driven-side contacts are in contact with each other to conduct the normal load current. Upon interrupting a short circuit current, the drive-side main contact linked to the operating mechanism is pulled in the opening direction, and the contacts separate. Briefly, after the main contact separation, the arcing contacts separate. At this time, an arc discharge is generated between the arcing contacts. The arc is extinguished by gas blowing due to the pressure rise resulting from mechanical compression and heating by an arc in the pressure build-up chamber. The dual motion mechanism drives the driven-side arcing contact in the opposite direction of the driving direction of the drive-side arcing contact. If the relative speeds of the drive-side and driven-side contacts are similar to the speed of the drive-side arcing contact only, the operating energy can be lowered and the cost of the operating mechanism can be reduced because of the reduction in the drive-side speed.

Figure 2 shows schematic drawings of the proposed dual motion mechanism and its motion. The displacement curves of the motion in Figures 2 (a)-(d) are shown in Figure 3. The operating mechanism pulls a drive rod with a grooved cam (the movable grooved cam). Until the drive pin reaches the curve of the movable grooved cam, in the period between the beginning of the drive-side motion and timing (a), the lever does not rotate because the drive pin does not apply any driving force to the lever. Therefore, the driven rod is motionless. When the curved surface of the movable grooved cam comes into contact with the drive pin, the lever rotates and the driven rod starts to move. While the drive pin moves along the stationary grooved cam, which is cut into a guide plate, the lever continues to rotate, changing the lever ratio of the driven-side arm length to drive-side arm length. When the drive pin reaches the end of the stationary grooved cam, the motion of the driven rod is finished because the lever stops rotating. After timing (d), only the drive rod moves because the drive pin does not apply any driving force to the lever.

This dual motion mechanism with double grooved cams has three advantages. The first one is that an arbitrarily shaped displacement curve can be expressed. The second one is that the driving force of the drive pin can be converted efficiently into rotational force of the lever because the drive pin's moving direction matches the vertical direction of the lever for adjusting the shape of the cam curves. The third one is that the overall size of the dual motion mechanism can be reduced because the path

length of the pin movement is shortened due to the lever ratio changing during the lever rotation.

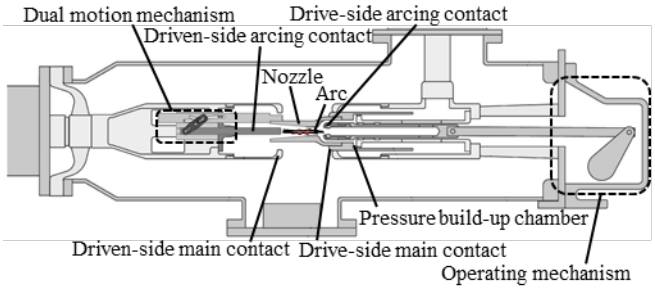


Figure 1: Overview of high-voltage gas circuit breaker with dual motion

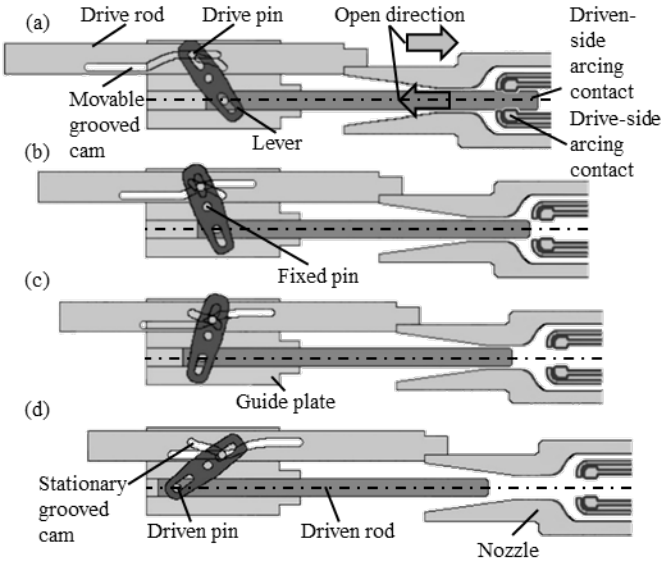


Figure 2: Schematics of the proposed dual motion mechanism

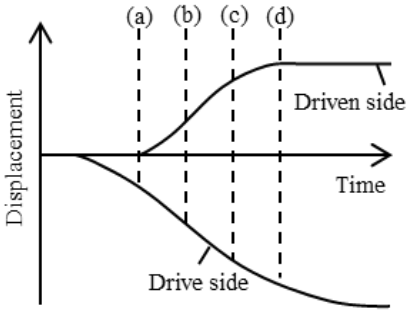


Figure 3: Drive-side and driven-side Displacement curves

3. Optimization of Driven-side Displacement Curve

To minimize the operating energy and reduce the mechanical load while keeping CCS performance high, it is desirable for optimizing the driven-side displacement curve.

3.1. Method of Designing Driven-side Displacement Curve

The steps in the design method for the driven-side displacement curve is shown in the flowchart of Figure 4 and described in detail here.

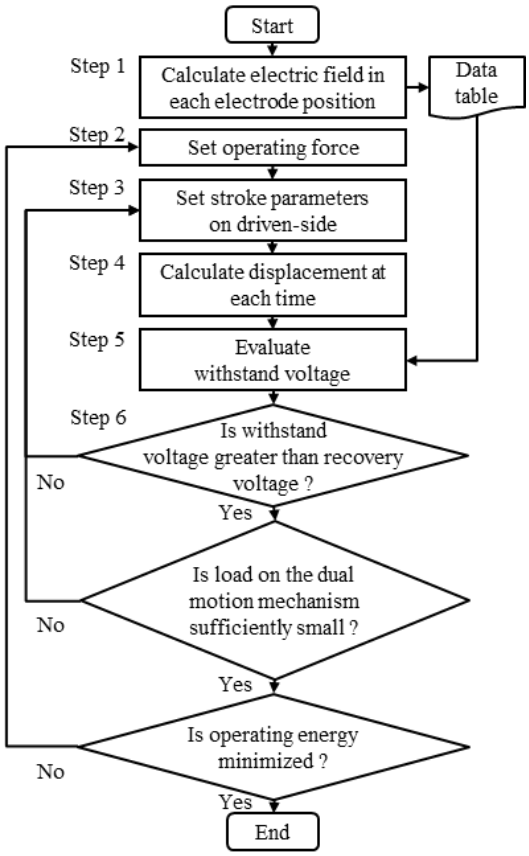


Figure 4: Design flow of displacement curve

Step 1: Calculate the electric field with the specified pitch from the position of the contact separation to the position of the half-cycle of applied voltage with power frequency, based on the electrode shape needed to withstand the lightning impulse voltage in the open position. The calculation results are saved as a data table.

Step 2: Set the operating force of the operating mechanism, which is defined by a spring force at the start position of the drive-side motion. The operating energy is calculated from the spring forces at the beginning and the end of the drive-side motion and the drive-side maximum displacement. The operating energy is proportional to the operating force which is calculated with the spring constant.

Step 3: Set the parameters of the displacement curve on the driven side in Figure 5 (a). Here, an intermittent motion with straight sliding parts of the movable grooved cam enables the start timing of the driven-side motion to be adjusted. The start timing and the end timing of the driven-side motion are the first and n-th motion points, respectively. The start timing and the end timing of the drive-side motion are the 0-th and $(n+1)$-th motion points, respectively. To set an arbitrarily time-varying displacement curve, the ratio of the speed of the driven side to that of the drive side is expressed with a Catmull-Rom spline of degree 3 [18] in Figure 5 (b). The control points of the spline curve are the design parameters. The speed ratio $l(t)_j$ between the j-th and $(n+1)$-th control points with the Catmull-Rom spline curve is expressed as follows,

$$l(t)_j = \left(-\frac{1}{2}l_{j-1} + \frac{3}{2}l_j - \frac{3}{2}l_{j+1} + \frac{1}{2}l_{j+2}\right)t^3 + \left(l_{j-1} - \frac{5}{2}l_j + \frac{5}{2}l_{j+1} + \frac{1}{2}l_{j+2}\right)t^2 + \left(-\frac{1}{2}l_{j-1} + \frac{1}{2}l_{j+1}\right)t + l_j, \quad (1)$$

where l_j is the j-th control point of the speed ratio and t is a real number in the range from 0 to 1. When the width of the drive-side displacement is ΔX_d, the i-th speed ratio L_i between l_4 and l_5 in Figure 5 (c) can be expressed as follows,

$$L_i = \alpha \frac{(x(t_{k+1})_4 - X_i)l(t_k)_4 + (X_i - x(t_k)_4)l(t_{k+1})_4}{x(t_{k+1})_4 - x(t_k)_4}, \quad (2)$$

where X_i is the i-th drive-side displacement, $x(t_k)_4$ is the k-th drive-side displacement between x_4 and x_5, $l(t_k)_4$ is the k-th speed ratio between l_4 and l_5. α is a coefficient to adjust the total amount of driven-side displacement, which is divided by the sum of L from the starting to the end of the displacement. The expression for α is

$$\alpha = \frac{X_{dn,n}}{\sum_{i=1}^{n} L_i \Delta X_d}. \quad (3)$$

Step 4: Evaluate the time-variant displacement on the drive side and driven side given by the operating force F from the law of conservation of energy,

$$\frac{1}{2}M_d {V_d}^2 + \frac{1}{2}M_{dn}{V_{dn}}^2 = \int F \cdot V_d dt\,, \quad (4)$$

where V_d and V_{dn} are the drive-side speed and driven-side speed, respectively. M_d and M_{dn} are the drive-side mass and driven-side mass, respectively. V_{dn} is the product of V_d and L. The load acting on the dual motion mechanism is calculated from the driven-side mass and the acceleration obtained from Equation (4).

Step 5: Evaluate the withstand voltage V_{BD} at each contact position with the displacements of both sides obtained from Step 3 and the electric field at each position with the data table made in Step 1. V_{BD} is calculated using Equation (5) [19],

$$V_{BD} = 100 \cdot \frac{AP^n}{E_0}, \quad (5)$$

where E (%/mm) is the calculated value, P (MPa) is absolute pressure, and A and n are coefficients that vary in accordance with the gas condition and the shape of the gas circuit breaker.

Step 6: Optimize the parameters of the displacement curve on the driven side until the withstand voltage is greater than the recovery voltage and the load acting on the dual motion mechanism is sufficiently small.

Step 7: Minimize the operating force.

3.2. Design Example of Displacement Curve

The design parameters are the controlled points of the speed curve on the driven side. The calculation conditions are as follows:

- Frequency: 60 Hz

- Gas pressure: 0.6 MPa

- Maximum displacement ratio of driven- to drive-side: 0.28

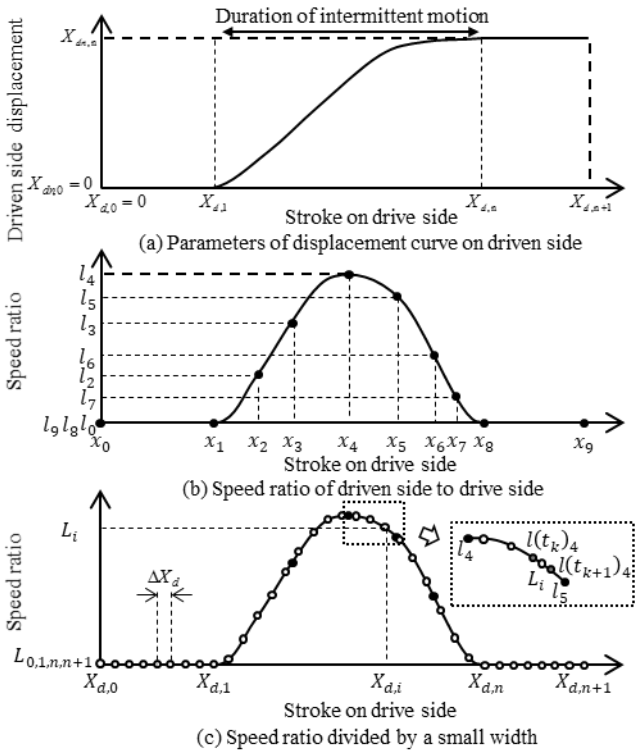


Figure 5: Expression of driven-side displacement curve

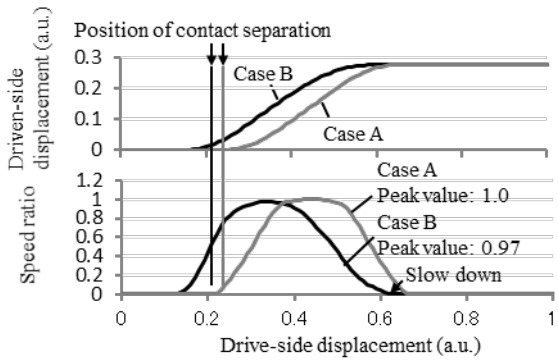


Figure 6: Input conditions for case A and case B: displacement curve (upper graph) and speed ratio (lower graph).

Figure 6 shows the input conditions for case A and case B. In case A, the movement starts at the same time as contact separation. In case B, the movement starts before contact separation, and the maximum speed ratio and acceleration in the latter half of the driven-side displacement decrease in comparison with case A. Figure 7 shows the results of the calculation. The operating energy in case B is lower than that in case A. The withstand voltages against recovery voltage for CCS in case A and case B are similar. This is because the driven-side speed at contact separation in case B is higher than in case A with the same operating energy. The mechanical load acting on the driven side in case B is lower than in case A because of the lower operating energy. To show that case B is the optimum solution, Figure 8 illustrates the relationship between the operating energy and the mechanical load. The driven-side speeds of the plotted squares are 1.2 times and 1.5 times faster than in case A, with the condition of the same timing of the motion. As the driven-side speed increases, the operating energy decreases

to have the same CCS performance. However, the mechanical load increases. In contrast, both operating energy and mechanical load are compatibly low in case B.

An optimized solution to improve CCS performance can be obtained by adjusting the driven-side displacement. The driven-side displacement of case B is assumed in section 4 to describe the design of the dual motion mechanism with the double grooved cams.

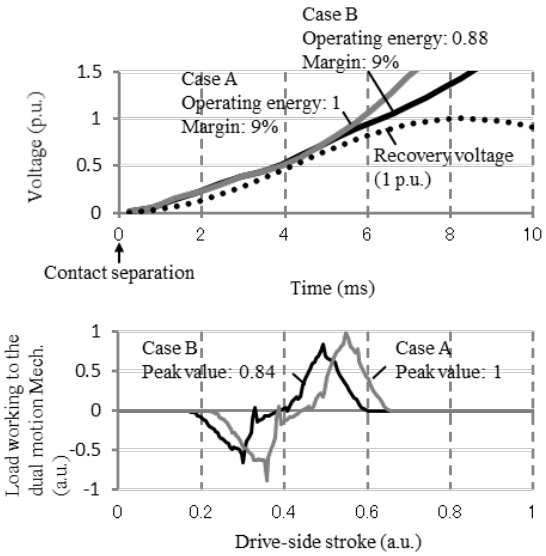


Figure 7: Results of calculation for case A and case B. The upper graph shows withstand voltages for CCS. The lower graph shows mechanical loads acting on the dual motion mechanism.

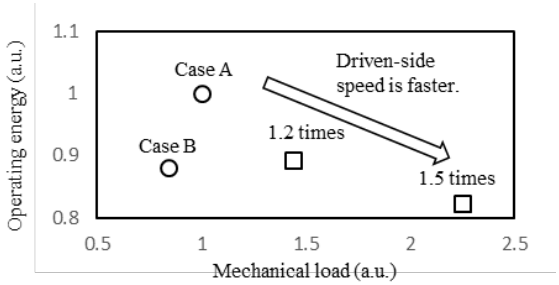


Figure 8: Relationship between operating energy and mechanical load acting on the dual motion mechanism

4. Shape Optimization of Double Grooved Cams

4.1. Method of designing the shape of double grooved cams

The dual motion mechanism with double grooved cams can be easily implemented in any shape because they have a high degree of design freedom. On the other hand, it is important to optimize the shape of the grooves because the motion mechanism must be compact and it must have mechanical strength and stability. In the following, the design method of the double grooved cams is described by referring to the flowchart shown in Figure 9.

Step 1: Input the drive-side and driven-side displacements shown in Figure 5.

Step 2: Set the dimensions for the dual motion mechanism with double grooved cams, which are the drive-side initial arm length of the link lever $L_{d,0}$, the initial lever angle θ_0, the driven-side arm length of the link lever L_{dn}, and the grooved cam width W shown in Figure 10. The end points P_0 and P_m of the stationary grooved cam are determined from these structural parameters and the driven-side maximum displacement $X_{dn,n}$.

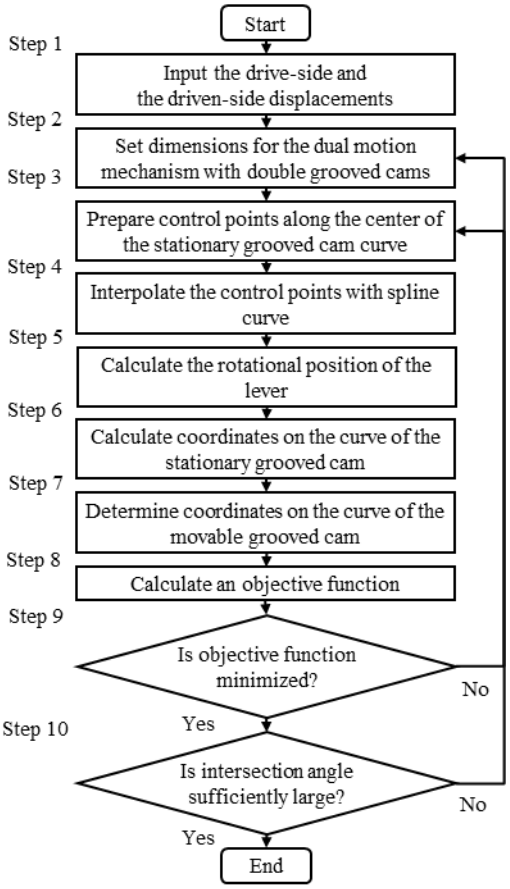


Figure 9: Design flow for shape of the dual motion mechanism with double grooved cams

Step 3: Prepare control points along the center line of the stationary grooved cam curve. P_i is the i-th control point of the stationary grooved cam ($0 \leq i \leq m$). To smooth the shape of the stationary grooved cam, the shape of the curve from P_0 to P_m must be a monotonic function. Therefore, the constraints between two neighbouring points are $P_{i,x} > P_{i+1,x}$ and $P_{i,y} > P_{i+1,y}$, where $P_{i,x}$ is the x-component of P_i and $P_{i,y}$ is the y-component of P_i.

Step 4: Interpolate the control points with a Catmull-Rom spline of degree 3. The point P between P_i and P_{i+1} is described in accordance with Equation (1), using P_{i-1}, P_i, P_{i+1} P_{i+2} and t which is a real number in the range of 0 to 1.

Step 5: Calculate the rotational positions of the link lever on the driven-side $Q_{dn,j}$, the rotational angle θ_j, and the driven-side displacement $X_{dn,j}$, shown on the left side of Figure 11 by using Equations (6) and (7),

$$Q_{dn,j,x} = S_{dn} \sin \theta_0 - X_{dn,j}, \quad (6)$$

$$Q_{dn,j,y} = -S_{dn} \cos \theta_j, \quad (7)$$

where θ_j is expressed as

$$\theta_j = \arcsin\left(\frac{S_{dn}\cos\theta_0 - X_{dn,j}}{S_{dn}}\right). \quad (8)$$

Step 6: Calculate the intersection point of the link lever and the stationary grooved cam $Q_{f,j}$ in accordance with the rotational points $Q_{dn,j}$, as shown on the right side of Figure 11. $Q_{f,j}$ is described using the k-th point $P(t_k)_i$ and k+1-th point $P(t_{k+1})_i$ between the control points P_i and P_{i+1} and a parameter α as follows,

$$Q_{f,j} = \alpha\{P(t_{k+1})_i - P(t_k)_i\} + P(t_k)_i, \quad (9)$$

$$\alpha = \frac{P(t_{k+1})_{i,x} - P(t_k)_{i,x}}{-P(t_k)_{i,x} + \tan\theta_j\, P(t_k)_{i,y}} - \tan\theta_j \frac{P(t_{k+1})_{i,y} - P(t_k)_{i,y}}{-P(t_k)_{i,x} + \tan\theta_j\, P(t_k)_{i,y}}. \quad (10)$$

Step 7: Determine the intersection point of the stationary grooved cam and the intersection point of the movable grooved cam $Q_{d,j}$, which is expressed as

$$Q_{d,j} = Q_{f,j} - X_{d,j}. \quad (11)$$

Step 8: Calculate an objective function to minimize the contact forces. In this dual motion mechanism, the drive pin is inserted at the intersection point of the stationary grooved cam and the movable grooved cam shifts while the operating mechanism pulls the movable grooved cam. Therefore, contact forces act at the points of contact between the drive pin and the surface of the grooved cams. For mechanical reliability, it is desirable for the contact force to be low. This mechanism has two possible ways to increase the contact force. The first is that the intersection angle between the double grooved cams becomes too small. The smaller the intersection angle is, the larger the contact force becomes, resulting in the possibility of a pin lock occurring during operation. This is called the wedge effect, as explained in section 4.2. The second possibility is that the contact stress becomes too large. The smaller the local curvature radius is, the larger the contact stress becomes because of the decreasing contact area. The intersection angle φ is determined to be the acute angle between the tangential line of the stationary grooved cam and the tangential line of the movable grooved cam. The local curvature radius R_s of the stationary grooved cam is determined as the radius of a circle calculated from three adjacent control points. The local curvature radius R_m is determined in the same way. φ, R_s, and R_m are shown in Figure 12. The dimensionless objective function O is expressed as

$$O = G_1\left(\frac{\varphi_t}{\min\varphi}\right) + G_2\left(\frac{R_t}{\min R_f}\right) + G_2\left(\frac{R_t}{\min R_m}\right), \quad (12)$$

where φ_t and R_t are criteria and G_1 and G_2 are weight coefficients. It is desirable for $G_1 >> G_2$ because a small intersection angle has the severe effect of increasing the contact force. The design described in section 4.3 sets G_1 to 100 and G_2 to 1. In addition, G_1 is set to zero if the intersection angle is larger than the criterion. The criterion of the intersection angle is also evaluated in section 4.2.

Step 9: Iterate the procedures of Step 3 to Step 7 to minimize the objective function. The design variables are the x and y coordinates of the control points P on the stationary grooved cam. The Rosenbrock method [20] is used for optimization, which is a type of direct search. This method leads quickly to a convergent and stable solution, but there is a possibility that a local minimum solution is reached. To avoid a local minimum solution, the best one is selected from several solutions in accordance with several initial conditions.

Step 10: Perform the calculation by changing the initial dimensions if the criterion of the intersection angle is not exceeded.

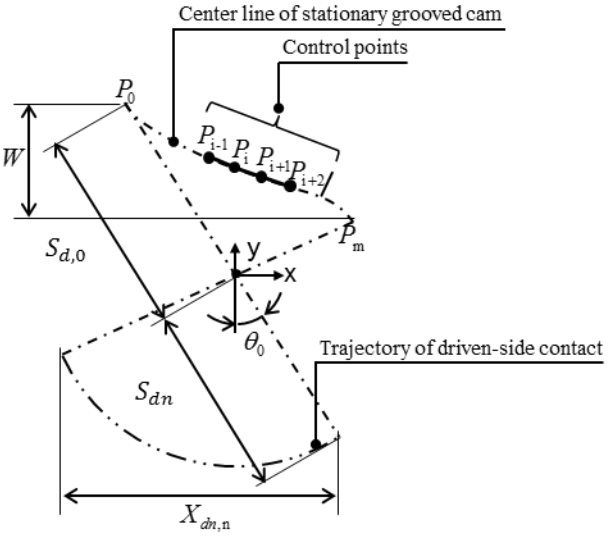


Figure 10: Structural parameters and control points for stationary grooved cam

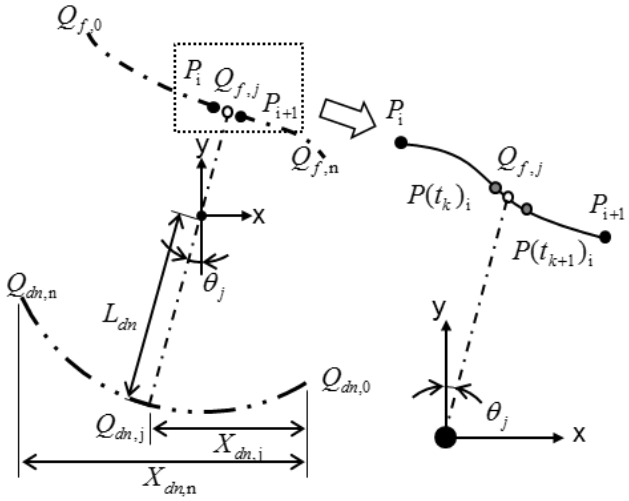


Figure 11: Moving positions on driven side and angles

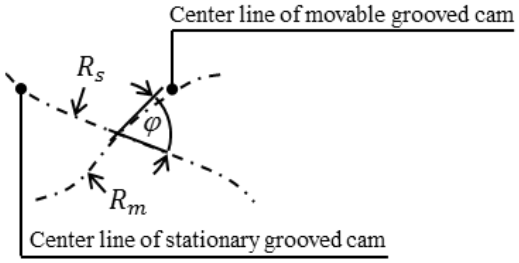


Figure 12: Intersection angle and curvature radii

4.2. Criterion of intersection angle between grooved cams

To design reliable double grooved cams, a criterion for the intersection angle of the stationary grooved cam and the movable grooved cam has to be established. The contact positions of the grooved cams and the drive pin generate nominal forces. The smaller the intersection angle, the larger the friction force induced by the nominal force. A large friction results in a pin lock due to the wedge effect during operation. This requires that the intersection angle be larger than a certain value.

Figure 13 defines the nominal forces and friction forces around the drive pin. When the lever (not illustrated) pushes the drive pin, the friction force acts on the contact position between the drive pin and the movable grooved cam. When the lever pushes the drive pin with a force F_0 (smaller than the friction force), which is the non-slip condition, the drive pin presses the upper surface of the stationary grooved cam with a force $F_0{}'$ (= F_0). Therefore, a nominal force $F_1{}'$ (=F_1) acts on the contact position between the drive pin and the stationary grooved cam, as described by the following equation,

$$F_1 = \frac{F_0}{\sin\varphi}. \tag{13}$$

Furthermore, a nominal force $F_2{}'$ (=F_2) acts on the lower surface of the movable grooved cam,

$$F_2 = \frac{F_0}{\sin\varphi}\cos\varphi = \frac{F_0}{\tan\varphi}. \tag{14}$$

The friction force F_3 is

$$F_3 = \mu F_2{}' = \mu\frac{F_0}{\tan\varphi}, \tag{15}$$

where μ is the friction coefficient. If $F_3 > F_0$, pin lock occurs during operation. Therefore, the locking condition is:

$$\tan\varphi < \mu. \tag{16}$$

If μ is 0.3, φ must be more than 17 degrees. The criterion φ_t is decided with some margin for mechanical reliability.

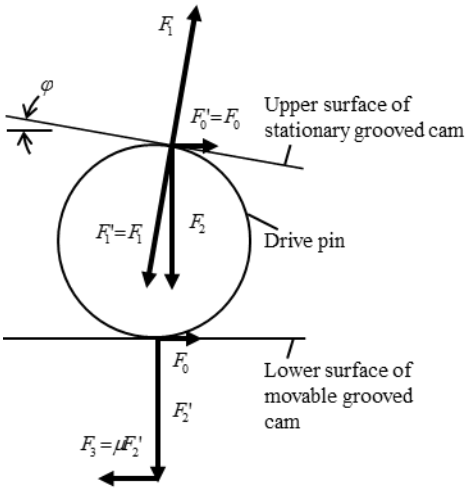


Figure 13: Contact forces and nominal forces on each contact surface

4.3. Method of structural design of double grooved cams

A method for designing the pin position holding part is presented for stabilizing the dual motion of the cam mechanism described in section 4.1. Figure 14 is a schematic drawing for distortion of the lever position caused by inclination of the drive pin. Pin rotation around the vertical axis in relation to the front view of Figure 14 caused by backlash in the grooved cams is considered. The inclination of the drive pin is defined by δ. A small inclination of the drive pin occurs due to backlash between the other pins and holes. However, when δ is greater than the backlash described above, a force resulting from the inclination of the drive pin is transmitted to the lever. Accordingly, the torque between the drive pin and the lever distorts the fixed pin and the driven pin and increases their stress. Finally, when the stress greatly exceeds the yield point, the pins deform plastically, resulting in destruction of the parts of the mechanism. Therefore, the inclination of the drive pin must be reduced. To reduce the inclination, pin holding parts are assembled at both ends of the pin. Even if the drive pin is inclined, the mechanism will not be destroyed because the pin holding parts touch the surface of the lever and force it to an ordinary position. δ and the distortion between the pin holding part and the surface of the lever ΔL_p are described as

$$\delta = L_p\sin\psi\,, \tag{17}$$

$$\Delta L_p = \frac{L_p}{2}(1-\cos\psi) + \frac{D_h}{2}\sin\psi\,, \tag{18}$$

where ψ is the drive-pin inclination angle, D_h is the outer diameter of the pin holding part, and L_p is the distance between the ends of the pin holding parts. Once ψ is derived from Equations (17) and (18), ΔL_p is described as follows:

$$\Delta L_p = \frac{L_p}{2} - \sqrt{\left(\frac{L_p}{2}\right)^2 - \left(\frac{\delta}{2}\right)^2} + \frac{D_h\delta}{2L_p}. \tag{19}$$

The threshold of the inclination of the drive pin δ_t is

$$\delta_t = \delta_f + \frac{L_d}{L_{dn}}(\delta_f + \delta_n)\,, \tag{20}$$

where δ_f and δ_n are inclinations of the fixed pin and the driven pin, respectively. L_d is the distance between the centers of the fixed pin and the drive pin. L_{dn} is the distance between the centers of the fixed pin and the driven pin. From Equation (19), the dimension D_h is determined to be a range such that δ is less than δ_t.

4.4. Design Example of Double Grooved Cams

The double grooved cams were designed to mitigate the contact forces acting on the contact positions between the grooved cams and the drive pin while keeping the overall size as small as possible so that the decided driven-side displacement curve in case B of Figure 6 was realized. Table 1 shows the design examples. Case I is the minimum-sized structure under a constraint such as on the diameter of the drive pin. One end of the stationary grooved cam is determined from the initial rotational angle θ_0 of the lever and the initial drive-side lever length $S_{d,0}$. The other end is determined from the width of the grooved cams W and the driven-

side maximum displacement $X_{dn,n}$. Ten control points between the ends of the stationary grooved cam are prepared. Compared with the initial intersection angle, the optimized intersection angle is 25 degrees, and it does not satisfy the criterion. To increase the intersection angle, θ_0 should be smaller and W should be wider in order to meet the requirement of inclining the stationary grooved cam curve steeply against the x-axis (Case II). Figure 15 shows the convergence process. The intersection angle after convergence is 40 degrees, much larger than the criterion, and the curvature radii are larger than the initial ones. The contact forces acting on the contact positions between the grooved cams and the drive pin are sufficiently small.

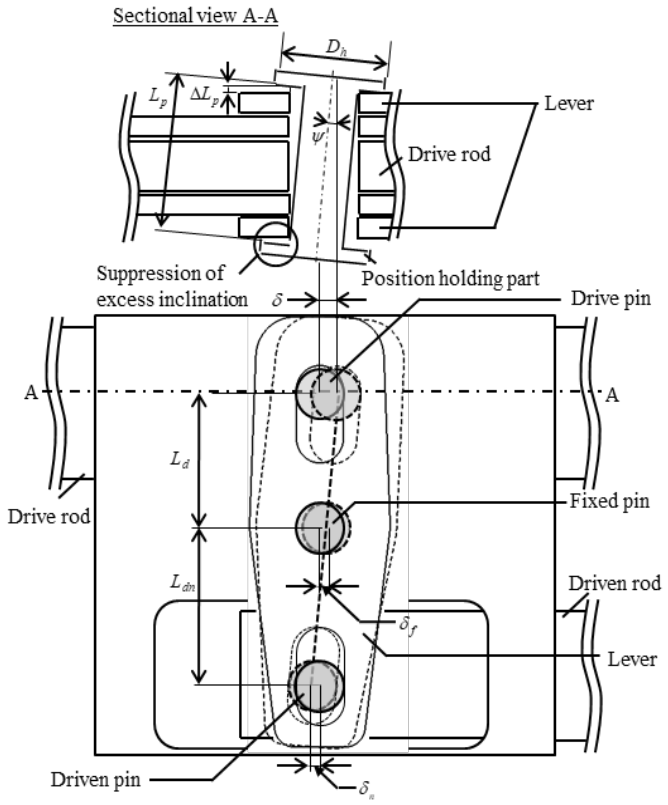


Figure 14: Schematic drawing of the double grooved cams

Table 1: Design results

Dimensions	Values			
	Case I		Case II	
	Init.	Opt.	Init.	Opt.
Lever ratio of driven-side length to drive-side length	1		1.2	
Initial angle of ling lever (degrees)	45		28.5	
Width of grooved cam (a.u.)	1		1.36	
Minimum intersection angle (degrees)	11	25	21	40
Minimum curvature radius in movable grooved cam (mm)	1	4	17	29
Minimum curvature radius in stationary grooved cam (mm)	2	2	7	13

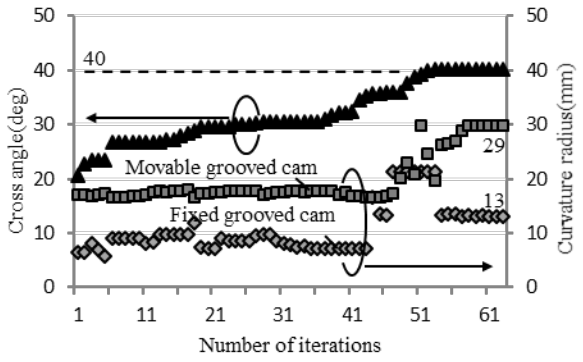


Figure 15: Convergence process

4.5. Design of pin holding parts

Figure 16 shows the relationship between the diameter of the pin position holding part and the drive-pin inclination expressed by Equation (19). As the threshold of the drive-pin inclination δ_t is 0.0033 due to the fit tolerances of the fixed pin and the driven pin, the diameter of the position holding part is 0.157, i.e., 1.96 times larger than the drive pin.

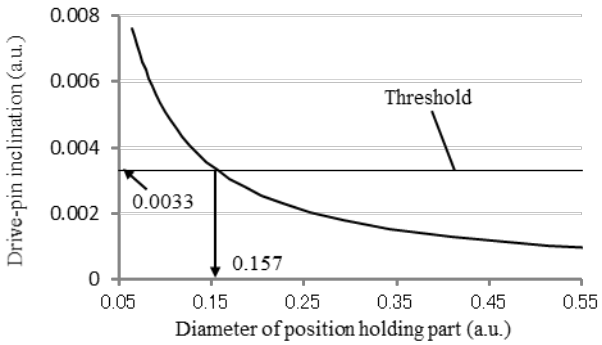


Figure 16: Drive-pin inclination versus diameter of pin position holding part

5. Motion verification of double grooved cams

Motion and endurance tests were performed to measure the mechanical reliability of the designed dual motion mechanism.

5.1. Experimental setup

Figure 17 shows the experimental setup for the motion test. An interrupter unit is set vertically. The dual motion mechanism is assembled on the top of the interrupter unit. The operating device drives the drive-side moving parts of the interrupter unit vertically. When the drive-side moving parts move downwards, i.e., in an opening motion, the driven-side moving parts move upwards. It should be noted that the direction of the downward motion is the positive direction. The drive-side maximum displacement and driven-side maximum displacement are 150 mm and 50 mm, respectively. The operating device is a combination of a 4.4-kW servomotor and a link mechanism. The maximum speed is about half of the real-scaled high-voltage circuit breaker. The upward moving speed for returning to the initial position of the opening motion is much slower than the downward moving speed. Linear

potentiometers are set on the top of the dual motion mechanism to measure the drive-side displacement and the driven-side displacement. A strain gauge is attached on the guide plate to measure strain near the fixed pin where the maximum load acts.

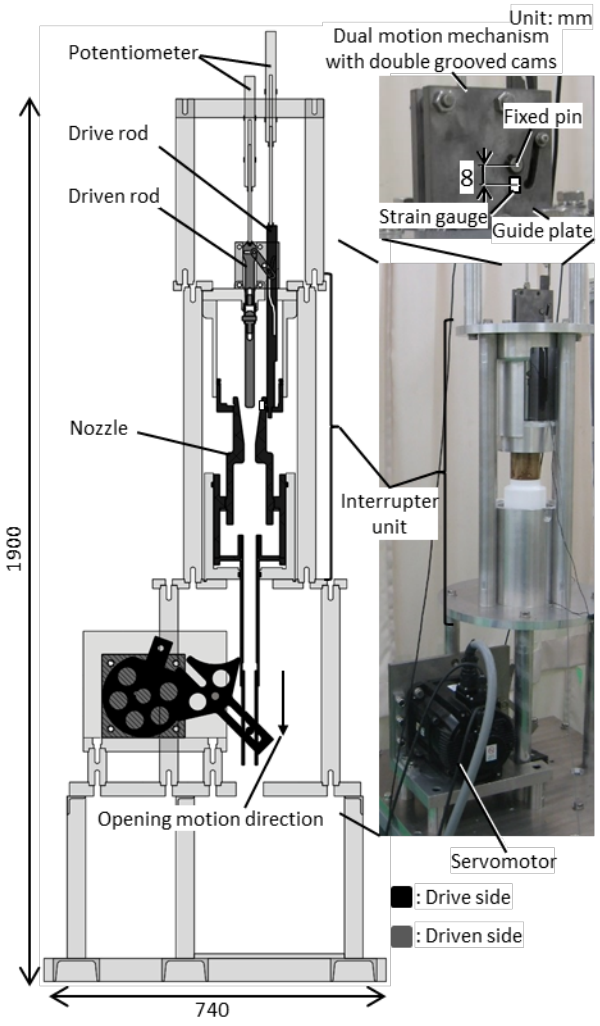


Figure 17: Experimental setup of motion test

5.2. Motion test results

Figure 18 compares the measured and calculated displacement curves. The results indicate that the displacement curve is in good agreement with the ideal displacement curve from 0 mm to -30 mm on the driven-side. In this period, the drive pin moves smoothly. After that, the deviation from the ideal curve is large, especially around the end of the driven-side motion. This is because the drive pin or the driven pin bends with elastic deformation due to inertia during the driven-side motion. Figure 19 shows the measured strain. Compressive strain occurred due to the reaction force at the start of operation on the driven-side. In the final stage of the operation on the driven-side, tensile strain occurs as a result of the upward force by inertia, and then compressive strain occurs as a result of downward force by gravity. The strains at the beginning and the end of the drive-side operation are almost 0%. This means that there is no residual strain, therefore the movement is not locked and it moves smoothly. Figure 20 shows deviations from the mean value of the maximum compressive strain on the guide plate in a 2,000-motion test. Since all values fall within the range of 5%, the operation is stable and the dual motion mechanism with the double grooved cams is mechanically reliable.

In addition, an endurance test was performed on a full-scale prototype of the high-voltage circuit breaker. The prototype successfully passed a 10,000-motion test.

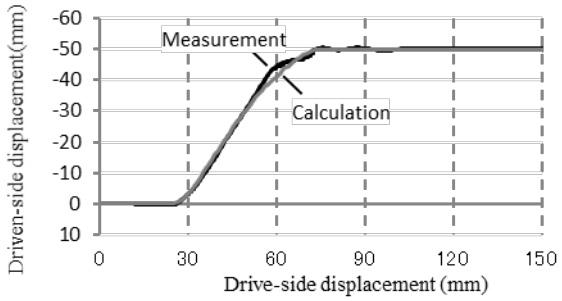


Figure 18: Comparison of measured and designed displacements

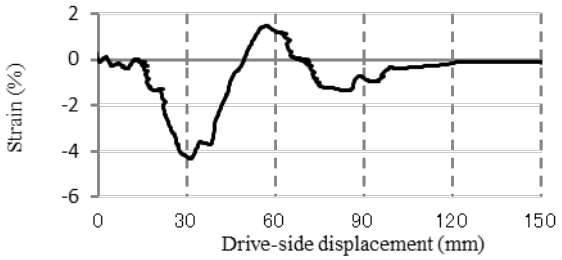


Figure 19: Tensile and compressive stress on the guide plate

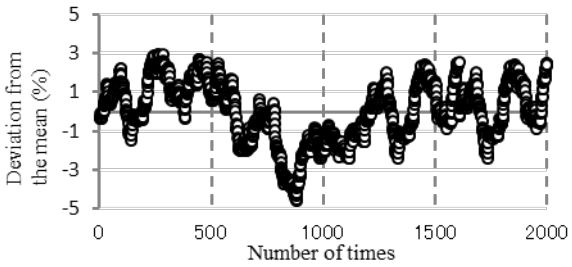


Figure 20: Deviation from the mean value of maximum compressive stress

6. Conclusion

The dual motion mechanism with double grooved cams that cross each other was proposed. A pin positioned at the intersection point of the grooved cams rotates a lever linked to both contacts while changing the lever ratio that shortens the path length of the pin movement. The proposed mechanism has three advantages. The first one is that an arbitrarily shaped displacement curve can be expressed. The second one is that the driving force of the pin can be converted efficiently into rotational force of the lever. The third one is that the overall size of the dual motion mechanism can be reduced. The following conclusions were obtained about: optimization of a displacement curve for the high-voltage circuit breaker, the shape of the grooved cams, and the mounting structure for the dual motion mechanism with double grooved cams.

- A displacement curve of the driven-side contact on the opposite side of the drive-side contact linked to the operating mechanism for the high-voltage circuit breaker was optimized,

as the driven-side contact operates quickly only during the capacitive current switching (CCS) period and starts to move before contact separation. The displacement curve can minimize the operating energy and mechanical load acting on the mechanism while keeping the CCS performance high.

- The optimal shape of the double grooved cams designed using the direct search algorithm can minimize the local contact forces acting on the contact positions between the grooved cams and the pin. Also, a position holding part designed in accordance with the gaps between the grooved cams and the pin can maintain the stability of the pin in motion.
- The measured displacement curve was in good agreement with the calculated displacement curve. In addition, since the variation of each measured strain in the 2,000-motion test is small, the operation is stable and the mechanism has high mechanical reliability. Furthermore, a full-scale prototype of the high-voltage circuit breaker successfully passed the 10,000-motion test.

References

[1] M. Terada, Y. Nakai, H. Hashimoto, D. Ebisawa, H. Urai, Y. Yokomizu, "Design of Dual Motion Mechanism Moving along Optimized Stroke Curve to Improve Capacitive Current Switching Performance for Gas Circuit Breaker" in 2019 5th International Conference on Electric Power Equipment - Switching Technology (ICEPE-ST), Kitakyushu, Japan, 2019. https://doi.org/10.1109/ICEPE-ST.2019.8928703

[2] A. Ahmethodzic, R. P. P. Smeets, V. Kertesz, M. Kapetanovic, K. Sokolija, "Design Improvement of a 245-kV SF6 Circuit Breaker With Double-Speed Mechanism Through Current Zero Analysis," IEEE Transaction on Power Delivery, 25(4), October, 2010. https://doi.org/10.1109/TPWRD.2010.2057262

[3] O. Hunger, "Circuit breaker with a gear having a dead point," U.S. Patent 8415578, 2013.

[4] H. Dienemann, V. Lehmann, H. Marin, "High Voltage Circuit Breaker with Two Arcing Contacts Which Can Be Actuated in an Opposite Direction," U.S. Patent 6271494, 2001.

[5] J. Y. Trepanier, M. Reggio, Y. Lauze, R. Jeanjean, "Analysis of the Dielectric Strength of an SF6 Circuit Breaker," IEEE Transactions on Power Delivery, 6, No. 2, April, 1991. https://doi.org/10.1016/j.aasri.2014.05.029

[6] H. Joshi, A. Pandharker, G. Patil, "Optimization of High Voltage Arc Assist Interrupters," International Journal of Scientific & Engineering Research, 4(3), March, 2013. https://doi.org/10.1109/TPWRD.2003.817536

[7] Q. Zhang, J. Liu, J. D. Yan, "Flow Structure Near Downstream Electrode of a Gas-Blast Circuit Breaker," IEEE Transactions on Plasma Science, 42(10), October, 2014. https://doi.org/10.1109/TPS.2014.2309174

[8] J.D. Mantilla, C.M. Franck, M. Seeger, "Measurements and Simulations of Cold Gas Flows in High Voltage Gas Circuit Breakers Geometries," Electrical Insulation 2008. ISEI 2008. Conference Record of the 2008 IEEE International Symposium, 720-723, 2008. https://doi.org/10.1109/ELINSL.2008.4570431

[9] H. K. Kim, K. Y. Park, C. H. Im, H. K. Jung, "Optimal Design of Gas Circuit Breaker for Increasing the Small Current Interruption Capacity," IEEE Transactions on Magnetics, 39(3), May, 2003. https://doi.org/10.1109/TMAG.2003.810176

[10] H. K. Kim, J. K. Chong, K. Y. Park, "Approximation Model-Assisted Optimization Technique to Improve Capacitive Current Interrupting Performance of Gas Circuit Breaker," IEEE Transactions on Magnetics, 45, No. 3, March, 2009. https://doi.org/10.1109/TMAG.2009.2012746

[11] M. T. Dhotre, X. Ye, F. Linares, P. Skarby, S. Kotilainen, "Multiobjective Optimization and CFD Simulation for a High-Voltage Circuit Breaker," IEEE Transaction on Power Delivery, 27(4), October, 2012. https://doi.org/10.1109/TPWRD.2012.2207133

[12] N. Yamada, S. Taki, N. Osawa, Y. Yoshioka, "Investigation of an Automatic Design Method of the Optimum GCB Insulating Structure and Necessary Opening Speeds for Successful Capacitive Current Switching," Proceeding of the ICEE 2004, Sapporo, Japan, 7(2), 901-906, 2004. https://www.researchgate.net/publication/267371487

[13] S. Delic, D. Beslija, D. Gorenc, A. Hajdarovic, M. Kapetanovic, "Capacitive current breaking capability estimation for a 145kV 40kA GIS circuit breaker," 2015 3rd International Conference on Electric Power Equipment – Switching Technology (ICEPE-ST), October 25-28, 2015. https://doi.org/10.1109/ICEPE-ST.2015.7368412

[14] M. Kang, K. H. Kim, T. Ohk, Y. T. Yoon, "Analysis of Insulation Coordination for Dual Motion Mechanism Gas Circuit Breaker," 3rd International conference on Electric Power Equipment – Switching Technology (ICEPE-ST), 2015. https://doi.org/10.1109/ICEPE-ST.2015.7368319

[15] D. M. Tsay, C. O. Huey, "Cam Motion Synthesis Using Spline Functions," Trans. ASME, Journal of Mechanisms, Transmissions, and Automation in Design, 110(2), 161-165, June, 1988. https://doi.org/10.1115/1.3258921

[16] J. S. Jang, J. J. Lee, J. H. Sohn, H. W. Kim, B. T. Bae, W. S. Yoo, "A Cam Profile Design of a Circuit Breaker by Using Multibody Dynamics Analysis," 2015 3rd International Conference on Electric Power Equipment – Switching Technology (ICEPE-ST), October 25-28, 2015. https://doi.org/10.1109/ICEPE-ST.2015.7368440

[17] A. Pourghodrat, C.A. Nelson, "A case study of designing a cam-follower mechanism with cycloidal motion," 13th World Congress in Mechanism and Machine Science, 2011. https://www.dmg-lib.org/dmglib/handler?docum=22419009

[18] E. Catmull, R. Rom, A class of local interpolating splines, Computer Aided Geometric Design, Academic Press, New York, 317-326, 1974.

[19] F. Endo, M. Sato, M. Tsukushi, Y. Yoshioka, K. Saito, K. Hirasawa, "Analytical prediction of transient breakdown characteristics of SF_6 gas circuit breakers," IEEE Transaction on Power Delivery, 4(3), July, 1989. https://doi.org/10.1109/MPER.1989.4310811

[20] D. M. Himmelblau, Applied Nonlinear Programming, McGraw Hill, New York, 1972.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 118-126 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Impact Assessment of the Errors in Determining the Mass and Zero Lift-Drag Coefficient on the Aircraft's Performance Data

Klyagin Viktor Anatolievich, Laushin Dmitry Andreevich*

Department Aircraft Design and Certification, Moscow Aviation Institute (National Research University), Moscow, 125993, Russia

ARTICLE INFO

Article history:
Received: 30 July, 2020
Accepted: 01 October, 2020
Online: 08 November, 2020

Keywords:
Uncertainties
Risks
Flying qualities and performances
Influence coefficients

ABSTRACT

The adoption of technical decisions in the development of a new aircraft is carried out under the condition of insufficient information on the mass - dimensional and aerodynamic characteristics that will be achieved during the project. Insufficient information on the input parameters (weight and aerodynamic characteristics) leads to uncertainties in the values of the flying qualities and performances. When creating a new aircraft, the chief designer and specialists of the design bureaus establish certain reserves (limits) for mass and aerodynamic characteristics. The allocation of reserves for the characteristics of aircraft elements is carried out, as a rule, on the basis of experience gained from previous projects, and to a certain extent subjective. At the same time, it is not possible to quantify the risk of non-fulfillment of the tactical and technical task in terms of performance characteristics. Knowing the contribution of the uncertainties of the input parameters (for the values of which a margin is allocated) to the uncertainty of the final values of the flight characteristics will allow the person who makes the decision to reasonably establish reserves in the design of the aircraft and take a more reasonable risk. In this paper, the authors describe an approach that allows for a quantitative assessment of the impact assessment of the errors in determining the aircraft's characteristics on the tactical and technical task in terms of the aircraft's performance data. Using the example of the mass and zero lift-drag coefficient, analytical dependencies are derived that allow a quantitative assessment of the effect. To verify the obtained analytical dependencies, the calculation of weight coefficients was made for several aviation complexes of operational-tactical aviation. The analysis of the sensitivity of the aircraft's performance data to the parameters under consideration is carried out.

1. Introduction

When developing new aircraft, the most important decisions that determine the future fate of the project are made at the initial stages of work in conditions of uncertainty about the impact of these decisions on the performance of the – tactical and technical task (TTT). Mistakes in the early stages of design tend to cost more money and time for their elimination at the subsequent stages of the aircraft, and sometimes on the correctness of decisions depends on the feasibility of the project as a whole.

The uncertainty of the performance of the tactical and technical assignment in terms of flight characteristics at the stage of internal and external design is associated with many factors. For example, a delay in the receipt from related enterprises of reliable data on the characteristics of aircraft systems (such as engine, target equipment, etc.), insufficient knowledge of some physical laws, which is most pronounced when new technical solutions are introduced, and others.

When predicting the level of performance characteristics of a new aircraft, they rely on indicators of its technical perfection, which will be achieved as a result of the project. These include the relative mass of the structure and systems, specific engine thrust, specific fuel consumption, aerodynamic quality, and others. With the development of aviation, these indicators are constantly improving.

The choice of a reasonable level of technical excellence, and therefore TTT in the design of a new aircraft, is an important task. Excessive overestimation of the requirements for its weight and

*Corresponding Author: Laushin Dmitry Andreevich, 125993, Moscow, Volokolamskoe shosse, 4, Moscow Aviation Institute (National Research University), Email: laushin.d@yandex.ru; maksmai33@gmail.com

aerodynamic characteristics can lead not only to non-fulfillment of the promising requirements for flight performance, but also deterioration of the flight characteristics of the designed aircraft relative to the current level already achieved due to the mismatch of the characteristics of its elements. At the same time, the use of previously developed technical solutions and technologies, in which there is no risk, does not make it possible to ensure the required increase in flight performance and, as a consequence, efficiency in the development of new generation aircraft. Therefore, at the initial design stage, when applying certain technical solutions, it is necessary to have a clear idea of the real possibilities of performing the performance characteristics specified in the TTT.

In accordance with the existing design methodology, when creating a new aircraft, the chief designer and specialists of the design bureaus establish certain reserves (limits) for mass and aerodynamic characteristics. The reason for the allocation of such reserves is that at the initial design stage there is no exact idea of the final value of mass and aerodynamic characteristics, which will be obtained upon completion of the development stage. Each aircraft has its own targets. Therefore, the characteristic for which the reserve is allocated depends on the specific type of aircraft. For example, the characteristics of a commercially available engine are known values. If the power plant is at the development stage, there is a risk of non-implementation of the TTT of the engine, which may affect the non-implementation of the TTT of the aircraft. The use of insufficiently mastered materials or technologies in the aircraft structure can lead to a significant deviation in the mass characteristics of the airframe units, etc. Therefore, the identification of the main elements that introduce uncertainty in the level of aircraft performance should be made in each case specifically.

The allocation of reserves for the characteristics of aircraft elements is carried out, as a rule, on the basis of experience gained from previous projects, and to a certain extent subjective. At the same time, it is not possible to quantify the risk of non-fulfillment of TTT in terms of performance characteristics. Knowing the contribution of the uncertainties of the input parameters (for the values of which a margin is allocated) to the uncertainty of the final values of the flight characteristics will allow the person who makes the decision to reasonably establish reserves in the design of the aircraft and take a more reasonable risk.

Approaches to the formation of the appearance of an aircraft with inaccurate determination of the initial data are reflected in many scientific works, for example [1-10]. The relevance of this issue is also confirmed by scientific works [11-15].

To assess the technical solutions adopted in the design, it is necessary to associate particular changes in various parameters (weight, aerodynamic characteristics) with a change in the aircraft flight characteristics.

2. Materials and Methods

Let's introduce the concept of the influence coefficient (IC) – a dimensionless coefficient showing the ratio of the change in function (f) to the change in input parameters (x_i) in the relative form:

$$K_{x_i\_f} = \frac{\Delta \hat{f}}{\Delta \hat{x}_i} \quad (1)$$

where $\hat{f} = \frac{\Delta f}{f_1}$, $\Delta \hat{x}_i = \frac{\Delta x_i}{x_{i_1}}$, Δf and Δx_i – uncertainty of the function and the input parameters, f_1 and x_{i_1} - expected value of the function and the input parameters.

In the absence of a relationship between the initial parameters, ICs are the coefficients of linear decomposition of the inaccuracy of determining the function from the errors in determining the parameters (in the relative form):

$$\frac{\Delta f}{f_{expect}} = \sum_{i=1}^{n} K_{x_i\_f} \frac{\Delta x_i}{x_{i_{expect}}} \quad (2)$$

The numerical value of the IC is equal to the percentage change of flying qualities and performances (FQP) when the parameter changes to 1%.

This paper assesses the influence of such parameters as mass (m) and zero lift-drag coefficient (Cx_0) on:
- maximum flying speed - V_{max};
- service ceiling - H_{max};
- specific excess power (max) - V_y^*;
- sustained load factor - n_y;
- acceleration time - t_p.

For a comprehensive study of the considered IC, the research is carried out in two directions:
1. Analytically, by deriving IC from the formula dependencies of the FQP;
2. Numerical calculations of the IC by calculating the changes in FQP with varying mass and zero lift-drag coefficient.

The analytical method makes it possible to establish the dependence of IC on specific aircraft parameters. These dependencies are derived from the flight dynamics formulas and can be applied to absolutely any aircraft with a similar flight principle. However, to represent the dependencies in a relatively simple and easy-to-use form, it is necessary to introduce some assumptions that can be used to perform the required mathematical transformations. The formulas obtained require verification, since one and the same introduced assumption can give different errors in magnitude, depending on the type of aircraft and its purpose.

The verification of the derived formulas for IC can be carried out by the calculation method. The computational method, in contrast to the analytical one, allows one to obtain the exact value of IC (without assumptions), but does not allow one to establish the dependence of IC on the aircraft parameters. The IC is calculated using the FQP calculation program detailed in section 2.

Thus, comparison of the analytical findings with numerical calculations allows to prove the accuracy of the first.

3. Calculation of the influence coefficient by analytically method

For various parameters, the degree of their uncertainty can differ significantly depending on the level of elaboration of certain elements of the aircraft. The accuracy of calculating the mass of

aircraft at the stage of the preliminary project is 6-10%, outline design 3-5%, working draft 1-2% [5].

The influence coefficient, as described above, is indicated $K_{x_{i\_f}}$. According to the IC mass for sustained load factor is indicated $K_{m\_n_y}$, IC C_{x_0} to specific excess power - $K_{C_{x_0}\_V_y^*}$, etc. We use the following notation:

$\Delta\widehat{m} = \frac{m_2 - m_1}{m_1}$ $\Delta\widehat{Cx}_0 = \frac{Cx_{0_2} - Cx_{0_1}}{Cx_{0_1}}$ – relative changes in mass and zero lift-drag coefficient.

3.1. The influence of parameters on maximum flying speed

The maximum flying speed can be expressed from the equality of thrust and drag:

$$V_{max} = \sqrt{\frac{2P_{max}}{(C_{x_0} + C_{x_i})\rho S}} \quad (3)$$

where C_{x_i} – inductive reactance coefficient, P_{max} – thrust at which maximum speed is reached, S – wing area, ρ – air density.

The inductive component of resistance when flying at maximum speed is small. In this case $C_{xi} \approx 0$ and the maximum flight speed can be calculated by the formula:

$$V_{max} = \sqrt{\frac{2P_{max}}{C_{x_0}\rho S}} \quad (4)$$

Let's get the formula for $K_{Cx_0 - V_{max}}$:

$$K_{Cx_0 - V_{max}} = \frac{\frac{V_{max2}}{V_{max1}} - 1}{\frac{Cx_{0_2}}{Cx_{0_1}} - 1} = \frac{\sqrt{\frac{Cx_{0_1}}{Cx_{0_2}}} * \sqrt{\frac{P_2}{P_1}} - 1}{\frac{Cx_{0_2}}{Cx_{0_1}} - 1}$$

Let $\tilde{c} = \frac{Cx_{0_2}}{Cx_{0_1}}$, and the coefficient $\tilde{p} = \frac{P_2}{P_1}$ will show the characteristics' flow of the altitude-speed performance (ASP) (with a change in V_{max}, the engine thrust will change). Then the influence coefficient Cx_0 to V_{max} will be equal to:

$$K_{Cx_0 - V_{max}} = -\frac{1 + \sqrt{\frac{\tilde{p}}{\tilde{c}}}}{\tilde{c} - 1} \quad (5)$$

The coefficient $\tilde{p}$ corresponds to a specific aircraft because its value is influenced by both the engine's ASP and the intake losses ($\Delta\bar{p}_{вх}$) and nozzle ($\Delta\bar{p}_c$):

$$\tilde{p} = f(\Delta\bar{p}_{вх}, \Delta\bar{p}_c, \tilde{c}) \quad (6)$$

The maximum flying speed is not always determined by the required and available thrusts. Limitations may include temperature, structural strength, stability and control of the aircraft.

As described above, it may be concluded that the biding of the coefficient $K_{Cx_0 - V_{max}}$ to the parameters of the aircraft is impractical, because it will not give a true picture of understanding the impact of the change Cx_0 to V_{max}.

3.2. The influence of parameters on service ceiling

The effect changes on the ceiling can be estimated by the formula [16]:

$$\Delta H_{max} = -6.3\Delta m\ [km] \quad (7)$$

In this case, IC will be equal to:

$$K_{m\_H_{max}} = \frac{\frac{Hmax_2 - Hmax_1}{Hmax_1}}{\frac{m_2 - m_1}{m_1}} = \frac{\frac{-6.3\Delta\widehat{m}}{Hmax_1}}{\Delta\widehat{m}} = \frac{-6.3}{Hmax_1}$$

$$\boldsymbol{K_{m\_H_{max}} = -\frac{6.3}{H_{max}}} \quad (8)$$

The value of the coefficient is related to the initial ceiling of the aircraft and does not depend on the percentage change in mass (Table 1).

Table 1: The dependence of $K_{m\_H_{max}}$ on the service ceiling.

Initial ceiling H_{max}, km	$\boldsymbol{K_{m\_H_{max}}}$	$\Delta\widehat{m}$	ΔH_{max}, %	ΔH_{max}, m
20	**-0.315**	0.01 (1%)	0.315	63
		0.05 (5%)	1.575	315
18	**-0.35**	0.01 (1%)	0.315	63
		0.05 (5%)	1.575	315
16	**-0.394**	0.01 (1%)	0.315	63

In order not to become attached to the initial ceiling, it makes sense to speak not about a relative change in H_{max}, but about the absolute. Therefore, with an increase in mass by 5% the ceiling decreases by 315 meters, 10% - 630 meters. An increase in the mass of 1% leads to a decrease in the ceiling of about 63 meters.

3.3. The influence of parameters on specific express power

In order not to become attached to the initial ceiling, it makes sense to speak not about a relative change in H_{max}, but about the absolute. Therefore, with an increase in mass by 5% the ceiling decreases by 315 meters, 10% - 630 meters. An increase in the mass of 1% leads to a decrease in the ceiling of about 63 meters. We write the specific excess power as follows [16]:

$$V_y^* = \frac{P - X}{G} V \quad (9)$$

where P – engine thrust, X – aircraft drag, G – aircraft weight, V – flight speed.

We assume that the flying speed at which the specific excess power is calculated is constant ($V_1 \approx V_2$), then $K_{m\_V_y^*}$ will be equal to:

$$K_{m\_V_y^*} = \frac{\frac{V_{y_2}^*}{V_{y_1}^*}-1}{\frac{m_2}{m_1}-1} = \frac{\frac{m_1}{m_2}\left(\frac{P-X_2}{P-X_1}\right)-1}{\frac{m_2}{m_1}-1} = \frac{\frac{m_1}{m_2}\left(\frac{P-X_1-\Delta X_i}{P-X_1}\right)-1}{\frac{m_2}{m_1}-1} = -\frac{1-\frac{m_1}{m_2}\left(1-\frac{\Delta X_i}{P-X_1}\right)}{\frac{m_2}{m_1}-1} = \left\langle t=\frac{m_2}{m_1}; K_{X_i}=1-\frac{\Delta X_i}{P-X_1}\right\rangle = -\frac{1-\frac{K_{X_i}}{t}}{t-1}$$

$$\boldsymbol{K_{m\_V_y^*}} = -\frac{\mathbf{1}+\frac{\boldsymbol{K_{X_i}}}{\boldsymbol{t}}}{\boldsymbol{t-1}} \quad (10)$$

where $t=\frac{m_2}{m_1}; K_{X_i}=1-\frac{\Delta X_i}{P-X_1}, \Delta X_i$ - increase in inductive component due to mass increase, $\Delta P = P - X$ – abundance of power.

Note that the higher ΔP, the smaller contribution to the value of the IC makes ΔX_i. With small changes in mass (in this case $\Delta X_i \to 0$) and a relativity high value of ΔP we can assume that $K_{m\_V_y^*} \approx -\frac{m_1}{m_2}$. Then at Δm=10% $K_{m\_V_y^*} \approx -0.91$, and the coefficient value without the assumptions described above will be even less.

We similarly take out $K_{C_{x_0}\_V_y^*}$:

$$K_{C_{x_0}\_V_y^*} = \frac{\frac{V_{y_2}^*}{V_{y_1}^*}-1}{\frac{Cx_{0_2}}{Cx_{0_1}}-1} = \frac{\left(\frac{P-X_2}{P-X_1}\right)-1}{\frac{Cx_{0_2}-Cx_{0_1}}{Cx_{0_1}}} = \frac{\left(\frac{P-X_1-\Delta X_0}{P-X_1}\right)-1}{\frac{Cx_{0_2}-Cx_{0_1}}{Cx_{0_1}}} = -\frac{1-\frac{\Delta X_0}{P-X_1}-1}{\frac{Cx_{0_2}-Cx_{0_1}}{Cx_{0_1}}} =$$

$$\left\langle \tilde{c}=\frac{Cx_{0_2}}{Cx_{0_1}}\right\rangle = -\frac{\Delta X_0\, Cx_{0_1}}{\Delta P\left(Cx_{0_2}-Cx_{0_1}\right)} = -\frac{\left(Cx_{0_2}-Cx_{0_1}\right)\frac{\rho V^2 S}{2} Cx_{0_1}}{\Delta P\left(Cx_{0_2}-Cx_{0_1}\right)} = -\frac{X_0}{\Delta P}$$

$$K_{C_{x_0}\_V_y^*} = -\frac{X_0}{\Delta P} \quad (11)$$

If we neglect the inductive component of the resistance, then the formula takes the form:

$$\boldsymbol{K_{C_{x_0}\_V_y^*}} = -\frac{\mathbf{1}}{\frac{\boldsymbol{P}}{\boldsymbol{X_0}}-\mathbf{1}} \quad (12)$$

The value of the coefficient does not depend on the percentage change C_{x_0} and is associated with a specific aircraft.

The max value of the specific excess power for fighters is achieved as a rule at M=0.8-0.9 near the ground. We calculate the IC value according to the formula (11) and (12) for several aircraft. At M=0.85 ram airflow is q=51250 H/m^2, the IC excluding and taking into account the inductance is presented in Table 2.

Table 2: The values $K_{C_{x_0}\_V_y^*}$ are calculated by the formula (4) and (5)

Air-craft	M	Cx_0	Cx_i	S, m^2	P, H	$\boldsymbol{K_{C_{x_0}\_V_y^*}}$ **excluding** $\boldsymbol{Cx_i}$	$\boldsymbol{K_{C_{x_0}\_V_y^*}}$ **taking into account** $\boldsymbol{Cx_i}$
Mig-29	0.9	0.025	0.0005	38	200000	**-0.398**	**-0.401**
Su-24 (69°)	0.8	0.0263	0.0029	51	230000	**-0.367**	**-0.383**
Mig-23 (72°)	0.9	0.0225	0.0026	34.2	150000	**-0.438**	**-0.460**

From Table 2 it is seen that the inductive component of the resistance makes an insignificant contribution to the value of the IC. When Cx_0 is changed by 10% the differences in the scatter of V_y^* excluding and taking into account Cx_i are about 0.2%, which can be considered a negligible value at the preliminary design stage. Therefore, to simplify, you can use the formula (5).

3.4. The influence of parameters on sustained load factor

The sustained load factor can be expressed through the formula [16]:

$$n_y = \sqrt{\frac{(P-X_0)qS}{A(mg)^2}} \quad (13)$$

where P – engine thrust, X_0 – non-inductive resistance, q – velocity head, A – drag-due-to-lift factor, g – acceleration of gravity, then:

$$K_{m\_n_{y\,уст}} = \frac{\frac{n_{y_2}}{n_{y_1}}-1}{\frac{m_2}{m_1}-1} = \frac{\frac{m_1}{m_2}-1}{\frac{m_2}{m_1}-1} = \left\langle t=\frac{m_2}{m_1}\right\rangle = \frac{\frac{1}{t}-1}{t-1} = \frac{1-t}{t(t-1)} = -\frac{1}{t}$$

$$\boldsymbol{K_{m\_n_y}} = -\frac{\boldsymbol{m_1}}{\boldsymbol{m_2}} \quad (14)$$

The value of the coefficient depends on the percentage change in mass (Table 3).

Table 3: Dependence of $K_{m\_n_y}$ on Δm

Mass change Δm, %	$\boldsymbol{K_{m\_n_y}}$	Change of the sustained load factor Δn_y, %
5	**-0.952**	*4.8*
10	**-0.909**	*9.1*
15	**-0.869**	*13*

Similarly, we get $K_{Cx_0\_n_y}$:

$$K_{Cx_0\_n_{y\,уст}} = \frac{\frac{n_{y_2}}{n_{y_1}}-1}{\frac{Cx_{0_2}}{Cx_{0_1}}-1} = \frac{\sqrt{\left(\frac{P-X_2}{P-X_1}\right)}-1}{\frac{Cx_{0_2}-Cx_{0_1}}{Cx_{0_1}}} = \frac{\sqrt{\left(\frac{P-X_1-\Delta X_0}{P-X_1}\right)}-1}{\frac{Cx_{0_2}-Cx_{0_1}}{Cx_{0_1}}} = \langle Cx_{0_2} - Cx_{0_1} > 0\rangle = \sqrt{\left(\frac{P-X_1-\Delta X_0}{P-X_1}\right)\frac{(Cx_{0_1})^2}{(Cx_{0_2}-Cx_{0_1})^2}} - \frac{1}{\Delta\widehat{Cx}_0} =$$

$$\sqrt{\left(1-\frac{\Delta X_0}{P-X_1}\right)\frac{{Cx_{0_1}}^2}{(Cx_{0_2}-Cx_{0_1})^2}} - \frac{1}{\Delta\widehat{Cx}_0} =$$

$$\sqrt{\frac{{Cx_{0_1}}^2}{(Cx_{0_2}-Cx_{0_1})^2} - \frac{qS(Cx_{0_2}-Cx_{0_1})}{qS\left(\frac{P}{Sq}-Cx_{0_1}\right)}\frac{{Cx_{0_1}}^2}{(Cx_{0_2}-Cx_{0_1})^2}} - \frac{1}{\Delta\widehat{Cx}_0} =$$

$$\sqrt{\frac{1}{(\Delta\widehat{Cx}_0)^2}-\frac{{Cx_{0_1}}^2}{(\frac{P}{Sq}-Cx_{0_1})(Cx_{0_2}-Cx_{0_1})}}-\frac{1}{\Delta\widehat{Cx}_0}=$$

$$\sqrt{\frac{1}{(\Delta\widehat{Cx}_0)^2}-\frac{Cx_{0_1}}{\left(\frac{P}{Sq}-Cx_{0_1}\right)(\Delta\widehat{Cx}_0)}}-\frac{1}{\Delta\widehat{Cx}_0}=$$

$$\sqrt{\frac{1}{(\Delta\widehat{Cx}_0)^2}-\frac{Cx_{0_1}\Delta\widehat{Cx}_0}{\left(\frac{P}{Sq}-Cx_{0_1}\right)}\frac{1}{(\Delta\widehat{Cx}_0)^2}}-\frac{1}{\Delta\widehat{Cx}_0}=\frac{1}{\Delta\widehat{Cx}_0}\sqrt{1-\frac{Cx_{0_1}\Delta\widehat{Cx}_0}{\frac{P}{Sq}-Cx_{0_1}}}-$$

$$\frac{1}{\Delta\widehat{Cx}_0}=\frac{1}{\Delta\widehat{Cx}_0}\left(\sqrt{1-\frac{Cx_{0_1}\Delta\widehat{Cx}_0}{\frac{P}{Sq}-Cx_{0_1}}}-1\right)$$

$$K_{Cx_{0\_}n_y}=\frac{1}{\Delta\widehat{Cx}_0}\left(\sqrt{1-\frac{Cx_{0_1}\Delta\widehat{Cx}_0}{\frac{P}{Sq}-Cx_{0_1}}}-1\right) \qquad (14)$$

The coefficient value slightly depends on the percentage change Cx_0 (differences – in thousandths). When Cx_0 changes by 10%, the formula takes the form:

$$\boldsymbol{K_{Cx_{0\_}n_y}=10\left(\sqrt{1-0.1\frac{Cx_{0_1}}{\frac{P}{Sq}-Cx_{0_1}}}-1\right)} \qquad (15)$$

For example, we calculate the IC for several AK. At M=0.8 at a higher H=1000m ram-air flow is equal q=41200 H/m^2, (Table 4).

Table 4: The values $K_{Cx_{0\_}n_y}$

Aircraft	Cx_0	S, m^2	P (H=1000, M=0.8), H	$\boldsymbol{K_{Cx_{0\_}n_y}}$
Mig-29	0.024	38	170000	**-0.16**
Su-24 (69°)	0.026	51	206000	**-0.21**
Mig-23 (72°)	0.022	34.16	130000	**-0.18**

3.5. The influence of parameters on acceleration time

Acceleration time in horizontal flight in within speed range is [17]:

$$t_p=\frac{1}{g}\int_{V_н}^{V_к}\frac{dV}{n_{xa}}\approx\frac{V_к-V_н}{g*(n_{xa})_{cp}} \qquad (16)$$

Where $V_н$ and $V_к$ – initial and final speed, $(n_{xa})_{cp}$ – averaged longitudinal g load for within speed range.

Let's get the formula for $K_{m\_t_p}$:

$$\Delta(n_{xa})_{cp}=(n_{xa})_{cp2}-(n_{xa})_{cp1}=\frac{(P_{cp}-X_{2_{cp}})}{G_2}-\frac{(P_{cp}-X_{1_{cp}})}{G_1}=$$

$$\frac{\frac{m_1}{m_2}P_{cp}-\frac{m_1}{m_2}X_{1_{cp}}-\frac{m_1}{m_2}\Delta X-P_{cp}+X_{1_{cp}}}{G_1}=-\frac{P_{cp}\left(1-\frac{m_1}{m_2}\right)-X_{1_{cp}}\left(1-\frac{m_1}{m_2}\right)+\Delta X\frac{m_1}{m_2}}{G_1}=$$

$$-\frac{(P_{cp}-X_{1_{cp}})\left(1-\frac{m_1}{m_2}\right)+\Delta X\frac{m_1}{m_2}}{G_1}\text{, then:}$$

$$K_{m\_t_p}=\frac{\frac{t_{p_2}}{t_{p_1}}-1}{\frac{m_2}{m_1}-1}=\left(\frac{t_{p_2}}{t_{p_1}}-1\right)\frac{1}{\Delta\hat{m}}=\left(\frac{(n_{xa})_{cp1}-(n_{xa})_{cp2}}{(n_{xa})_{cp2}}\right)\frac{1}{\Delta\hat{m}}=$$

$$\left(\frac{1}{\frac{(n_{xa})_{cp2}}{-\Delta(n_{xa})_{cp}}}\right)\frac{1}{\Delta\hat{m}}=\left(\frac{1}{-1-\frac{1}{\frac{\Delta(n_{xa})_{cp}}{(n_{xa})_{cp1}}}}\right)\frac{1}{\Delta\hat{m}}=$$

$$\left(\frac{1}{\frac{1}{\frac{(P_{cp}-X_{1_{cp}})\left(1-\frac{m_1}{m_2}\right)G_1}{(P_{cp}-X_1)G_1}+\frac{\Delta X*m_1*G_1}{m_2(P_{cp}-X_1)G_1}}-1}\right)\frac{1}{\Delta\hat{m}}=$$

$$\left(\frac{1}{\frac{1}{1-\frac{m_1}{m_2}+\frac{m_1}{m_2}\frac{\Delta X}{(P_{cp}-X_{1_{cp}})}}-1}\right)\frac{1}{\Delta\hat{m}}=\left(\frac{1}{\frac{1}{1-\frac{m_1}{m_2}\left(1-\frac{\Delta X}{(P_{cp}-X_{1_{cp}})}\right)}-1}\right)\frac{1}{\Delta\hat{m}}$$

$$\boldsymbol{K_{m\_t_p}=\left(\frac{1}{\frac{1}{1-\frac{m_1}{m_2}\left(1-\frac{\Delta X_{i_{cp}}}{\Delta P_{cp}}\right)}-1}\right)\frac{1}{\Delta\hat{m}}} \qquad (17)$$

where ΔP_{cp} and $\Delta X_{i_{cp}}$ – the average engine thrust margin and the average increase of the induced drag for the speed range acceleration.

The value of the coefficient slightly depends on the percentage change in the mass. Assuming $\Delta X_{i_{cp}}/\Delta P_{cp}=0$ (ΔX_i is significantly less than the average available thrust ΔP_{cp}) $K_{m\_t_p}$=1, and taking into account the influence of increased inductance (due to the increase in mass) the coefficient value will be slightly larger. For example, for Mig-29 during acceleration 600 – 1100 km/hour near the ground $K_{m\_t_p}=1{,}12$ (Table 5).

Table 5: The values $K_{m\_t_p}$ for Mig-29 (600 – 1100 km/hour on the deck)

m_1, kg	Δm, %	ΔC_y	ΔCx_i	P_{cp}, kgf	X_{cp}, kgf	ΔX_i, kgf	$\boldsymbol{K_{m\_t_p}}$
15000	10	0.011	0.001	16000	4000	130	**1.12**

If accept $\Delta Cx_{0_{cp}}\approx\Delta Cx_0$ and $Cx_{0_{1cp}}\approx Cx_{0_1}$, $K_{Cx_{0\_}t_p}$ takes the following form:

$$K_{Cx_{0\_}t_p}=\frac{\frac{t_{p_2}}{t_{p_1}}-1}{\frac{Cx_{0_2}}{Cx_{0_1}}-1}=\left(\frac{t_{p_2}}{t_{p_1}}-1\right)\frac{1}{\Delta\widehat{Cx}_0}=\left(\frac{1}{-1-\frac{(n_{xa})_{cp1}}{\Delta(n_{xa})_{cp}}}\right)\frac{1}{\Delta\widehat{Cx}_0}=$$

$$\left\langle\Delta(n_{xa})_{cp}=\frac{-\Delta X_0}{G_1}\right\rangle=\left(\frac{1}{\frac{P_{cp}-X_{1_{cp}}}{\Delta X_{0_{cp}}}-1}\right)\frac{1}{\Delta\widehat{Cx}_0}=$$

$$\left(\frac{1}{\frac{P_{cp}-X_{1_{cp}}-\Delta X_{0_{cp}}}{\Delta X_{0_{cp}}}}\right)\frac{1}{\Delta\widehat{Cx}_0}=\left(\frac{\Delta Cx_{0_{cp}}qS}{qS\left(\frac{P_{cp}}{qS}-Cx_{0_{1cp}}-\Delta Cx_{0_{cp}}\right)}\right)\frac{Cx_{0_1}}{\Delta Cx_0}=$$

$$\left\langle\begin{matrix}\Delta Cx_{0_{cp}}\approx\Delta Cx_0\\ Cx_{0_{1cp}}\approx Cx_{0_1}\end{matrix}\right\rangle=\frac{Cx_{0_1}}{\frac{P_{cp}}{qS}-Cx_{0_{1cp}}-\Delta Cx_{0_{cp}}}=\frac{1}{\frac{P_{cp}}{X_{0_{1cp}}}-1-\Delta\widehat{Cx}_0}$$

$$K_{Cx_0\_t_p} = \frac{1}{\frac{P_{cp}}{X_{0_{1cp}}} - 1 - \Delta\widehat{Cx_0}} \quad (18)$$

The coefficient value slightly depends on the percentage change Cx_0 and is associated with a specific aircraft. For Mig-29 $K_{Cx_0\_t_p}$ = 0.27 (Table 6).

Table 6: The values $K_{Cx_0\_t_p}$ for Mig-29 (600 – 1100 km/hour near the ground)

Cx0	Δ Cx0, %	Pcp, kgf	$X_{0_{1cp}}$, kgf	$K_{Cx_0\_t_p}$
0.025	10	16000	3300	**0.27**

4. Calculation of the influent coefficients by numerical method

To confirm the derived dependencies (8), (10), (12), (14), (15), (17), (18), we calculate the IC using the "Program for preparing the initial data and calculating the FQP". This software product was developed by Sukhoi Design Bureau and verified by many years of the application practice. The program uses weight, aerodynamic characteristics (Cx_0(M), Cx_i (C_y, M), Cy_{max}(M), C_y^a) and propulsive characteristics P (H, M, Dr) Qc(H, M, Dr), as well as some limitations (q_{max}, ny_{max}) and constants. The graphs Cx_0(M), Cx_i (C_y, M), Cy_{max}(M) C_y^a, P (H, M), Qc(H, M) are digitized from books of practical aerodynamics [18-21]. ICs are calculated according to the algorithm shown in the Figure 1.

We calculate IC for the following aircraft of the class of operational - tactical aviation: fighter 4th generation Su-27; frontline bomber Su-24; strike-fighter Su-25; fighter 4th generation Mig-29; fighter 3rd generation Mig-23; interceptor jet 3rd generation Mig-25.

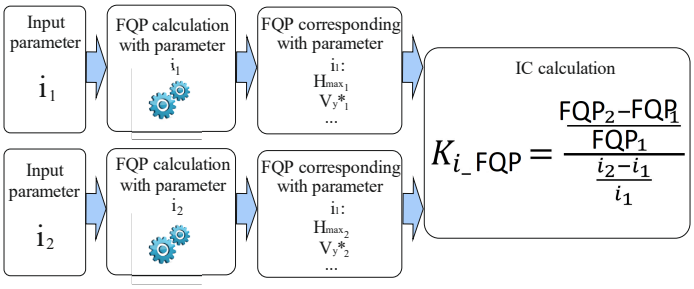


Figure 1: IC calculation scheme

Table 9 and Figures 2-11 show the results of calculating the IC programmatically. ICs were obtained when the mass changed by 10%, Cx_0 changed for M<1 by 10% and for M>1 by 18%. Such differences in the spread of Cx_0 are taken on the assumption that the inaccuracy of determining this parameter for the supersonic domain of flight is higher. Aircraft characteristics are taken in the calculations are presented in Table 7.

The effect on the service ceiling is shown in Figures 2 and 3, on the specific excess power - in Figures 4 and 5, on the sustained load factor - in Figures 6 and 7, for the acceleration time - in Figures 8 and 9. Figures 2-9 also show the values obtained from the above formulas.

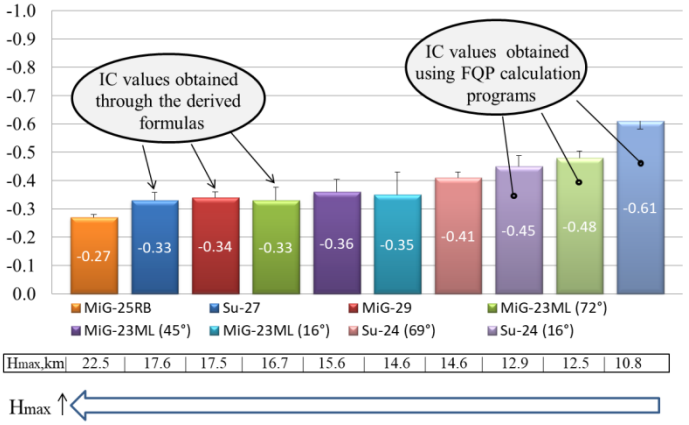


Figure 2: $K_{m\_H_{max}}$ values

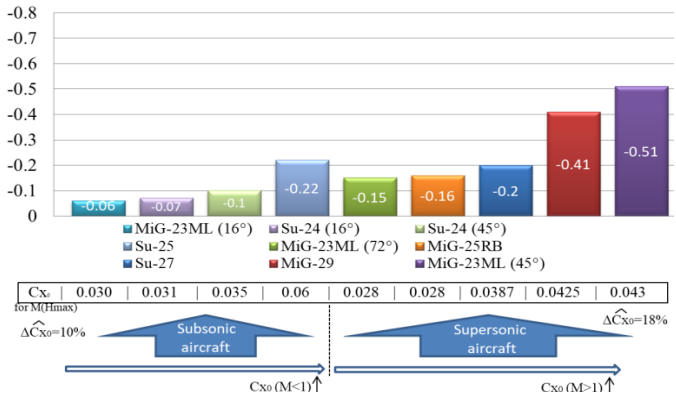
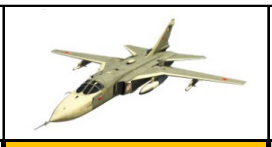


Figure 3: $K_{Cx_0\_H_{max}}$ values

Table 7: Aircraft characteristics, used in the calculations

Aircraft	MiG-29	MiG-25RB	MiG-23ML (72°)	MiG-23ML (45°)	MiG-23ML (16°)	Su-27	Su-25	Su-24 (69°)	Su-24 (45°)	Su-24 (16°)
Weight										
Mass in calculations, kg	15000	24000	18000	18000	18000	21200	13000	25000	25000	25000
Wing parameters										
Wing area, m^2	38	61.5	34.16	35.3	37.27	62	30.1	51	53	55.18
Wing extension	3.39	2.94	1.77	3.43	5.26	3.5	6	2.11	3.9	5.64
Leading edge a sweep angle, deg.	42	41	74	47	18	42	20	69	45	16
Power plant specifications										
Jet engine	2xRD-33	2xR-15-300	R-35	R-35	R-35	2xA1-31F	2xR-95 SH	2xA1-21 F-3	2xA1-21 F-3	2xA1-21 F-3
Thrust (H=0, M=0), kgf	16000	22400	12500	12500	12500	25000	8200	22400	22400	22400
Aircraft design parameters										
Starting thrust-to weight ratio	1.07	0.93	0.69	0.69	0.69	1.18	0.63	0.90	0.90	0.90
Wing load, kg/m^2	395	390	527	510	483	342	432	490	472	453

Table 8: IC for the studied aircrafts

										
	Su-27		**MiG-29**		**MiG-23ML (72°)**		**MiG-23ML (45°)**		**MiG-23ML (16°)**	
FQ&P	**m**	**Cx0**	**m**	**Cx0**	**m**	**Cx0**	**m**	**Cx0**	**m**	**Cx0**
Vmax	-0.07	-0.81	-0.09	-0.71	-	-	-0.14	-0.85	-0.01	-0.17
Hmax	-0.33	-0.2	-0.34	-0.41	-0.33	-0.15	-0.36	-0.51	-0.35	-0.06
Vy*	-0.91	-0.37	-0.91	-0.47	-1.02	-0.47	-0.98	-0.53	-0.91	-0.45
ny	-0.91	-0.09	-0.91	-0.09	-0.91	-0.11	-0.91	-0.1	-0.91	-0.09
tp	1.09	0.26	1.08	0.27	1.22	0.24	1.11	0.28	1.02	0.35
										
	MiG-25RB		**Su-25**		**Su-24 (69°)**		**Su-24 (45°)**		**Su-24 (16°)**	
FQ&P	**m**	**Cx0**	**m**	**Cx0**	**m**	**Cx0**	**m**	**Cx0**	**m**	**Cx0**
Vmax	-	-	-0.02	-0.24	-	-	-0.02	-0.18	-0.02	-0.14
Hmax	-0.27	-0.16	-0.61	-0.22	-0.41	-	-0.48	-0.1	-0.45	-0.07
Vy*	-0.96	-0.43	-0.94	-0.46	-0.98	-0.39	-0.96	-0.45	-0.94	-0.41
ny	-0.91	-0.21	-0.91	-0.26	-0.91	-0.19	-0.91	-0.14	-0.91	-0.18
tp	1.06	0.35	1.05	0.35	1.19	0.31	1.11	0.22	1.06	0.30

- subsonic aircraft

Sustained load factor on high 1000 m M=0.8 (for Su-25 M=0.6)

Acceleration time on the deck 600-1100 km/hour (for Su-25 500-800 km/hour)

Aircrafts images courtesy of Google images

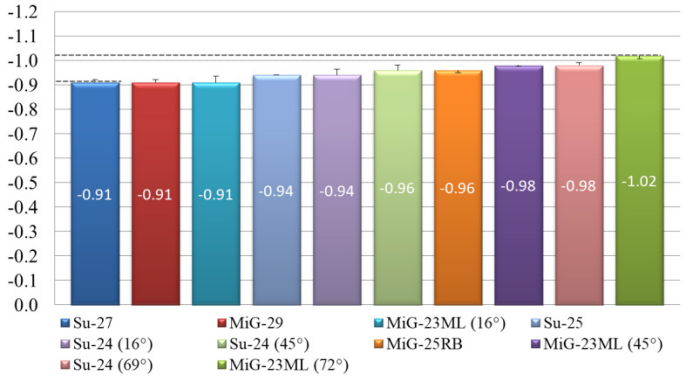


Figure 4: $K_{m\_V_y^*}$ values

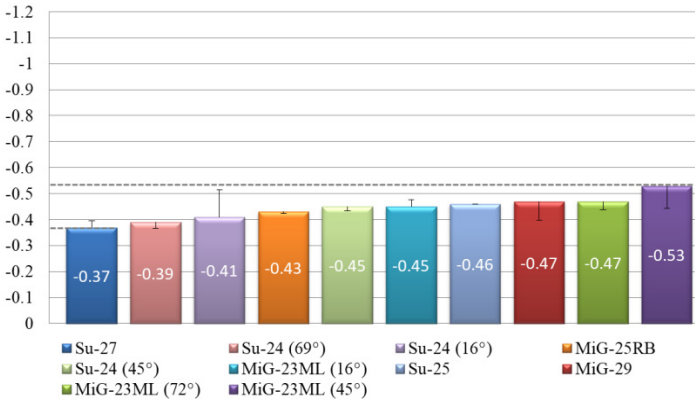


Figure 5: $K_{C_{x_0}\_V_y^*}$ values

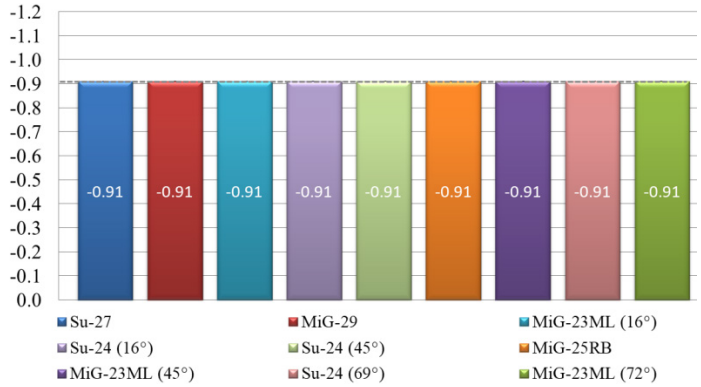


Figure 6: $K_{m\_n_y}$ values (H=1000m, M=0.8)

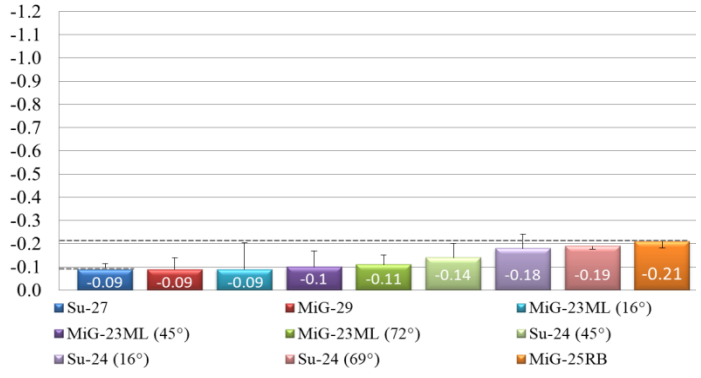


Figure 7: $K_{C_{x_0}\_n_y}$ values (H=1000m, M=0.8)

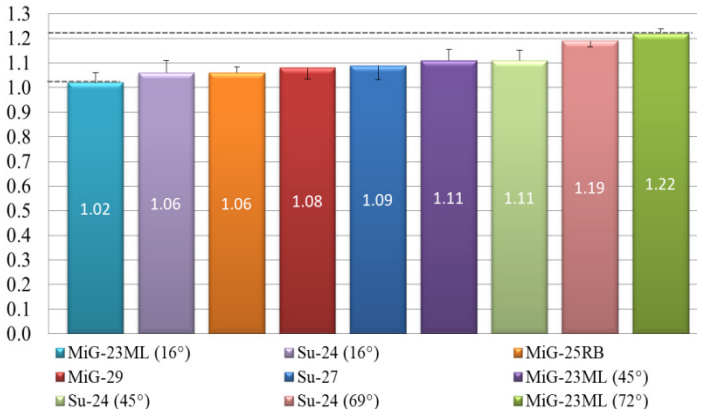


Figure 8: $K_{m\_t_p}$ values

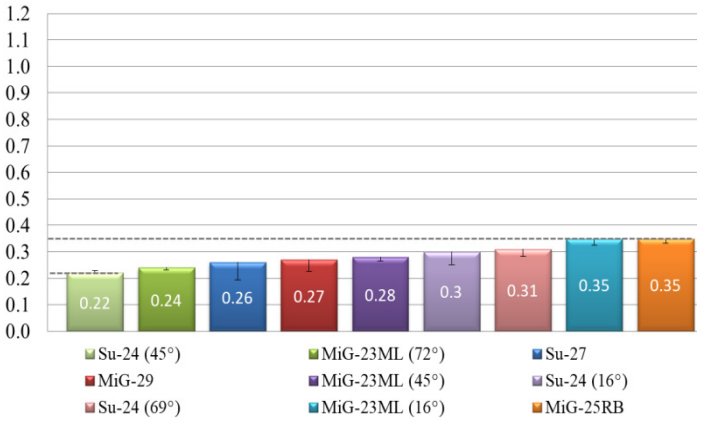


Figure 9: $K_{C_{x_0}\_t_p}$ values

5. Results and Discussions

As can be seen from Figures 2-10, the analytically obtained ICs are confirmed by numerical calculations. The absolute error when using the derived formulas is less than 0.1 of the IC value, which in terms of FQP values gives an error of less than 10%. At the initial stage of aircraft design (when the error of other constituent elements is comparable in magnitude), this is an acceptable indicator.

Summarize the calculated IC in Table 9, the percentage change in FQP is presented in Table 10.

Table 9: IC ranges for the considered OTA aircrafts

FQP	K_{m_FQP}	$K_{Cx_0\_FQP}$	
		$M_{max}<1$	$M_{max}>1$
V_{max}	-0.1	-0.1…-0.25	-0.7…-0.85
H_{max}	0.25…0.6	-0.05…-0.2	-0.15…-0.5
$V_y^* max$	-0.9….-1.05	-0.35…-0.55	-
n_y (H=1000m, M=0.8)	-0.91	-0.05…-0.2	-
t_p (600-1100 km/h, H=200m)	1…1.25	-0.2…-0.35	-

Table 10: Percent change in the FQP

FQP	With an increase in mass by 10%, the FQP deteriorates by:	With an increase in Cx_0 in the subsonic area by 10% and by 18% in supersonic the FQP deteriorate by:	
		$M_{max}<1$	$M_{max}>1$
V_{max}	1%	1…2.5%	12…15.5%
H_{max}	decreases by 630 m	0.5…2%	3…8%
$V_y^* max$	9….10.5%	3.5…5.5%	-
n_y (H=1000m, M=0.8)	9.1%	0.5…2%	-
t_p (600-1100 km/h, H=200m)	10…12.5%	2…3.5%	-

From Tables 9 and 10 we can conclude that the considered FQP, such as the maximum specific excess power, sustained load factor, and acceleration time, are most sensitive to a change in mass. The inaccuracy in determining the mass in 1% gives a spread of these FQP in about 1% (more accurate values are given earlier). The IC of the mass on the practical ceiling depends on the absolute value of the ceiling, so it makes sense to talk about its absolute change. With an inaccuracy of mass of 1%, the ceiling is determined with an accuracy of about 60-70 meters.

To inaccuracy C_{x_0} is most sensitive to the maximum flying speed (for M>1). The numerical value of the IC on V_{max} is determined by the curve crossing nature of the required and available thrust (Figure 11). Graphs Cx_0 (M) for the studied aircrafts are shown in Figure 10.

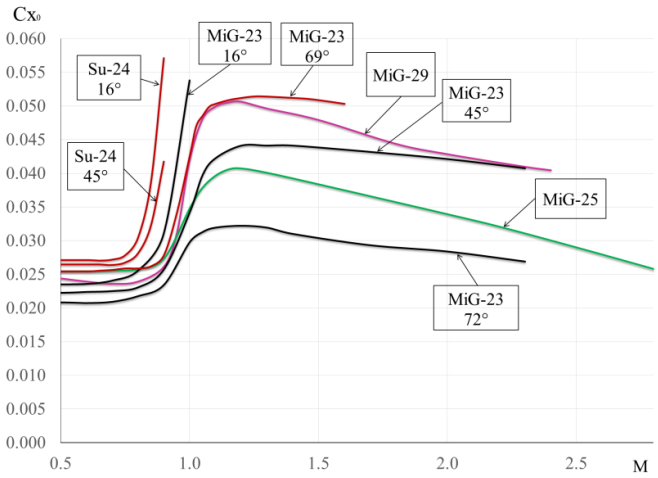


Figure 10: Graphs Cx0 for the studied aircrafts.

Figure 10 shows that in the area of maximum flight speed, the slope of the curves C_{x0} (M) is different for different aircraft.

The power-available curve depends on the engine unit's altitude speed performances, the required one on the drag run, therefore, the IC depends strongly on the specific aircraft (on the flow pattern of Cx_0 and altitude-speed performances), therefore,

linking this coefficient to the aircraft parameters is impractical. However, as the research showed, the IC value for subsonic aircraft is approximately 0.1…0.2 (Table 9), for supersonic: 0.7…0.9.

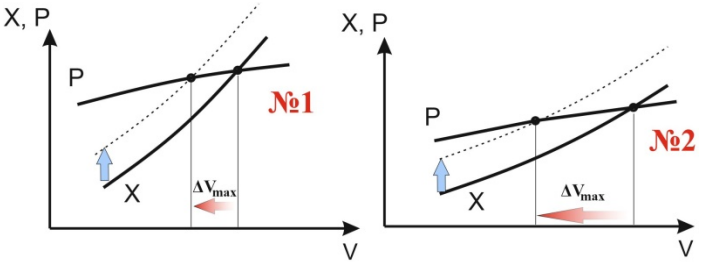


Figure 11. The different nature of Zhukovsky curve crossing. P – engine thrust curve, X – airplane drag curve.

6. Conclusions

The results of this work allow us to estimate the contribution of mass uncertainties and drag coefficient at zero lifting force to the range of variation of the FQP values.

The formulas obtained make it possible to determine the quantitative influence of the parameter on the value of the FQP. Verification of the formulas for some aircraft from the class of operational-tactical aviation showed that the error in determining the spread of flight characteristics according to the formulas obtained is within 10%.

The research showed that the mass influence coefficients Cx_0 on maximum specific excess power, sustained load factor and acceleration time practically do not depend on the parameters of the considered aircrafts of the same class. The IC of mass and Cx_0 on the maximum speed and practical ceiling associated with specific aircraft parameters. The sensitivity of the maximum speed to the zero lift-drag coefficient is determined by the curve crossing nature of the required and available thrust.

The inaccuracy in determining the mass in 1% gives a spread of maximum specific excess power, sustained load factor and acceleration time in about 0.9…1.3%. For the studied aircrafts the inaccuracy in determining the zero lift-drag coefficient of 1% gives a spread in the maximum speed within 0.7 ... 0.9% for supersonic aircraft and 0.1 ... 0.2% for subsonic aircraft.

References

[1] V.E. Denisov, V.K. Isaev, A.M. Ryabov, L.M. Shkadov, "Statisticheskaya ocenka harakteristik proektiruemogo samoleta s pomoschyu metoda Monte_Karlo" [Statistical evaluation of the characteristics of the designed aircraft using the Monte Carlo method], Uchenie zapiski CAGI [Scientific notes of TsAGI], **4**(2), 137–142, 1973.

[2] O.K. Yugov, O.D. Selivanov, Osnovi integracii samoleta i dvigatelya [Fundamentals of the integration of the aircraft and engine], Mashinostroenie [Mechanical Engineering], 1989.

[3] G.M. Skopec, "Vneshnee proektirovanie aviacionnih kompleksov: Metodologicheskie aspekti" [External design of aviation complexes: Methodological aspects], LENAND, 2017.

[4] M. Ravachol, Y. Deremeaux, Q.V. Dinh, E. Herbin, "Uncertainties at the conceptual stage: multilevel multidisciplinary design and optimization approach", ICAS 2008 26TH International Congress of the Aeronautical Sciences, Anchorage, Alaska; September 14-19, 2008.

[5] L.V. Myshkin, Prognozirovanie razvitiya aviacionnoi tehniki [Forecasting the development of aviation technology], Izdatelskii Dom Nauka [Publishing House Science], 2017.

[6] T. Pfeiffer, E. Moerland, D. Böhnke, B. Nagel, V. Gollnick, «Aircraft configuration analysis using low-fidelity, physics based aerospace framework under uncertainty considerations», ICAS 2014 29TH Congress of the International Council of the Aeronautical Sciences, St. Peterburg, Russia; September 7-12, 2014.

[7] D. Neufeld , J. Chung, K. Behdinan, "Aircraft Conceptual Design Optimization with Uncertain Contributing Analyses", AIAA Modeling and Simulation Technologies Conference, Chicago, Illinois; August 10–13, 2009. doi: 10.2514/6.2009-6237

[8] M.R. Kirby, D.N. Mavris, "Forecasting Technology Uncertainty in Preliminary Aircraft Design", AIAA, World Aviation Conference, San Francisco, CA; October 19-21, 1999.

[9] L.L. Green, H.Z. Lin, M.R. Khalessi, "Probabilistic methods for uncertainty propagation applied to aircraft design", AIAA, PredictionProbe, Inc. Newport Beach, CA 92660, 2002. doi:10.2514/6.2002-3140

[10] M. Hosseini, M. Nosratollahi, H. Sadati, "Multidisciplinary Design Optimization of UAV under Uncertainty", J. Aerosp. Technol. Manag., São José dos Campos, **9**(2), 169–178, 2017. doi: 10.5028/jatm.v9i2.725

[11] D.N. Mavris, D.A. DeLaurentis , D.S. Soban, "Probabilistic Assessment of Handling Qualities Characteristics in Preliminary Aircraft Design", AIAA 98-0492, 36th Aerospace Sciences Meeting & Exhibit Copyright, 1998.

[12] D.N. Mavris, D.A. DeLaurentis, D.P.Schrage, "System Synthesis in Preliminary Aircraft Design Using Statistical Methods", the 20th International Council of the Aeronautical Sciences (ICAS) Congress, Sorrento, Italy, September 8-13, 1996.

[13] G.S. Veresnikova, L.A. Pankova, V.A. Pronina, O.V. Ogorodnicov, I.I. Ikryanov, "Determining maneuverable aircraft parameters in preliminary design under conditions of uncertainty", International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Marseille, France; September 6–8, 2017. doi: 10.1016/j.procs.2017.08.143

[14] D.A. Laushin, V.A. Klyagin, "K voprosu o vliyanii faktorov neopredelennosti na oblik samoleta" [To the question of the influence of uncertainty factors on the appearance of the aircraft], v sbornike tezisov XLIV Mejdunarodnoi molodejnoi nauchnoi konferencii Gagarinskie chteniya [In the collection of theses of the XLIV International Youth Scientific Conference Gagarin Readings], Moscow Aviation Institute, 2018.

[15] D.A. Laushin, V.A. Klyagin, "Uchet riskov, svyazannih s dostijeniem zadannih letno_tehnicheskih harakteristik samoleta" [Accounting for the risks associated with achieving the specified flight performance of the aircraft], Obscherossiiskii nauchno_tehnicheskii jurnal POLET [All-Russian scientific and technical journal POLET], 1, 28–32, 2019.

[16] A.M. Tarasenkov, V.G. Braga, V.T. Taranenko, Dinamika poleta i boevogo manevrirovaniya letatelnih apparatov [Dynamics of flight and combat maneuvering of aircraft, VVIA im. prof. N.E. Jukovskogo [Zhukovsky VVIA], 1984.

[17] G.S. Busgens, Aerodinamika, ustoichivost i upravlyaemost sverhzvukovih samoletov [Aerodynamics, stability and controllability of supersonic aircraft], Fizmatlit, 1998.

[18] "Prakticheskaya aerodinamika samoleta MiG-29" [Practical aerodynamics of the MiG-29], Moskva, slujebnoe izdanie VVS [Moscow, official publication of the Air Force], 1987. Available Online at: http://bookre.org/reader?file=543564

[19] "Prakticheskaya aerodinamika samoleta MiG-23ML i MiG-23UB" [Practical aerodynamics of the MiG-23ML and MiG-23UB], VVIA im. Jukovskogo [Zhukovsky VVIA], 1978. Available Online at: http://bookre.org/reader?file=543565

[20] "Prakticheskaya aerodinamika samoleta MiG-25RB" [Practical aerodynamics of the MiG-25RB], Voennoe izdatelstvo Ministerstva oboroni SSSR [Military Publishing House of the Ministry of Defense of the USSR], 1978. Available Online at: http://bookre.org/reader?file=543563

[21] "Prakticheskaya aerodinamika Su-24" [Practical aerodynamics of the Su-24]. Available Online at: http://bookre.org/reader?file=540492

ASTES

Advances in Science, Technology and Engineering Systems Journal 5, No. 4, 119-129 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design of an EEG Acquisition System for Embedded Edge Computing

Kanishk Rai[1,*], Keshav Kumar Thakur[1], Preethi K Mane[1], Narayan Panigrahi[2]

[1]*Department of Electronics and Instrumentation, BMS College of Engineering, 560019, India*

[2]*Centre for Artificial Intelligence and Robotics, DRDO, 560093, India*

ARTICLE INFO

Article history:
Received: 17 April, 2020
Accepted: 17 June, 2020
Online: 12 July, 2020

Keywords:
Electroencephalogram (EEG)
Internet of Things (IOT)
Brain Computer Interface (BCI)
Acquisition system
Edge Computing

ABSTRACT

The human brain is one of the most complex machines on the planet. Being the only method to get real-time data with high temporal resolution from the brain makes EEG a highly sought upon signal in the neurological and psychiatric domain. However, recent developments in this field have made EEG more than just a tool for medical professionals. The decreasing size and increasing complexity of EEG acquisition systems have brought it out of the lab and into the field where it is used for varied applications like neurofeedback, person recognition and other recreational activities. Amalgamation of the EEG signal with new developing standards of Industry 4.0 to control basic IOT devices using edge computing techniques marks the next step in the design and development our low-cost yet robust Brain Computer Interface (BCI); which is just one of the many applications that a versatile and well-built EEG acquisition system can be used for.

1. Introduction

What makes the human brain so complex is not just the 100 billion neurons that it is constituted of but also the 100 trillion plus synapses or the unique ways in which those neurons connect and communicate with each other. Even with more than 7.5 billion people on the planet, no two brains are alike. Therefore, designing a unified device that can capture the diverse Electroencephalogram (EEG) signals emanating from different brains is a challenging task. EEG signals can be processed to extract the characteristic patterns of the brain to study its spatio-temporal characteristics.

Though other methods of measuring brain activity like Magnetic Resonance Imaging (MRI) and Functional Magnetic Resonance Imaging (fMRI) have more spatial resolution yet, they lack temporal resolution and are often more costly to setup and operate. EEG on the other hand offers better temporal resolution and is a cost-effective solution which can directly measure brain activity. EEG can approximate spatial activity with the use of greater number of electrodes placed on the scalp, providing better resolution and thus, making it a scalable choice.

A low cost yet robust design of an EEG acquisition system was presented at the *IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS),* 2019 [1]. Though the design was primarily focused on reduction of cost of manufacturing so that the system becomes affordable, yet there were many limitations of the design. We studied the base design and propose a refined design which will overcome the followings drawbacks of the initial design:

a) Overcome the limitations of low sampling rate

b) To perform filtering and feature extraction from acquired EEG signal on the board itself

c) To reduce and eliminate the noise during the EEG signal acquisition process by reducing wires in the system

d) Better processing mechanism to extract and interpret the EEG data acquired, in real-time

The whole system has now been fabricated on a Printed Circuit Board (PCB) and the primary mode of communication between the acquisition system and the host computer is now wireless to minimize artefacts introduced due to wire movements and other Electromagnetic Interference (EMI) / Electromagnetic Compatibility (EMC) interference. The detailed design update is enumerated in the system design section.

The next section discusses current research trend in the field of BCI (Brain Computer Interface) and HMI (Human Machine Interface) using EEG, followed by the Problem Definition. Next, we briefly look at the testing results from our initial design and highlight its shortcomings. We then move on to introduce the updated block diagram and directly compare the initial and the updated components. This serves the double purpose of

*Kanishk Rai, +91-8853389016, kr2252@gmail.com

introducing the initial design while highlighting the updates in the improved model. We next discuss the results obtained during testing of the improved EEG acquisition system and finally, we illustrate one of the many applications of our system - which is to automatically control IoT devices such as the television, lights, wheel chair and fan depending upon the state of the subject e.g. active, semi active, asleep in real time.

2. Literature Review

2.1. Background

Since the first experiments of electroencephalography on humans in 1929, the electroencephalogram (EEG) of the human brain has been used mainly to evaluate neurological disorders in the clinical environment and to investigate brain functions in the laboratory. An idea that brain activity could be used as a communication channel between the subject and an IoT device has gradually emerged [2-6]. The possibility of recognizing a single message or command, considering the complexity, distortion, and variability of brain signals appeared to be extremely remote. Yet EEG demonstrates direct correlations with user intentions, thereby enabling a direct brain-computer communication channel.

Brain-Computer Interface (BCI) technology, as it is known, is a communication channel that enables users to control devices and applications without the use of muscles. The development of cognitive neuroscience field has been instigated by recent advances in brain imaging technologies such as electroencephalography, magneto encephalography and functional magnetic resonance imaging. The first BCI prototype was created by Dr. Vidal in 1973. This system was intended to be used as a promising communication channel for persons with severe disabilities, such as paralysis, Amyotrophic Lateral Sclerosis (ALS), stroke or cerebral paralysis.

Since then BCI research has been successfully used not only for helping the disabled, but also as being an additional data input channel for healthy people. It can be exploited as an extra channel in game control, augmented reality applications, household device control, fatigue and stress monitoring and other applications. BCI design represents a new frontier in science and technology that requires multidisciplinary skills from fields such as neuroscience, engineering, computer science, psychology and clinical rehabilitation.

2.2. Review of EEG for BCI Applications

Controlling real-world objects with nothing more than our mind is no longer science fiction. Jaime, et al [7] have proved the same by demonstrating the fact that the steering of a tractor by use of surface EOG and EEG signals is just as accurate as manual guidance or GPS guidance with a minimal variation in response time.

Furthermore, Yuan, et al [8] have conclusively established the feasibility of using a consumer-level EEG headset to realize an online steady-state visual-evoked potential (SSVEP)-based BCI during human walking using a 14 channel commercially available EEG headset to implement a four-target online SSVEP decoding system which has in-turn helped bridge the gap between laboratory BCI demonstrations and real-life application of consumer-level EEG headsets.

Starting with some simple applications, Gerardo Rosas-Cholula, et al [9] have successfully controlled a mouse pointer using a commercially available EEG headset. They have implemented Empirical Mode Decomposition (EMD) to reduce noise and simulated mouse clicks with the help of blinks. Next we come across articles [10-14] that focus on controlling robotic arms or wheelchairs for amputees and paraplegics using different computational techniques for identification and classification of various signals in the EEG like Event-related Potential (ERP), P300, Mu rhythm, etc. and demonstrate the accuracy and ease with which these EEG based BCIs can control simple devices.

Taking this use of EEG one step forward multiple articles [15-21] discuss integrating EEG with the Internet of Things (IOT) to control different devices in a smart home using protocols like Message Queue Telemetry Transport (MQTT) and Extensible Messaging and Presence Protocol (XMPP). These interfaces can further be used to improve healthcare and the general living standard of differently-abled people and in-turn make everyone's life easier. However, there is much work still to be done in this domain considering the new threats that continuously emerge in the realm of cyber security.

Researchers have worked upon and built multiple EEG acquisition systems [22-28] but have not used these systems to demonstrate an application. Others have demonstrated the use of EEG for various applications be it in terms of person recognition [29-32] or in the field of healthcare [33-38] or the ones mentioned above, but they have either used commercially available EEG acquisition systems or pre-existing datasets. With our research we aim to bridge this gap by not only building a robust yet low-cost EEG acquisition system but also using the same system to control IoT devices in real-time, and in doing so we not only design the hardware to acquire EEG data but also develop the software for data representation and signal processing which showcase the edge computing capabilities of our device.

3. Problem Definition (Limitations of Existing Designs of EEG Acquisition Systems)

The research trends towards problems addressed in the EEG and BCI/HMI domain [1] show us that only 3 out of 30 publications focus on data acquisition and 5 out of 30 discuss EEG signal processing. However, this can paint a slightly misleading picture. To fully understand the need to design and develop a low-cost yet robust EEG acquisition system capable of edge processing we need to look at the research trends towards adaptation of technology where only 2 out of 21 publications [1] use custom-built EEG acquisition systems, while others rely on expensive alternatives available in the market or pre-existing datasets [1].

The above gaps in research are the major motivation behind this research, but the problems encountered while designing a robust yet low-cost EEG acquisition system are two-fold:

Domain specific problems are: -

- To ensure that the pickup, amplification and transmission of EEG signal is done while introducing minimum noise in the signal.
- To ensure that the sampling rate of the acquisition system is at lest 2.5 times the highest frequency of input EEG signal, but at the same time not too high so as to avoid the problem of oversampling.

- To ensure that the resolution of the system is enough to register minute changes in the peak-to-peak amplitude and the frequency of the EEG signal.

Problems specific to the design of our acquisition system are:

- To build a Graphical User Interface (GUI) that accurately represents the EEG signal picked up by the acquisition system in real-time.
- To perform real-time processing and feature extraction from EEG signal in order to build a device capable of edge processing.

4. System Design

After careful examination of EEG signals and other medical grade systems available in the market, we have arrived at the following design-constraints to design our EEG acquisition system.

Design constraints:

- EEG voltage range: 10 µV to 100 µV
- Useable EEG frequency range: 0.3 Hz to 60 Hz
- Max variation of input signal per unit time: 50µV
- Sampling Rate: 256 to 1k SPS

The subsequent sections talk about the step by step design of each component in the acquisition system. Beginning with an introduction to the initial system design, its results and drawbacks.

4.1. Initial Block Diagram, Testing and Observations

For the initial design [1] as shown in Figure 1, gold plated, cup shaped electrodes (placed on the scalp using the international 10-20 system) were connected to the instrumentation amplifiers (INA333) which were then directly connected to the analog input pins of Arduino Mega. We only used 3 input channels; two for data and one for reference. Arduino IDE however, is neither capable of plotting data from multiple channels nor is it capable of performing complex algorithms on the acquired data.

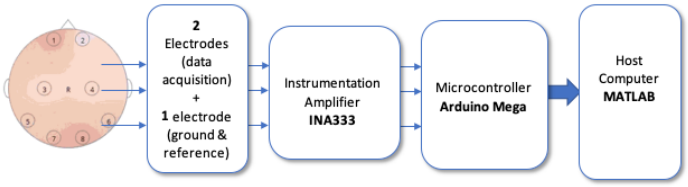


Figure 1: Initial block diagram

Therefore, to overcome the abovementioned drawbacks, we interfaced Arduino with MATLAB, which is used to plot real time frequency domain and time domain spectrum of the incoming EEG signal with further possibility of manipulating data and representing it in different formats.

Figure 2 shows the acquired EEG data from a single channel connected to analog port A0 of the Arduino at a baud rate of 9600. Subplot 1 shows the real time RAW EEG signal amplified 101 times whereas subplot 2 shows the real time Frequency plot (FFT).

A brief description of output from the initial design of our EEG acquisition system is as follows:

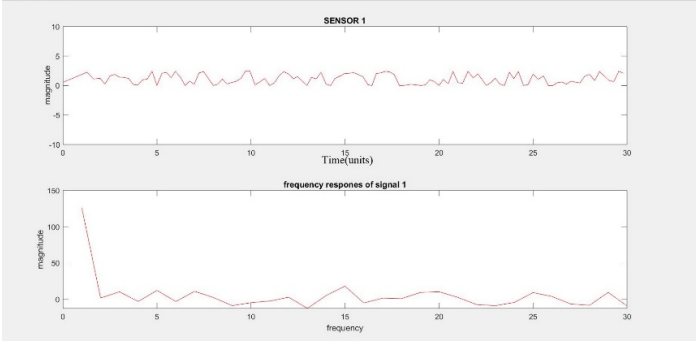


Figure 2: Data Acquisition in MATLAB interfaced with Arduino

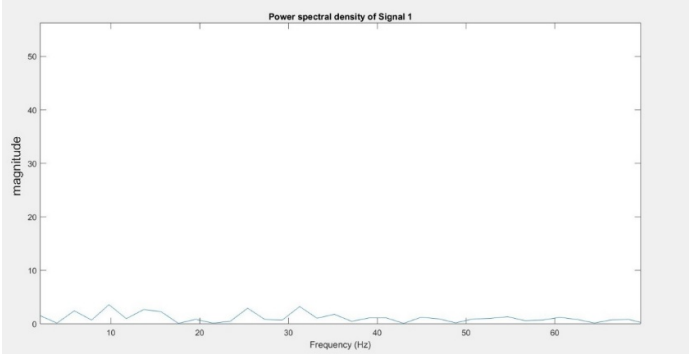


Figure 3: Power spectral density of data acquired from single sensor coming from instrumentation amplifier with a gain of 101

Once the data acquisition is complete, MATLAB then computes the power spectral density of the acquired EEG signal as observed in Figure 3. Major peaks seen around 5, 10, 25 and 32 Hz correspond to Alpha, Beta and Theta waves which contribute majorly to awake EEG patterns.

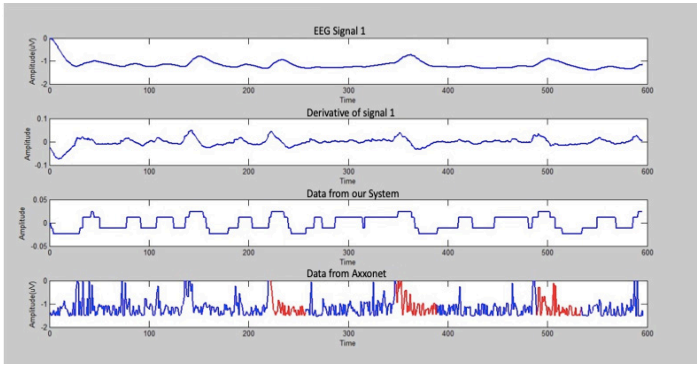


Figure 4: Comparison of data acquired from a medical grade system and our proposed system

Figure 4 shows the comparison of saccade, fix and blink data. Subplot 1 shows the data from our system, while Subplot 2 shows the data from the medical grade system. Similar for Subplot 3 and 4. It is evident from the above figure that our proposed system captures the correct data as all peaks match between both signals. The drawback here is the low sampling rate because of which the reconstructed signal does not have as much detail as is required for complex analysis – this can be seen when comparing Subplot 3 and 4. The maximum rate at which the data is captured by our initial system is about 5 to 20 Hz [1].

4.2. Drawbacks of the Initial System

Due to the low sampling rates which are in-turn caused by the hardware limitations of the 10-bit ADC on Arduino Mega2560, we were only able to capture data at a maximum rate of 20 Hz. But the Practical Nyquist Sampling Theorem states that to properly capture and reconstruct incoming signal, the sampling frequency of the acquisition system must be at-least 2.5 times the maximum frequency component of the incoming signal, therefore to capture brain waves with a maximum useable frequency component of 60 Hz, the acquisition system must have a sampling rate of 150 Hz or higher. To overcome this issue of lower sampling rate, we propose the use of an ADC (ADS1299) introduced after the instrumentation amplifier and before the microcontroller as shown in the block diagram of our initial design (Figure 1).

The microcontroller Arduino Mega does not have the inherent capability to transmit data wirelessly meaning more noise is introduced in the signal due to the movement of the wires. This is overcome by replacing Arduino with NodeMCU that has an inbuilt wireless transmitter.

Moreover, the use of MATLAB for signal processing and representation poses its own set of limitations including the need for a MATLAB license to use our system, hence the same is replaced by a GUI built using open-source code in Python.

4.3. Improved Block Diagram

The initial components of the system like the Electrodes (gold-plated, cup shaped, copper, non-invasive), Electrode Placement (international 10-20 system), Electrode Cap and the Instrumentation Amplifier (INA333) remain the same and are carried over from the initial design (Figure 1).

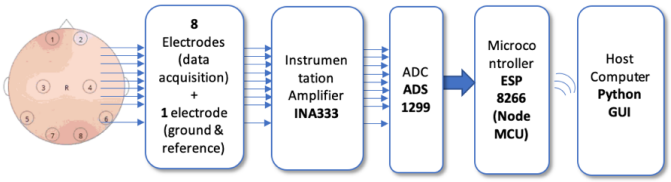


Figure 5: Updated Block Diagram

Figure 5 shows the block diagram of the updated hardware with an Analog to Digital Converter introduced in the system after the instrumentation amplifier to handle the conversion of analog data to digital domain at a much faster rate. The ADS1299 connects to the NodeMCU using SPI, which then transmits the data wirelessly to the host computer that runs a Python GUI for data processing, filtration and signal representation.

4.4. Analog to Digital Converter (ADC)

The ADC chosen for our application is the TI analog front-end chip ADS1299, which is specifically designed to capture biopotential signals like EEG, ECG, EOG, EMG, among others. The chip has a 24-bit Sigma-Delta based architecture, with a maximum data rate of 16 kSPS. It boasts of a -110 dB CMRR and a Signal to Noise Ratio of 121 dB. It is a low power, low noise, high precision chip. It can further be multiplexed to use more channels and is available in 4, 6 and 8 channel alternatives.

The chip communicates with the microcontroller using SPI (Serial Programming Interface) which can be used to configure the chip to read data in Continuous mode or Single Shot mode and also to read the output data from the inbuilt registers.

4.5. Microcontroller

Arduino Mega from the initial design has now been replaced with ESP 8266 NodeMCU as the choice of microcontroller for our system. The NodeMCU has a L106 32-bit RISC microprocessor running at 80 MHz. It supports 32 KB instruction RAM and 80 KB user-data RAM. It comes with integrated IEEE 802.11 b/g/n Wi-Fi connectivity and matching network WEP or WPA/WPA2 authentication. It has 16 GPIO pins and also supports SPI communication – all of which is ideal for our application. The NodeMCU does not only have more GPIO pins, better clock speed of the processor and more flash and SRAM than the Arduino, but also operates at a lower 3.3 V which translates to less power consumption for battery powered wireless use of our EEG acquisition system.

4.6. Initial Observations with ADS1299

To set up and test the SPI connection between ADS1299 and NodeMCU, ADS is soldered on an adaptor (Figure 6) and the connections are made on a bread-board (Figure 7). It must be noted that for this initial testing, ADS is connected to Arduino. Once the connections are verified, the final PCB is interfaced with NodeMCU.

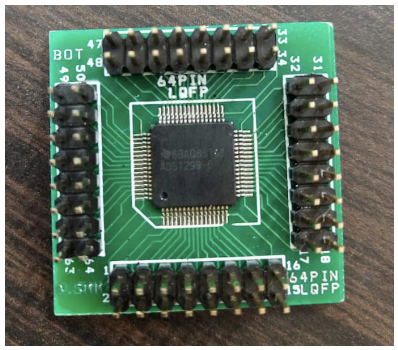


Figure 6: ADS1299 soldered on an adaptor

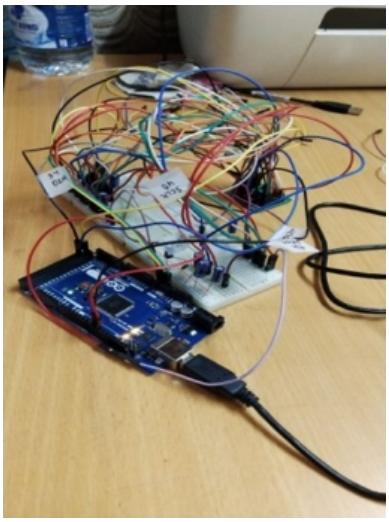


Figure 7: ADS1299 connected to Arduino Mega using SPI on a bread board

After running multiple tests Arduino Mega is able to read the correct Device ID (Figure 8). The device ID reads 1111 1100 where the last four bits correspond to the 4 channel ADS1299 chip. This means that the SPI communication was successful, but due to the inherent noise of the system which is set-up on a bread-board, the output is not consistent.

Therefore, the next logical step is to design a PCB to acquire the signals. This is carried out in the next subsection.

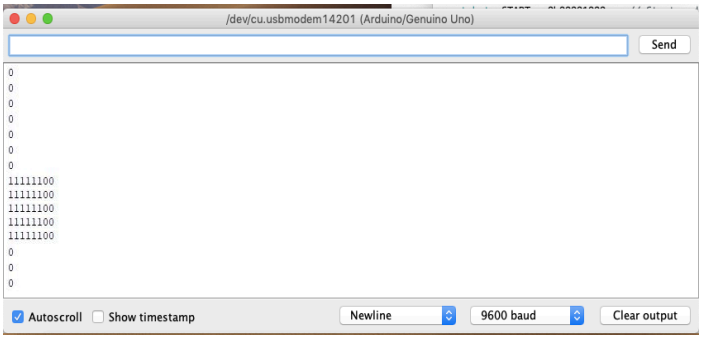


Figure 8: Correct device ID of ADS1299-4PAG displayed on serial monitor of Arduino IDE connected to Arduino

4.7. Design for PCB Fabrication

The PCB is designed using an open-source software called PCB Artist. The schematic for the design is shown in Figure 10. The design comprises of a power supply unit which steps down the incoming 220V AC to 5V and 3.3V DC. Both the 5V and 3.3V lines have been taken out as well to provide power to the microcontroller NodeMCU and for other purposes. Internally, the ADC has multiple pins that require 5V and 3.3V to power up the Analog and Digital circuits in the chip respectively.

For the input electrodes, each channel is first passed through the non-inverting pin of the instrumentation amplifier INA333 for first stage amplification. The signal is then passed through an RC high-pass filter to remove DC offsets and fed to the ADC via referential montage with one reference electrode and an internal bias. The example Schematic using the ADS1299 in an EEG Data Acquisition Application in Referential Montage from its datasheet is shown in Figure 9. The advantage of using the referential montage is that all electrodes are referenced to a single reference electrode instead of being referenced to each-other as is the case with sequential montage.

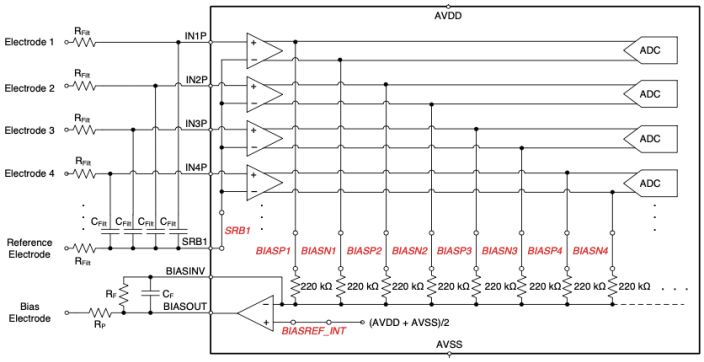


Figure 9: Use of ADS1299 in referential montage

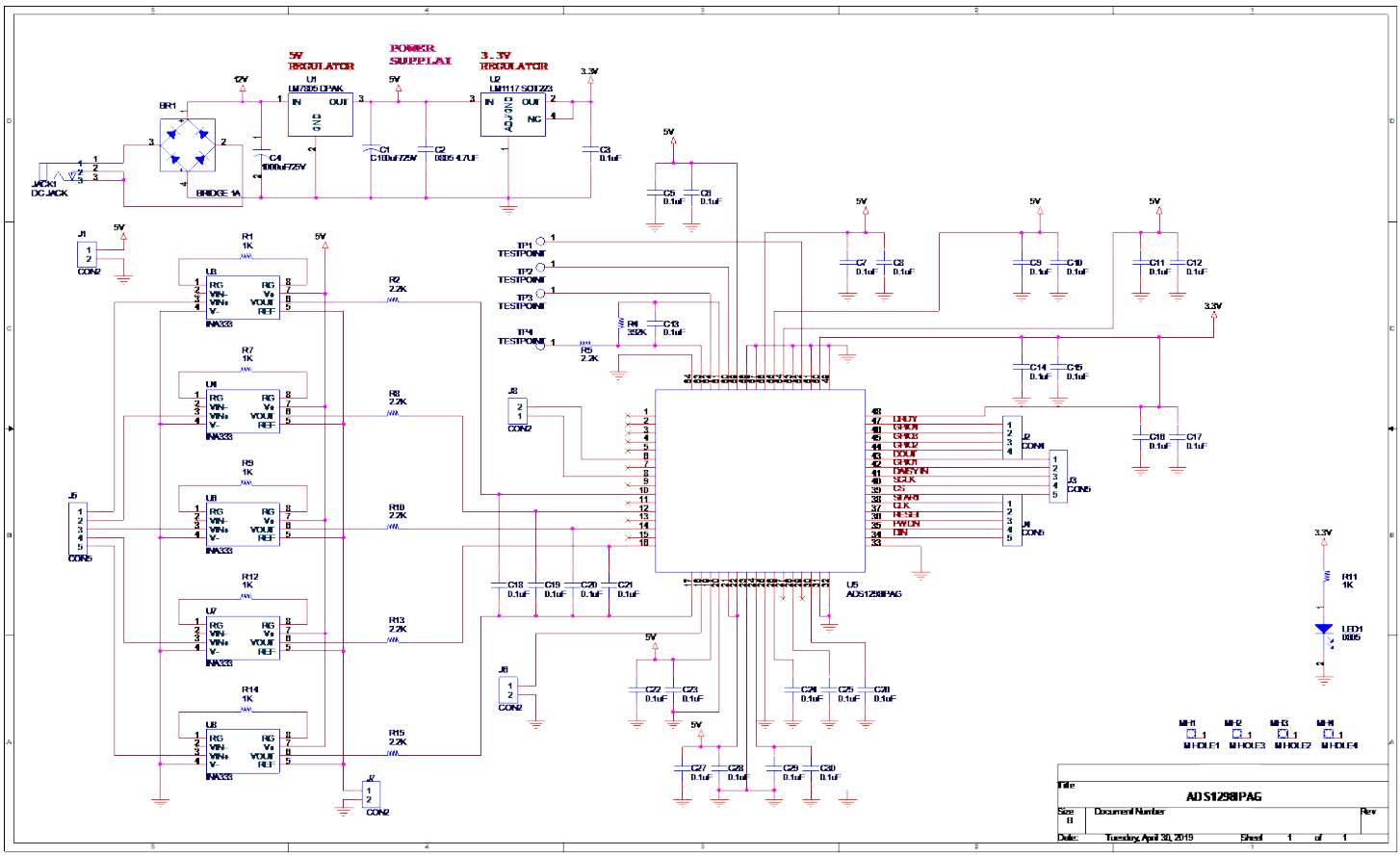


Figure 10: Schematic for PCB design

The final fabricated, two-layer PCB is shown in Figure 11 and Figure 12.

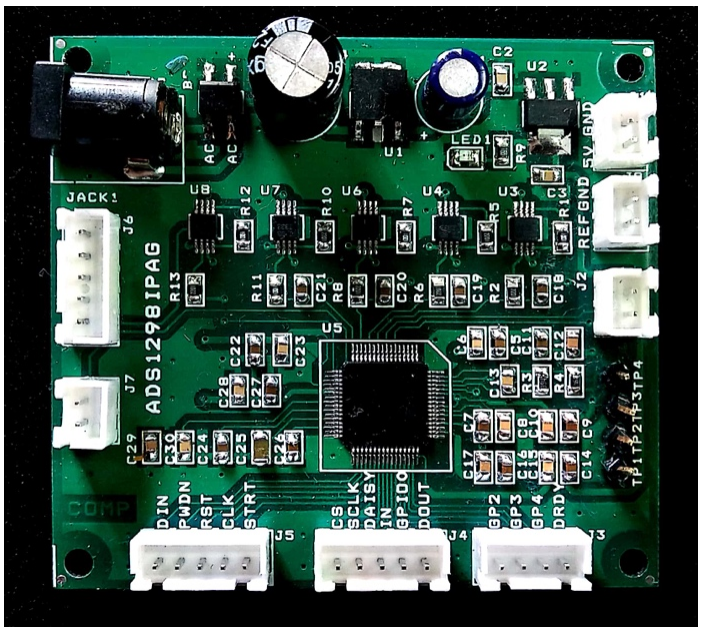


Figure 11: Fabricated PCB (top)

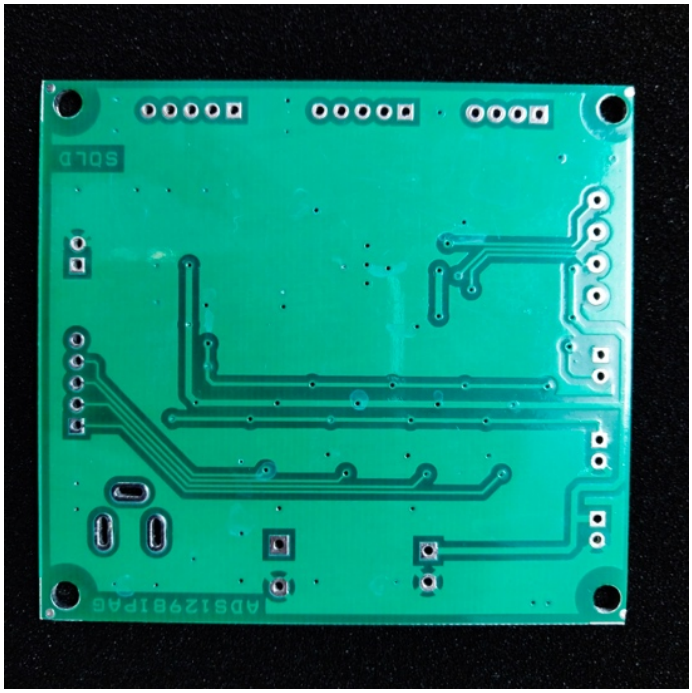


Figure 12: Fabricated PCB (bottom)

5. Output from PCB

Once the PCB is fabricated, it is connected to the microcontroller using SPI interface. The ADS1299 is then configured to transmit continuous data with 24-bit resolution at 250 SPS. The register values corresponding to the above configuration are shown in Figure 13 with the correct device ID. Even though the chip is capable of acquiring data at speeds up to 16 kSPS, we have not used these higher sampling rates to avoid the problem of oversampling.

```
-----------------------------------------
Register 0x1 modified.
-----------------------------------------
ID, 0x00, 0x3C, 0, 0, 1, 1, 1, 1, 0, 0
CONFIG1, 0x01, 0x0A, 0, 0, 0, 0, 1, 0, 1, 0
CONFIG2, 0x02, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
CONFIG3, 0x03, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
LOFF, 0x04, 0x00, 0, 0, 0, 0, 0, 0, 0, 0
CH1SET, 0x05, 0x07, 0, 0, 0, 0, 0, 1, 1, 1
CH2SET, 0x06, 0x07, 0, 0, 0, 0, 0, 1, 1, 1
CH3SET, 0x07, 0x8F, 1, 0, 0, 0, 1, 1, 1, 1
CH4SET, 0x08, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
CH5SET, 0x09, 0x00, 0, 0, 0, 0, 0, 0, 0, 0
CH6SET, 0x0A, 0x07, 0, 0, 0, 0, 0, 1, 1, 1
CH7SET, 0x0B, 0x08, 0, 0, 0, 0, 1, 0, 0, 0
CH8SET, 0x0C, 0x6A, 0, 1, 1, 0, 1, 0, 1, 0
BIAS_SENSP, 0x0D, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
BIAS_SENSN, 0x0E, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
LOFF_SENSP, 0x0F, 0x00, 0, 0, 0, 0, 0, 0, 0, 0
LOFF_SENSN, 0x10, 0x07, 0, 0, 0, 0, 0, 1, 1, 1
LOFF_FLIP, 0x11, 0x08, 0, 0, 0, 0, 1, 0, 0, 0
LOFF_STATP, 0x12, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
LOFF_STATN, 0x13, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
GPIO, 0x14, 0x00, 0, 0, 0, 0, 0, 0, 0, 0
MISC1, 0x15, 0x07, 0, 0, 0, 0, 0, 1, 1, 1
MISC2, 0x16, 0x08, 0, 0, 0, 0, 1, 0, 0, 0
CONFIG4, 0x17, 0xC0, 1, 1, 0, 0, 0, 0, 0, 0
-----------------------------------------
```

Figure 13: Register configuration of ADS1299

```
Millis: 224
0, C0000F, AC8A8, 1FC3E8, 3EB5AD, 50BDAD, FFFFFF, FFFFFF, FFFFFF, FFFFFF
1, C00007, 845F0, 1E69C4, 3CB9F6, 4F82C3, FFFFFF, FFFFFF, FFFFFF, FFFFFF
2, C0000F, B89FF, 1F7716, 3ECD0A, 50173E, FFFFFF, FFFFFF, FFFFFF, FFFFFF
3, C00007, 99422, 1F0B09, 3CC192, 4EF1B3, FFFFFF, FFFFFF, FF8000, 0
4, C0000F, B3454, 1F2A17, 3F23D8, 4F877B, FFFFFF, FFFFFF, FFFFFF, FFFFFF
5, C00007, 8E698, 1E9DF4, 3D8FD5, 4FB02F, FFFFFF, FFFFFF, FFFFFF, 800000
6, C0000F, AB00E, 1F0CC4, 3F1755, 50AA95, FFFFFF, FFFFFF, FFFFFF, FFFFFF
7, C00007, 8717F, 1E3768, 3CA722, 502684, FFFFFF, FFFFFF, FFFFFF, 800000
8, C0000F, A3709, 20393F, 3F3B78, 4FFEC1, FFFFFF, FFFFFF, FFFFFF, FFFFFF
9, C00007, 8A828, 1EF6D1, 3C5B48, 4ED02A, FFFFFF, FFFFFF, FFFF00, 0
10, C0000F, BC64C, 1FB876, 3F9BD5, 50B430, FFFFFF, FFFFFF, FFFFFF, FFFFFF
11, C00007, 8DADC, 1F3574, 3D3CA8, 4FC687, E00000, 0, 0, 0
12, C0000F, ACA48, 1FB81A, 40EC78, 4FE043, FFFFFF, FFFFFF, FFFFFF, FFFFFF
13, C00007, 8F07E, 1F75F8, 3D061D, 4F7899, E00000, 0, 0, 0
14, C0000F, B4FD9, 1F1D42, 3F045C, 50662E, FFFFFF, FFFFFF, FFFFFF, FFFFFF
15, C00007, 8CFF4, 1E9DF6, 3CB969, 4FEDED, E00000, 0, 0, 0
16, C0000F, AAF94, 1F864C, 3F1044, 50CE7D, FFFFFF, FFFFFF, FFFFFF, FFFFFF
17, C00007, 92530, 1E0DAC, 3C0099, 5001C8, E00000, 0, 0, 0
18, C0000F, B6B34, 203E2C, 401502, 502315, FFFFFF, FFFFFF, FFFFFF, FFFFFF
19, C00007, 883F2, 1EF0D1, 3CE39D, 4F2403, E00000, 0, 0, 0
20, C0000F, BCFC9, 1FBE40, 3F7BC7, 5018AE, FFFFFF, FFFFFF, FFFFFF, FFFFFF
21, C00007, 8E369, 1F1AC0, 3CD92A, 4FC144, 800000, 0, 0, 0
22, C0000F, B7E03, 1EF62A, 3FED85, 50079A, FFFFFF, FFFFFF, FFFFFF, FFFFFF
23, C00007, 8B601, 1EEC70, 3D2930, 4F48F9, E00000, 0, 0, 0
24, C0000F, AC008, 20DBA1, 3FC0C6, 507312, FFFFFF, FFFFFF, FFFFFF, FFFFFF
25, C00007, 8681F, 1E8947, 3C9FCA, 4F449B, E00000, 0, 0, 0
26, C0000F, A6233, 1F64D7, 3FA866, 50BF51, FFFFFF, FFFFFF, FFFFFF, FFFFFF
27, C00007, 8AD20, 1EB9A6, 3CB8EF, 4F90C1, E00000, 0, 0, 0
28, C0000F, ABC6A, 1FE306, 3DF771, 4FDBA8, FFFFFF, FFFFFF, FFFFFF, FFFFFF
29, C00007, 8DF90, 1F09C9, 3CE496, 4EC212, 0, 0, 0, 0
30, C0000F, B7E05, 1FE20E, 3F3FFC, 50C352, FFFFFF, FFFFFF, FFFFFF, FFFFFF
31, C00007, 89906, 1E0C32, 3C1DF7, 4FFEF8, 0, 0, 0, 0
32, C0000F, B3618, 1F7934, 3F4059, 513716, FFFFFF, FFFFFF, FFFFFF, FFFFFF
33, C00007, 95F40, 1F32D4, 3D600F, 5023CA, 0, 0, 0, 0
34, C0000F, B8109, 1F8DAD, 3FB9D9, 50AEAA, FFFFFF, FFFFFF, FFFFFF, FFFFFF
35, C00007, 85471, 1EB142, 3CC3AD, 4EFD71, E00000, 0, 0, 0
36, C0000F, B8823, 1FCEFB, 40C8B1, 502A43, FFFFFF, FFFFFF, FFFFFF, FFFFFF
37, C00007, 9181B, 1E7FC9, 3D7A7A, 4FBF21, 800000, 0, 0, 0
38, C0000F, B1161, 1F81DA, 4084F3, 501EE0, FFFFFF, FFFFFF, FFFFFF, FFFFFF
39, C00007, 8AD14, 1F2981, 3CE11B, 4F9E57, 800000, 0, 0, 0
40, C0000F, B725D, 1EC921, 3EF51A, 503866, FFFFFF, FFFFFF, FFFFFF, FFFFFF
41, C00007, 93F25, 1F35AA, 3E734A, 4FE0C3, 0, 0, 0, 0
```

Figure 14: Incoming data from EEG electrodes in ADS1299

Figure 14 shows the filtered data acquired from the EEG sensors connected to the input channels of ADS1299. The RAW data is in Hexadecimal format. This is converted to Decimal format and then scaled down to convert into Volts. Since we have used a 4 channel ADS1299 chip, the data format as observed in the output is <start bit> <channel 1> <channel 2> <channel 3> <channel 4>. The last 4 bits in each row are of no use to us because those channels don't exist on an ADS1299-4PAG chip.

6. Data Representation using Python GUI

6.1. Design

The code for the GUI is burnt on the NodeMCU using the Arduino IDE after installing relevant drivers. First off, this code checks the device ID coming from the attached ADS1299 chip and if it is correct, modifies the register values to acquire data. In our case the ADS is set to acquire data at 250 SPS in continuous mode. It then creates the GUI window as shown in Figure 15.

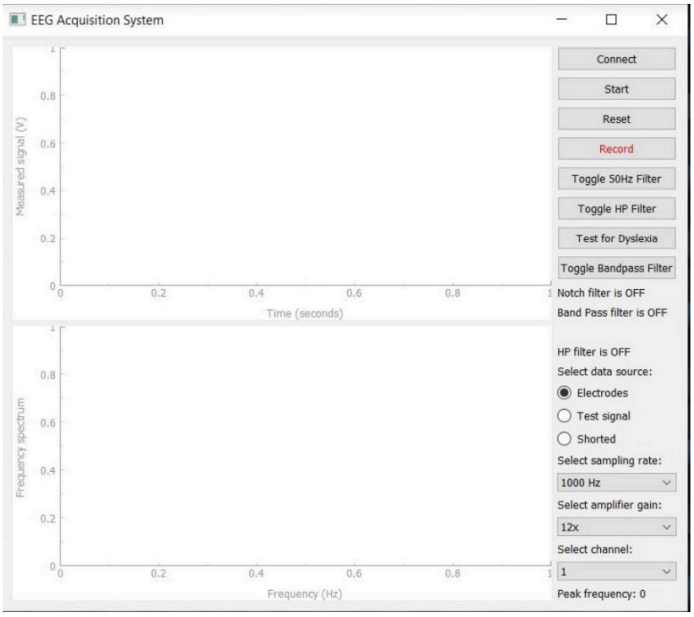


Figure 15: Python GUI window

This GUI window has options to toggle the 50 Hz filter, Low Pass filter and the band pass filter all of which are deployed in the python code using digital filtering. No hardware filters are deployed because they clip waveforms; due to the cutoff from op-amp power supply and induce more noise and drift in the system.

6.2. Output

A simple test to capture EEG data is run on the subject for 60 seconds. The real-time time domain plot and the frequency plot are observed in Figure 16.

The GUI has a continuous rolling time domain plot with a 4 second window (customizable) and frequency domain plot that accumulates frequency responses over the course of time that the experiment is run.

As observed in Figure 16, our acquisition system performs filtering in real-time on the PCB itself with the use of hardware high-pass filter implemented using the RC filter and low-pass & notch filter implemented in software in the code burned on to the microcontroller. Furthermore, the splitting of acquired EEG signal into its component waves based on their frequency ranges and feature classification & identification in different frequency ranges is also implemented on the microcontroller.

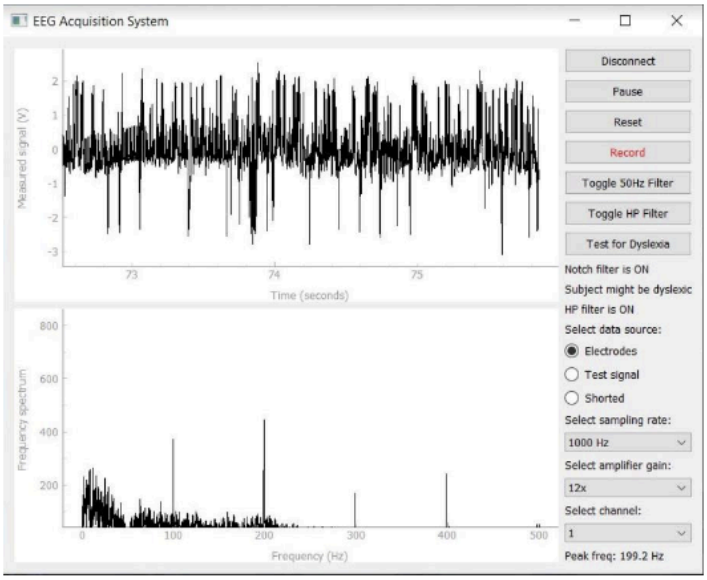


Figure 16: Real-time frequency and time domain plot of acquired EEG signal

The major peaks observed in the frequency domain plot of Figure 16 are in the range of 7 to 30 Hz which comprise mainly of Alpha (7-12 Hz) and Beta (12-30 Hz) waves that constitute a major portion of awake EEG.

7. Use of EEG for IOT application

Now that we have successfully designed and built a plug and play device for EEG acquisition and a GUI for data representation it is time to use it for an IOT application to highlight its edge computing capabilities.

We know in hospitals; the staff has to keep an eye on patients that cannot move to ensure that they are comfortable. The lights are off when they sleep and the fan is turned on as and when needed. Keeping this equipment running all the time does not only cause discomfort to the patient but also a higher electricity bill and waste of energy.

To tackle this issue, we use our EEG acquisition system to monitor the Alpha and Beta waves of the patient. We know that in awake EEG the Alpha waves' peak-to-peak amplitude is low when our eyes are open and high when our eyes are closed. This variation can be used to turn off lights in a room when the patient closes his/her eyes to sleep. Similar interactions are observed in the Beta waves when an individual moves or even thinks of moving their limbs to perform a certain function or action. These variations in frequency and peak-to-peak amplitude can be monitored and with the right threshold can be used to turn on or off a fan.

7.1. Block Diagram

In Figure 17 our acquisition system is now represented as a black-box which wirelessly transmits data to the host computer where the GUI displays EEG waves in both time and frequency domain in real-time.

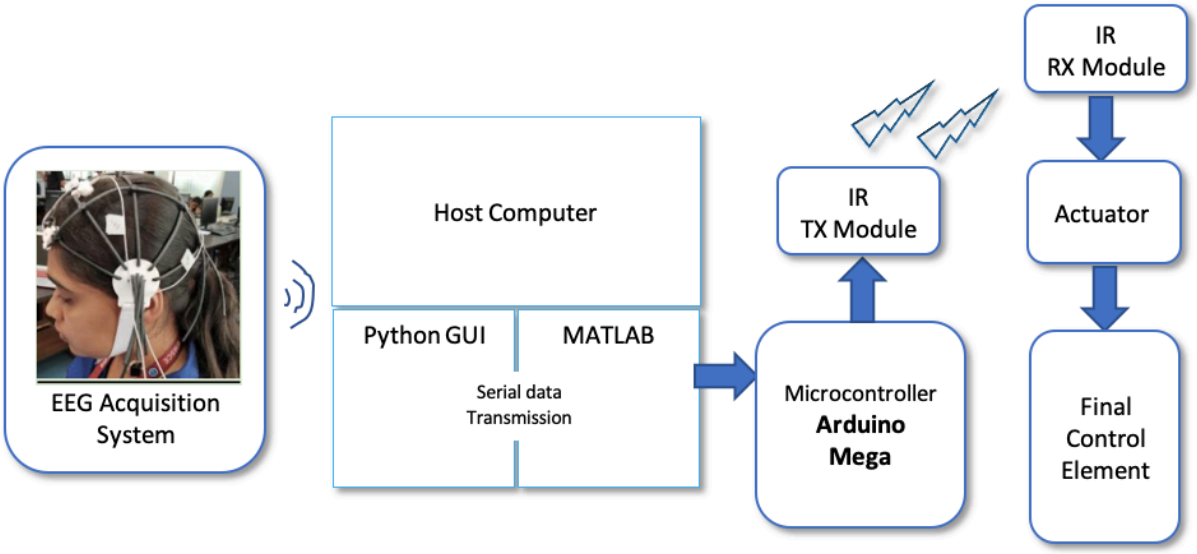
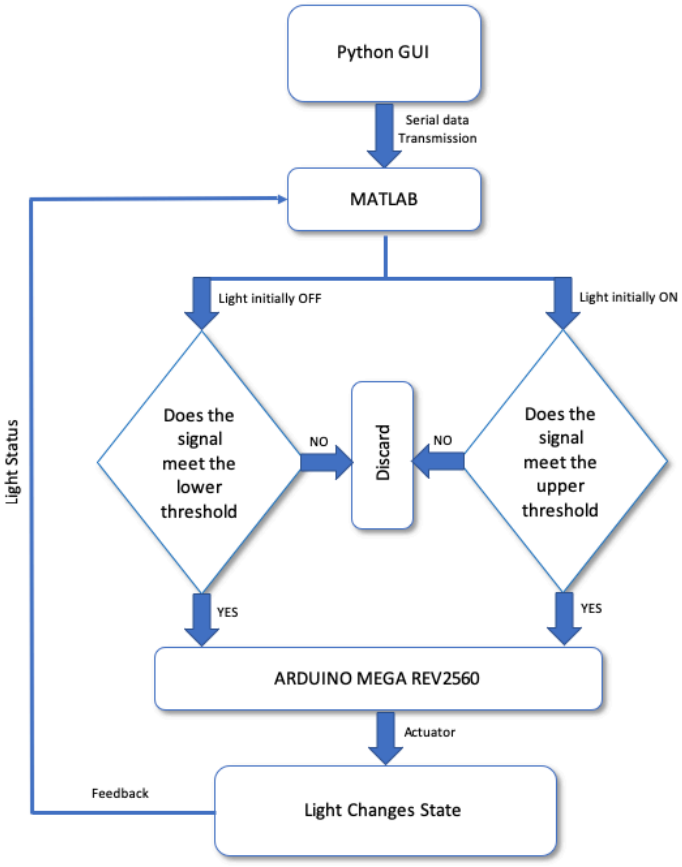


Figure 17: Block diagram for IOT application using EEG

Once the data is received by the GUI, it is transmitted to MATLAB using serial communication. MATLAB checks the peak-to-peak amplitude of the incoming signal and matches it against a set threshold in real-time. Once the threshold is met; in case of Alpha waves the peak-to-peak amplitude crosses the threshold, the program sends a command to Arduino Mega which in-turn transmits it wirelessly using an IR module to actuate or control the final control element which in this case are the lights.



7.2. Algorithm Work-flow

As shown in Figure 18 the incoming EEG data from the Python GUI is serially transmitted to MATLAB where an algorithm continuously scans this data and based upon the initial value of the light (on or off) compares it with the upper (light on) or lower (light off) threshold. When the EEG peak-to-peak amplitude crosses the respective threshold, the Arduino sends a command wirelessly to the actuator to change the state of the light.

A similar algorithm is applied to control the fan, however in that case, changes in a combination of frequency and amplitude; in the Beta range of frequency (12 – 30 Hz), meeting the threshold result in the change of state of the fan. Based on further analysis of the EEG signal in various frequency ranges, multiple such IOT devices can be controlled. However, for such an application the processing module will have to apply AI, machine learning and deep learning algorithms to improve efficiency and the user will also have to be trained enough to concentrate in certain ways in order to control different objects.

8. Results and Discussion

Figure 19 shows the output plotted from channels F1 and F2. The x-axis is time and y-axis is amplitude in milli-volts. The experiment in this case lasted for 20 seconds, where the subject's eyes were open for the first 10 seconds and then closed for the next 10 seconds. As observed in the figure, the EEG peak-to-peak variation is less when the subject's eyes were open except for a few spikes which correspond to blinks. The variation in peak-to-peak voltage increases dramatically half way through the experiment when the subject's eyes were closed. This corresponds to the increase in Alpha Wave activity in Awake EEG when our eyes are closed.

At the 12 second mark in the experiment when the subject's eyes were closed, the lights turned off after detection of increase in Alpha wave activity at the 10 second mark. The 2 second delay is introduced to reduce the chance of false positives.

The above experiment, quite simply demonstrates the accuracy of the data coming from our EEG acquisition system and its ability

to control an IOT device (the lights) precisely in real-time. It is generally observed that the filtering of noise from the EEG signal is done post its collection. Same goes for the removal of artifacts from the EEG signal. Also, the analysis of the signal such as selection of features, classification and training of the classifier is carried out after the data is acquired and stored in a secondary device making the entire flow of denoising, analysis and classification of EEG a delayed process.

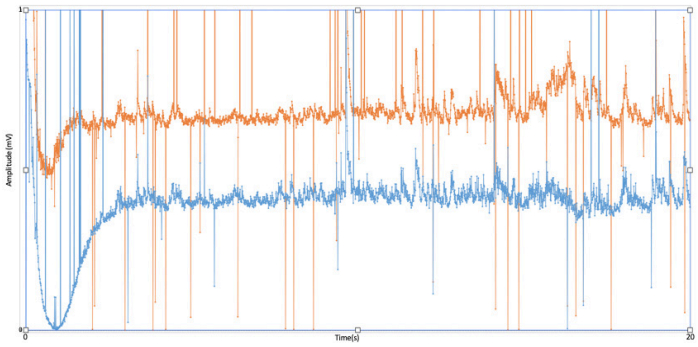


Figure 19: Plotted output from serial data communicated to MATLAB

However, our recommended design can be used to remove the noise in the EEG acquisition system and detect the features from the signal at the time of the acquisition in the device itself. Therefore, moving the denoising of the signal and classification in the EEG acquisition system to a real time system (Edge computing).

Table 1 shows a comparison of our EEG acquisition system with other commercially available systems.

Table 1: Cost comparison of EEG acquisition systems

S.no.	Product	Cost
1.	NeuroSky (1 or 3 channels)	INR 6999 (1 channels) INR 18,000 (3 channels)
2.	EMOTIV EPOC (5 or 14 channels)	USD 299 (5 channels) {INR 21,220} USD 799 (14 channels) {INR 56,707}
3.	OpenBCI (8 – 16 Channels)	USD 249.99 (4 channels) {INR 17,742} USD 499.99 (8 channels) {INR 35,485} USD 949.99 (16 channels) {INR 67,423}
4.	Neuro Insight (4,6,8 can be daisy-chained) – **our acquisition system**	INR 15,821 (4 channels) INR 16,242 (6 channels) INR 16,604 (8 channels)

The cost of our EEG acquisition system is much less compared to other commercially available EEG acquisition systems because of the fact that we have stripped down the design to a bare minimum and have only retained the most essential components that make it work efficiently and effortlessly. It must be noted that the cost mentioned for our system is applicable when single parts were ordered, the cost will further come down by about 15-20% when ordering parts and fabricating PCBs in bulk.

For our particular application we have used the ADS1299 – 4PAG chip that has 4 input EEG channels and one reference channel. For other applications that require more input channels, more ADS1299 chips can be daisy-chained to achieve a 16, 32 or even a 64 channel EEG acquisition system.

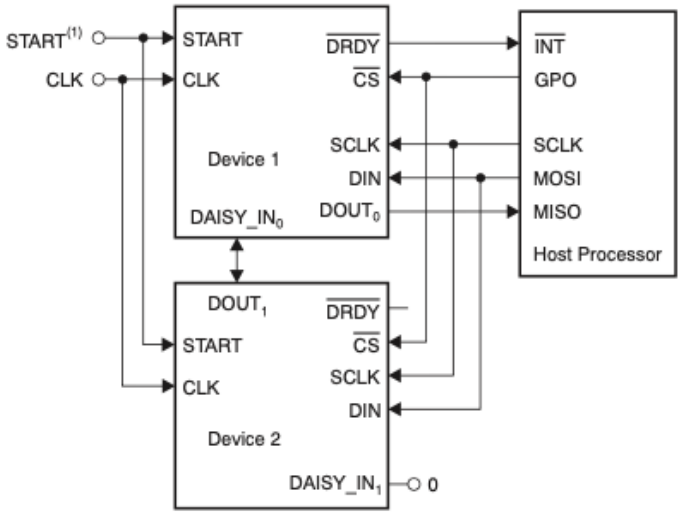


Figure 20: Daisy chain configuration of two ADS1299 chips

Figure 20 shows how two ADS1299 chips can be daisy-chained to achieve up-to 64 input channels.

Furthermore, Table 2 shows a broad specification comparison of our EEG acquisition system with others available in the market. The performance metrics considered are sampling rate, bandwidth, number of EEG acquisition channels, mode of transmission between the acquisition system and the host computer, placement of reference electrodes and the medium of conduction used between the electrode and the scalp. It is worth mentioning that all devices considered, use non-invasive electrodes for data acquisition. A direct comparison of data is also shown in Figure 4.

As evident in table 2, our EEG acquisition system is not only the most cost effective, but also happens to be superior when comparing hardware specifications with those of other commercially available products. The use of Wi-Fi for wireless transmission gives our system the maximum range of communication. Our system also gives the user the ability to choose the number of channels (by daisy-chaining ADS1299, ADS1299-4PAG and ADS1299-6PAG chips), bandwidth (by manipulating the cut-off frequency of high-pass and low-pass filters) and sampling rate (by changing the register value on ADS1299 to be as high as 16kSPS). None of these are offered by any other product in the market right now.

Table 2: Comparison of EEG acquisition systems

S. no.	EEG System	Sampling rate (Hz)	Bandwidth (Hz)	# of EEG Channels	Transmission	Reference	Conductive Mechanism
1.	NeuroSky	512	100	1 or 3	Bluetooth	Earlobe	None
2.	EMOTIV EPOC	256	43	5 or 14	Bluetooth	Left / right mastoid	Saline based
3.	OpenBCI	1000	60	8 to 16	RFD Bluetooth	Earlobe	Gel / Dry
4.	Neuro Insight (**our system**)	1000 (configurable)	100	4,6,8 (can be daisy-chained up to 64)	Wi-Fi	Earlobe	Gel

9. Conclusion

From the above design and application, the followings can be concluded:

- The proposed system captures accurate EEG signal despite the commercially available off the shelf electronic components used in designing the overall system - making it a low-cost yet robust instrument for capturing EEG data.
- In order to improve the robustness of the signal captured the ADC with the instrumentation amplifier is soldered on the same board with the microcontroller, resulting in a single smaller unit, making the overall device more portable.
- To increase the number of EEG signal channels to be captured through the proposed system more ADS1299 chips must be daisy-chained, this will further make the system modular and robust, where more input channels can be added on demand based on the needs of the user.

Further research can be carried out to compare the power consumption v/s performance of various components like the instrumentation amplifier, ADC and microcontroller to design a truly power efficient, battery powered EEG acquisition system capable of collecting data over long periods of time. The reason for moving from an AC powered device to a battery powered one is to get rid of the noise at 50 Hz and its harmonics.

On the software front, the python code can be improved to display more than one EEG channel at a time. Further computational capabilities and the ability to connect with more than one device at a time can be added to completely eliminate the use of MATLAB and finally, the software must also have the ability to export recorded data in popular EEG formats like European Data Format (EDF), CSV, etc. for further analysis and storage.

References

[1] K. Rai et al., "Designing Low Cost Yet Robust EEG Acquisition System" in 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Rourkela, India, 2019 https://doi.org/10.1109/iSES47678.2019.00096

[2] K.D. Nielsen, A.F. Cabrera, O.F. do Nascimento, "EEG based Brain Computer Interface - towards a better control Brain computer interface research at Aalborg university," IEEE Transactions on Neural Systems and Rehabilitation Engineering., 14(2), Article ID 1642769, 202–204, 2006 https://doi.org/10.1109/TNSRE.2006.875529

[3] A. Caplier, S. Charbonnier, A. Picot, "On-Line Detection of Drowsiness using Brain and Visual Information," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, 42(3), 773-774, May 2012 https://doi.org/10.1109/TSMCA.2011.2164242

[4] S. Roy, A. De, N. Panigrahi, "Saccade and Fix Detection from EOG Signal" in 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Rourkela, India, 2019 https://doi.org/10.1109/iSES47678.2019.00099

[5] N. Panigrahi, A. De, S. Roy, "A Method to Detect Blink from The EEG Signal" in Intelligent Computing and Communication, ICICC 2019, Advances in Intelligent Systems and Computing, 1034, Springer, Singapore, 2019. https://doi.org/10.1007/978-981-15-1084-7_24

[6] N. Panigrahi, K. Lavu, S.K. Gorijala, P. Corcoran, S.P. Mohanty, "A Method for Localizing the Eye Pupil for Point-of-Gaze Estimation" in IEEE Potentials 38(1), 37-42, 2019. https://doi.org/10.1109/MPOT.2018.2850540

[7] J. Gomez-Gil, I. San-Jose-Gonzalez, L.F. Nicolas-Alonso, S. Alonso-Garcia "Steering a Tractor by Means of an EMG-Based Human-Machine Interface", MDPI Sensors, 1424-8220, 2011. https://doi.org/10.3390/s110707110

[8] Y. Lin, Yijun Wang and Tzyy-Ping Jung "Assessing the feasibility of online SSVEP decoding in human walking using a consumer EEG headset", Journal of NeuroEngineering and Rehabilitation, Published online 2014 Aug 9 https://doi.org/10.1186/1743-0003-11-119

[9] G. Rosas-Cholula, J.M. Ramirez-Cortes, V. Alarcon-Aquino, P. Gomez-Gil, J.J. Rangel-Magdaleno, C. Reyes-Garcia, "Gyroscope-Driven Mouse Pointer with an EMOTIV® EEG Headset and Data Analysis Based on Empirical Mode Decomposition" Sensors, 13, 10561-10583, 2013. https://doi.org/10.3390/s130810561

[10] T. A. Mariya Celin & B. Preethi "Brain Controlled and Switch Controlled Robotic Leg For The Paraplegics" ITSI Transactions on Electrical and Electronics Engineering (ITSI-TEEE), 2320 – 8945, Volume -1, Issue -1, 2013.

[11] H.A. Shedeed et al., "Brain EEG Signal Proccessing for Controlling a Robotic Arm", IEEE, 152-157, 2013. https://doi.org/10.1109/ICCES.2013.6707191

[12] H. Kamlesh et al., "Brainwave Controlled Robot" International Research Journal of Engineering and Technology (IRJET), 02, 609-612, July 2015

[13] Sun Chuan, Wu Chaozhong, Chu Duanfeng, Tian Fei, "A Design of Brain-computer Interface Control Platform for Intelligent Vehicle", The 3rd International Conference on Transportation Information and Safety, June 25 – June 28, 2015. https://doi.org/10.1109/ICTIS.2015.7232147

[14] B.J.A. Rani, A. Umamakeswari "Electroencephalogram-based Brain Controlled Robotic Wheelchair" Indian Journal of Science and Technology, Vol 8(S9), 188–197, May 2015. https://dx.doi.org/10.17485/ijst/2015/v8iS9/65580

[15] E. Mathe and E. Spyrou, "Connecting a consumer brain-computer interface to an internet-of-things ecosystem," in Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, ser. PETRA '16, 90:1–90:2, 2016. https://doi.org/10.1145/2910674.2935844

[16] C. P. Brennan, P. J. McCullagh, L. Galway, and G. Lightbody, "Promoting autonomy in a smart home environment with a smarter interface" in Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2015. https://doi.org/10.1109/EMBC.2015.7319522

[17] B. Sujatha, G. Ambica "Eeg Based Brain Computer Interface For Controlling Home Appliances" International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 09, Dec-2015.
[18] N. Abdel Ilah et al., "EEG-based Brain-computer Interface for Automating Home Appliances" JOURNAL OF COMPUTERS, 9, NO. 9, SEPTEMBER 2014. http://dx.doi.org/10.4304/jcp.9.9.2159-2166
[19] K.E. Anu et al., "Non Invasive Electroencephalograph Control for Smart Home Automation" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 5, Special Issue 3, March 2016.
[20] K. Alhalaseh et al., "Home Automation Application Using Eeg Sensor" Proceedings of 98th The IRES International Conference, Antalya, Turkey, 21st-22nd January, 2018.
[21] H. Mohammad et al., "Automated Classification of L/R Hand Movement EEG Signals using Advanced Feature Extraction and Machine Learning" (IJACSA) International Journal of Advanced Computer Science and Applications, 4, No. 6, 2013. http://dx.doi.org/10.14569/IJACSA.2013.0406
[22] T. Uktveris, V. Jusas "Development of a Modular Board for EEG Signal Acquisition" MDPI Sensors, Published: 3 July 2018. https://doi.org/10.3390/s18072140
[23] B. Luan, M. Sun, W. Jia "Portable Amplifier Design for a Novel EEG Monitor in Point-of- Care Applications" Proc IEEE Annu Northeast Bioeng Conference, 2014. https://doi.org/10.1109/NEBC.2012.6207127
[24] A.J. Bhagawati, R. Chutia "Design of Single Channel Portable Eeg Signal Acquisition System For Brain Computer Interface Application" International Journal of Biomedical Engineering and Science (IJBES), 3(1), January 2016. http://dx.doi.org/10.5121/ijbes.2016.310
[25] C. Jaganathan, A. Amudhavalli, T. Janani, M. Dhanalakshmi, Nirmala Madian "Automated Algorithm for Extracting α, β, δ, θ Of A Human Eeg" International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), April 2015.
[26] J. Tian, W. Song "LabVIEW for EEG Signal Processing" Saudi Journal of Engineering and Technology 1(4), Oct-Dec, 2016. DOI: https://doi.org/10.21276/sjeat.2016.1.4.10
[27] M. Rajya Lakshmi, Dr. T. V. Prasad, Dr. V. Chandra Prakash "Survey on EEG Signal Processing Methods" International Journal of Advanced Research in Computer Science and Software Engineering 4(1), 84-91, January 2014.
[28] N.R. Miss et al., "Review: Wavelet transform based electroencephalogram methods" International Journal of Trend in Scientific Research and Development (IJTSRD) 2(3), Mar-Apr Page: 1777, 2018.
[29] T. Pham et al., "A Study on the Feasibility of Using EEG Signals for Authentication Purpose" ICONIP 2013, Part II, LNCS 8227, 562–569, 2013.
[30] T. Wilaiprasitporn, A. Ditthapron, K. Matchaparn, T. Tongbuasirilai, N. Banluesombatkul, E. Chuangsuwanich "Affective EEG-Based Person Identification Using the Deep Learning Approach" Journal of Latex Class Files, 14, August 2018. https://doi.org/10.1109/TCDS.2019.2924648
[31] M. Byoung-Kyong, H. Suk, A. Min-Hee, L. Min-Ho, L. Seong-Whan, "Individual Identification Using Cognitive Electroencephalographic Neurodynamics" IEEE Transactions on Information Forensics And Security, 12(9), September 2017. https://doi.org/10.1109/TIFS.2017.2699944
[32] G. Matthews, "Metrics for individual differences in EEG response to cognitive workload: Optimizing performance prediction" Personality and Individual Differences 118, 22–28, 2017 https://doi.org/10.1016/j.paid.2017.03.002
[33] D. Gajic, Z. Djurovic, S.D. Gennaro, F. Gustafsson "Classification of EEG Signals for Detection of Epileptic Seizures Based on Wavelets and Statistical Pattern Recognition" Biomedical Engineering: Applications, Basis and Communications, 26(21), 2014. https://doi.org/10.4015/S1016237214500215
[34] S. Mantri, V. Patil, R. Mitkar "EEG Based Emotional Distress Analysis – A Survey" International Journal of Engineering Research and Development 4(6), 24-28, 2012.
[35] F. Al-shargie, T.B. Tang, N. Badruddin, M. Kiguchi "Mental Stress Quantification Using Eeg Signals" ResearchGate Publication 2016, from book International Conference for Innovation in Biomedical Engineering and Life Sciences: ICIBEL2015, Putrajaya, Malaysia, (15-19), 6-8 2015. https://doi.org/10.1007/978-981-10-0266-3_4
[36] R. Khosrowabadi "Stress and Perception of Emotional Stimuli: Long-term Stress Rewiring the Brain" Basic and Clinical Neuroscience, 9(2), 107-120, 2017. https://dx.doi.org/10.29252%2FNIRP.BCN.9.2.107
[37] M. Jeffrey et al., "Acute EEG Patterns Associated with Transient Ischemic Attack" Clinical EEG and Neuroscience, Article first published online; 2019. https://doi.org/10.1177/1550059418790708
[38] M. Nami, S. Mehrabi, S. Derman "Employing Neural Network Methods to Label Sleep EEG Micro-Arousals in Obstructive Sleep Apnea Syndrome" Journal of Advanced Medical Sciences and Applied Technologies 3(4):221-226, 2017. http://dx.doi.org/10.32598/jamsat.3.4.221

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 127-132 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Infrared Uplink Implementation for Software Defined Visible Light Communication Systems

Oswaldo René Banda-Sayco*

Universidad Nacional de San Agustín de Arequipa, 04001, Peru

ARTICLE INFO

Article history:
Received: 30 July, 2020
Accepted: 01 October, 2020
Online: 08 November, 2020

Keywords:
Infrared Communication
Analog front-end
Visible Light Communication
Software Defined Radio
USRP
GNU Radio

ABSTRACT

The main advantage of software-defined visible light communication (VLC) systems is that they allow us to reuse modulation, coding, synchronization and medium access techniques used in conventional radio systems. This article focuses on the implementation of an uplink alternative using an infrared (IR) light communication system. For this goal, analog front-ends are designed and implemented. Therefore, they convert a radio frequency signal into an optical signal. In addition, a Universal Software Radio Peripheral (USRP) is used as a software-defined radio platform. The obtained bit error rate is 5.33×10^{-5} *using QPSK modulation with a bit rate of 4 Mbps at a distance of 160 cm between transmitter and receiver.*

1. Introduction

This paper is an extension of work originally presented in *2019 IEEE Colombian Conference on Communications and Computing (COLCOM)* [1]. This article aims to focus into the design process followed to implement analog front ends for the infrared (IR) communication uplink. On the other hand, an explanation is given about the blocks that make up the program to measure the bit error rate (BER) of the IR system. Finally, the signal-to-noise ratio (SNR) measurement is made and compared with the bit error rate (BER) obtained at different distances between transmitter and receiver.

Visible Light Communication (VLC) is currently being developed as a solution that enables the use of standard illumination light-emitting diodes (LEDs) to transmit data wirelessly. This type of communication is an alternative to mitigate the effects of saturation of the radioelectric spectrum in densely populated areas. Furthermore, it is known that the number of wireless devices connected to the internet is growing at an annual rate of 47% until 2021 [2]. This is mainly due to the greater access of terminal devices such as smartphones, tablets, laptops and recently to the deployment of the Internet of Things (IoT). For this reason, VLC is a technology that can serve as a complement of the currently deployed communication systems and provide access to the network. An important feature of VLC is that it can be used as a ubiquitous communication medium because LED lighting lamps are widely used in existing lighting infrastructure. On the other hand, VLC operates in the visible light band, which is not licensed and there is no electromagnetic interference generated by other communication equipment [3].

Information is sent and received through one or more LEDs on the transmitting side and a photodiode on the receiving side using a technique known as intensity modulation and direct detection (IM-DD) [4]. In such a way that the information that is sent produces variations in the level of illumination of the LED stage, these variations are imperceptible to the human eye. After that, the receiving photodiode converts the optical signal into an electrical signal for further processing.

In [5], the authors focused on the design and implementation of Analog Front-Ends, which would allow easy replication of the system through the use of low-cost components. The achieved bit rate was 20 kbps with a BER of 4.89×10^{-10} using OOK modulation and Manchester encoding. In [6], a similar paper was done, but it did not specify the components of the designed interfaces, a BER of 9.77×10^{-5} was measured at a bit rate of 2.5 Mbps and at a distance of 1.6 meters using BPSK modulation technique. In [7], an uplink based on infrared communication is proposed, the simulation result shows a value of 3.57 Gb / s using a transmitter with beam stearing and an angle diversity receiver. However, the experimental implementation stage of the system was not reached. In [8], it was possible to implement an infrared uplink by using an off-the-shelf receiver module (TSOP 1730) and the PPM modulation technique, however, the data transfer rate of

*Corresponding Author: Oswaldo René Banda Sayco, Arequipa, Peru, obanda@unsa.edu.pe

the infrared link reaches up to 1200 bps. In [9], a surveillance camera is used as a receiver for an uplink based on infrared light. In [10], an IR link was implemented using a silicon integrated beam steering device, the achieved bit rate was 12.5 Gb/s and the bit error rate (BER) was less than 10^{-9}.

This article shows the design and implementation process of Analog Front-Ends used on the transmitter and receiver side, achieving a bit rate of 4 Mbps and a BER of 5.33×10^{-5} using QPSK modulation at a distance of 1.6 m. These interfaces have the advantage of being coupled to Software Defined Radio (SDR) systems, in such a way that the modulation, coding and synchronization of the data is carried out within a computer connected to a Universal Software Radio Peripheral (USRP). This allows for rapid prototyping and measurement of the IR communication system.

Section two presents the methodology used to carry out this work, section three delves into the design and implementation of Front-Ends, section 4 shows the results of the BER and SNR. Finally, section 5 shows the conclusions of this research work.

2. Methodology

Figure 1 shows the methodology followed to design and implement the Analog Front-Ends. First, an analysis of the communication system to be implemented is made. This analysis allows us to know all the elements that are required to make the VLC system work correctly. In addition, it is necessary to define the design parameters that allow the development of Analog Front-Ends. After that, the second procedure is Development of the Front-Ends. To carry out this task, it has been divided into two threads: design and implementation.

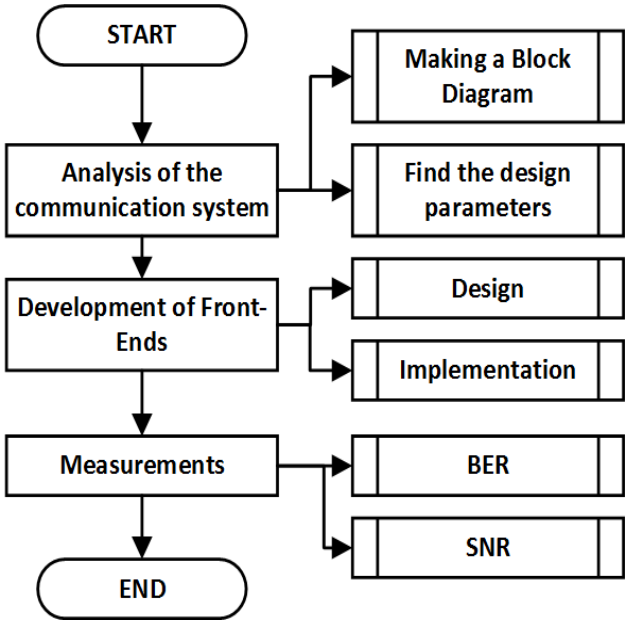


Figure 1: Methodology diagram

Finally, the last step is to carry out the corresponding measurements of the implemented communication system. To do this, a BER measurement is made as a function of the separation distance between transmitter and receiver. On the other hand, the measurement of the signal-to-noise ratio (SNR) is also carried out and a comparison is made with the previously measured BER.

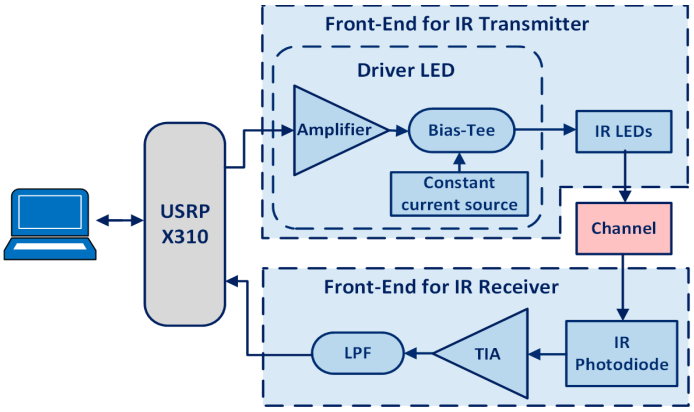


Figure 2: Block diagram of the IR communication system for BER measurement

3. Analysis of the Communication System

3.1. Block Diagram of the System

Figure 2 shows the block diagram of the IR communication system implemented for the purpose of making the BER measurement. This diagram is based on the research article previously made in [1] . It is observed that a laptop has a software stage where the modulation and synchronization of the information is carried out. After that, a USRP X300 is connected, which is a software-defined radio platform that allows to generate a modulated signal in the RF band. In this case, a central frequency of 5MHz has been used to carry out this test. Then, the signal goes to the transmitter front-end. This front-end is composed by an amplifier, a bias-tee, a constant current source and the IR LEDs. After that, the signal travels towards the receiver through the channel. In this case, the channel is the air in the indoor environment. On the other hand, the receiver front-end includes an IR photodiode, a transimpedance amplifier (TIA) and a low pass filter (LPF).

Table 1: Design Parameters for the IR link

Parameter	Description
Application	Indoor
Coverage range	1.6 m
Configuration	Direct LOS link
Bitrate	4 Mbps
Modulation	QPSK / BPSK
Transmitter	LED Specification (each one) - Radiant Intensity: 550 mW/sr - View angle: 10° - Wavelenght: 800 nm – 875 nm
Receiver	Photodiode specifications: - Code: SFH 203 FA - Spectral response: 800 nm – 1100 nm - Sensitive area: 1 mm^2
OP-AMPs	LT6268-10 and LT6268 with dual supply ±2.5 V_{DC}

3.2. Design Parameters

The design parameters of the electronic circuits for the transmitter and receiver analog interfaces are shown in Table 1. It is observed that the configuration used in this paper requires a line of sight (LOS), this is due to the fact that the angle of view of the LEDs arrangement is 10°. Which allows maximum energy transfer in only one direction. In this way, the use of an optical lens that amplifies the optical signal at the receiver is avoided. In addition to this, the other parameters taken into account to implement this communication system can be read.

4. Development of Front-Ends

Front-ends are developed based on the design parameters mentioned in the previous section. First, the design of the front-ends is done. After that, the physical implementation of them is made.

4.1. Transmitter Design

The transmitter front-end is composed by four blocks. The first block is an amplifier that must take the signal from the USRP and feed a LED array. The second block is a bias-tee that superimposes the RF signal coming from the amplifier on a constant current level coming from the third block called "Constant current source". The three previous blocks make up the LED driver. Finally, the last block is an array of LEDs, it is composed of eight high-power LEDs placed in series, which produce an optical signal that will travel through free space until it reaches the receiver front-end.

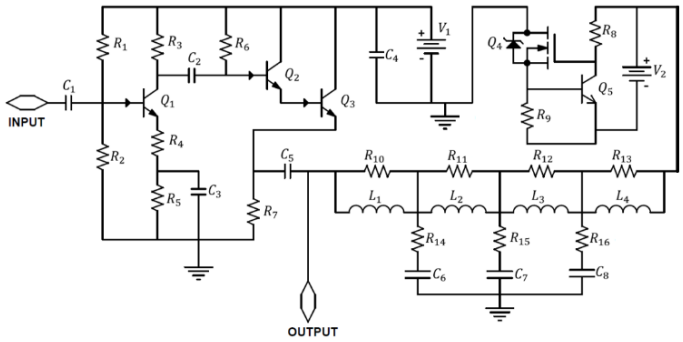


Figure 3: Schematic circuit of the Transmitter Front-end

Figure 3 shows the schematic circuit of the transmitter front-end, it can be seen that it includes a common emitter amplifier, that includes the transistor Q_1. After that, the circuit is followed by a Darlington amplifier, which is made of the transistors Q_2 and Q_3. The common emitter amplifier was designed following the procedure described in [11]. The design equations are as follows:

$$I_b = \frac{I_c}{\beta} \tag{1}$$

$$V_{ce} = \frac{V_1}{3} \tag{2}$$

$$R_3 = R_4 + R_5 \tag{3}$$

$$I_c(R_3 + R_4 + R_5) + V_{ce} = V_1 \tag{4}$$

$$R_3 = R_4 + R_5 = \frac{V_1 - V_{ce}}{2I_c} \tag{5}$$

$$R_{bb} = R_1 // R_2 \tag{6}$$

$$R_{bb} \ll \beta(R_4 + R_5) \tag{7}$$

$$Z_{in} = R_{bb} // \beta(r_e + R_4) \tag{8}$$

$$Z_{out} = R_3 \tag{9}$$

From (1) the base current is obtained from the value of the collector current, which in this case is 10 mA, in addition β has a value of 95. In (2) and (4), V_1 has a value of 12 V. The values of the resistors R_3, R_4 and R_5 are calculated from (4) and (5). On the other hand, (6) and (7) allow the calculation of R_1 and R_2. Next, (8) and (9) allow us to know the input and output impedances. Finally, a Darlington amplifier is inserted to reduce the output impedance and increase the output power.

In Figure 3, Q_4 is a MOSFET that provides a constant current source whose value is defined by (10). In this case, 140 mA was chosen as the I_D value because it allows the LEDs to keep turned on.

$$I_D = \frac{V_{BE}}{R_2} \tag{10}$$

I_D: drain current of MOSFET Q_4

V_{BE}: Base-Emitter Saturation Voltage of Transistor Q_5

LED's must be turned on to modulate their light intensity without turning them off. For this purpose, it is necessary to superimpose the AC signal coming from the amplifier with the DC level coming from the constant current source. The bias-tee is in charge of this task. Finally, this LED driver is connected to an array of eight LEDs in series. In this way, the Front-end transmitter has been designed.

4.2. Receiver Design

Figure 4 shows the schematic circuit of the receiver front-end, which is composed of a photodiode, a transimpedance amplifier and a low-pass filter. The photodiode makes it possible to transform an optical signal into an electrical one. In this project, the SFH 203 FA photodiode is used, which has a switching time of 5 ns, it is inversely polarized with a voltage of 12 V. Furthermore, R_3 and C_5 constitute a low-pass RC filter in order to eliminate noise from source V_1.

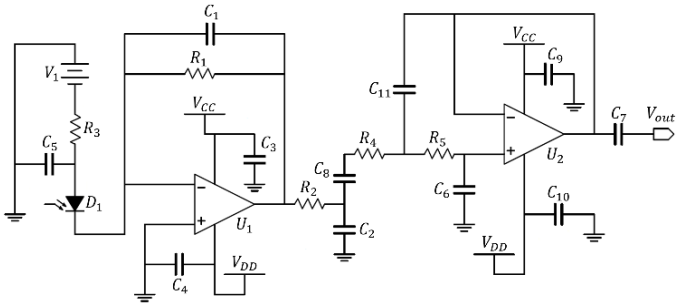


Figure 4: Schematic circuit of the Receiver Front-end

The current at the output of the photodiode enters a transimpedance amplifier. The gain of the transimpedance amplifier is equal to the feedback resistance R_1. After a voltage is obtained at the output of the transimpedance amplifier (TIA), the signal passes through a two-stage low-pass filter. The first stage is

made up of an RC filter composed of R_2 and C_2 and whose cutoff frequency can be calculated using (11). After that, the coupling capacitor C_8 is connected, this capacitor allows only the alternating component of the signal to pass. In this way, the voltage produced by the ambient light is minimized. Finally, a Chebyshev low-pass filter is responsible for eliminating high-frequency noise and adapting the output impedance for connection to the USRP equipment. The design of this filter was done following the procedure described in [12], the calculations were made using (12), (13), (14) and (15).

$$f_{c1} = \frac{1}{2\pi(R_2)(C_2)} \tag{11}$$

$$R_4 = R_5 = \frac{a_1}{2w_c C_6} \tag{12}$$

$$C_{11} = \frac{b_1}{{w_c}^2 R_4 R_5 C_6} \tag{13}$$

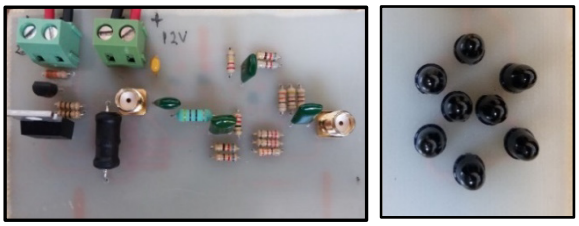
$$a_1 = 1.1813 \tag{14}$$

$$b_1 = 1.7775 \tag{15}$$

4.3. Implementation of Front-ends

The transmitter front-end was implemented as shown in Figure 5. The physical implementation of this circuit can be shown on two single printed circuit boards (PCB). Figure 5.a shows the LED driver. On the other hand, Figure 5.b is the IR-LED array. The reason why the LED array is not included in the same PCB is because it allows to do further tests with different LED configurations. In this way, it is avoided to build new front-ends in order to speed up the tests and improve the transmitting analog interface. Both circuits are connected by a coaxial cable in the experimental setup.

Figure 6 shows the physical circuit of the receiver front-end for the IR uplink. It is observed that the output connection is made through a female SMA connector, to which an RG58 coaxial cable is connected.



a) b)

Figure 5: Physical implementation of the transmitter Front-end.

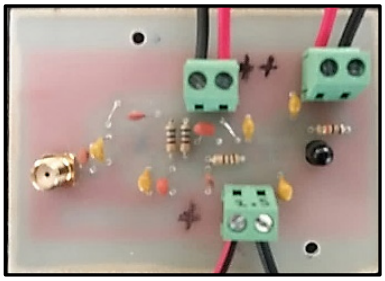


Figure 6: Physical implementation of the receiver Front-end

5. Measurements

Once the Front-ends have been implemented, the Bit Error Rate (BER) measurement is performed. To do this, it is necessary to establish the experimental setup shown in Figure 7. This setup is based on the block diagram described in the third section of this paper. The connection between the laptop and the USRP is made through a gigabit ethernet cable. The connections between the USRP and the front-ends are made through RG58 coaxial cable.

To carry out the BER measurement, the program shown in Figure 8 is used. This program was developed in GNU Radio from the information available in [6], with the difference that now QPSK modulation and the rate are used. sampling rate is changed to 4 MS/s.

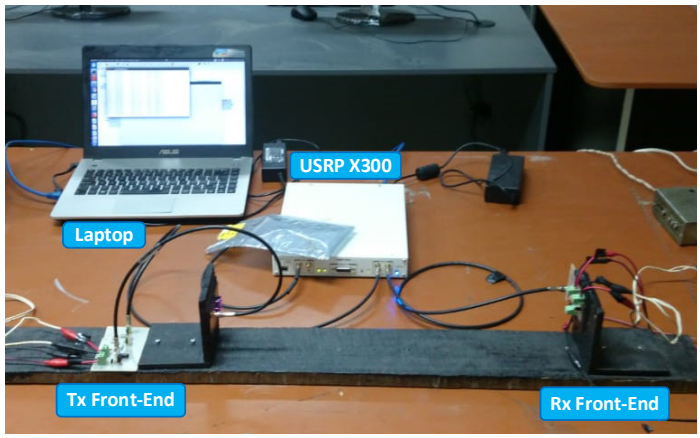


Figure 7: Experimental setup for BER measurement

The explanation of the blocks that make up this program is as follows:

- The "PRBS Source B" block is responsible for generating a random sequence of bits, which is repeated indefinitely.
- “Preamble Insert” is a block that inserts a bit preamble, which is used to indicate the moment at which the BER measurement starts.
- "Root Raiced Cosine Filter" is a block that is used in transmission and reception to minimize intersymbolic interference.
- “Costas Loop” is responsible for the frequency recovery of the carrier signal.
- Subsequently, the "Decode using Preamble Synch" block takes care of the demodulation and detection of the preamble to start the BER measurement.
- The “Binary Slicer” block transforms the demodulated signal into bits. These bits can have two values 0 or 1.
- Finally, the "PRBS Sink B" block is in charge of comparing the received bits with the original sequence and showing the bit error rate (BER).

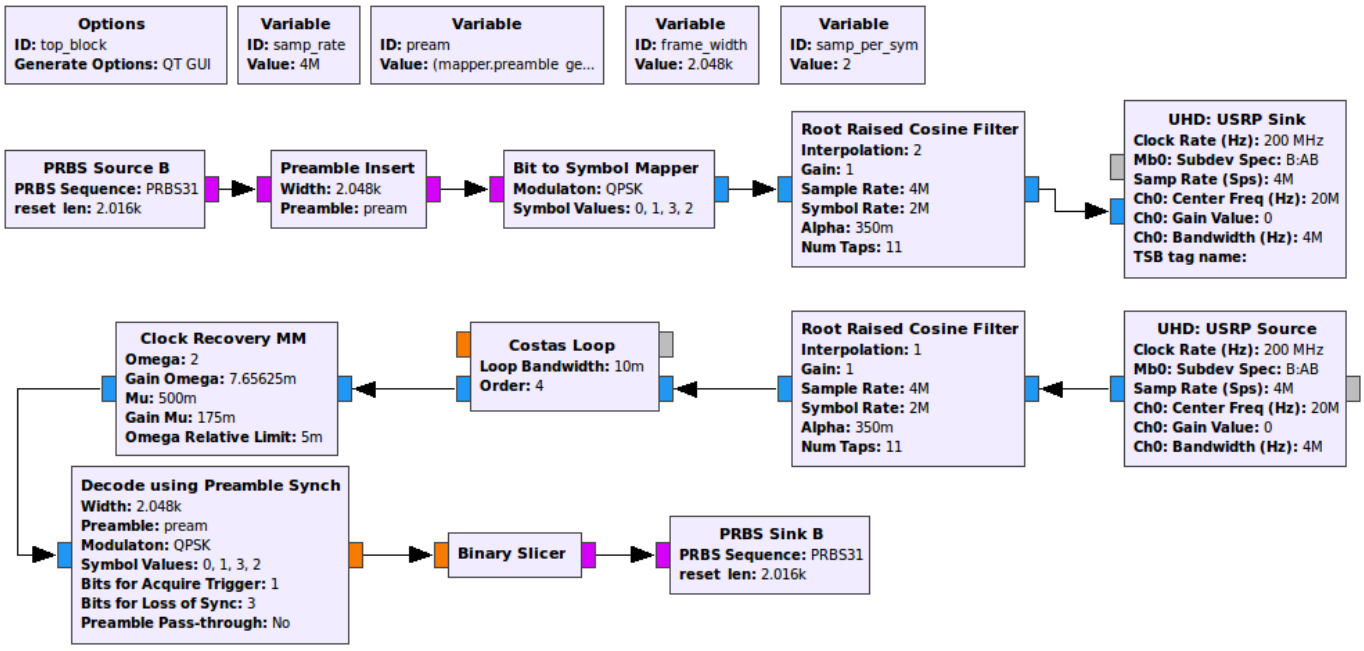


Figure 8: Program in GNU Radio for BER measurement using QPSK modulation

Figure 9 shows the results obtained from the BER measurement as a function of the separation distance between transmitter and receiver. The red graph indicates the BER measured using QPSK modulation at a bit rate of 4 Mbps. A BER of 5.33×10^{-5} was obtained at a distance of 160 cm between transmitter and receiver. The blue graph shows the BER obtained using BPSK modulation, which has a value of 1.35×10^{-5} at a distance of 160 cm and the bit rate is reduced to 2 Mbps.

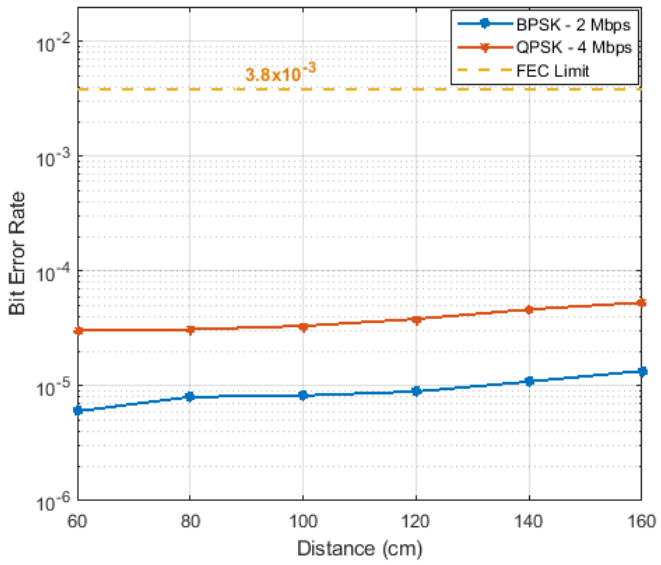


Figure 9: BER with respect to distance

Finally, signal-to-noise ratio (SNR) is measured at different distances between the transmitter and receiver using BPSK modulation. From the obtained data, it has been possible to make the graph of Figure 10, in which the BER is observed as a function of the SNR. With a SNR of 36.2 dB, a BER of 1.35×10^{-5} was measured. In the other hand, with a SNR of 46.4 dB, a BER of 8.032×10^{-6} was obtained. It can be seen that, if the SNR increases, a lower BER is measured. In this way, having an SNR greater than 45 dB, the BER seems to converge, however this is due to the fact that a finite number of packets were sent. Therefore, it is convenient to have the least amount of noise in the system or, failing that, to increase the power of the transmitted signal.

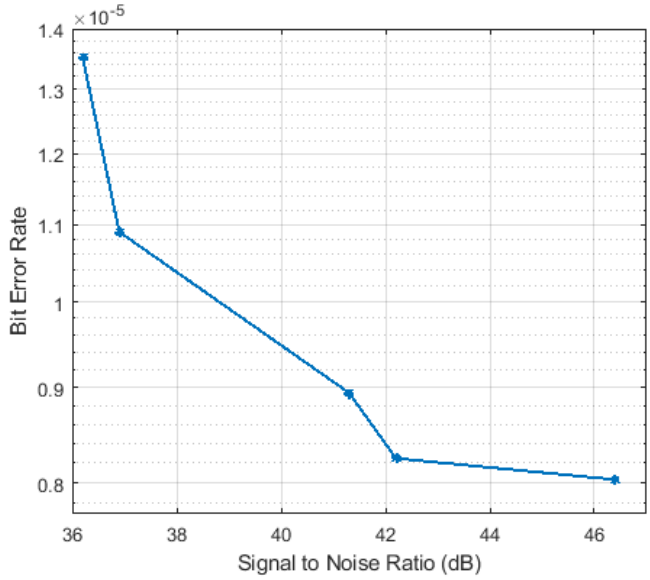


Figure 10: Measured BER vs SNR

6. Conclusions

- A bit error rate of 5.33×10^{-5} has been measured when transmitting at a bit rate of 4 Mbps using QPSK modulation at a distance of 160 cm between transmitter and receiver for an infrared light communication system.

- Signal to noise ratio decreases when the distance between transmitter and receiver is greater. In turn, the bit error rate increases as there is a lower signal-to-noise ratio.
- The transmitter front-end provides a constant current level of 140 mA, this keeps the LED stage turned on, while at the same it is modulated the intensity of the optical signal.
- The receiver front-end is composed by a transimpedance amplifier and a two-stage low-pass filter. The output voltage does not exceed 2 Vpp, which is the input limit for a USRP X300.
- A software program was developed in GNU Radio to measure the bit error rate of the infrared communication link. The used sampling rate is 4 MS/s and the central operating frequency is 5 MHz.
- Front-ends have been designed and implemented for infrared light communication that can be used in Software Defined Radio systems as an alternative for VLC uplink. Taking into account the requirements of this type of links.
- The main advantage of this system is that, as it is based on software defined radio, the modulation, coding and synchronization stages are implemented within a personal computer. Therefore, to implement it, it is necessary to use analog front-ends that allow the conversion of RF signals to optical signals.
- The main disadvantage is that the bit rate is limited by the capacity of the computer processor where the transmission and reception programs are executed, it was necessary to carry out tests with different personal computers and the one with the best performance was chosen.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author thanks to the National University of San Agustin of Arequipa for providing access to the laboratories to carry out the tests of this paper. On the other hand, the author also thanks to professors Alex Cartagena Gordillo, Pablo Lizardo Pari Pinto and Ruben Matheos Herrera.

References

[1] O.R. Banda Sayco, A.C. Gordillo, "Development of an IR system as an alternative for VLC uplink," in 2019 IEEE Colombian Conference on Communications and Computing (COLCOM), IEEE: 1–5, 2019, doi:10.1109/ColComCon.2019.8809131.

[2] M. Uysal, "Visible Light Communications: From Theory to Industrial Standardization," in Optical Fiber Communication Conference (OFC) 2019, OSA, Washington, D.C.: Th3I.4, 2019, doi:10.1364/OFC.2019.Th3I.4.

[3] Z. Wang, Q. Wang, W. Huang, Z. Xu, Visible Light Communications Modulation and Signal Processing, 2017.

[4] P. Kamsula, Desing and implementation of a bi-directional visible light communication testbed, University of Oulu, 2015.

[5] M. Basha, M.J.N. Sibley, P.J. Mather, A. Makama, "Design and Implementation of a Tuned Analog Front-End for Extending VLC Transmission Range," in 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), IEEE: 173–177, 2018, doi:10.1109/iCCECOME.2018.8658815.

[6] O.R.B. Sayco, A.C. Gordillo, "Design and Implementation for a USRP – Based Visible Light Communications Transceiver," in 2019 UNSA International Symposium on Communications (UNSA ISCOMM), IEEE: 1–5, 2019, doi:10.1109/UNSAISC.2019.8712826.

[7] O.Z. Alsulami, M.T. Alresheedi, J.M.H. Elmirghani, "Infrared Uplink Design for Visible Light Communication (VLC) Systems with Beam Steering," in 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), IEEE: 57–60, 2019, doi:10.1109/CSE/EUC.2019.00020.

[8] T. Adiono, A. Pradana, R.V.W. Putra, W.A. Cahyadi, Y.H. Chung, "Physical layer design with analog front end for bidirectional DCO-OFDM visible light communications," Optik, **138**, 103–118, 2017, doi:10.1016/j.ijleo.2017.03.046.

[9] T.C. Bui, S. Kiravittaya, "Demonstration of using camera communication based infrared LED for uplink in indoor visible light communication," in 2016 IEEE 6th International Conference on Communications and Electronics, IEEE ICCE 2016, 2016, doi:10.1109/CCE.2016.7562615.

[10] K. Wang, Z. Yuan, E. Wong, K. Alameh, H. Li, K. Sithamparanathan, E. Skafidas, "Experimental demonstration of indoor infrared optical wireless communications with a silicon photonic integrated circuit," Journal of Lightwave Technology, 2019, doi:10.1109/JLT.2018.2889252.

[11] Hunter College - City University of New York, 15. Transistor Amplifier Design and Measurement, Http://Www.Hunter.Cuny.Edu/Physics/Courses/Physics222/Repository/Files/Pdf/ElectronicsLab15.Pdf, 1–13.

[12] T. Kugelstadt, Active Filter Design Techniques, Elsevier: 365–438, 2009, doi:10.1016/B978-1-85617-505-0.00020-X.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 1, 128-133 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Downlink Indoor Coverage Performance of Unmanned Aerial Vehicle LTE Base Stations

Mahmut Demirtaş*, Kerem Çağdaş Durmuş, Gülçin Tanış, Caner Arslan, Metin Balcı

Ulak Telecommunications Inc., Istanbul, 34906, Turkey

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 25 December, 2020
Online: 15 January, 2021

Keywords:
Unmanned air vehicle
Mission critical
Air-to-ground communication
Indoor coverage

ABSTRACT

In this study, we study on the downlink indoor coverage performance of unmanned air vehicle (UAV) base stations. We consider a probabilistic expression for air-to-ground (ATG) path loss, and a deterministic one for additional indoor losses in order to provide a practical model. One of our important assumptions is that the UAV base station operates at the frequencies reserved for 4th Generation (4G) – Long Term Evaluation (LTE) based mission critical services in Turkey –around 2.6 GHz–. Therefore, we are able to neglect intercell interference issue, and we may rely on signal-to-noise ratio (SNR) for our coverage definition. We consider four different SNR requirement throughout our performance evaluation, and investigate the effect of UAV altitude and other related parameters on the radius of service area. We first observe that rural coverage performance is always better than urban conditions –this result is fully compatible with the fact that attenuation levels significantly arise in urban regions–. In addition, we show that increasing quality of service (QoS) requirement and/or using a more directive antenna unit decrease the coverage radius as they are expected. Thereupon, we conclude that it is possible to obtain the optimum altitude level –by employing the framework proposed here– in order to satisfy certain service criteria.

1 Introduction

In this paper, we extend one of our earlier studies presented in 27th Signal Processing and Communications Applications Conference (SIU) [1]. As we argue in [1], UAV base station idea seems to be a promising solution to provide cellular service for both the locations which cannot be covered by employing terrestrial networks and the times in which terrestrial networks may not be serving at all (e.g. natural disasters). In [2] and [3], earlier studies related to the UAV base station concept are summarized, and possible use cases are investigated. In addition, topics open for further studies are discussed in these publications as well. Among several other topics to discuss, modelling channel conditions of such a non-terrestrial network in a practical and also tractable way is an important requirement [4]–[11].

In [4, 5, 6], coverage performance of UAV networks is investigated for outdoor users, and the effect of UAV altitude on the coverage radius is discussed in detail. In the studies mentioned, ATG channel is modelled in two parts. First part corresponds to the free-space losses, and assumed to be a deterministic function which depends on the radiation distance. Second part basically aims the losses in where man-made structures exist –urban environment–, and modelled as a Normal distributed random variable. In this paper, we follow this piecewise model to express the path loss related to ATG channel, and examine urban and rural conditions separately. The reason of the using deterministic approaches is that reliable characterization of the ATG propagation for large scale fading statistics. In addition, it is important to note that fast fading effect of the channel is ignored throughout this study.

Some earlier publications related to the topic utilize a practical UAV to form a testbed, and generate empirical channel models to figure out ATG propagation characteristics [9]–[11]. All three of those are funded by the same institution, and aims to model over-water, mountainous and near-urban environments, respectively. As the next step of this work, we also target to assemble a functional and realistic testbed.

Besides, some previous studies utilize stochastic processes to concentrate on outdoor coverage performance of UAV base stations [12, 13]. Authors of the former assume that both UAV and terrestrial base stations are deployed in an overlapping manner, and coordinate with each other to provide a continuous service. Such an overlapping structure of UAV and terrestrial base stations is not cov-

*Corresponding Author: Mahmut Demirtaş, Turkey, +903122869487, mahmut.demirtas@ulakhaberlesme.com.tr

ered in the scope of this paper. The latter one considers only UAV base stations distributed according to a Poisson point process (PPP). Then, authors provide approximate expressions for both coverage ratio and average data rate. Here, we emphasize that we assume there exists only one UAV base station that operates on mission critical frequency bands, and therefore such stochastic processes are not required to model UAV base station density around. Yet, as a future concern, investigating networks consist of multiple UAV base stations is already scheduled.

Another significant aspect of this work is considering additional losses to serve indoor users. In literature, these indoor losses are generally studied for terrestrial networks rather than UAV ones [14]–[17]. In [14, 15], indoor losses are evaluated as a function of number of floors penetrated. Authors of [16, 17] offer a more complicated system model which accounts the number of internal/external walls and indoor propagation distance. Since only the latter one of these supports sub-6 GHz frequency bands, we follow the model proposed in [17]. Once we build a practical testbed, generating a novel indoor propagation model will be another important objective in our agenda.

Rest of this paper is organized as follows: Section 2 summarizes the system model that we employ, Section 3 proposes a parametric expression for indoor coverage ratio, Section 4 calculates coverage radius for several SNR demands, quality-of-service (QoS) requirements and antenna directivity levels, and finally Section 5 discusses the results obtained throughout our extensive simulations.

2 System Model

The aim of this study is to show that UAV base stations can be provide a fast and effective coverage in areas not covered by terrestrial networks or in emergency situations –by using frequency bands reserved for mission critical services–. Considering that UAV base station users are not exposed to interference, SNR is employed as performance criterion. The altitude range of UAV base station is accepted as 0.5-3 km. Maximum velocity of the UAV is 180 km/h. The movement of the UAV can be neglected because displacement on the air within the channel stationarity duration is too small compared to altitude of the UAV. Figure 1 shows an example of UAV base station layout.

The study assumes that UAV carries an outdoor base station which operates on Band 7 (2.6 GHz) which is defined by European Telecommunications Standards Institute (ETSI) as an LTE frequency band [18]. In this system model, UAV base station works on Single Input Single Output (SISO) mode and has single sector which is 120 °.

2.1 Propagation Model and Antenna Gain

In the literature, the propagation model of UAV base stations is examined in two parts as free space loss and urban losses [5]. It is accepted that free space loss has a deterministic character and urban losses have a probabilistic one. The attenuation for cases with line of sight (LOS) and non line of sight (NLOS) are as follows:

$$PL_i = 20\log\left(\frac{4\pi f_c d}{c}\right) + \lambda_i \quad [\text{dB}] \tag{1}$$

where f_c, d, c and λ_i, $i \in \{\text{LOS}, \text{NLOS}\}$ refers to carrier frequency, propagation distance, speed of light and urban losses respectively. In literature, these urban losses are considered to have a Normal distribution: $\lambda_{\text{LOS}} \sim \mathcal{N}(\mu_{\text{LOS}}, \sigma^2_{\text{LOS}})$ and $\lambda_{\text{NLOS}} \sim \mathcal{N}(\mu_{\text{NLOS}}, \sigma^2_{\text{NLOS}})$. μ_i and σ^2_i, $i \in \{\text{LOS}, \text{NLOS}\}$ express the mean and variance values of corresponding Normal distributions and numerical values can be calculated with respect to the propagation angle.

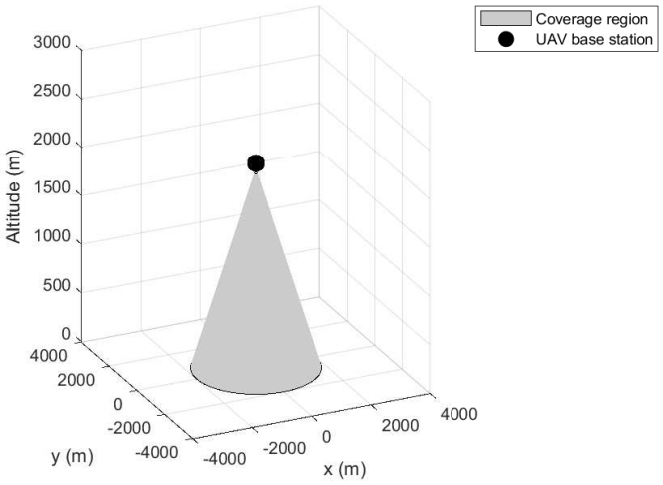


Figure 1: UAV base station coverage region.

The LOS and NLOS possibilities required to complete the propagation model are also functions of the propagation angle and denoted as follows:

$$\begin{aligned} \Pr\{\text{LOS}\} &= \frac{1}{1 + a(-b(\theta - a))} \\ \Pr\{\text{NLOS}\} &= 1 - \Pr\{\text{LOS}\} \end{aligned} \tag{2}$$

Since calculating the constants a and b above is not in the scope of this study, numerical values are given in Chapter 4. The average loss of external environment obtained using Equations (1) and (2) can be expressed as follows:

$$\begin{aligned} PL_{\text{out}} &= \sum_i PL_i \Pr\{i\} \quad [\text{dB}] \\ &= 20\log\left(\frac{4\pi f_c d}{c}\right) \\ &\quad + \lambda_{\text{LOS}}\Pr\{\text{LOS}\} + \lambda_{\text{NLOS}}\Pr\{\text{NLOS}\} \quad [\text{dB}] \end{aligned} \tag{3}$$

In addition, a model defined by ETSI is used to evaluate base station antenna gain [19]. According to this model, the antenna gain is given as a function of the beam angle:

$$G = G_{max} - \min\left[12\left(\frac{\alpha}{\alpha_{3\text{dB}}}\right)^2, A_{max}\right] \quad [\text{dB}] \tag{4}$$

where $\alpha_{3\text{dB}}$, is the beam width corresponding to a 3 dB loss and its value basically determines the directivity of the antenna. Also, $G_{max} = 15$ dB and $A_{max} = 20$ dB represent the maximum antenna gain and the maximum attenuation, respectively. Variable of the expression above, namely beam angle is indeed a function of penetration angle: $\alpha = \pi/2 - \theta$.

2.2 Indoor Model

In literature, there exist studies which expresses indoor penetration (internal and external walls) and free space loss due to the indoor propagation by using empirical models [16, 17]. Within the model given in [17], total path loss for indoor users is given as follows:

$$PL_{\text{in}} = md_{in} + nL_{in} + kL_{out} \quad [\text{dB}] \tag{5}$$

where d_{in} and m denote the indoor propagation distance and the indoor path loss constant. n, L_{in}, k and L_{out} represent the number of penetrated internal walls, penetration loss of one internal wall, number of external walls penetrated and penetration loss of one external wall, respectively. The values of these parameters at the carrier frequency of 2.6 GHz are given as $m = 0.49$ [dB/m], $L_{in} = 4.9$ [db] and $L_{out} = 24.8$ [dB]. In this study, it was assumed that one external and one inner wall is penetrated to provide indoor coverage, and there is an indoor propagation distance of five meters. Therefore, the additional loss due to indoor coverage is calculated as $PL_{\text{in}} = 32.15$ dB.

2.3 Performance Criterion

As it is stated above, we assume that UAV base station utilizes the frequencies reserved for mission critical services. Thereby, we do not consider any interference effect, and employ SNR as our performance criterion. More precisely, we define the coverage incident by the probability of SNR is greater than a given threshold. By using (3), (4) and (5), SNR can be calculated as follows:

$$\text{SNR} = P_{tx} + G - PL_{\text{out}} - PL_{\text{in}} - P_N \quad [\text{dB}] \tag{6}$$

Here, P_{tx} ve P_N represent transmitted signal strength and noise power, respectively. We note that total noise power is basically given as a function of bandwidth (BW). In the following section, by using the system model given, the probability of indoor coverage of ATG communication systems will be expressed.

3 Probability of Indoor Coverage

Indoor coverage probability of a geographical location is defined as the probability of SNR exceeds a certain value. Corresponding expression is given as

$$\begin{aligned}\Pr\{\text{c}\} =& \Pr\{P_{rx} \geq s\} \\ =& \Pr\{PL_{\text{out}} \leq P_{tx} + G - PL_{\text{in}} - P_N - s\}\end{aligned} \tag{7}$$

where $s \in \{0.76, 4.7, 10.4, 15.9\}$ refer to the different SNR limits considered in this study. The reason for choosing values given is that these values are used in the literature as SNR levels to determine certain channel quality indices (CQIs) [20].

As given in (3), PL_{out} is basically the summation of the two different random variables with Normal distribution multiplied by two coefficients, namely Pr{LOS} and Pr{NLOS}. Thereupon, we conclude that PL_{out} is also a Normally distributed random variable

$$PL_{\text{out}} \sim \mathcal{N}(\mu_{PL}, \sigma^2_{PL}) \tag{8}$$

where μ_{PL} and σ^2_{PL} represent the mean and variance parameters of the distribution, respectively. Numerical values of mean and variance can be calculated as follows:

$$\begin{aligned}\mu_{PL} =& 20\log\left(\frac{4\pi f_c d}{c}\right) \\ &+ \mu_{\text{LOS}}\Pr\{\text{LOS}\} + \mu_{\text{NLOS}}\Pr\{\text{NLOS}\}\end{aligned} \tag{9}$$

$$\sigma^2_{PL} = \sigma^2_{\text{LOS}}(\Pr\{\text{LOS}\})^2 + \sigma^2_{\text{NLOS}}(\Pr\{\text{NLOS}\})^2 \tag{10}$$

Once we compute the distribution parameters in (9) and (10) –and according to the fact that outdoor path loss is modelled as a Gaussian random variable–, coverage probability given in (7) can be evaluated by using the tail probability which is defined as $\mathbb{Q}$ function [21]. Thus, (7) is given as

$$\Pr\{\text{c}\} = 1 - \mathbb{Q}\left(\frac{P_{tx} + G - PL_{\text{in}} - P_N - s - \mu_{PL}}{\sigma_{PL}}\right) \tag{11}$$

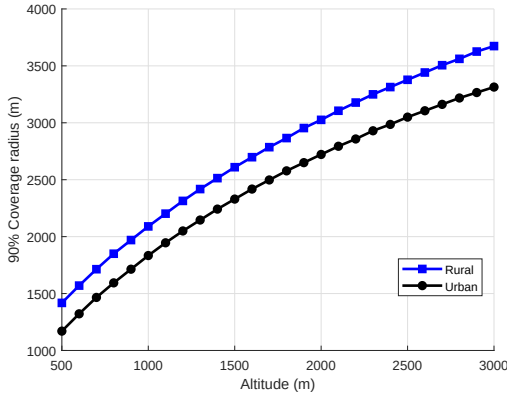
As a result of $\mathbb{Q}$ function exists in (11), it is not easy to express coverage probability in closed form. Yet, a fast, numerical solution exists. In the next section, we demonstrate our comprehensive simulation results in order to support our theoretical findings.

4 Numerical Analysis

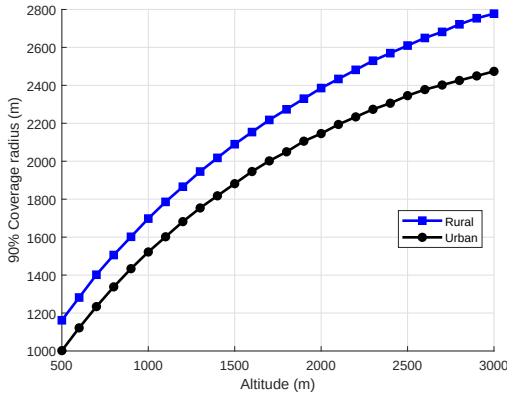
In this section, we investigate the coverage radius for different SNR demands by utilizing the coverage expression above. The system parameters used in the simulations are: $f_c = 2.6$ GHz, $P_{tx} = 46$ dBm, $\alpha_{3\text{dB}} = 65°$, $BW = 10$ MHz, $P_N = -95$ dB, $a_{\text{rural}} = 4.88$, $a_{\text{urban}} = 9.61$, $b_{\text{rural}} = 0.43$ and $b_{\text{urban}} = 0.16$. The center frequency and bandwidth of the base station are commonly used numerical values in LTE systems. Transmit power is maximum base station power value for given bandwidth standardised by ETSI [19]. Effective noise power is calculated according to formulation stated in [22]. 3dB beam width corresponds to 65° when UAV base station has 120° sector wide [19]. Other parameters used in simulations are empirical values from earlier researches.

In Figure 2a, we plot 90% coverage radius for an SNR requirement of 0.76 dB with respect to the UAV altitude. The first important observation obtained from the figure is that the coverage radius in the rural area is greater for the whole altitude interval. This result is intuitive since the channel conditions are rougher for urban areas. Another important result is that the coverage radius increases with increasing altitude. Therefore, in cases where the SNR requirement is as low as in this scenario, the highest feasible altitude value provides the best coverage radius.

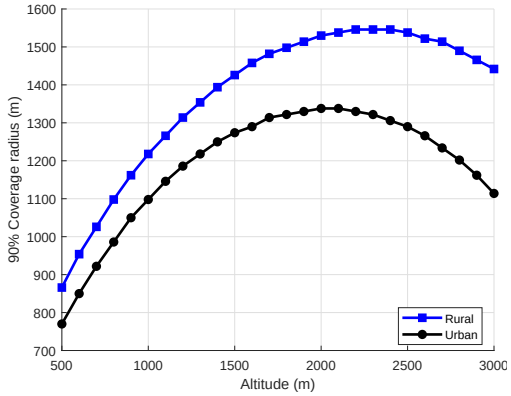
In Figures 2b, 2c and 2d, the SNR requirements are chosen as 4.7 dB, 10.4 dB and 15.9 dB, respectively. The first major finding regarding to the figures is that the coverage radius is getting smaller with the increasing SNR requirement. As a matter of fact, in the highest SNR requirement scenario (Figure 2d), after the altitude of 2000 m in urban scenario and 2200 m in rural scenario, the coverage is completely lost. In addition, when the SNR requirement is high, the coverage radius becomes a concave function rather than a monotone increasing one. Therefore, it is seen that optimum altitude is not the greatest feasible value anymore.



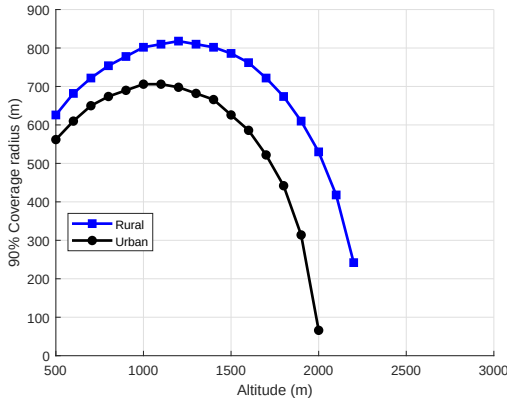
(a) s=0.76 dB.



(b) s = 4.7 dB.



(c) s=10.4 dB.



(d) s=15.9 dB.

Figure 2: 90% coverage requirement, $\alpha_{3\text{dB}} = 65°$.

As seen in Figure 2c, the optimum altitude values for s = 10.4 dB in rural and urban scenarios are 2300 m and 2000 m, respectively. In the case of s = 15.9 dB (Figure 2d), the optimum altitude values decrease further and become 1200 m for rural scenario and 1100 m for urban scenario. Under the results achieved, we conclude that it is possible to obtain an optimum altitude value for providing a certain QoS through the largest service area.

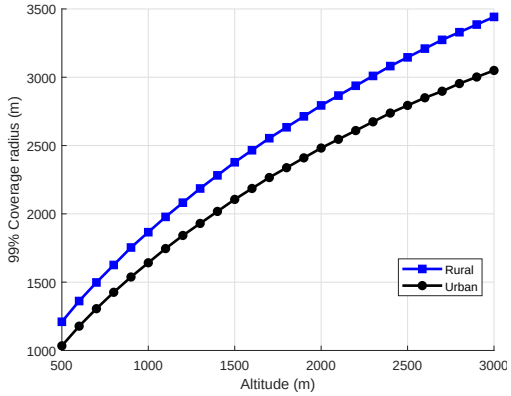
Then, we increase the coverage ratio requirement to 99% to understand the effect of QoS expectation on coverage radius. In Figure 3, we observe that for each SNR level, coverage area gets slightly smaller with increasing coverage ratio requirement. This result is quite intuitive regarding to both theoretical coverage expression proposed in (11) and the general understanding of wireless communications. We also emphasize that maximum urban area radius that can be served with the largest SNR level decreases when we require a coverage ratio of 99%.

Finally, we analyze the effect of antenna directivity on the coverage radius. In Figure 4, we fix s=15.9 dB and consider a variable beam angle. For each configuration, it is observed that coverage radius enlarges with the increasing beam angle (decreasing directivity) whereas function characteristics do not change much. However, it is important to emphasize that a narrower coverage area would be preferred for certain mission critical services. In addition, for a multi-UAV base station network, narrow beams may be required to limit the intercell interference from time to time. Thereupon, a configurable directivity pattern seems to be a plus for UAV networks.

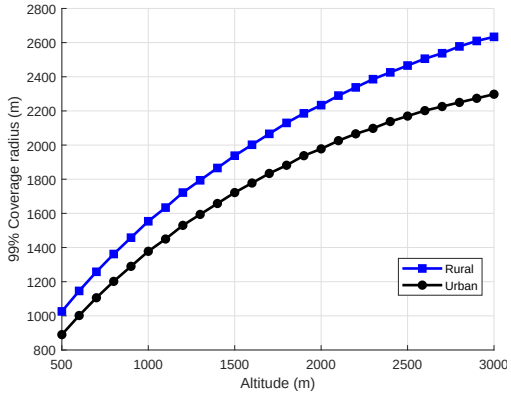
5 Conclusion

In this paper, we study on downlink indoor coverage performance of UAV base stations. According to the system model proposed, we accept SNR as the primary performance criterion. For the channel, we employ a two-staged model in which outdoor and indoor losses are assumed to be probabilistic and deterministic, respectively. Then, by employing the system model specified, we express the coverage ratio in terms of UAV altitude, cell radius and SNR requirement. In addition, to support our theoretical findings, we share our extensive simulation results. Throughout our simulations, we investigate the effect of UAV altitude, SNR and coverage ratio requirements, and antenna directivity on the coverage radius. According to our simulation results and observations, we conclude that it is possible to choose the UAV altitude level that optimizes the coverage radius. Finally, we note that we aim to build a practical test setup in order to build our own air-to-ground channel model to utilize in our future studies.

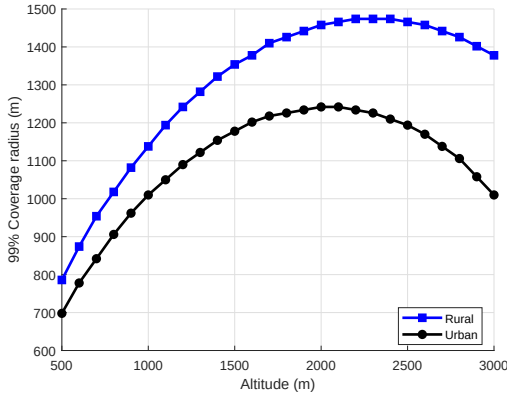
Conflict of Interest The authors declare no conflict of interest.



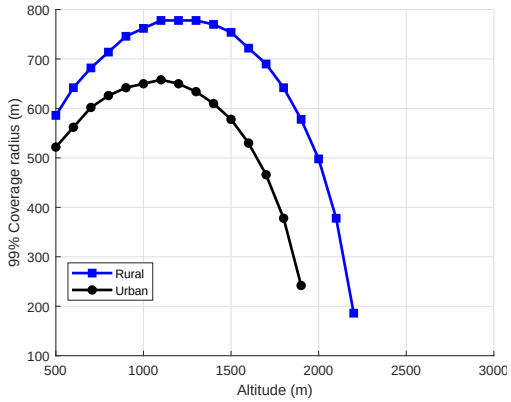
(a) s=0.76 dB.



(b) s = 4.7 dB.

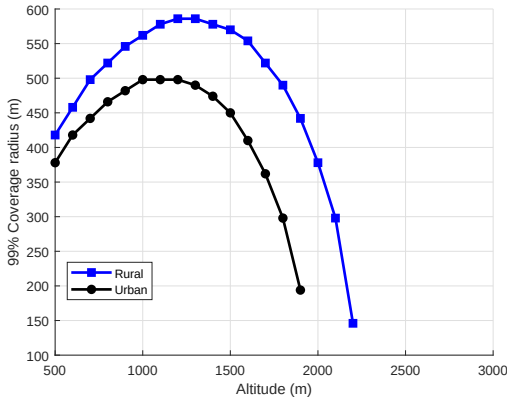


(c) s=10.4 dB.

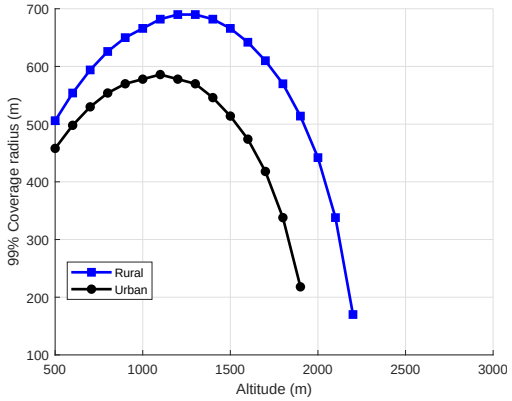


(d) s=15.9 dB.

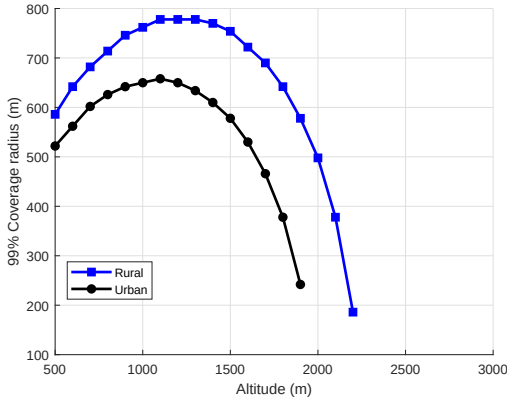
Figure 3: 99% coverage requirement, $\alpha_{3\text{dB}} = 65°$.



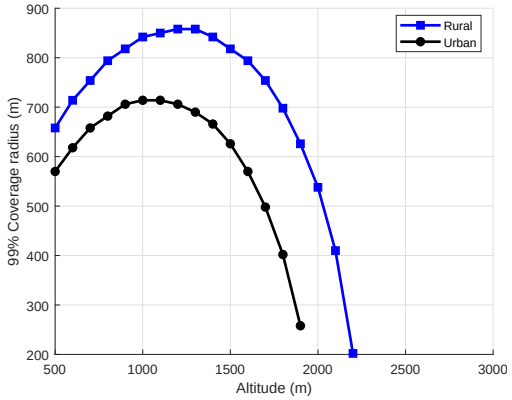
(a) $\alpha_{3\text{dB}} = 45°$.



(b) $\alpha_{3\text{dB}} = 55°$.



(c) $\alpha_{3\text{dB}} = 65°$.



(d) $\alpha_{3\text{dB}} = 75°$.

Figure 4: 99% coverage requirement, s=15.9 dB.

References

[1] M. Demirtaş, F. Aydın, G. Tanış, C. Arslan, "Indoor Coverage Analysis for Unmanned Aerial Vehicle Base Stations," in 2019 27th Signal Processing and Communications Applications Conference (SIU), 1–4, 2019, doi: 10.1109/SIU.2019.8806575.

[2] T. C. Tozer, D. Grace, "High-altitude platforms for wireless communications," Electronics Communication Engineering Journal, **13**(3), 127–137, 2001, doi: 10.1049/ecej:20010303.

[3] I. Bor-Yaliniz, H. Yanikomeroglu, "The New Frontier in RAN Heterogeneity: Multi-Tier Drone-Cells," IEEE Communications Magazine, **54**(11), 48–55, 2016, doi:10.1109/MCOM.2016.1600178CM.

[4] M. Alzenad, A. El-Keyi, F. Lagum, H. Yanikomeroglu, "3-D Placement of an Unmanned Aerial Vehicle Base Station (UAV-BS) for Energy-Efficient Maximal Coverage," IEEE Wireless Communications Letters, **6**(4), 434–437, 2017, doi:10.1109/LWC.2017.2700840.

[5] A. Al-Hourani, S. Kandeepan, S. Lardner, "Optimal LAP Altitude for Maximum Coverage," IEEE Wireless Communications Letters, **3**(6), 569–572, 2014, doi:10.1109/LWC.2014.2342736.

[6] A. Al-Hourani, S. Kandeepan, A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in 2014 IEEE Global Communications Conference, 2898–2904, 2014, doi:10.1109/GLOCOM.2014.7037248.

[7] J. Holis, P. Pechac, "Elevation Dependent Shadowing Model for Mobile Communications via High Altitude Platforms in Built-Up Areas," IEEE Transactions on Antennas and Propagation, **56**(4), 1078–1084, 2008, doi: 10.1109/TAP.2008.919209.

[8] Q. Feng, J. McGeehan, E. K. Tameh, A. R. Nix, "Path Loss Models for Air-to-Ground Radio Channels in Urban Environments," in 2006 IEEE 63rd Vehicular Technology Conference, volume 6, 2901–2905, 2006, doi: 10.1109/VETECS.2006.1683399.

[9] D. W. Matolak, R. Sun, "Air–Ground Channel Characterization for Unmanned Aircraft Systems—Part I: Methods, Measurements, and Models for Over-Water Settings," IEEE Transactions on Vehicular Technology, **66**(1), 26–44, 2017, doi:10.1109/TVT.2016.2530306.

[10] R. Sun, D. W. Matolak, "Air–Ground Channel Characterization for Unmanned Aircraft Systems Part II: Hilly and Mountainous Settings," IEEE Transactions on Vehicular Technology, **66**(3), 1913–1925, 2017, doi:10.1109/TVT.2016.2585504.

[11] D. W. Matolak, R. Sun, "Air–Ground Channel Characterization for Unmanned Aircraft Systems—Part III: The Suburban and Near-Urban Environments," IEEE Transactions on Vehicular Technology, **66**(8), 6607–6618, 2017, doi: 10.1109/TVT.2017.2659651.

[12] A. V. Savkin, H. Huang, "Deployment of Unmanned Aerial Vehicle Base Stations for Optimal Quality of Coverage," IEEE Wireless Communications Letters, 1–1, 2018, doi:10.1109/LWC.2018.2872547.

[13] M. Alzenad, H. Yanikomeroglu, "Coverage and Rate Analysis for Unmanned Aerial Vehicle Base Stations with LoS/NLoS Propagation," in 2018 IEEE Globecom Workshops (GC Wkshps), 1–7, 2018.

[14] B. Jadhavar, T. Sontakke, "2.4 GHz propagation prediction models for indoor wireless communications within building," International Journal of Soft Computing and Engineering (IJSCE), **2**(3), 108–113, 2012.

[15] ITU, "Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 900 MHz to 100 GHz," ITU-R Recommendation, 1238, 2012.

[16] E. Semaan, F. Harrysson, A. Furuskär, H. Asplund, "Outdoor-to-indoor coverage in high frequency bands," in 2014 IEEE Globecom Workshops (GC Wkshps), 393–398, 2014, doi:10.1109/GLOCOMW.2014.7063463.

[17] I. Rodriguez, H. C. Nguyen, I. Z. Kovács, T. B. Sørensen, P. Mogensen, "An Empirical Outdoor-to-Indoor Path Loss Model From Below 6 GHz to cm-Wave Frequency Bands," IEEE Antennas and Wireless Propagation Letters, **16**, 1329–1332, 2017, doi:10.1109/LAWP.2016.2633787.

[18] T. ETSI, "136 101,"," LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (3GPP TS 36.101 version 14.8. 0 Release 14), **10**, 0, 2018.

[19] T. ETSI, "136 942,"," LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (3GPP TR 36.942 version 10.2. 0 Release 10), **10**, 0, 2011.

[20] M. Taranetz, T. Blazek, T. Kropfreiter, M. Muller, S. Schwarz, M. Rupp, "Runtime Precoding: Enabling Multipoint Transmission in LTE-Advanced System-Level Simulations," IEEE Access, **3**, 725–736, 2015, doi:10.1109/ACCESS.2015.2437903.

[21] R. Sheldon, et al., A first course in probability, Pearson Education India, 2002.

[22] T. 3GPP, "36.888,"," Technical Specification Group Radio Access Network; Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE (3GPP TS 36.888 version 12.0. 0 Release 12), **06**, 0, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 130-143 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Composition of Methods to Ensure Iris Liveness and Authenticity

Ali Al-Rashid*

Division of ICT, College of Science and Engineering, Hamad Bin Khalifa University, 00974, Doha-Qatar

ARTICLE INFO

Article history:
Received: 08 June, 2020
Accepted: 26 June, 2020
Online: 18 July, 2020

Keywords:
Authentication
Iris Spoofing Detection
Iris Recognition

ABSTRACT

In a biometric system technology, a person is authenticated based on processing the unique features of the human biometric signs. One of the well known biometric systems is iris recognition, this technique being considered as one of the most secure authentication solutions in the biometric field. However, several attacks do exist that are able to spoof iris. In this paper, we propose a novel approach for securing the iris recognition system by eye liveness detection technique. The proposed system detects the eye liveness, and recognises the iris. This process includes multiple steps. As per the first step, the person opens his eye and the system reads remotely the changes in pupil size as a result of the response to the ambient illumination. Then, the system starts matching the iris with a database. The second step: the person closes his eye and the system remotely detects the heartbeats signals under the skin of the eyelid. As per the third step, the person opens his eyes again and the system reads the pupil size again and compares the results of the pupil size, and then the system matches again the iris with the above database. For the iris recognition to be validated, all the above checks have to be passed. We have conducted several experiments with our proposed system, based on a brand new dataset comprised of 40 subjects. In addition, we also used public datasets: CASIA-Interval, CASIA-Twin and Ubiris.V1. The achieved results show the quality and viability of our proposal.

1 Introduction

Among all human physical biometrics, iris recognition systems are considered the most secure biometric systems that can be operated at a low false acceptance rate (FAR). The applications of iris recognition comprise personal identification cards, border controls, and other government applications [1]. Moreover, the distinct feature of the iris recognition system that makes it so popular in security is its uniqueness; even the iris patterns of twins are different.

However, iris is still subject to attacks. The most well known attacks are called spoofing—whose detection is still an open challenge. Iris recognition technology is considered as a robust authentication technology because of the difficulty of counterfeiting the human iris. But adversaries have found numerous manners by tricking the iris recognition systems through forging the iris [2]. So, researchers began to analyze the structure of the human eye to find effective countermeasures versus the adversary.

An attack technique is to utilize an artificial human iris to be placed in front of the camera sensor to grant authentication to the adversary as the legitimate person [3]. While a popular attack is by using a printed photo. The adversary captures a photo of the targeted user to impersonate him as an authorized user [4]. A countermeasure is liveness recognition, that is employed to identify whether the person is alive or not. One of the common approaches to recognize life signs is the signal of the heartbeats. When the heartbeats occurs, then we cannot ignore the assumption that alive human is facing the sensor. The existing work in recognizing the heartbeats is remote photoplethysmography (rPPG) [5]. The rPPG detects the signals of blood flooding underneath the human skin. Furthermore, various ways to detect rPPG— the widespread technique is probably by sensing remotely the signals of the heartbeats through the face.

The Author in [6] demonstrated the possibility to remotely acquire a heart pulse through one particular camera. The experiment of capturing rPPG is achieved by detecting a certain area of the face skin and calculating a liveness score in the region of interest (ROI) within a short period [6].

In [7], the remote capturing of heartbeats is based on the detection of the color variations in the face skin of the person and recognizes the irregular bio signs. If the indication of the heartbeats under the skin does not occur, it would indicate that the analyzed face is a fake.

Contribution We present in our work a new contribution to detect the eye-liveness, integrated with iris recognition. The proposed methodology is articulated over a series of steps: eye liveness, iris recognition, and again eye liveness and iris recognition. The eye-liveness detection depend on

*Ali Al-Rashid, Doha-Qatar & e-mail: alialrashid@mail.hbku.edu.qa

the reaction of pupil diameter size during the existence of the ambient light, integrated with the remote reading of the heartbeat of the eyelid; while iris recognition is based on matching the iris—with the one previously stored in a database—two times, in real-time.

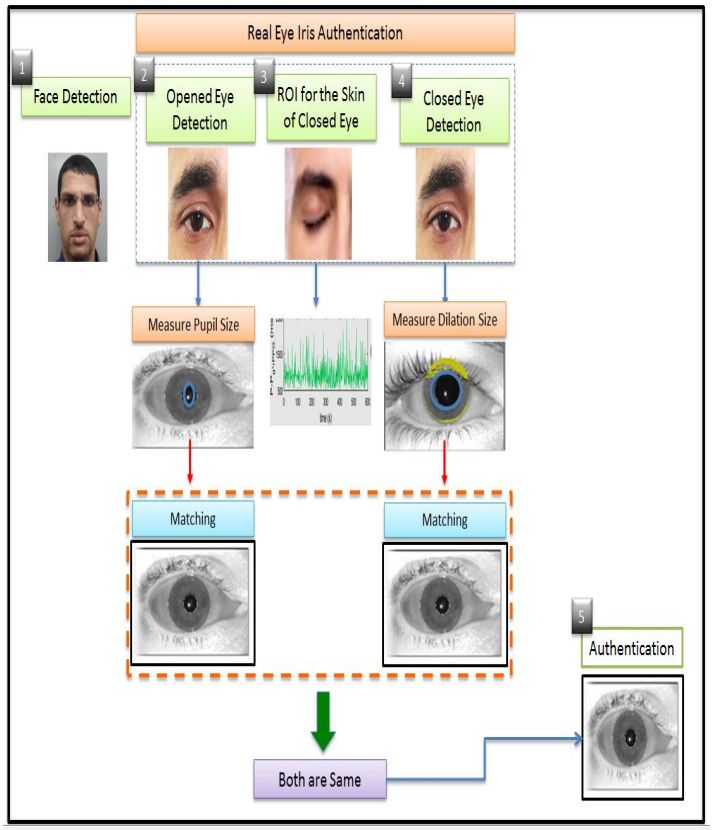


Figure 1: Proposed system

The system starts capturing the face of the supplicant, and runs a time counter. Next, the system recognizes the eye as opened during opening the eye and then reading the actual measurement of the size of pupil diameter. After that, the system initiates matching the person's iris with the database as the first match. Then, the user closes his eyes and the system identifies the eyes as closed and then starts detecting the signals of blood flowing under the eyelid skin as heart pulses. The user subsequently opens the eyes, and the system reads the diameter size of the pupil again to recognize the dilation. The system matches again the iris with the database as a second match. Consequently, if the iris first match is equivalent to the second match, the dilation of the pupil exists, the heartbeat is recognized, and the time counter of the face did not interrupt, then eye-liveness is validated and iris is authenticated.

Roadmap We have structured our proposal as follows: Section 2 reviews the prior art in the field. Section 3 introduces the adversary models and illustrates the methods of spoofing that the can be performed by the adversary. Section 4 presents our system and how to grant a secure authentication discriminating between an authentic and a spoofed eye. Section 5 is the iris recognition and explains the process of identifying the iris from the database. Section 6 illustrates the results of experiment. Section 7 is a discussion of the proposed system. Section 8 reports the limitations of our work, while conclusions are reported in Section 9.

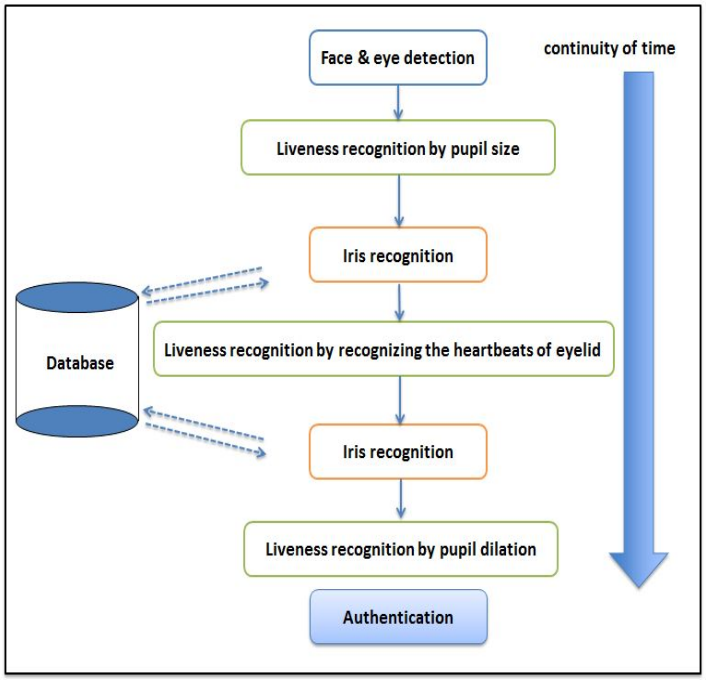


Figure 2: Flowchart for the proposed sytem

2 Related work

The concept of the iris recognition system started initially in 1987 by Flom and Safir [8]. This was demonstrated as a basic matching approach for the imaging system by using mathematical theories.

One of the famous methods for matching iris is the Daugman method. Daugman proposed an automatic segmentation approach, processes for normalization and 2D Gabor filter feature extraction, and match the data to get iris matching [9]. Where [10] made a comparison by comparing the iris recognition algorithm between Principal Component Analysis (PCA), Independent Component (ICA) and Gabor Wavelet for getting iris code.

As countermeasures for spoofing attacks, the previous proposals concentrated on liveness detection. The principle of the liveness is to guarantee that the iris of the eye has a life sign by examining the biometric characteristics of the eye [1]. Basically, illumination technique is one of the popular techniques in forged iris detection, where the eye obtains illumination by focusing light to the eye. So, the pupil reacts with the light by expanding or contraction. The change in the size of pupil confirms the eye is real.

The contributions that depend on checking the pupil radius after centering the illumination have certain restrictions. For instance, if the room is lightened by strong light, then the pupil response could be not satisfactory. Park [2] demonstrated a mechanism utilizing a texture feature to determine the countrified the human iris, by the pupillary light response.

Besides, the illumination method has the ability to identify the corneal whether as alive or fake. The corneal usually

reflects when directly illuminated, and creates a red circle in the center of the pupil. By examining the color that occurs in the corneal [6], an alive eye is recognized. Therefore, the attention concentrated on the reaction of the eye to the light, to discriminate between the alive eye and spoofed eye [11]. The author in [4] proposed Pupil Centre Corneal Reflection (PCCR) method relied on gaze estimation to discriminate between the signals of live and spoof eye. In addition, the vein of the sclera used to identify the liveness of the eye. The sclera is the white region inside the eye and the pattern can be extracted and compare it with the pre-stored database, such as in [8].

Regarding the rPPG, the earlier works were focused on analyzing the face skin of the human [6]. The used method is called a color-based method, it measures the heartbeat from the reflection of the colors that occur on the skin— those colors are the blood streaming under the skin [7]. This method is one of the cornerstone for detecting spoofing attacks.

The work in [12] presented a contactless heart pulse system to detect the fabricated mask of the victim. This approach has been achieved by measuring the heart pulse of the face. Therefore, when the heart pulse of the face is detected, then the face is considered alive.

Finally, the authors in [13] proposed to preventing 3D mask PAD (presentation attack detection) by using rPPG pulse detection.

3 Adversary models

The adversary relies on spoofing to obtain the identity of the legitimate user. In this section, we introduce the techniques of spoofing that have been used to tamper iris recognition systems.

In general, the iris system captures the iris of the person and then matches it with stored images. The major purpose of the adversary is to fool iris recognition systems by using forged iris models of the targeted user in iris recognition devices to grant an authentication. The iris spoofing attacks can be classified as follows:

- Photo Attack;
- Video Attack;
- Contact Lens Attack; and,
- Artificial Attack.

Photo Attack: This type of attack is one of the popular attacks in iris recognition systems due to its simplicity. The attacker uses printed photo one a piece of paper to impersonate a victim. Also, the attacker can use a stored photo in the smartphone to implement the same attack [14].

Video Attack: It is a recorded short video of the victim's eye that encompasses all the parts of the eye such as: iris and pupil. The attacker strives to make the video very clear to view the iris in the screen looks real by the adjusting graphics and the resolution.

Contact lens attack: This attack is considered a development of a photo attack. The adversary prints the iris drawing of the victim on the contact lens in a special manner. After that, he wears a fake lens to authenticate himself and bypass the system. Normally, contact lens attack is not easy to be detected, unless the attacker shows unusual behavior in front of the iris camera.

Artificial Eye: The adversary uses plastic or glass to fabricate a spoofed eye. After that, the attacker place this artificial eye in iris systems to get authentication.

4 Proposed Method

This section details the framework of our system. The system authenticates the eye of the user and identifies spoofing attacks by recognizing a real eye from a fake eye. The proposed system in Figure 1 depicts the liveness of eye recognition. As shown in Figure 1, our system is structured over several stages, described in the following.

1. The system detects the face and starts a timer (the face must be presented facing the camera sensor and the timer will be stopped in case any interruption in the detection of the face.
2. The person opens his eyes and the system identifies the opened eye. Next, begin capturing the pupil diameter and it must be small (because of surrounding illumination) and then matches the iris with the database.
3. The person closes his eyes and the system identifies a closed eye. After that, the system extracts the vein of the eyelid skin.
4. Detecting the heart signals of the eyelid skin.
5. The person opens his eyes again and pupil size will be dilated. Then, the system immediately identifies the dilated pupil and matches the iris again with the database.

Furthermore, the process should be implemented while maintaining its continuity. If any step of the process is stopped or interrupted, then the system detects this abnormal behavior as a spoofing attack. Therefore, when the pupil size is changed before and after closing the eye and the signals are recognized as a heartbeat of the eyelid, and at the same time the iris is matched twice, the system recognizes the eye as real and then iris matching process will initiate.

4.1 Pupil Diameter Detection with Heartbeat Signals Recognition from the Eyelid

As an initial step, the system captures the user's face and a temporal window will set with length T [15]. The window can be configured to execute time-dependent analysis.

Subsequently, pupil diameter detection and rPPG of the eyelid will be performed into various phases, as follows: face and eye detection, pupil diameter size readings, and Region of Interest (ROI)detection with rPPG readings recognition.

4.1.1 Face Detection

Capturing the face precisely is necessary to capture the eye in clear view and to determine good quality candidate regions of the eye. Consequently, the false positive rate will be reduced and also the computation time. The Haar-like feature

is typically the used approach for detecting the face due to its fast processing and accuracy [16].

4.1.2 Eye Detection

The eye-detection implemented by using Haar-like features and a cascade classifiers algorithm. As explained in the previous section, if face detection is performed accurately, the efficiency of detecting the eye will be raised as well [16].

4.1.3 Pupil Detection

Pupil detection is implemented by using the Hough transform algorithm [17]. There is a pre-processing step that must be achieved before implementing the Hough transform [17]. First, we have transformed the image into a grayscale. Next, we got the major data of the eye contour. To implement this stage, we used the 'Prewitt filter' to obtain the contour image [18]. It gives two types of edges Gx and Gy on the image, where Gx detects edges in horizontal direction and Gy detects edges in vertical direction as demonstrated below:

$$Gx = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \; Gy = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

And the contour image as shown in Figure 3:

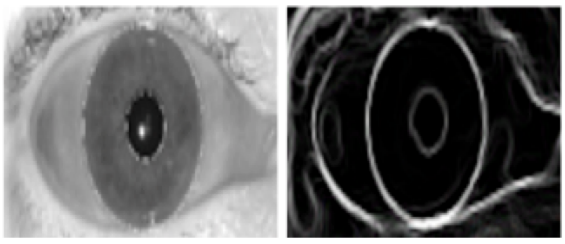


Figure 3: Prewitt filter to get the contour of the eye image

When the circle is detected on the image, the pattern of the pupil is extracted and stored. To get the circle of the pupil after this stage, we apply Hough Transform.

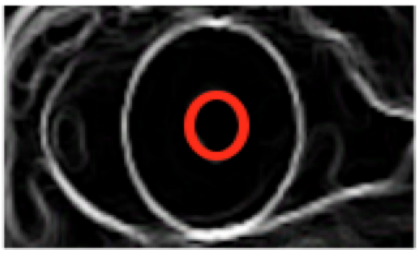


Figure 4: Hough circle

4.1.4 Region of Interest (ROI) Selection

The ROI is a significant element in the captured image of the eye. It locates the skin portion of the eye [19, 20]. Furthermore,the ROI determines the external layer of the eyelid which includes the vein (inside the vein is the blood flowing, used for the detection of the heart pulses).

4.1.5 rPPG Extraction

Once the ROI of the eyelid is determined in the image, we begin measuring the rPPG based on the average of the pixels which are located inside the eyelid skin [6, 7]. The measurements are accomplished independently, over three separate colored channels: Red, Green, and Blue (RGB). Accordingly, rPPG measurements appear on RGB image at every frame and the output appears on the temporal window [7].

4.2 Feature Extraction

4.2.1 rPPG signal pre-processing

The signals generated from the rPPG contain the light variations from the heart-beats and the light surrounding the environment [12]. So, signals of the rPPG are affected by the light of the environment, and compounded with the noise from other sources. Hence, a pre-processing step is important to be performed before getting the rPPG signals. The step consists of three stages as follows [6, 7, 12, 13]:

- Detrending filter: it operates as a temporal filter that is utilized to reduce the static part of the signals that produced from the rPPG. For instance: reducing the signals that are not part of the expected heartbeats and removing the interference of light environment.
- Moving-average filter: this type of filter used to remove the random noise of rPPG signals. This random noise could occur due to incorrect parameters which are gained during the image capturing.
- Band-pass filter: commonly, the human heart rate ranges between 40-240 bpm (beats per minute), which is equivalent to a frequency between 0.6 and 4 Hz. Therefore, the frequency outside this range will be be removed (because they are not heart pulses).

4.2.2 Transformation of the eyelid skin to frequency

This phase has a special input signal, which is, filtered rPPG signals. The filtered rPPG is the extracted signals from the estimation of the changes in eyelid skin tone caused by the blood streaming under the eyelid skin. Now, the extracted signals will be converted to the frequency domain by using fast Fourier transform (FFT) [6, 13].

4.2.3 Extracting the heart pulse

The extracted domain after the pre-processing step shows a range of frequencies; the highest frequency peaks are the heart pulses. In our state, we need to detect the heart-beat to identify the real eye or fake eye.

The method applied to the features is called transformation. The signals are converted from the domain to the frequency by applying fast Fourier transforms (FFT), and later its power spectral density (PSD) distribution is computed [12]. The patterns of PSD present the discrimination of the eye—whether it is a real or a spoofed one. An alive eye has an obvious peak in the PSD, while a spoofed eye has random noise peaks in PSD [21].

Subsequently, we extracted for every color channel two features for detecting the real eyelid.

In the feature extraction stage, PSD is visible and its maximum power response is specified as first feature P. The frequency range (expected between 0.6 - 4Hz), is the second feature R which is the ratio of P to the total power. The pair of features (P, R) are the outcome of three channels of RGB video [21].

5 Iris Recognition

This section explains the stages of processing the iris image and then matching the database.

5.1 Segmentation

The segmentation method is used to get the contours of iris precisely, which is an inner boundary comprising the pupil with the iris and an external boundary comprises the iris with the sclera.

These boundaries are used to discriminate the pixels whether there is the iris. So, this outcome produces a binary mask that consists of on-pixels (belong to the iris) and off-pixels (do not belong to the iris) [22]. Subsequently, the Hough Transform Circle is used to determine the circular iris area. The segmentation step generates two contours (iris and pupil) that will be used in the next step which is the normalization, as depicted in Figure 5.

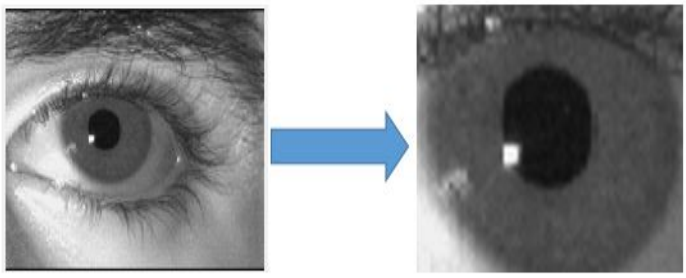


Figure 5: The left image is the original eye for a subject and the right image is the eye after segmentation

5.2 Normalization

The normalization converts the area of iris into a straight strip by using the Daugman's rubber-sheet technique, as depicted in Figure 7.

Daugman's rubber-sheet manner is a popular technique for iris normalization which transform the circle of iris area to a rectangular block with a fixed size [22].

$$I[x(r,\theta), y(r,\theta)] \rightarrow I \tag{1}$$

Based on equation (1), the equation converts the pixels in the circle of the iris into an equivalent location on the polar axes (r, θ), where r is the radial distance and θ is the rotated angle at the corresponding radius.

5.3 Encoding

With encoding, the iris texture will be extracted through filtering the normalized image by using a Gabor filter method, as shown in Figure 8. The Gabor filter has the ability to generate an optimum conjoint representation of a signal in space and spatial frequency. [23] Moreover, a Gabor filter is structured by modulating a sine/cosine wave with a Gaussian.

So, using the Daugman method we can later use the 2D Gabor filter to encode the iris data. The equation of the 2D Gabor filter over the image domain as (2):

$$G(x,y) = e^{-\pi[(x-x0)^2/a^2+(y-y0)^2/\beta^2]} e^{-2\pi[\mu(x-x0)/v0(y-y0)^2]} \tag{2}$$

where, $(x0, y0)$ determines the position on the image, (α, β) determines the actual width and length, and $(u0, v0)$ determines the modulation, which has spatial frequency t $(u0^2 + v0^2)$.

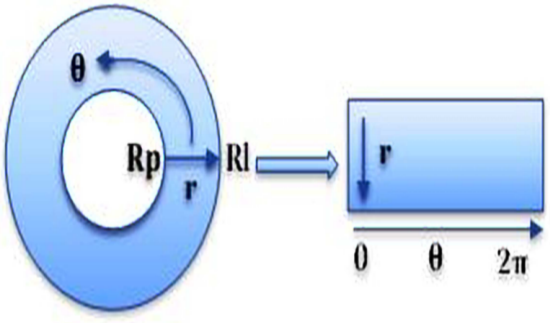


Figure 6: Daugman's model used for iris normalization.

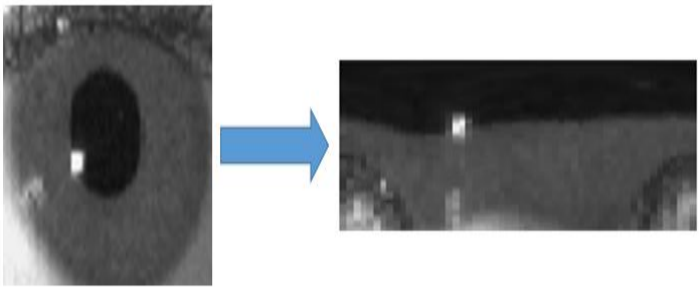


Figure 7: The subject's iris converted to strip by using normalization method.

5.4 Matching

The matching section is the final stage. It compares two iris codes using the Hamming distance (HD) between the binary codes corresponding to the selected points within the iris templates.

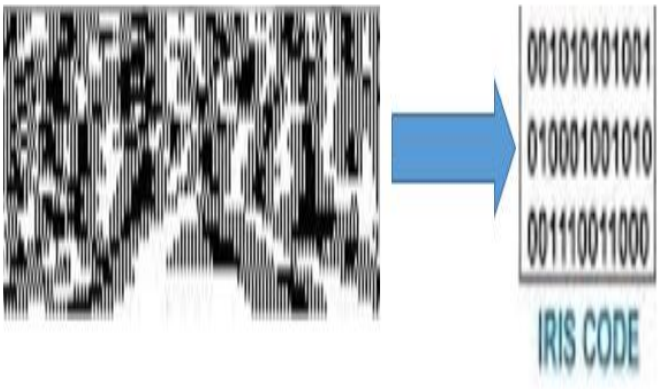
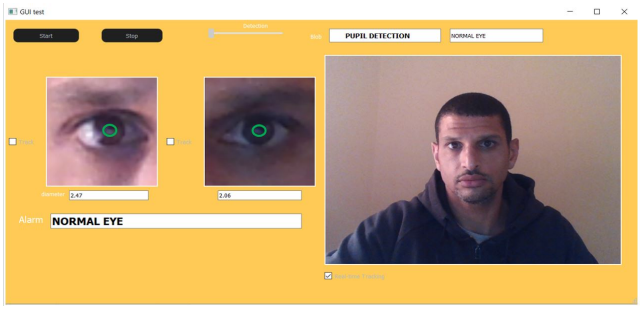
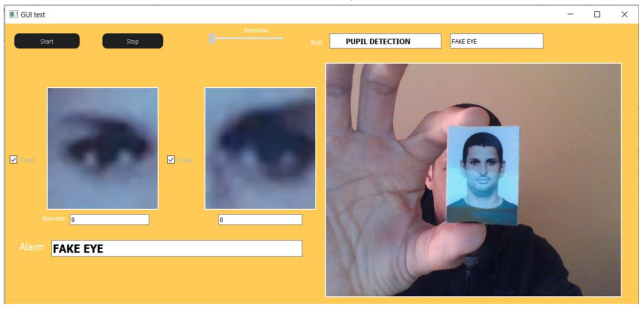


Figure 8: Encoding stage by converting the normalized iris to iris code

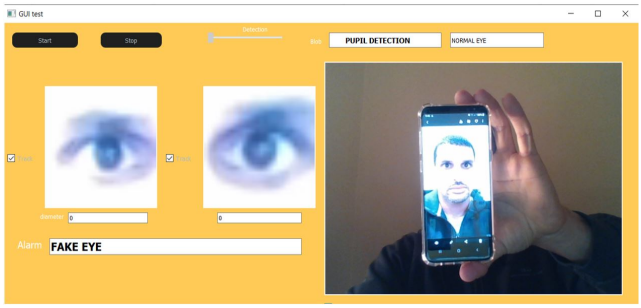
Indeed, if two patterns are derived from the same iris, the HD between them will be close to 0. In this stage, we match the iris two times to make sure the same user is physically present in front of the sensor. The first check, when the user opens his eye, and the second check when the user finishes from measuring the heartbeats through the eyelid while closing the eye and then open again his eye.



(a) Real eye



(b) Photo attack



(c) Video attack

Figure 9: (a) Real eye. (b) Photo attack. (c) Video attack.

The principle of matching the iris two times with running a timer is to make sure that it is the same user present in front of the sensor, otherwise the attacker could first pass the liveness process, and then fool the system by using counterfeit iris.

6 Experiment and Results

6.1 Experiment Setup

The experiments were conducted with the resources reported in the following.
The hardware: 1. PC with OS Win10 ; and, 2. two cameras sensors: one infrared camera type DingDangSmart 2MP 1080P OV2710 Mini IR Webcam and one cameras type Logitec C930e. We fix the two cameras on one stand. The first camera (infrered camera) was used for detecting the pupil diameter size and iris of the eye and the second camera (logitec C930e) for detecting the eyelid.

The application software: we designed GUI (Graphical User Interface) by using QT program, C++ for pupil diameter detection and iris matching and python for heart pulse detection through the eyelid. The first GUI is configured to camera 1 and captures the pupil from a 1-meter distance and shows on the screen the actual diameter size of the pupil depicts as in Figure 10 and matches the iris with stored database as in Figure 11. Furthermore, it determines the difference between the closed and opened eye through displaying a message.

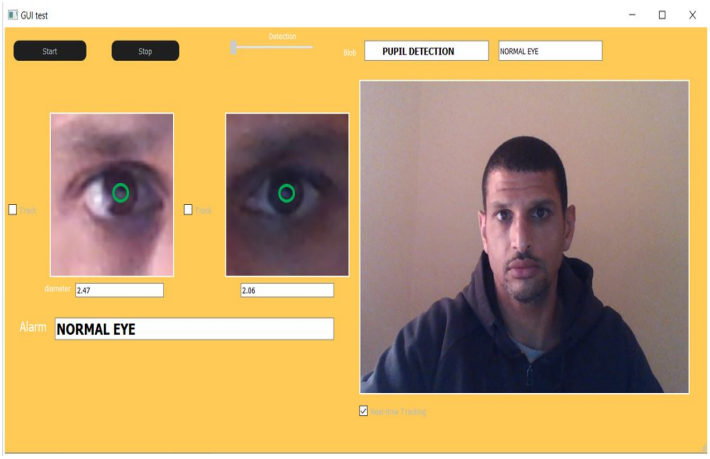


Figure 10: Pupil size recognition

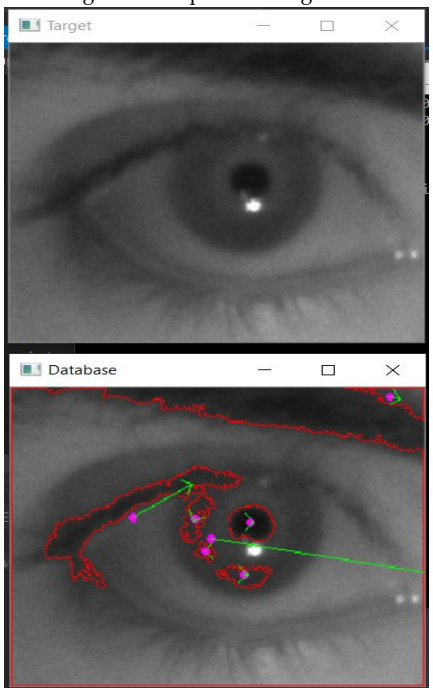


Figure 11: The first image is the scanned eye from the proposed system and the second image is recognized eye with the database.

Also, camera 1 is configured for scanning the iris and matching it with the Database. To detect the heart pulses of the eyelid, we designed another software and configured to camera 2 and measures remotely the user heart pulses of the eyelid as shown in Figure 12. The SW also displays in

Table 1: The pupil size values for 40 subjects

(a) The values of the initial state, dilation, and final state of the pupil sizes for first 20 subjects

Subjects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Initial State of Pupil Size	2.88	2.34	2.06	2.04	2.55	2.88	2.47	2.53	2.62	2.11	2.06	2.67	2.33	2.26	2.72	2.06	2.42	2.57	2.57	2.31
Pupil Dilation	3.29	2.67	2.88	2.56	3.29	3.31	3.09	3.29	2.77	2.96	2.88	3.19	2.68	2,57	2.88	3.09	2.67	3.29	3.11	2.56
Final State of Pupil Size	2.88	2.34	2.06	2.04	2.55	2.88	2.47	2.53	2.62	2.11	2.06	2.67	2.33	2.26	2.72	2.06	2.42	2.57	2.57	2.31

(b) The values of the initial state, dilation, and final state of the pupil sizes for second 20 subjects

Subjects	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Initial State of Pupil Size	2.87	2.13	2.26	2.72	2.06	2.42	2.57	2.06	2.04	2.55	2.88	2.47	2.53	2.62	2.11	2.06	2.77	2.38	2.46	2.31
Pupil Dilation	3.47	2.67	2.67	3.09	3.91	2.67	3.11	2.67	2.47	3.11	3.21	3.09	3.09	2.88	2.22	2.67	3.67	2.79	3.07	2.56
Final State of Pupil Size	2.87	2.13	2.26	2.72	2.06	2.42	2.57	2.06	2.04	2.55	2.88	2.47	2.53	2.62	2.11	2.06	2.77	2.38	2.46	2.31

real-time the heart pulse on the screen as in Figure 13. In addition, we used certified medical oximeter device type TATRIX Medical fingertip pulse oximeter device for comparing the heartbeats signals with our system.



Figure 12: Eyelid detection

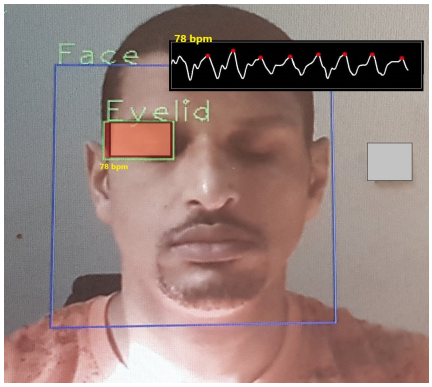
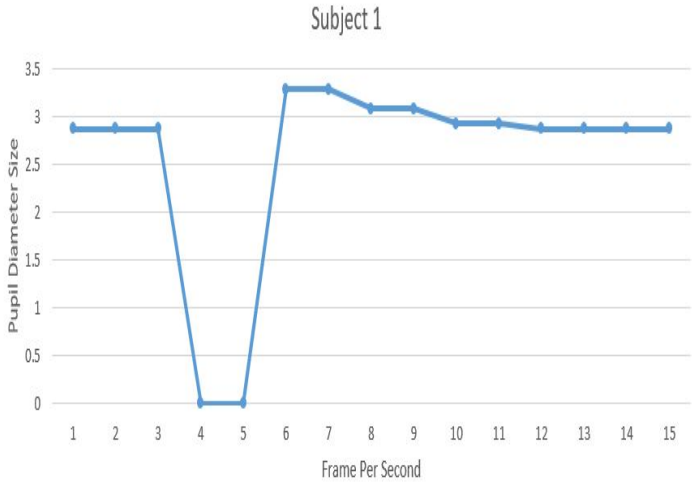


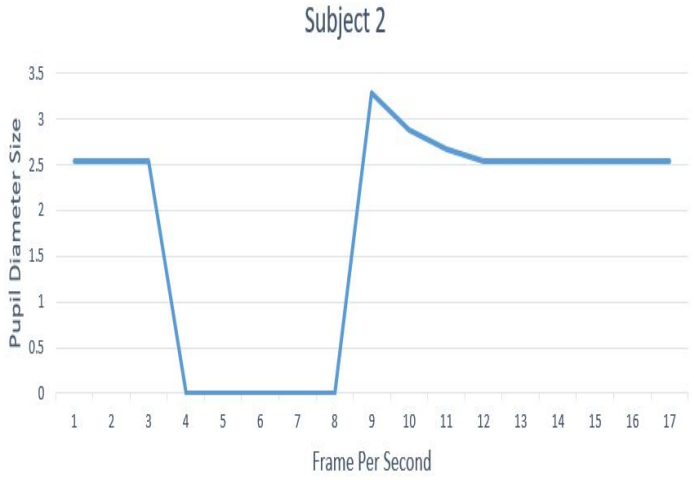
Figure 13: Heart pulse detection by eyelid

6.2 Data

Since, to the best of our knowledge, there is no available dataset for heartbeats signals of eyelid and pupil diameter size changes before and after closing the eye, we have created three data-sets composed of 40 subjects. The first data set is the diameter size of the eye pupil.



(a) Pupil diameter size changes for user 1



(b) Pupil diameter size changes for user 2

Figure 14: Pupil diameter size changes

The second one stores the heartbeats parameters of the eyelid. The third dataset is the iris dataset. In addition, we used external datasets from Chinese Academy of Sciences Institute

of Automation (CASIA): CASIA-IrisV3-Interval contains 2655 samples with 320*280 image size , CASIA-IrisV3-Twins contains 3183 samples with 640*480 image size and UBIRIS.V1 contains 1214 samples 200*150 image size to evaluate the performance for the iris recognition [24].

Algorithm 1: The proposed system for eye liveness detection

Input :Capture the face
While the same face on the screen **do**
Run the continuation timer
Begin Stage 1
The system asks to open the eye for 3 seconds
Detect the face
Detect the eye
End of Stage 1
Begin Stage 2
if *The eye is opened* **then**
Detect the pupil and measure its diameter size;
Match the iris with the database
else
Termination in case the eye is closed;
end
End of stage 2
Begin Stage 3
The system asks to close the eye for 3 seconds...
if *The eye is closed* **then**
Detect the eyelid;
Extract the vein of the eyelid
else
Termination in case the eye is opened;
end
End of Stage 3
Begin Stage 4
if *The vein is detected and the blood streaming detected* **then**
Read the heart pulse signals that come through the eyelid
else
Termination in case not detected
end
End of Stage 4
Stage 5 (Last Stage)
The systems asks to open the eye again
if *The eye is still closed* **then**
Termination
else
Detect the pupil and measure the changes in diameter size
Pupil size must be dilated and starts to shrink gradually until it becomes smaller like the first time in step 2
if *The pupil size is different* **then**
Match the iris with the database
Output: The eye is real and the iris authentication is triggered
end
end

6.3 Methods

We have divided the proposed method into 3 stages: stage 1 for opening the eye, stage 2 for closing the eye and stage 3 for opening the eye again. We created dataset for pupil diameter sizes and iris of every subject. The system captures the pupil diameter with the iris during opening the eye in stage 1 and 3. In the meanwhile the system compares the changes in pupil diameter size and matches the iris two times .

The concept behind the two matches of the iris during the stages is to confirm the physical presence of the subject. In case the iris matched just one time, it means an error occurred in the system or the attacker tried to bypass the camera sensor by forging the iris. As per stage 2, we created a dataset for the heartbeats of the eyelid and measured it for the same 40 subjects. We let the subjects to look at camera sensor (which is dedicated for measuring the heartbeats signals) for 20 seconds. Next, the system detects and stored the heartbeats of the eyelid in database1. Furthermore, we used certified oximeter device to gather the parameters of heartbeats of the 40 subjects and stored them in database2.

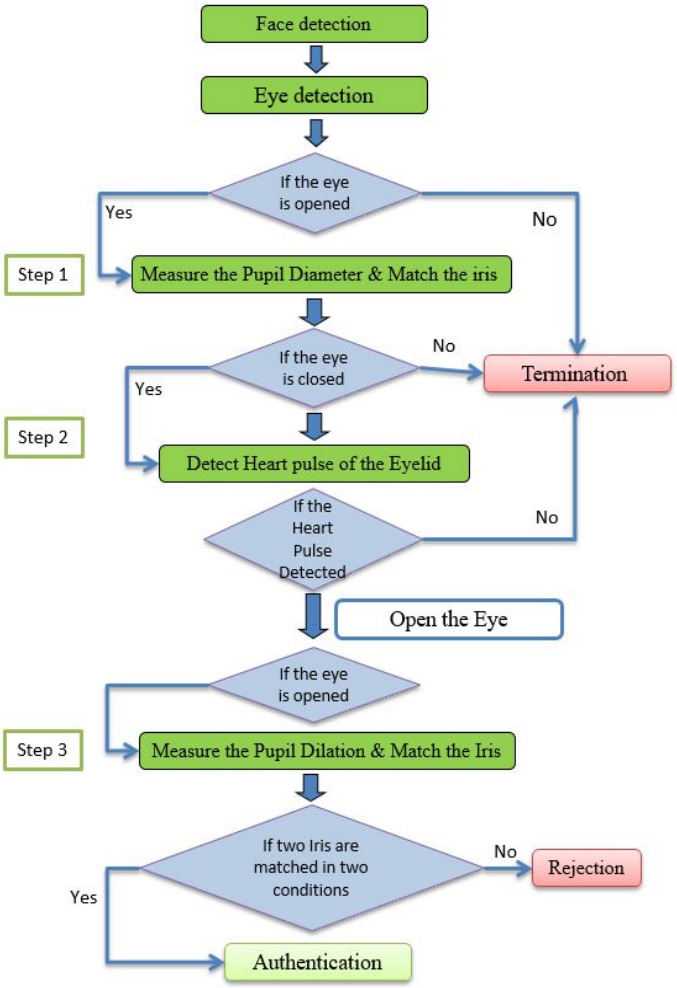


Figure 15: Proposed system

6.4 Results

In this section, we report the evaluation of the results for our proposal.

Pupil diameter detection

We calculated the pupil diameter size while the subject's eye was opened and the diameter size was smaller than while the eye was closed for a short time and then opened again. As shown in Figures 14a and 14b, we observe the curve of the pupil in a static state while the eye is opened. After that, the

curve moves toward "0" during closing the eye and then the curve increased higher than the static state. Then, the curve returns to the static state again because of the response to the surrounding light.

Moreover, we measured the average of the initial state of the pupil diameter size (when the user looks to the camera), dilation state of the pupil diameter size, and the final state of pupil diameter size (when the user closes and opens his eye again) for all the 40 subjects, as depicted in tables 1a and 1b.

Iris recognition In this stage we have examined 40 subjects by matching their irises two times: one before and one after closing the eye. The result of matching the same iris two times—using our algorithm—resulted in a 100% accuracy rate.

Moreover, we used outsource datasets CASIA and UBIRS to evaluate the recognition performance. In table 2 we report the parameters for evaluating the proposed system which checks the same person two times, namely: Correct Recognition Rate (CRR), false accept rate (FAR)—it calculates the probability of the user being incorrectly recognized as another user—, false reject rate (FRR)—it calculates the probability of registered users not being recognized, and Equal Error Rate (EER)—it is the value where FAR and FRR cross. Decision Threshold: when the hamming distance (HD) of two irises is lower than the threshold, then we consider the two iris the same (authorized user). However, if HD of two irises is bigger than the threshold, we consider them as different (unauthorized user). The ROC (receiver-operating characteristic) shows the relationship between the FRR and FAR.

Eyelid signals

We conducted a comparison between the eyelid dataset (database1) and the medical oximeter dataset (database2). Each subject was scanned for 20 seconds by our proposed system along with a medical oximeter device which was attached physically on subject's finger.

The results that we found from database1 were very near to database2 as demonstrated in Figures 16 and 17.

According to the Figures 16 and 17, the curves depict that the parameters are close to each other. We observe that the curve of of the eyelid's heartbeat is very close to the one related to the oximeter, till some variance appears between 13 to 16 seconds because of some delay in the proposed system during the capture of the heartbeats of the eyelid. This delay considers regular because the size of the vein is small which is exists inside in the eyelid and difficult to capture this vein. Besides, the vein of the eyelid differs from subject to subject. However, the delay occurs for few seconds only (not more than 3 seconds). Overall, the parameters of the proposed system are almost the same as the certified oximeter device.

Table 2: Evaluation of our dataset, CASIA-interval, CASIA-twin and UBIRIS.v1

Database	FAR %	FRR %	CRR	ERR
Our dataset (40 irises)	1 0.1 0.01	0.06 3.14 5.67	99.25%	0.74
CASIA-Interval	1 0.1 0.01	1.27 4.27 12.34	98.85%	1.15
CASIA-twin	1 0.1 0.01	1.76 4.32 8.76	98.25	1.75
UBIRIS.V1	1 0.1 0.01	3.67 9.67 19.4	97.57%	2.43

Moreover, with reference to the heart pulse parameters, we measured the mean, the median, and the standard deviation for all the subjects, as shown in tables 3a and 3b.

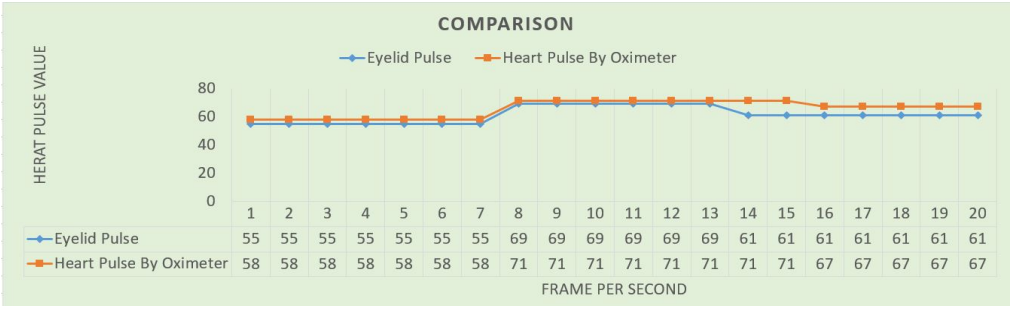


Figure 16: Comparison between heart pulse by eyelid and oximeter for user 1

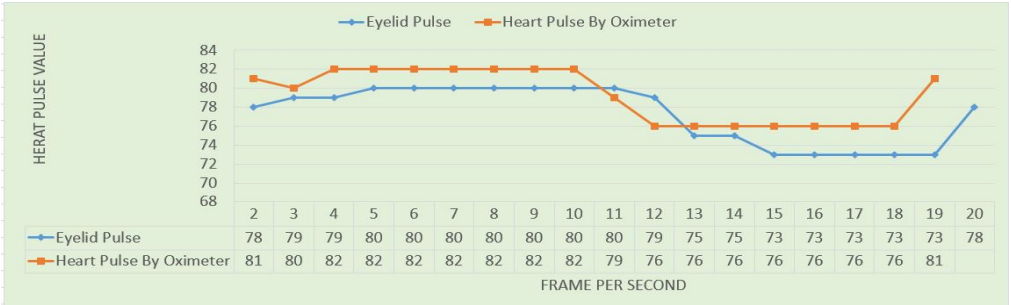


Figure 17: Comparison between heartbeats signals of eyelid and oximeter for user 2

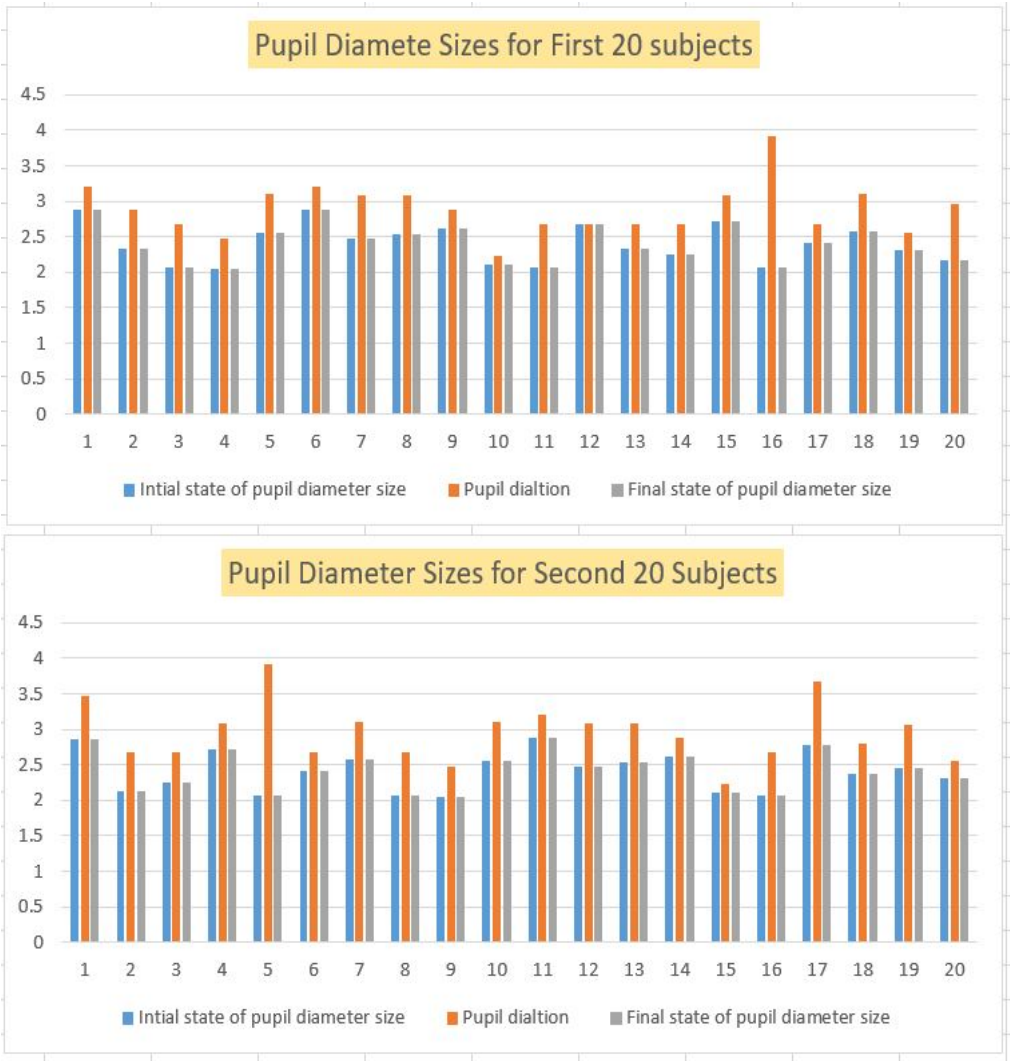


Figure 18: All subjects are grouped, and their pupil diameter sizes is reported to highlight the differentiation between the initial state and dilation state of the pupil diameter size changes.

Table 3: The heart beats parameters for 40 subjects

(a) Mean, median, and standard deviation heart pulse values for all subjects

Subjects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Mean	55	65	55	70	71	78	55	60	78	55	53	60	59	62	57	61	59	58	72	78
Median	55	65	55	69	71	78	56	60	78	56	53	58	59	63	57	60	60	58	71	78
Standard Deviation	1	1	2	2	2	1	1	1	1	1	1	2	1	2	1	2	5	1	2	1

(b) Mean, median, and standard deviation heart pulse values for all subjects

Subjects	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Mean	55	60	78	70	55	53	60	59	53	60	59	62	59	57	61	59	58	72	65	55
Median	56	60	78	56	53	58	59	53	58	59	63	57	60	60	58	71	65	55	69	71
Standard Deviation	1	1	1	1	1	1	1	1	2	1	2	1	2	5	1	2	1	2	2	2

6.5 Alarms

According to our experiments, we got the average of the pupil diameter change as follows:
- The size of pupil diameter in the initial state (before closing the eye) is **2.4mm**.
- The size of pupil diameter in dilation is **2.9mm**.
- The size of pupil diameter in the final state is **2.4mm**.

Therefore, if the sizes detected at run time are not the same as the above sizes, then an exception is raised an the test would fail—that is, we detected either a spoofing attack or a system error.

The curves in Figures 19a and 19b demonstrate how the fake eyes shape differs from the expected curve. For the first fake eye in Figure 19a, we notice that the curve is in a straight line without any changes in pupil diameter.

As per the second fake eye in Figure 19b, there is no dilatation in the curve after closing the eye. While for the Figures 19c and 19d, we notice that the curve is lower than the initial state, which can be considered as an error, due to incorrect positioning of the user during the scan of the eye.

7 Discussion

In the following, we will discuss the significance of our proposal. Our proposed system is designed to recognize the iris and at the same time the liveness of different parts of the eye: the pupil and the eyelid.

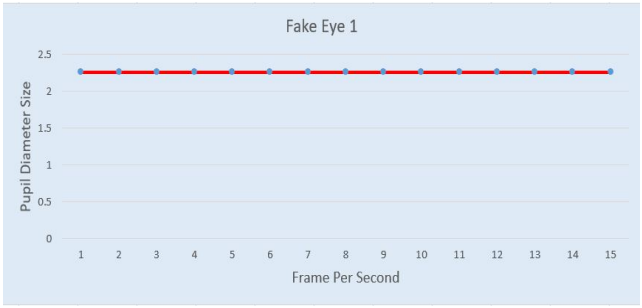
7.1 Pupil analysis detection

As demonstrated in the previous section, we are concentrating on the response of the pupil to the surrounding light without using LED flashlight to decrease the pupil diameter size. The physiological characteristics of the eye make the pupil dilates and shrinks, according to on the surrounding illumination. Therefore, with just ambient illumination, we are able to capture the diameter size of the pupil during all process of our proposal.

According to the Figures 14a and 14b of the experiment, we notice that the curve for the pupil of the subjects was stable when the eye was opened and changed to 'zero' during closing the eye. Next, with opening the eye again the curve raised and then it stabilizes at around the same level of when the eye was initially opened. This experiment proves the pupil's response to surrounding light.

7.2 Iris recognition

We have conducted numbers of experiments to optimize the performance quality of the iris recognition system. These experiments were focused on finding the best parameters for the Gabor filter in order to raise the performance of the proposed system. As depicted in Figures 20a, 20b, 20c, and 20d, we evaluated FAR and FRR of 4 datasets: our dataset (40 irises), CASIA interval, CASIA-Twin and UBiris.V1. In our dataset, as mentioned, we matched the iris of every subject two times: before and after closing the eye. The two irises that are matched two times must be similar to each other. consequently, we enhanced the recognition accuracy in every time and obtained better results as per the FAR and the FRR, as shown in 20a.
The external datasets (CASIA-Interval, CASIA-Twin, Ubiris.V1) show different results due to their huge datasets compared with our dataset and their image quality and noise as shown in Figures 20b, 20c, and 20d.



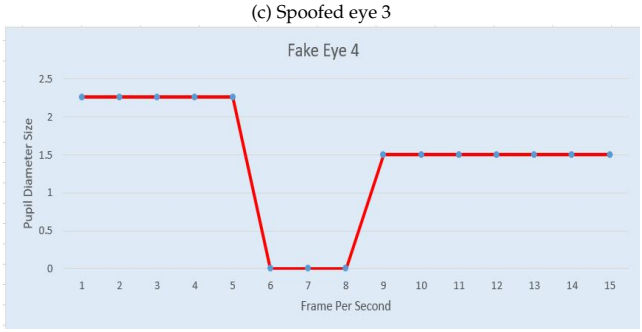
(a) Spoofed eye 1



(b) Spoofed eye 2



(c) Spoofed eye 3



(d) Spoofed eye 4

Figure 19: (a) Straight line. (b) The curve in the initial state is equal to dilation state . (c) The curve remains static line without any changes after the initial state. (d) The curve does not raise after closing.

7.3 Eyelid analysis detection

As per eyelid analysis, we provide a novel contribution by measuring the heartbeats of the eyelid. This technique implements liveness recognition of the human eye. Indeed, our

system detects the vein which is located inside the eyelid skin and then our system begins reading the signals of heart pulse.

In the experiment for the eyelid heartbeat, we have matched the results gathered by the proposed system with the results gathered from a certified oximeter machine. Subsequently, the results obtained shown to be very close to each other, while scanning for 20 seconds every subject by both systems, as depicted in Figures 16 and 17.

7.4 Assessment of the proposed work

In this section we show our proposed solution thwarts the previously identified attacker capabilities.

- *Pupil detection:* With the proposed method in pupil detection, the attackers will likely fail in bypassing the checks. For example: if the attacker makes a 3D model of the victim's eye, the pupil size will not change by expanding or contracting as a live eye and, subsequently, the proposed system will detect this 3D model as a fake eye. Note that the attackers could attempt to fake the pupil motion by presenting filmed video. However, these efforts will not succeed. The proposed system captures the diameter size of the pupil and its actual size changes in relation to the period of opening and closing the user's eye.

- *Eyelid detection:* By detecting the heart pulse of the eyelid, the attackers will be faced with another challenge to discover a method to create a fake heart pulse.

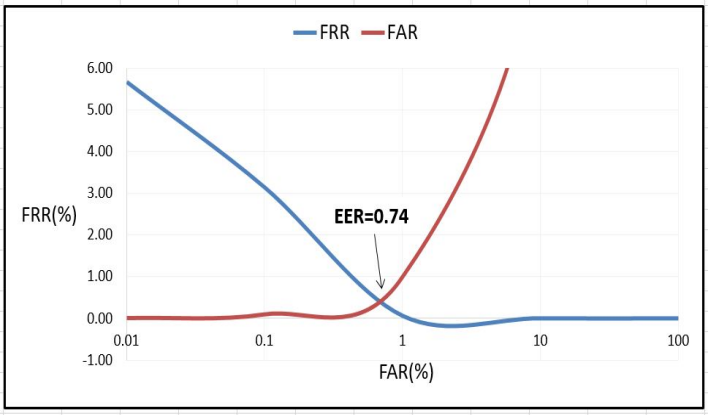
Thus, merging eyelid heart-beat and pupil identification into our solution will enhance the security of iris recognition. As an outcome, we have compared in table 4 our work with prior works in terms of spoofing detection capabilities.

8 Limitations of the proposed system

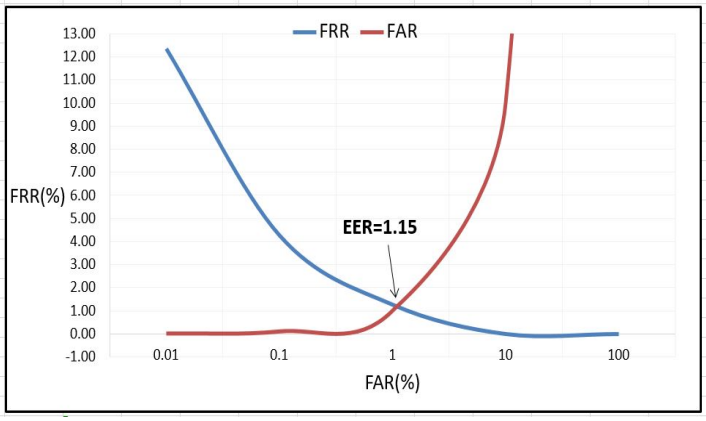
One possibility for an attacker to overcome our solution would be to adopt a new approach in spoofing attacks: eye robot. The attack could be deployed as illustrated in the following.

- *Pupil detection:*
  If the adversary designs an eye robot—including a synthetic pupil—, the robot eye can be configured to have the pupil of normal size, similar to a real pupil, and after a period the pupil size of the robot eye may dilate and contract as a normal pupil. So, there is a probability for the adversary to bypass the proposed system.

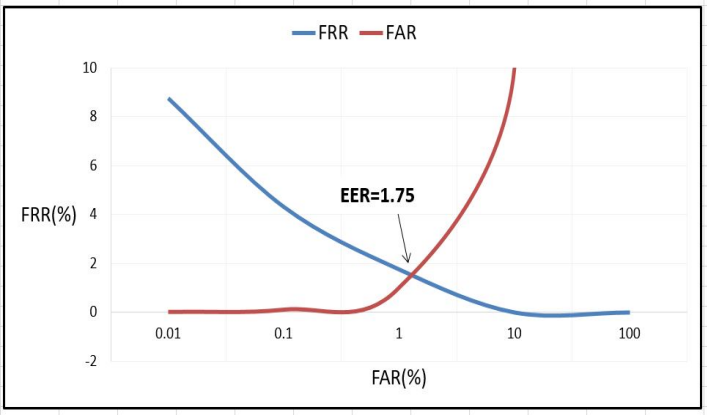
- *Eyelid detection:*
  The eye robot could also have synthetic eyelid and the adversary can add a particular fluid inside the synthetic skin that could mimic the bloodstream—with support of a pump system. Therefore, the proposed system could be triggered into identifying the spoofed pumped-fluid as a real bloodstream.



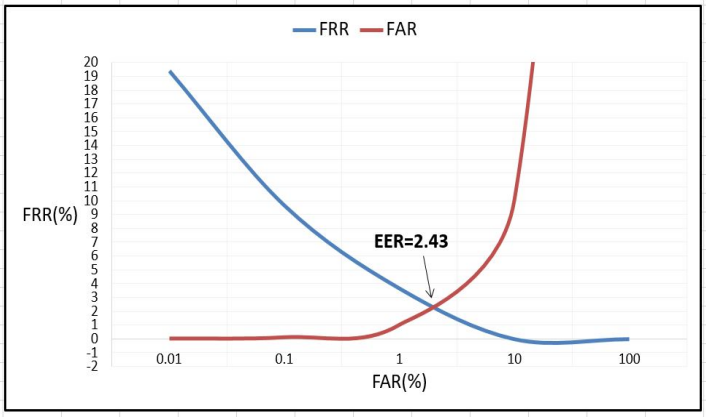
(a) Our dataset



(b) CASIA-interval



(c) UBIRIS



(d) CASIA-twin

Figure 20: The evaluation results of FAR and FRR.

Table 4: Comparison of the proposed work with previous works in terms of detecting spoofing attacks

Author	Features and Methodology	Photo attack	Video attack	Artificial Eye	Contact Lens	Fake Eyelid
Czajka [25]	Pupil dynamics	Detected	Detected	Detected	Detected	N.A.
Raghavendra [26]	Pupil dynamics	Detected	Detected	Detected	Detected	N.A.
Zhang [27]	Fake iris detection by using LBP and statistical features	Detected	Detected	Detected	Detected	N.A.
The Proposed system	Pupil size detection with heart pulse recognition by eyelid	Detected	Detected	Detected	Detected	Detected

Moreover, the ambient illumination can impact the heart pulse signals of the eyelid. For example, the adversary concentrates LED light on the synthetic eye robot to make noise in the heart beats signals that are captured by the camera sensor for the eyelid. The noise will give wrong readings of the heart pulses to the Webcam. Hence, even if unlikely, the adversary has a potential chance to bypass the check.

9 Conclusion

In this paper, we have provided a solution to eye biometric spoofing. Our solution relies on multiple factors. First, we examine the pupil movements during opening and closing the eye with support of just ambient illumination. Later, we recognize whether the eye is closed, and then we start measuring remotely a heartbeat signals from the eyelid to validate the liveness feature. During these checks, we match the iris against a pre-loaded database two times: the first time before closing the eye, and the second time after closing the eye. A third contribution is a database composed of 40 subjects. The same 40 subjects have contributed to test the proposed methodology: by using a camera sensor with a software program that includes all the algorithms for liveness and iris recognition. The results have shown the quality and viability of our proposal. We believe that the proposed methodology, together with the experimental results reported in this paper, other than being interesting on their own, will also encourage researchers to conduct future research in this field.

References

[1] A. Al-Rashid, "A Three Steps Eye-Liveness Validation System," 2019 International Conference on Cyber Security for Emerging Technologies (CSET), Doha, Qatar, 2019, pp. 1-8, doi: 10.1109/CSET.2019.8904884.

[2] K.R. Park Robust Fake Iris Detection. In: Perales F.J., Fisher R.B. (eds) Articulated Motion and Deformable Objects. AMDO 2006. Lecture Notes in Computer Science, **4069**. Springer, Berlin, Heidelberg, 2006. doi.org/10.1007/11789239.2

[3] J. Connell, N. Ratha, J. Gentile and R. Bolle, "Fake iris detection using structured light," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8692-8696, doi: 10.1109/ICASSP.2013.6639363.

[4] Rigas, Ioannis, and Oleg V. Komogortsev. "Eye movement-driven defense against iris print-attacks." Pattern Recognition Letters. Elsevier, 2015. doi.org/10.1016/j.patrec.2015.06.011

[5] A. George and A. Routray, "Fast and accurate algorithm for eye localisation for gaze tracking in low-resolution images," in IET Computer Vision, **10**(7), pp. 660-669, 10, 2016. doi.org/10.1049/iet-cvi.2015.0316

[6] M. Poh, D. J. McDuff and R. W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," in IEEE Transactions on Biomedical Engineering, **58**(1), pp. 7-11, Jan. 2011, doi: 10.1109/TBME.2010.2086456.

[7] M. P. Tarvainen, P. O. Ranta-aho and P. A. Karjalainen, "An advanced detrending method with application to HRV analysis," in IEEE Transactions on Biomedical Engineering, **49**(2), pp. 172-175, Feb. 2002, doi: 10.1109/10.979357.

[8] Wayman, James, Book review: Handbook of Iris Recognition. Biometrics, IET. 3. 41-43, 2014. DOI: 10.1049/iet-bmt.2014.0003.

[9] J. Daugman, "Biometric personal identification system based on iris analysis," U.S. Patent 5 291 560, Mar. 1, 1994.

[10] J. Shi and X. Gu, "The comparison of iris recognition using principal component analysis, independent component analysis and Gabor wavelets," 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, pp. 61-64, 2010. doi: 10.1109/ICCSIT.2010.5563947.

[11] J. Galbally and M. Gomez-Barrero, "A review of iris anti-spoofing," 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, pp. 1-6, 2016. doi: 10.1109/IWBF.2016.7449676.

[12] X. Li, J. Komulainen, G. Zhao, Pong-Chi Yuen and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, pp. 4244-4249, 2016. doi: 10.1109/ICPR.2016.7900300.

[13] C. Wang, T. Pun, G.A. Chanel, "Comparative Survey of Methods for Remote Heart Rate Detection From Frontal Face Videos". Front Bioeng Biotechnol, **6**(33).doi: 10.3389/fbioe.2018.00033

[14] S. Thavalengal, T. Nedelcu, P. Bigioi and P. Corcoran, "Iris liveness detection for next generation smartphones," in IEEE Transactions on Consumer Electronics, **62**(2), pp. 95-102, May 2016, doi: 10.1109/TCE.2016.7514667.

[15] Królak, A., Strumiłło, P. Eye-blink detection system for human–computer interaction. Univ Access Inf Soc 11, 409–419, 2012. https://doi.org/10.1007/s10209-011-0256-6

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 2001. doi: 10.1109/CVPR.2001.990517.

[17] R. G. Bozomitu, A. Păsărică, V. Cehan, C. Rotariu and C. Barabaşa, "Pupil centre coordinates detection using the circular Hough transform technique," 2015 38th International Spring Seminar on Electronics Technology (ISSE), Eger, pp. 462-465, 2015. doi: 10.1109/ISSE.2015.7248041.

[18] A. R. Azar and F. Khalilzadeh, "Real time eye detection using edge detection and Euclidean distance," 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, pp. 43-48, 2015. doi: 10.1109/KBEI.2015.7436019.

[19] M. Heshmat, M. Girgis, W. M. Abd-Elhafiez and S. Elaw, "An efficient scheme for face detection based on contours and feature skin recognition," 2015 Tenth International Conference on Computer Engineering and Systems (ICCES), Cairo, pp. 255-260, 2015. doi: 10.1109/ICCES.2015.7393056.

[20] W. Wang, S. Stuijk and G. de Haan, "Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG," in IEEE Transactions on Biomedical Engineering, **62**(2), pp. 415-425, Feb. 2015, doi: 10.1109/TBME.2014.2356291.

[21] J. Hernandez-Ortega, J. Fierrez, A. Morales and P. Tome, "Time Analysis of Pulse-Based Face Anti-Spoofing in Visible and NIR," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, pp. 657-6578, 2018. doi: 10.1109/CVPRW.2018.00096.

[22] J. Daugman, "How iris recognition works. Proceedings of 2002 International Conference on Image Processing, **1**, 2002. doi: 10.1109/TCSVT.2003.818350

[23] O. Koç, L. Tosku, J. Hoxha, A. O. Topal, M. Ali and A. Uka, "Detailed Analysis of IRIS Recognition Performance," 2019 International Conference on Computing, Electronics and Communications Engineering (iCCECE), London, United Kingdom, pp. 253-258, 2019. doi: 10.1109/iCCECE46942.2019.8941784.

[24] CASIA-Iris, Chinese Academy of Sciences–Institute of Automation. http://www.cbsr.ia.ac.cn/english/IrisDatabase.asp

[25] A. Czajka, "Pupil Dynamics for Iris Liveness Detection," in IEEE Transactions on Information Forensics and Security, **10**(4), pp. 726-735, April 2015, doi: 10.1109/TIFS.2015.2398815.

[26] R. Raghavendra and C. Busch, "Robust Scheme for Iris Presentation Attack Detection Using Multiscale Binarized Statistical Image Features," in IEEE Transactions on Information Forensics and Security, **10**(4), pp. 703-715, April 2015, doi: 10.1109/TIFS.2015.2400393.

[27] H. Zhang, Z. Sun, and T. Tan. Contact lens detection based on weighted lbp. In Proc. of ICPR, pages 4279–4282, 2010. doi: 10.1109/ICPR.2010.1040

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 1, 134-145 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Balancing Exploration-Exploitation in the Set Covering Problem Resolution with a Self-adaptive Intelligent Water Drops Algorithm

Broderick Crawford[*,1], Ricardo Soto[1], Gino Astorga[2], José Lemus-Romani[1], Sanjay Misra[3], Mauricio Castillo[1], Felipe Cisternas-Caneo[1], Diego Tapia[1], Marcelo Becerra-Rozas[1]

[1]*Pontificia Universidad Católica de Valparaíso, Valparaíso, 2340000, Chile*

[2]*Universidad de Valparaíso, Valparaíso, 2340000, Chile*

[3]*Covenant University, Ota, 100001, Nigeria*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 23 December, 2020
Online: 15 January, 2021

Keywords:
Metaheuristics
Autonomous Search
Combinatorial Optimization
Set Covering Problem
Intelligent Water Drops

ABSTRACT

The objective of the metaheuristics, together with obtaining quality results in reasonable time, is to be able to control the exploration and exploitation balance within the iterative processes of these methodologies. Large combinatorial problems present ample search space, so Metaheuristics must efficiently explore this space; and exploits looking in the vicinity of good solutions previously located. The objective of any metaheuristic process is to achieve a "proper" balance between intensive local exploitation and global exploration. In these processes two extreme situations can occur, on the one hand an imbalance with a bias towards exploration, which produces a distributed search in the search space, but avoiding convergence, so the quality of the solutions will be low, the other case is the bias towards exploitation, which tends to converge prematurely in local optimals, impacting equally on the quality of the solutions. To make a correct balance of exploration and exploitation, it is necessary to be able to control adequately the parameters of the Metaheuristics, in order to infer in the movements taking advantage of the maximum capacity of these. Among the most widely used optimization techniques to solve large problems are metaheuristics, which allow us to obtain quality results in a short period of time. In order to facilitate the use of the tools provided by the metaheuristic optimization techniques, it is necessary to reduce the difficulties in their configuration. For this reason, the automatic control of parameters eliminates the difficult task of obtaining a correct configuration. In this work we implemented an autonomous component to the Intelligent Water Drops algorithm, which allows the control of some parameters dynamically during the execution of the algorithm, achieving a good exploration-exploitation balance of the search process. The correct functioning of the proposal is demonstrated by the Set Covering Problem, which is a classic problem present in the industry, along with this we have made an exhaustive comparison between the standard algorithm and the autonomous version that we propose, using the respective statistical tests. The proposal presents promising results, along with facilitating the implementation of these techniques to industry problems.

1 Introduction

This paper is an extension of work "An Adaptive Intelligent Water Drops Algorithm for Set Covering Problem", originally presented in 19th International Conference on Computational Science and Its Applications (ICCSA) [1].

The post-pandemic economic recovery brings with it a number of challenges for the industry, ranging from rethinking business models to making good use of every available resource. In this sense, optimization is an important tool to achieve the desired recovery [2]–[6].

There are various economic sectors where there are problems that need to be optimised such as the airport sector where, given the environmental restrictions, it is becoming increasingly difficult to build new airports and for this reason it is necessary to optimise each of the tasks of the current installations [7]–[9]. Another ex-

[*]Corresponding Author: Broderick Crawford, broderick.crawford@pucv.cl

ample is the educational sector, in particular the universities where it is necessary to optimize the use of rooms and the curriculum of students [10, 11]. In summary, we can find a series of problems that can be treated through optimization techniques that have a positive impact on economic recovery.

In this sense the metaheuristics that are supervised heuristics [12], are a real option to give solution to the industrial optimization problems since their attraction is that they allow to deliver a high quality solution that can even be the optimal one in limited computation times. There is a great variety of metaheuristics which have been used to solve NP-Hard problems [13]–[16].

In general, metaheuristics have a great amount of parameters, which allow controlling the high and low level strategies of these, having direct relation with the performance of the techniques in the different problems [17], the optimal configuration of parameters constitutes in itself an optimization problem [18, 19]. The importance of a good parameter assignment to metaheuristic algorithms impacts on the quality of solutions, but on the other hand it is not possible to consider transversal parameters for all the problems to be solved, since the assignment of values depends directly on the problem and the instances to be solved [20].

The values that can assume the parameters can have a great amount of combinations, for that reason we can occupy different strategies to approach the optimal configuration of parameters, that according to the literature are divided in two great groups, off-line configuration and on-line control. The off-line configuration of parameters corresponds to the determination of the metaheuristic algorithm outside the run, that is, before its execution begins [21], while the on-line control dynamically updates the values of the algorithm during the execution [22].

Given the current context of using metaheuristic techniques to solve the problems of the industry, we must facilitate the use of these techniques to non-expert users, so the use of autonomous techniques for the control of parameters, is a great contribution to bring this type of tools to all users of the industry.

The present work, solves a classic problem of the real world, as it is it Set Convering Problem (SCP) [23], which has diverse applications in the industry, using metaheuristic techniques to solve problems of great dimension. We have used the metaheuristic technique Intelligent Water Drops (IWD), which is inspired by the physical behavior of water droplets in a river bed [24], Along with this, we have incorporated autonomous elements for the on-line control of its parameters, which has two main components, the first is the obtaining of external information on the problems to be solved, and the second is the obtaining of internal information on the behaviour of the algorithm. With this we can better combine the information of each problem to be solved, along with the internal behavior of the algorithm, which generates that our proposal is adapted to the various problems that are solved.

The work is structured in the following form, in Section 2 what has been done regarding parameter tuning is presented, in Section 3 the functioning of IWD is explained, while in Section 4 the adaptive implementation proposed for IWD is detailed, in Section 5 set covering problem is shown, in Section 6 the corresponding experiments and statistical analyses are presented, ending in Section 7 with the conclusions and future work.

2 Parameter Tuning

In this section we will review the various techniques of parameterization of algorithms in order to find a correct configuration.

The proper selection of parameter values for the algorithms is not an easy task and has an important impact. Usually a great deal of time is spent, using previous experiences and expert knowledge for the correct assignment of values which also constitutes a not easy task for non-expert users. In consideration of the latter, the ideal condition from the point of view of the non-expert user is that he or she should give general information about the problem and with the least possible interaction receive an adequate response (Fig. 1). This concept, called Autonomous Search, was proposed by [18], using indicators to determine the best strategy to use.

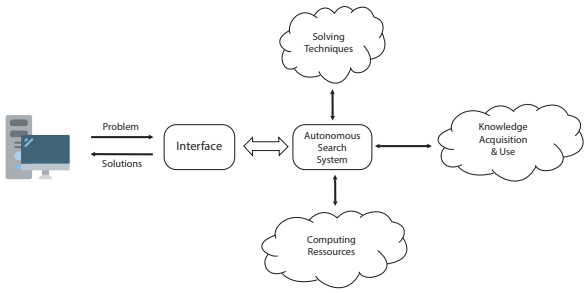


Figure 1: Autonomous Search.

The techniques for finding a good configuration for the parameters were initially grouped into two categories: offline configuration and online control. However, a new proposal is made in [21]: Simple Generate-Evaluate Methods, Iterative Generate-Evaluate Methods and High-Level Generate-Evaluate Methods.

2.1 Offline configuration

Finding a convenient parameter configuration previous to the execution of the algorithm is known as offline configuration.

This is primarily a trial and error process and can consume considerable time in research. The efficiency of this depends mainly on the insight and knowledge of the researcher or author of the algorithm. These processes are typically undocumented and are not reproducible, often driving to an unequal adjustment of different algorithms.

The following techniques exist within this group:

- F-race is an inspired method of racing algorithms in automatic learning, in particular, races in [25, 26]. This algorithm was presented in [27] and studied in detail in Birattari's PhD thesis [28]. The purpose of these methods is to iteratively evaluate a set of candidate configurations for a given set of instances. A set is removed when there is sufficient statistical evidence and there are survivors within the continuous race. This method is used after each evaluation of the candidate configurations, in which nonparametric Friedman two-way variance analysis [29] determines if there is evidence that at less than one of the configurations is significantly different

from the other ones. If the hypothesis is null and there is no difference, it is rejected; Friedman's subsequent tests are applied to remove these candidate configurations that are significantly worse than the best. This approach had been used by the ACO in [30].

- ParamILS was introduced by [31] as a versatile stochastic local search approach for automatic algorithm configuration. ParalILS includes methods to optimize the performance of an algorithmic, deterministic or stochastic object in a particular class of problems by switching a set of ordinal and/or categorical parameters. This proposal is based on an iterated local search algorithm that uses a combination of configurations chosen, a priori or at random, at initialization. The underlying idea of local search is an iterative "best first" method, which uses a variety of random movements to disturb the solution and avoid getting into local minima. The method always accepts configurations that improve or match the performance of the best configuration and restarts the search from a random configuration under a certain probability. The local search method moves through the search space by modifying only one parameter each time[32]. This approach has been used to configure ACO to solve transportation planning problems in [33].

- The other approach to optimization, focused on the sequential paradigm, includes improving the parameter initial values by the alternation of experimental design and the parameter recognition.

  In this paradigm, statistical significance takes a preponderant role, since each new experiment contributes with information about the performance of the parameters used, which are then referenced to the new stages.

  In the case of a parallel method, different experiments are formulated concurrently (all of them taking the same parameter nominal values); experiments are then carried out. The parameters are calculated and their suitability tested using the data obtained in all parallel experiments. If the data obtained from the parallel experiments are inadequate after the parameter estimation the procedure can be replicated [34, 35].

- In [36], the author used the graphic radial technique. The four basic metrics are employed for this method: worst case, best case, average case and average run-time. The region under the radar plot curve is obtained with these four metrics to set the best setting.

- The meta-optimization method was initially studied by [37] and is characterized by using an algorithm to optimize the parameters of another optimization algorithm, that is, there is an algorithm at a higher level that optimizes the parameters of a lower-level algorithm. This method has been used for covering problems in facility locations in [38], stand management in [39], and parameter selection in machine learning in [40].

2.2 Online configuration

One of the areas of research that has taken great strength in recent years is the online configuration, as it gives the algorithms the ability to adapt to the characteristics of a particular instance for better performance. For this online configuration to be useful, the exploration and exploitation phases must be clearly identified because the internal adjustments may be different for each of the phases. Thus, the online configuration can improve the results when the algorithms are used in situations very different from those that were built.

Different techniques exist, and the most simple is to define the parameter variation rules before executing the algorithm.
One possibility is the use of parameter adaptation, where the parameter modification scheme is defined for some behavior statistics of the algorithm.

- Absolute evidence: An example is the threshold of the distance between the solutions.

- Relative evidence: Considers that the relative difference in the performance with parameters of different values adapt to those with the best behavior.

2.3 Simple Generate-Evualate Methods

In this method the principle of generating and evaluating is used. Candidate configurations are occupied and then each of the parameters are evaluated to find the best configuration (Figure 2).

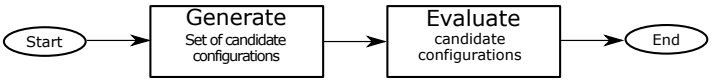


Figure 2: Simple Generate-Evualate Methods.

2.4 Iterative Generate-Evualate Methods

Unlike the first method, this one repeatedly performs the generation and evaluation steps. In this method the historical information allows to guide the generation, so that the search space for parameters is better explored and is more efficient in large search spaces (Figure 3).

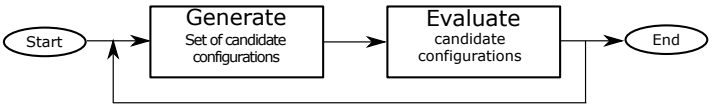


Figure 3: Iterative Generate-Evaluate Methods.

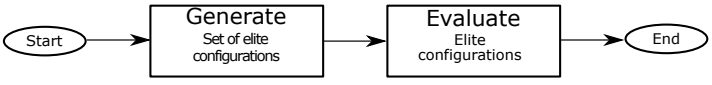


Figure 4: High-Level Generate-Evualate Methods.

2.5 High-Level Generate-Evualate Methods

This approach generates candidate configurations to then evaluate them and select the best configuration. The idea is to generate a set of high quality configurations quickly using few computational resources and select the best one. The idea is to greatly reduce the resources used when exploring candidate configurations and use them to thoroughly evaluate these configurations (Figure 4).

3 Intelligent Water Drops Algorithm

During the year 2007 in [24] he presented the algorithm that is classified as constructive since it builds a solution from 0. Its inspiration is based on nature and corresponds to the displacement that water drops make through the flow of a river and its friction with the earth, considering that when a drop is moved from one point to another it displaces earth from the river bed making the friction less and less which makes the drops increase their speed and in turn incorporate part of the extracted earth.

In the Figure 5, we can see how the soil of the river behaves, on the one hand during the passage of the drop is removed soil from the bottom of the river and on the other hand is incorporated into the drop which directly influences the speed that is acquired, this is demonstrated in that two drops of similar size but with different speeds at the end of its movement are with different size because of the soil that is incorporated. This abstraction of the behavior of water drops in river soil, allows us to build diversified solutions at the beginning, while during the execution of the algorithm, these solutions converge to promising search spaces, which allows to intensify the search.

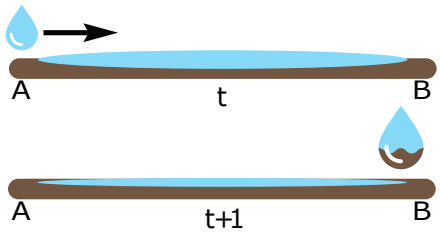


Figure 5: Water drop removing riverbed soil and adding to its own soil.

This algorithm has been used to solve mainly scheduling problems: Cooperative Search Path Optimization problem [41], Work Flow Scheduling Algorithm for Infrastructure as a Service (IaaS) cloud [42], Capacitated Vehicle Routing Problem [43], Trajectory Planning of Unmanned Combat Aerial Vehicle (UCAV) is a rather complicated Global Optimum Problem in UCAV Mission Planning [44], Method for Optimal Location and Sizing of Distributed Generation (DG) [45].

This algorithm has an initialization phase where both the static parameters number of drops, number of iterations, initial soil, initial speed and constants a_s, b_s, c_s, a_v, b_v, c_v are initialized and also the dynamic parameters Speed of drop k and soil value of drop k are initialized.

Then there is a construction phase where every drop builds a solution by visiting the nodes. The Eq. 1 is used to determine which node to visit:

$$p_i^k(j) = \frac{f(soil(i,j))}{\sum_{\forall l \notin VC^k} f(soil(i,l))} \tag{1}$$

where $soil(i,j)$ corresponds to the amount of soil between i and j and is calculated according to the following equation (Equation 2):

$$f(soil(i,j)) = \frac{1}{\varepsilon + g(soil(i,j))} \tag{2}$$

where ε is a small positive value to avoid division by zero.

The function g(soil(i,j) always obtains a positive soil value (Equation 3):

$$g(soil(i,j)) = \begin{cases} soil(i,j) & \text{if } \min_{\forall l \notin VC^k} soil(i,l) \geq 0, \\ soil(i,j) - \min_{\forall l \notin VC^k} soil(i,l) & \text{otherwise} \end{cases} \tag{3}$$

Once the drop has selected the new node, its velocity should be updated (equation 4):

$$vel^k(t+1) = vel^k(t) + \frac{a_v}{b_v + c_v \cdot soil^2(i,j)} \tag{4}$$

where a_v, b_v and c_v correspond to static parameters, and $soil(i,j)$ represents the amount of soil between i and j.

Next, it is necessary to update the soil after the step of the drop; for this, we use the following equation 5:

$$soil(i,j) = (1-\rho) \cdot soil(i,j) - \rho \cdot \Delta soil(i,j) \tag{5}$$

where ρ is a small positive value between 0 and 1, and $\Delta soil(i,j)$ is the amount of soil removed from paths i and j.

The value of $\Delta soil(i,j)$ is obtained with the following equation (6):

$$\Delta soil(i,j) = \frac{a_s}{b_s + c_s \cdot time(i,j : vel^k(t+1))} \tag{6}$$

where a_s, b_s and c_s are static parameters, and $time(i,j : vel^k(t+1))$ is calculated as follows (7):

$$time(i,j : vel^k(t+1)) = \frac{HUD(i,j)}{vel^k(t+1)} \tag{7}$$

where HUD is a heuristic that measures the degree of undesirability of the drop to jump from one node to another.

The Reinforcement phase is responsible for updating the global soil with the best drop of the iteration, and the following equation is used (8):

$$soil(i,j) = (1 + p_{iwd}) \cdot soil(i,j) - p_{iwd} \cdot \frac{1}{q(T^{IB})} \tag{8}$$

where p_{iwd} is a small positive value between 0 and 1, and $q(T^{IB})$ is the fitness value.

The best solution for each iteration is compared with the best global solution and updated according to the following equation (9):

$$T^{TB} = \begin{cases} T^{IB} & \text{if } q(T^{IB}) < q(T^{TB}) \\ T^{TB} & \text{otherwise} \end{cases} \tag{9}$$

Finally, the Termination phase is responsible for the completion of the algorithm process when the stop condition is met, which can be the number of iterations.

In 6 the original algorithm of the metaheuristics is shown.

Algorithm 1 Intelligent Water Drops Algorithm

{**Initialization phase**}
Initialize : Amount iterations, Amount drops.
Initialize : a_s, b_s, c_s
Initialize : a_v, b_v, c_v
repeat
  for $k = 1$ to N **do**
    Initialize VC^k list as empty.
    Initialize $soil^k$ value as zero.
    Initialize vel^k value as $InitVel$.
    Create the kth water drop (IWD^k).
    Select randomly a node for IWD^k.
    Update VC^k.
  end for
  {**Construction phase**}
  for $k = 1$ to N **do**
    Choose a path for IWD^k.
    Update velocity (vel^k).
    Compute the amount of soil ($\Delta soil$) to be carried by IWD^k
    Remove $\Delta soil$ from the path and add it to IWD^k
  end for
  for $k = 1$ to N **do**
    Calculate $fitness^k$
  end for
  {**Reinforcement phase**}
  Update Paths of the Best Solution
until $i = MAX\_ITERATION$
{**Termination phase**}

Figure 6: Intelligent Water Drops Algorithm.

4 Adaptative Intelligent Water Drops

The configuration of initial parameters is a costly task, carried out by an expert user and performed prior to the execution of the algorithm. We can reduce the cost of this task based on the concepts from Autonomous Search [18], which obtains information from two different sources, internal, corresponding to the operation of the algorithm and external, corresponding to the problem we are solving. The use of these data sources allows us to compare the quality of the solution in the current iteration (f_{IB}) with the quality of the best found solution (f_{TB}).

The general scheme for obtaining information to guide the adjustment of parameters is presented in Figure 7.

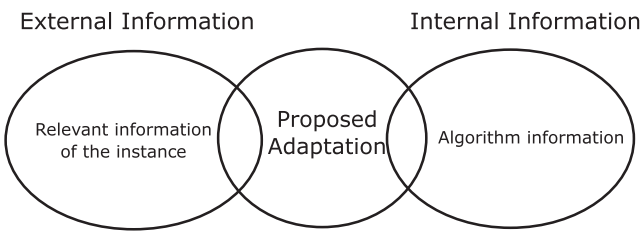


Figure 7: Information required for the proposal.

The internal information source considers the performance of the algorithm. In our proposal, we collect the number of iterations in which no improvement in the quality of the solutions is detected. The external information source comes from the data corresponding to the instance of the problem we are solving. We consider the number of columns and the density of the instance, which allows us to differentiate them. A weight is assigned to the columns of the instance that allows us to assign importance to them, multiplying it by the density, given in the Figure 8).

SET	**Instance Set**	m	n	Cost Range	Density	Optimum Solution
4	10	200	1000	[1,100]	2.00	Known
5	10	200	1000	[1,100]	2.00	Known
6	5	200	1000	[1,100]	5.00	Known
A	5	300	1000	[1,100]	2.00	Known
B	5	300	1000	[1,100]	5.00	Known
C	5	400	1000	[1,100]	2.00	Known
D	5	400	1000	[1,100]	5.00	Known
NRE	5	500	1000	[1,100]	10.00	Known
NRF	5	500	1000	[1,100]	20.00	Known

Figure 8: Standard description of the benchmark OR-library.

The algorithm has two initial parameters to which we assign values, these correspond to the Soil and Initial Speed ($Soil_{Initial}$ and $Velocity_{Initial}$ respectively).

In the first case we assign the value considering the number of columns of the instance and its density, which is done as follows:

To determine the first value we consider information of the instance considering on one hand the number of columns available and on the other hand the density that is presented. This value is obtained in the following way:

$$Soil_{Initial} = SoilInitial \cdot nColumn \cdot Density \tag{10}$$

Where $Soil_{Initial} = 1$, $nColumn$ corresponds to the number of columns in the instance group and $Density$ corresponds to the percentage of nonzeros in the matrix.

$$Velocity_{Initial} = 15 \tag{11}$$

For the case of the initial speed $Velocity_{Initial}$ the value is determined by a series of 10 previous executions of the original algorithm.

For the adaptation, in this work, it is considered the quality of the solution for which a weight is calculated that will allow us to update the local velocity, avoiding that the solution remains trapped in an optimal local. This percentage will be calculated by comparing the solution delivered in each iteration with the best solution obtained.

It has been experimentally determined that a low initial drop velocity impacts the convergence of solutions, making it more pronounced. This characteristic is taken into account in the parameter configuration.

Initial soil and velocity parameters update rule is given by the following equation:

$$Percentage = (f_{IB} \cdot 100)/f_{TB} \tag{12}$$

where f_{IB} is the best solution obtained at the current iteration, and f_{TB} is the best global solution found.

$$Soil_{Initial} = Soil_{Initial} \cdot Percentage \tag{13}$$

where $Soil_{Initial}$ is the available initial soil at the current iteration and $Percentage$ is the percentage difference between f_{IB} and f_{TB}.

$$Velocity_{Initial} = Velocity_{Initial} \cdot Percentage \cdot \beta \tag{14}$$

where $Velocity_{Initial}$ is the available initial velocity at current iteration and β is a random number in the interval $[0, 1]$.

Figure 9 shows the proposed adaptation for the local soil parameter, aiming to improve the metaheuristic performance.

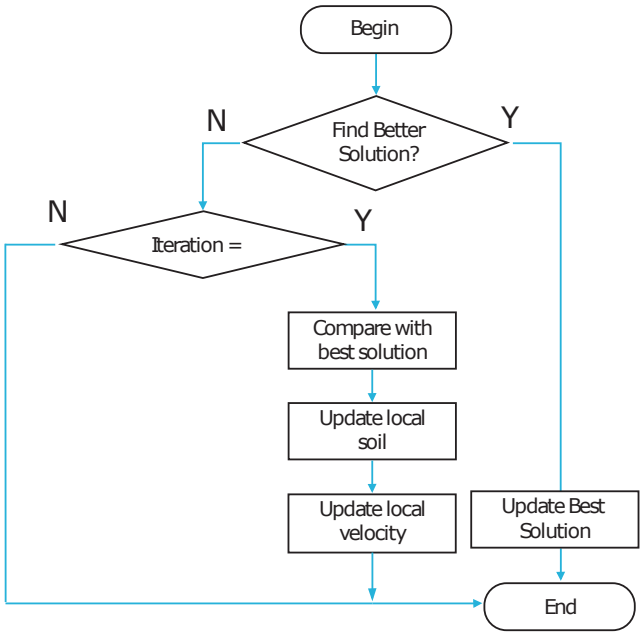


Figure 9: Parameter adaptation proposal.

At each iteration the improvement of the best solution found is evaluated. If an improvement is detected, the previous best solution found is replaced. Otherwise, a counter for not improved iterations is increased. An α value is set for the maximum number of iterations without improvement. If the counter is equal compared with the α parameter, the percentage difference of the current solution compared to the best solution is obtained and used to increase the local soil for the next iteration.

To obtain the value for the α parameter we perform 10 executions of the standard algorithm for each instance. The elapsed iteration number until a fitness improvement was evaluated. The analysis of this number allows us to determine the value for the α parameter to 5. A greater α value causes the fitness improvement probability to decrease, and the computation cost increases. The results of this analysis is shown in Figure 10, the column % represents the occurence percentage for each α, and *%Acum* column shows the accumulated percentage for each α. Figure 11 shows the fitness change for two instances.

α	Occurrences	%	% Acum
1	41	59%	59%
2	10	14%	74%
3	6	9%	83%
4	4	6%	88%
5	1	1%	90%
6	2	3%	93%
7	0	0%	93%
8	0	0%	93%
9	1	1%	94%
10	1	1%	96%
11	1	1%	97%
12	0	0%	97%
13	1	1%	99%
14	0	0%	99%
15	0	0%	99%
16	1	1%	100%
17	0	0%	100%
18	0	0%	100%
19	0	0%	100%
20	0	0%	100%
20+	0	0%	100%

α	Occurrences	%	% Acum
1	33	40%	40%
2	21	26%	66%
3	8	10%	76%
4	3	4%	79%
5	7	9%	88%
6	1	1%	89%
7	4	5%	94%
8	3	4%	98%
9	0	0%	98%
10	0	0%	98%
11	0	0%	98%
12	1	1%	99%
13	1	1%	100%
14	0	0%	100%
15	0	0%	100%
16	0	0%	100%
17	0	0%	100%
18	0	0%	100%
19	0	0%	100%
20	0	0%	100%
20+	0	0%	100%

Figure 10: Experiments to determine the alpha value.

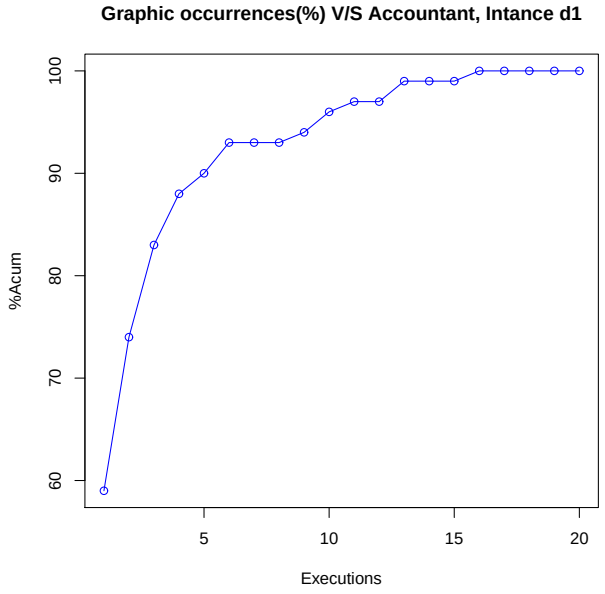


Figure 11: Chart of fitness change based on the number of iterations.

After a series of tests, we show the convergence for different initial velocities in Figure 12. When the velocity is low, the convergence shows to be premature. For this case, the initial velocity was adjusted, multiplying it by a β value in range $[0, 1]$.

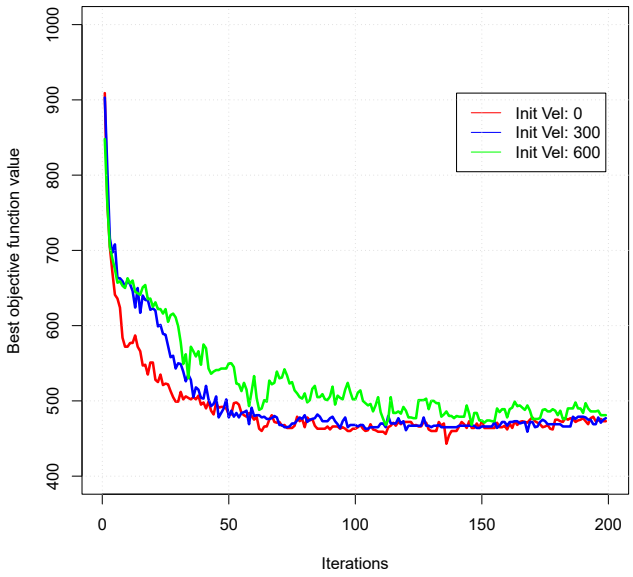


Figure 12: Convergence chart of different velocities.

5 Problem to solve

SCP is defined as a binary matrix (A) of size m-rows and n-columns, where $a_{i,j} \in \{0, 1\}$ is the value of each cell in the matrix A; i and j are of the size m-rows and n-columns, respectively:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & ... & a_{1,n} \\ a_{2,1} & a_{2,2} & ... & a_{2,n} \\ ... & ... & ... & ... \\ a_{m,1} & a_{m,2} & ... & a_{m,n} \end{pmatrix} \quad (15)$$

Defining column j satisfies a row i, if a_{ij} is equal to 1, which is a contrary case if a_{ij} is equal to 0. In addition, an associated cost $c \in C$ is incurred, where $C = \{c_1, c_2, ...c_n\}$, together with $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$, which are the sets of rows and columns, respectively.

The problem has an objective of minimizing the cost of the subset $S \subseteq J$, with the constraint that all rows $i \in I$ are covered by at least one column $j \in J$. When column j is in the subset of solution S, this is equal to 1 and is 0 otherwise.

The SCP can be defined as follows:

$$\text{Min} \quad Z = \sum_{j=1}^{n} c_j x_j \quad (16)$$

Subject to

$$\sum_{j=1}^{n} a_{ij} x_j \geq 1 \quad \forall i \in I \quad (17)$$

$$x_j \in \{0, 1\} \quad \forall j \in J \quad (18)$$

6 Experiments

6.1 Benchmark Instances

In order to validate the proposed approach, the 9 sets of the Beasleys library were used, executing 55 instances that are known and that allow a finished experimental evaluation (Figure 8).

The proposed algorithm was built using Java Language and was executed on an Intel(R) Core(TM) i7-6700 CPU with a speed of 3.40 GHz, with 16GB of memory and using a 64-bit Windows 10 operating system.

6.2 Parameter Setting

As far as the configuration of the parameters used by this algorithm is concerned, IWD, as it is generally known, is a very difficult task that requires a great use of resources. Our approach takes care of working two of the most important parameters in IWD operation which are the initial soil and the initial speed, IntSoil and InitVelocity respectively. To verify the impact of these two parameters, a series of experiments with different values for the selected parameters were carried out. The idea is to adapt these two parameters dynamically in order to use the quality of the solution as an indicator of change. The action occurs if the best overall solution does not change after a defined number of iterations.

The parameters used for the different experiments are shown in the following Figure 13.

Parameter	Detail	Value
N	Amount of drops.	5000
$MAX\_ITERATION$	Number ofiterations.	50
$InitialSoil$	Initial soil value	1000
a_s	Soil parameter “a"	900000
b_s	Soil parameter “b"	0.01
c_s	Soil parameter “c"	1.0
$InitialVel$	Initial velocity value	15
a_v	Velocity parameter “a"	10.0
b_v	Velocity parameter “b"	0.01
c_v	Velocity parameter “c"	1.0

Figure 13: Experimental parameters.

6.3 Experimental results

The results obtained with the adaptive test are shown in the tables 14 to 22 and the comparison between the original and adaptive algorithm is shown in the Figure 23.-.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
4.1	429	440	467.18	2.56
4.2	512	551	612.88	7.62
4.3	516	545	614.84	5.62
4.4	494	510	565.88	3.24
4.5	512	533	586.84	4.10
4.6	560	589	576.26	5.18
4.7	430	451	492.28	4.88
4.8	492	517	564.26	5.08
4.9	641	728	809.2	13.57
4.10	514	540	581.4	5.06

Figure 14: Results of Group 4.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
5.1	253	270	301.1	6.72
5.2	302	333	366.36	10.26
5.3	226	234	254.72	3.54
5.4	242	259	278.38	7.02
5.5	211	222	245.22	5.21
5.6	213	234	246.56	9.86
5.7	293	319	353.2	8.87
5.8	288	309	340.4	7.29
5.9	279	293	320.42	5.02
5.10	265	288	306.88	8.68

Figure 15: Results of Group 5.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
6.1	138	149	176.96	7.97
6.2	146	162	202.66	10.96
6.3	145	149	200.1	2.76
6.4	131	136	157.26	3.82
6.5	161	183	218.34	13.66

Figure 16: Results of Group 6.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
A.1	253	287	334.74	13.44
A.2	252	276	330.3	9.52
A.3	232	251	296.4	8.19
A.4	234	277	315.48	18.38
A.5	236	255	295.64	8.05

Figure 17: Results of Group A.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
B.1	69	79	113.52	14.49
B.2	76	96	130.84	26.32
B.3	80	89	141.92	11.25
B.4	79	90	145.16	13.92
B.5	72	81	121.38	12.50

Figure 18: Results of Group B.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
C.1	227	258	347.48	13.66
C.2	219	248	344.98	13.24
C.3	243	278	371.94	14.40
C.4	219	258	335.32	17.81
C.5	215	245	346.7	13.95

Figure 19: Results of Group C.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
D.1	60	77	134.72	28.33
D.2	66	81	149.44	22.73
D.3	72	89	179	23.61
D.4	62	75	162.26	20.97

Figure 20: Results of Group D.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
E.1	29	44	91.84	51.72
E.2	30	49	137	63.33
E.3	27	41	104.42	51.85
E.4	28	37	103.5	32.14
E.5	28	39	110.76	39.29

Figure 21: Results of Group E.

Instances	Z_{BKS}	Z_{min}	Z_{avg}	RPD
F.1	14	21	50.68	50.00
F.2	15	22	39.3	46.67
F.3	14	21	40.38	50.00
F.4	14	19	38.78	35.71
F.5	13	17	42.8	30.77

Figure 22: Results of Group F.

6.4 Statistical Analysis

Figure 24 shows a general scheme of the existing statistical techniques, where in this study the statistical analysis included the different tests:

- Kolmogorov-Smirnov-Lilliefors [46] is used to establish the independence of the samples.
- Wilcoxon's Signed Rank [47] is used to verify that the IWD algorithm with AD is better than the Standard IWD algorithm.

Both statistical tests considered a significance level of 0.05, so that values lower than 0.05 indicate that the null hypothesis cannot be assumed, i.e., H_0 is rejected.

		IWD_{STD}			IWD_{AD}			
Intances	Z_{BKS}	Z_{min}	Z_{avg}	RPD	Z_{min}	Z_{avg}	RPD	$diff_{rpd}$
4.1	429	454	503.48	5.83	440	467.18	2.56	3.26
5.1	253	285	335.28	12.65	270	301.1	6.72	5.93
6.1	138	177	217.96	28.26	149	176.96	7.97	20.29
A.1	253	310	350.68	22.53	287	334.74	13.44	9.09
B.1	69	92	124.44	33.33	79	113.52	14.49	18.84
C1	227	297	351.74	30.84	258	347.48	13.66	17.18
D.1	60	87	146.9	45.00	77	134.72	28.33	16.67
E.1	29	43	92.22	48.28	44	91.84	51.72	-3.45
F.1	14	21	50.68	50.00	21	50.68	50.00	0.00

Figure 23: Results IWD_{STD} AND IWD_{AD}.

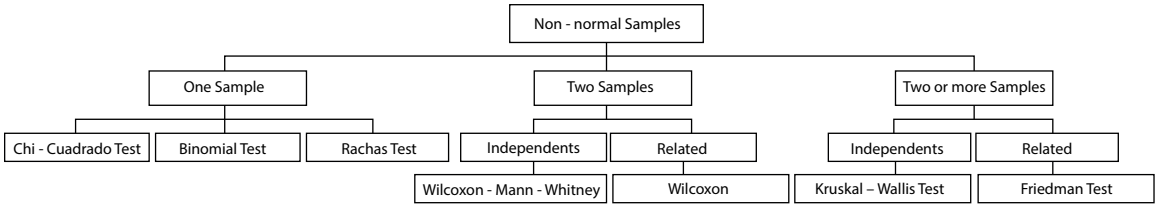


Figure 24: Statistical techniques.

For the determination of data independence the following hypothesis is used:

H_0 = The data follow a normal distribution.
H_1 = The data do not follow a normal distribution.

Given the P values obtained in the tests, the hypothesis is rejected. That is, the data do not follow a normal distribution.

H_0		IWD_{STD}	IWD_{AD}
4.1	IWD_{STD}	-	4.30^{-18}
	IWD_{AD}	>0.05	-
5.1	IWD_{STD}	-	6.45^{-18}
	IWD_{AD}	>0.05	-
6.1	IWD_{STD}	-	2.15^{-18}
	IWD_{AD}	>0.05	-
A.1	IWD_{STD}	-	2.15^{-18}
	IWD_{AD}	>0.05	-
B.1	IWD_{STD}	-	1.55^{-16}
	IWD_{AD}	>0.05	-
C.1	IWD_{STD}	-	9.88^{-16}
	IWD_{AD}	>0.05	-
D.1	IWD_{STD}	-	1.46^{-16}
	IWD_{AD}	>0.05	-
E.1	IWD_{STD}	-	>0.05
	IWD_{AD}	3.27^{-07}	-
F.1	IWD_{STD}	-	>0.05
	IWD_{AD}	1.33^{-09}	-

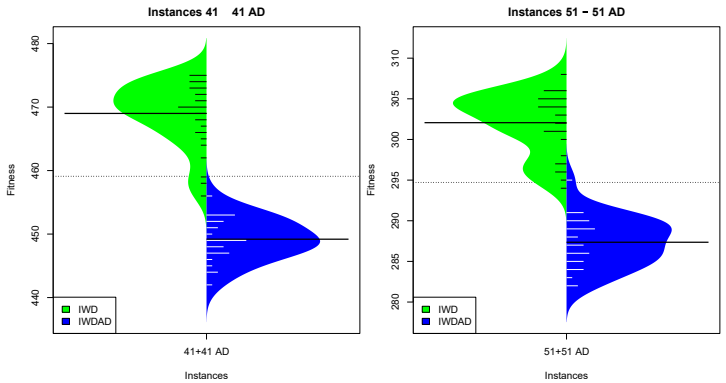
Figure 25: Statistical Analysis Results.

When it is obtained that the data does not follow a normal distribution, the Wilcoxon-Mann-Whitney [47] test is applied. This test is applied to verify that the version with AD is better, where the hypotheses are as follows:

H_0 = Standard IWD algorithm ≥ IWD algorithm with AD
H_1 = Standard IWD algorithm < IWD algorithm with AD

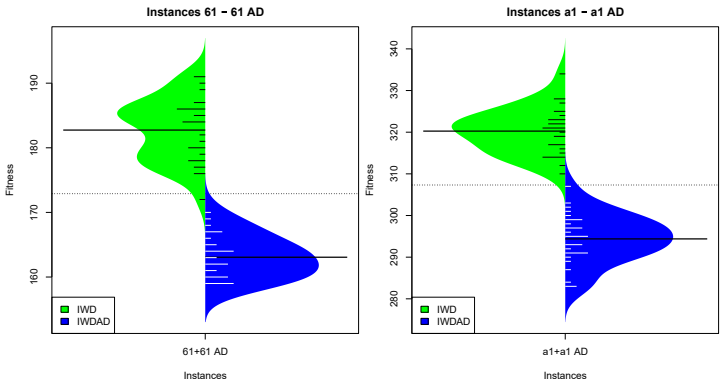
To obtain the p-values we use the programming language R, where obtaining a p-value < 0.05 implies rejecting H_0, and accepting H_1. The AD version is statistically better than the standard version, which extends to each instance of the benchmark (Figure 25. In addition, what is indicated by the statistical tests supports what has been obtained through verification by RPD.

The results of each instance can be seen graphically in the figure 26 to 30. IWD algorithm with AD has better results in all instances except for instances E and F.



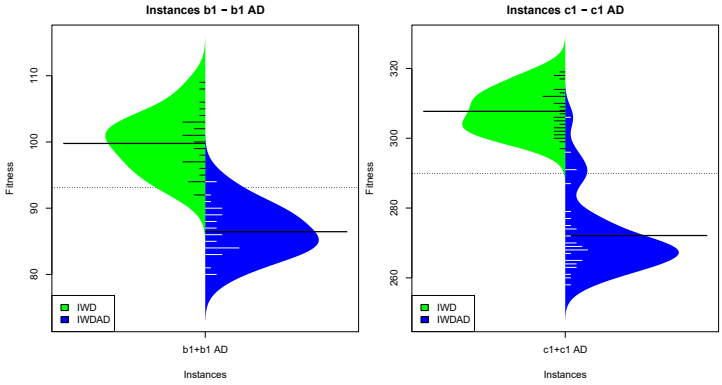
(a) Instance 4.1 (b) Instance 5.1

Figure 26: Instances 4.1 and 5.1.



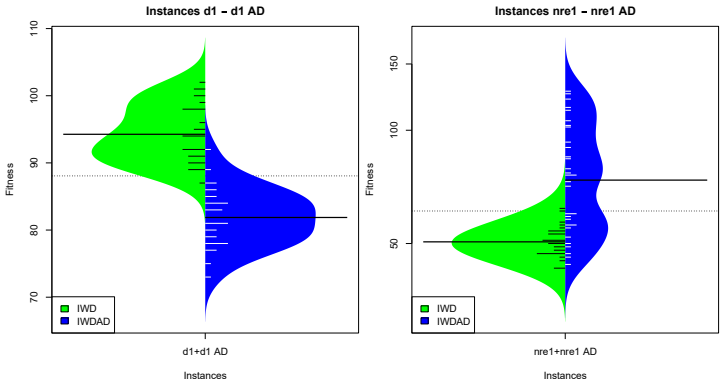
(a) Instance 6.1 (b) Instance A.1

Figure 27: Instances 6.1 and A.1.



(a) Instance B.1 (b) Instance C.1

Figure 28: Instances B.1 and C.1.



(a) Instance d.1 (b) Instance E.1

Figure 29: Instances D.1 and E.1.

7 Conclusion

In this work, we have presented an adaptive version of the Intelligent Water Drops algorithm, which is a tool that facilitates the use of metaheuristic techniques to non-expert users, reducing the difficulties in the configuration of parameters when implementing a metaheuristic in a real problem, a situation that will become increasingly common in the scenario of recovery from the economic crisis in post-pandemic times.

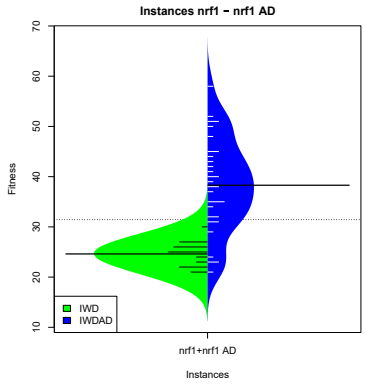


Figure 30: Instance F.1.

The adaptive element incorporated, obtains the necessary information to make decisions, based on external sources, coming from the problem and internal sources, typical of the normal functioning of the Intelligent Water Drops algorithm. With the correct combination of this information, it was possible to improve the balance of the exploration and exploitation of the search space of most of the solved instances, for Set Covering Problem, where the improvement was in average of 13.04%. In addition, the difference obtained by our adaptive proposal is significant in 77,78% of the compared instances, compared to the corresponding statistical comparisons.

The adaptive configuration is an interesting focus of future work, since there are a large number of static parameters that can be self-adjusting, in order to improve the performance of implementations, in order to decrease the influence of the off-line configuration in the performance when solving diverse instances. Along with this, it is necessary to replicate the implementation developed in this work, for other NP-Hard problems, validating the correct functioning of the incorporated adaptive components. Moreover, in the face of the growing line of research concerning the interaction between metaheuristics and machine learning techniques, a wide range of machine learning possibilities is presented in the decision making of the values to be taken by the adaptive parameters, adding to this, the option of choosing between different ways of calculating parameters according to internal and external information.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment Felipe Cisternas-Caneo and Marcelo Becerra-Rozas are supported by Grant DI Investigación Interdisciplinaria del Pregrado/VRIEA/PUCV/039.324/2020. Broderick Crawford is supported by Grant CONICYT/FONDECYT/REGULAR/1171243. Ricardo Soto is supported by Grant CONICYT/FONDECYT/REGULAR/1190129. José Lemus-Romani is supported by National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO NACIONAL/2019-21191692.

References

[1] B. Crawford, R. Soto, G. Astorga, J. Lemus-Romani, S. Misra, J.-M. Rubio, "An Adaptive Intelligent Water Drops Algorithm for Set Covering Problem," in 2019 19th International Conference on Computational Science and Its Applications (ICCSA), 39–45, IEEE, 2019, doi:10.1109/ICCSA.2019.000-6.

[2] A. Aloi, B. Alonso, J. Benavente, R. Cordera, E. Echániz, F. González, C. Ladisa, R. Lezama-Romanelli, Á. López-Parra, V. Mazzei, et al., "Effects of the COVID-19 Lockdown on Urban Mobility: Empirical Evidence from the City of Santander (Spain)," Sustainability, **12**(9), 3870, 2020, doi:10.3390/su12093870.

[3] J. B. Sobieralski, "COVID-19 and airline employment: Insights from historical uncertainty shocks to the industry," Transportation Research Interdisciplinary Perspectives, 100123, 2020, doi:10.1016/j.trip.2020.100123.

[4] M. Sigala, "Tourism and COVID-19: impacts and implications for advancing and resetting industry and research," Journal of Business Research, 2020, doi:10.1016/j.jbusres.2020.06.015.

[5] F. Hao, Q. Xiao, K. Chon, "COVID-19 and China's Hotel Industry: Impacts, a Disaster Management Framework, and Post-Pandemic Agenda," International Journal of Hospitality Management, 102636, 2020, doi:10.1016/j.ijhm.2020.102636.

[6] T. Laing, "The economic impact of the Coronavirus 2019 (Covid-2019): Implications for the mining industry," The Extractive Industries and Society, 2020, doi:10.1016/j.exis.2020.04.003.

[7] D. Wang, X. Zhao, L. Shen, Z. Yang, "Industry choice for an airport economic zone by multi-objective optimization," Journal of Air Transport Management, **88**, 101872, 2020, doi:10.1016/j.jairtraman.2020.101872.

[8] L. Adacher, M. Flamini, M. Guaita, E. Romano, "A model to optimize the airport terminal departure operations," Transportation Research Procedia, **27**, 53–60, 2017, doi:10.1016/j.trpro.2017.12.151.

[9] N. A. Ribeiro, A. Jacquillat, A. P. Antunes, A. R. Odoni, J. P. Pita, "An optimization approach for airport slot allocation under IATA guidelines," Transportation Research Part B: Methodological, **112**, 132–156, 2018, doi:10.1016/j.trb.2018.04.005.

[10] C. Castro, B. Crawford, E. Monfroy, "A quantitative approach for the design of academic curricula," in Symposium on Human Interface and the Management of Information, 279–288, Springer, 2007, doi:10.1007/978-3-540-73354-6_31.

[11] C. Castro, B. Crawford, E. Monfroy, "A genetic local search algorithm for the multiple optimisation of the balanced academic curriculum problem," in International Conference on Multiple Criteria Decision Making, 824–832, Springer, 2009, doi:10.1007/978-3-642-02298-2_119.

[12] C. Blum, A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," ACM computing surveys (CSUR), **35**(3), 268–308, 2003, doi:10.1145/937503.937505.

[13] C. Vásquez, B. Crawford, R. Soto, J. Lemus-Romani, G. Astorga, S. Misra, A. Salas-Fernández, J.-M. Rubio, "Galactic Swarm Optimization Applied to Reinforcement of Bridges by Conversion in Cable-Stayed Arch," in International Conference on Computational Science and Its Applications, 108–119, Springer, 2019, doi:10.1007/978-3-030-24308-1_10.

[14] B. Crawford, R. Soto, G. Astorga, J. Lemus, A. Salas-Fernández, "Self-configuring Intelligent Water Drops Algorithm for Software Project Scheduling Problem," in International Conference on Information Technology & Systems, 274–283, Springer, 2019, doi:10.1007/978-3-030-11890-7_27.

[15] R. Soto, B. Crawford, F. González, E. Vega, C. Castro, F. Paredes, "Solving the manufacturing cell design problem using human behavior-based algorithm supported by autonomous search," IEEE Access, **7**, 132228–132239, 2019, doi:10.1109/ACCESS.2019.2940012.

[16] S. Valdivia, R. Soto, B. Crawford, N. Caselli, F. Paredes, C. Castro, R. Olivares, "Clustering-Based Binarization Methods Applied to the Crow Search Algorithm for 0/1 Combinatorial Problems," Mathematics, **8**(7), 1070, 2020, doi:10.3390/math8071070.

[17] M. Birattari, J. Kacprzyk, Tuning metaheuristics: a machine learning perspective, volume 197, Springer, 2009, doi:10.1007/978-3-642-00483-4.

[18] Y. Hamadi, E. Monfroy, F. Saubion, "What is autonomous search?" in Hybrid Optimization, 357–391, Springer, 2011, doi:10.1007/978-1-4419-1644-0_11.

[19] M. López-Ibáñez, J. Dubois-Lacoste, L. P. Cáceres, M. Birattari, T. Stützle, "The irace package: Iterated racing for automatic algorithm configuration," Operations Research Perspectives, **3**, 43–58, 2016, doi:10.1016/j.orp.2016.09.002.

[20] F. Hutter, H. H. Hoos, K. Leyton-Brown, "Automated configuration of mixed integer programming solvers," in International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming, 186–202, Springer, 2010, doi:10.1007/978-3-642-13520-0_23.

[21] C. Huang, Y. Li, X. Yao, "A survey of automatic parameter tuning methods for metaheuristics," IEEE Transactions on Evolutionary Computation, **24**(2), 201–216, 2019, doi:10.1109/TEVC.2019.2921598.

[22] R. Battiti, M. Brunato, F. Mascia, Reactive search and intelligent optimization, volume 45, Springer Science & Business Media, 2008, doi:10.1007/978-0-387-09624-7.

[23] J. E. Beasley, "An algorithm for set covering problem," European Journal of Operational Research, **31**(1), 85–93, 1987, doi:10.1016/0377-2217(87)90141-X.

[24] H. Shah-Hosseini, "Intelligent water drops algorithm: A new optimization method for solving the multiple knapsack problem," International Journal of Intelligent Computing and Cybernetics, **1**(2), 193–212, 2008, doi:10.1108/17563780810874717.

[25] O. Maron, A. W. Moore, "Hoeffding races: Accelerating model selection search for classification and function approximation," in Advances in neural information processing systems, 59–66, 1994.

[26] O. Maron, A. W. Moore, "The racing algorithm: Model selection for lazy learners," in Lazy learning, 193–225, Springer, 1997, doi:10.1023/A:1006556606079.

[27] M. Birattari, T. Stützle, L. Paquete, K. Varrentrapp, "A racing algorithm for configuring metaheuristics," in Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation, 11–18, Morgan Kaufmann Publishers Inc., 2002, doi:10.5555/2955491.2955494.

[28] M. Birattari, M. Dorigo, "The problem of tuning metaheuristics as seen from a machine learning perspective," 2004, doi:10.1007/978-3-642-00483-4.

[29] W. Conover, "Statistics of the Kolmogorov-Smirnov type," Practical nonparametric statistics, 428–473, 1999.

[30] T. Liao, T. Stützle, M. A. M. de Oca, M. Dorigo, "A unified ant colony optimization algorithm for continuous optimization," European Journal of Operational Research, **234**(3), 597–609, 2014, doi:10.1016/j.ejor.2013.10.024.

[31] F. Hutter, H. H. Hoos, T. Stützle, "Automatic algorithm configuration based on local search," in Aaai, volume 7, 1152–1157, 2007, doi:10.5555/1619797.1619831.

[32] F. Hutter, H. H. Hoos, K. Leyton-Brown, T. Stützle, "ParamILS: an automatic algorithm configuration framework," Journal of Artificial Intelligence Research, **36**, 267–306, 2009, doi:10.1613/jair.2861.

[33] P. Lin, J. Zhang, M. A. Contreras, "Automatically configuring ACO using multilevel ParamILS to solve transportation planning problems with underlying weighted networks," Swarm and Evolutionary Computation, **20**, 48–57, 2015, doi:10.1016/j.swevo.2014.10.006.

[34] G. Franceschini, S. Macchietto, "Model-based design of experiments for parameter precision: State of the art," Chemical Engineering Science, **63**(19), 4846–4872, 2008, doi:10.1016/j.ces.2007.11.034.

[35] F. Hutter, H. H. Hoos, K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in International Conference on Learning and Intelligent Optimization, 507–523, Springer, 2011, doi:10.1007/978-3-642-25566-3_40.

[36] J. García, B. Crawford, R. Soto, C. Castro, F. Paredes, "A k-means binarization framework applied to multidimensional knapsack problem," Applied Intelligence, **48**(2), 357–380, 2018, doi:10.1007/s10489-017-0972-6.

[37] R. E. Mercer, J. Sampson, "Adaptive search using a reproductive meta-plan," Kybernetes, **7**(3), 215–228, 1978, doi:10.1108/eb005486.

[38] B. Crawford, R. Soto, E. Monfroy, G. Astorga, J. García, E. Cortes, "A meta-optimization approach for covering problems in facility location," in Workshop on Engineering Applications, 565–578, Springer, 2017, doi:10.1007/978-3-319-66963-2_50.

[39] X. Jin, T. Pukkala, F. Li, "Meta optimization of stand management with population-based methods," Canadian Journal of Forest Research, **48**(6), 697–708, 2018, doi:10.1139/cjfr-2017-0404.

[40] M. Camilleri, F. Neri, M. Papoutsidakis, "An algorithmic approach to parameter selection in machine learning using meta-optimization techniques," WSEAS Transactions on systems, **13**(2014), 202–213, 2014.

[41] X. Sun, C. Cai, S. Pan, Z. Zhang, Q. Li, "A cooperative target search method based on intelligent water drops algorithm," Computers & Electrical Engineering, **80**, 106494, 2019, doi:10.1016/j.compeleceng.2019.106494.

[42] M. Adhikari, T. Amgoth, "An intelligent water drops-based workflow scheduling for IaaS cloud," Applied Soft Computing, **77**, 547–566, 2019, doi:10.1016/j.asoc.2019.02.004.

[43] E. Teymourian, V. Kayvanfar, G. M. Komaki, M. Zandieh, "Enhanced intelligent water drops and cuckoo search algorithms for solving the capacitated vehicle routing problem," Information Sciences, **334**, 354–378, 2016, doi:10.1016/j.ins.2015.11.036.

[44] H. Duan, S. Liu, J. Wu, "Novel intelligent water drops optimization approach to single UCAV smooth trajectory planning," Aerospace science and technology, **13**(8), 442–449, 2009, doi:10.1016/j.ast.2009.07.002.

[45] D. R. Prabha, T. Jayabarathi, R. Umamageswari, S. Saranya, "Optimal location and sizing of distributed generation unit using intelligent water drop algorithm," Sustainable Energy Technologies and Assessments, **11**, 106–113, 2015, doi:10.1016/j.seta.2015.07.003.

[46] S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika, **52**(3/4), 591–611, 1965, doi:10.2307/2333709.

[47] H. B. Mann, D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," The annals of mathematical statistics, 50–60, 1947, doi:10.1214/aoms/1177730491.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 3, 142-149 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Performance Analysis of Joint Precoding and Equalization Design with Shared Redundancy for Imperfect CSI MIMO Systems

Bui Quoc Doanh[1,2], Ta Chi Hieu[1], Truong Sy Nam [3], Pham Thi Phuong Anh[4], Pham Thanh Hiep[*,1]

[1]*Faculty of Radio-Electronics Engineering, Le Quy Don Technical University, 100000, Vietnam*

[2]*Faculty of Professional Telecommunication, University of Telecommunications, 650000, Vietnam*

[3]*Center of Informations and Foreign Language, Ha Tinh medical College, 480000, Vietnam*

[4]*Missile Institute, Institute of military science and technology, 100000, Vietnam*

ARTICLE INFO

Article history:
Received: 20 March, 2020
Accepted: 26 April, 2020
Online: 10 May, 2020

Keywords:
Imperfect CSI
Cyclic prefix
Joint Precoding & Equalization
MIMO ISI channel
Protective interval
Unweighted MMSE

ABSTRACT

Analytical researches on a potential performance of multipath multiple-input multiple-output (MIMO) systems inspire the development of new technologies that decompose a MIMO channel into independent sub-channels on the condition of constrained transmit power. Moreover, in current studies of inter-symbol interference (ISI) MIMO systems, there is an assumption that channel state information (CSI) at receivers and/or transmitters is perfect. In this paper, we propose a hybrid design of precoding and equalization schemes based on the unweighted minimum mean square error criterion that not only eliminates the ISI but also improves the system performance. Additionally, the impact of imperfect channel knowledge at receivers on the system performance of MIMO ISI system is investigated. The simulation result shown that the proposed hybrid design of precoding and equalization with shared redundancy outperforms the conventional method in all considered scenarios. Furthermore, the proposed and the conventional schemes are extremely sensitive to the CSI factor, the performance of these systems is quickly deteriorated when the accuracy of channel estimation decreases.

1 Introduction

For satisfying the demand of advanced communications in terms of high data rates, relatively low costs and high-quality services, it is necessary to constantly improve the performance of wireless communication systems [1, 2]. However, protective intervals such as cyclic prefix or zero padding intervals are inserted to block transmission systems because of inter-symbol interference (ISI), which results in a decline in spectral efficiencies, especially in multiple-input multiple-output (MIMO) ISI channels having a long impulse response [3, 4]. In order to overcome the decrease of spectral efficiency and enhance the system performance, which attract lots of attention from the worldwide researchers. For example, redundancies are used instead of protector intervals throughout the signal processing to remove the ISI as in [5]-[9]. Moreover, various approaches based on the precoding scheme or the hybrid of precoding and equalization schemes of MIMO ISI channels were proposed [10]-[17].

In many approaches, it is assumed that channel status information (CSI) is able to be obtained perfectly. However, in practical scenarios, the perfect CSI is difficult to achieve in MIMO ISI systems because of the imperfection of channel estimation, feedback delay and finite rate channel quantization. Therefore, the research on a negative influences of imperfect CSI in MIMO ISI systems is necessary. A number of researches in this field developing precoders or connecting precoder and equalizer schemes with respect to the CSI at transmitters, receivers or transceivers have been published. There are many surveys expressed to analyze the performance of the MIMO system with imperfect CSI as follows.

Firstly, several surveys with the negative influences of imperfect CSI at transmitters were proposed and analyzed in [18]-[22]. To give a clear example, a MIMO system with joint decoding technique evaluated transmission rates which are affected by imperfect CSI [18]. Multi-user MIMO filterbank multicarrier systems employ-

[*]Corresponding Author: Pham Thanh Hiep, Faculty of Radio-Electronics Engineering, Le Quy Don Technical University, 236 Hoang Quoc Viet, Hanoi, 100000, Vietnam, Email: phamthanhhiep@gmail.com

ing the Zero Forcing technique are affected by imperfect CSI as demonstrated in [19]. The obtainable transmission rate of multipath channel systems with decoding technique was described in [20]. Analysis and investigation of MIMO systems occupying Tomlinson-Harashima precoding are mentioned in [21, 22].

Next, there have been also some papers studied the negative influences of imperfect CSI at the receiver [23]-[27]. For instance, the achievable transmission rate was evaluated in [23] by assumption of knowing the transmitted training symbols. The impact of CSI and feedback quantization error bring to MIMO systems in term of adaptive modulation was investigated in [24]. The effect of CSI to a capacity of MIMO systems was studied in [25]-[27].

Finally, an investigation about the combination of linear precoding and decoding to minimize the total mean-square error of MIMO systems was shown in [28, 29] when there is the imperfect CSI at transceivers. Furthermore, the negative influences of the imperfect CSI on a resource allocation of base stations by using several techniques such as transmit antennas selection, power allocation and beamforming was considered in [30]-[32].

In this work, we focus on combination of both precoding and equalization schemes, which employ redundancies for MIMO ISI systems. The proposed scheme, shared redundancies for both transmitters and receivers are used to perform a combination design of precoder and equalizer. In the next part, an investigation of the negative influence of imperfect CSI on the performance of these systems will be performed. Herein, the length of protective intervals generally defines redundancies as in [8].

The contribution of the research is summarized as follows.

- The proposed joint scheme of precoding and equalization schemes with shared redundancy is mathematical analyzed in both perfect and imperfect CSI.
- The conventional methods, such as training zero and leading zero, are investigated to compare with the proposed scheme.
- The impact of imperfect CSI on the system performance is discussed based on bit error rate and channel capacity.
- The proposed scheme is evaluated based on different parameters, such as the accuracy of channel estimation, the order of the finite impulse response and the transmission block size.

The order of other parts in this paper are as follows: In Section 2, the system model with imperfect CSI is introduced, while both precoding and equalization techniques are combined for the MIMO ISI systems, and then the combination of them are demonstrated. The negative influences of imperfect CSI on the performance of system is analyzed in Section 3. Simulation results and conclusions are shown in Section 4 and Section 5, respectively. Notations in this paper are used as follows: sets of complex numbers are represented by symbol $\mathbb{C}$, while $(\cdot)^H$ denotes the conjugate transpose, $(\cdot)^T$ is the transpose and boldface font is used for vector and matrix.

2 System Model

In this work, joint scheme of precoding and equalization for MIMO ISI channels of block transmission system models is proposed as illustrated in Fig.1. M_T and M_R denote the transmitting and receiving antennas, The channel model of this system is assumed to be frequency selective fading. Moreover, the finite impulse response (FIR) of channel has an order of D, in which the channel impulse response (CIR) is expressed by matrices $\bar{\mathbf{H}}[0], \bar{\mathbf{H}}[1], ..., \bar{\mathbf{H}}[D]$, herein, $\bar{\mathbf{H}}[d] \in \mathbb{C}^{M_T \times M_R}, (d = 0, ..., D)$.

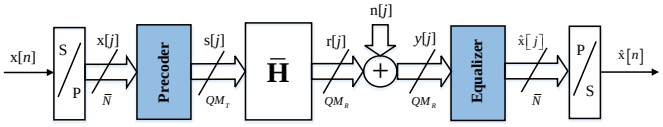


Figure 1: Hybrid scheme of precoder and equalizer in MIMO ISI systems

The system model of the MIMO ISI channel with joint scheme of precoder and equalizer in Figure 1 is expressed as follows. At the transmitter, a single input stream $\mathbf{x}[n]$ is transferred to a serial-to-parallel converter, and then converted into a block of vectors $\mathbf{x}[j]$ with a block size of $\bar{N} \times 1$. Then, the vector $\mathbf{x}[j]$ is sent to the precoder. The precoder performs signal processing procedure so as to perform symbol vector generation $\mathbf{x}[j]$ with the dimension of $QM_T \times 1$ from input vector $\mathbf{s}[j]$. Afterwards, the vector $\mathbf{s}[j]$ is divided into Q vectors, each of which has the size of $M_T \times 1$, and converted by the parallel-to-serial converter, then transmitted over the MIMO ISI channel. At the output side, the received symbol vector $\mathbf{y}[j]$ has two parts because of the affection of noise, including an information vector $\mathbf{r}[j]$ and a noise sample vector $\mathbf{n}[j]$ that is considered as an additive white Gaussian noise sample vector, and $\mathbf{n}[j] \backsim CN(0, 1)$.

The signal processing procedure is employed conversely at the receiver. $QM_R \times 1$ symbol vector $\mathbf{y}[j]$ is formed by the Q received vectors in the serial-to-parallel converter. Subsequently, the symbol vector $\mathbf{y}[j]$ is sent to the equalizer in order to regenerate the original symbol vector $\hat{\mathbf{x}}[i]$ with the dimension of $\bar{N} \times 1$. Finally, the parallel-to-serial converter reforms the output symbol stream $\hat{\mathbf{x}}[n]$ from the symbol vector $\hat{\mathbf{x}}[i]$.

Based on the operation and signal processing of system mentioned above, the terms defined as $\mathbf{x}[j]$, $\mathbf{s}[j]$, $\mathbf{y}[j]$, $\hat{\mathbf{x}}[i]$, $\mathbf{r}[j]$ and $\mathbf{n}[j]$ according to the input symbol stream $\mathbf{x}[n]$ and the sampled vector of received signal $\hat{\mathbf{x}}[n]$, can be mathematically shown in the (1) - (6) equations, respectively.

$$\mathbf{x}[j] = [x[j\bar{N}], x[j\bar{N}+1], \ldots, x[j\bar{N}+\bar{N}-1]]^T \quad (1)$$

$$\mathbf{s}[j] = [s[jQM_T], s[jQM_T+1], \ldots, s[jQM_T+QM_T-1]]^T \quad (2)$$

$$\mathbf{y}[j] = [y[jQM_R], y[jQM_R+1], \ldots, y[jQM_R+QM_R-1]]^T \quad (3)$$

$$\hat{\mathbf{x}}[j] = [\hat{x}[j\bar{N}], \hat{x}[j\bar{N}+1], \ldots, \hat{x}[j\bar{N}+\bar{N}-1]]^T \quad (4)$$

$$\mathbf{r}[j] = [r[jQM_R], r[jQM_R+1], \ldots, r[jQM_R+QM_R-1]]^T \quad (5)$$

$$\mathbf{n}[j] = [n[jQM_R], n[jQM_R+1], \ldots, n[jQM_R+QM_R-1]]^T \quad (6)$$

3 Performance Analysis

3.1 Joint Scheme of Precoder and Equalizer

In this section, the proposed joint scheme of precoder and equalizer based on MIMO ISI channels is explained. In addition, the channel

model is assumed to be narrow-band, therefore the Saleh-Valenzuela model [33] can be adopted. In the case of $Q \geq D$, the symbol $\hat{\mathbf{x}}[j]$ is expressed following the ideas in [6]

$$\hat{\mathbf{x}}[j] = \mathbf{G}\bar{\mathbf{H}}_0\mathbf{F}\mathbf{x}[j] + \mathbf{G}\bar{\mathbf{H}}_1\mathbf{F}\mathbf{x}[j-1] + \mathbf{G}\mathbf{n}[j] \tag{7}$$

where $\mathbf{n}[j] \in \mathbb{C}^{QM_R \times 1}$; $\mathbf{F} \in \mathbb{C}^{QM_T \times \bar{N}}$ and $\mathbf{G} \in \mathbb{C}^{\bar{N} \times QM_R}$ are respectively the precoding and the equalization matrices, and expressed by following equations

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_0[0] & \cdots & \mathbf{F}_{\bar{N}-2}[0] & \mathbf{F}_{\bar{N}-1}[0] \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{F}_0[QM_T-2] & \cdots & \mathbf{F}_{\bar{N}-2}[QM_T-2] & \mathbf{F}_{\bar{N}-1}[QM_T-2] \\ \mathbf{F}_0[QM_T-1] & \cdots & \mathbf{F}_{\bar{N}-2}[QM_T-1] & \mathbf{F}_{\bar{N}-1}[QM_T-1] \end{bmatrix} \tag{8}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_0[0] & \cdots & \mathbf{G}_{(QM_R-2)}[0] & \mathbf{G}_{(QM_R-1)}[0] \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{G}_0\begin{bmatrix}\bar{N}-2\end{bmatrix} & \cdots & \mathbf{G}_{(QM_R-1)}\begin{bmatrix}\bar{N}-2\end{bmatrix} & \mathbf{G}_{(QM_R-1)}\begin{bmatrix}\bar{N}-2\end{bmatrix} \\ \mathbf{G}_0\begin{bmatrix}\bar{N}-1\end{bmatrix} & \cdots & \mathbf{G}_{(QM_R-1)}\begin{bmatrix}\bar{N}-1\end{bmatrix} & \mathbf{G}_{(QM_R-1)}\begin{bmatrix}\bar{N}-1\end{bmatrix} \end{bmatrix}. \tag{9}$$

Additionally, $\bar{\mathbf{H}}_\mathbf{0}$ and $\bar{\mathbf{H}}_\mathbf{1}$ are utilized to give a definition of the Toeplitz matrices whose uniform dimension is $QM_R \times QM_T$, and expressed by following equations

$$\bar{\mathbf{H}}_0 = \begin{bmatrix} \bar{\mathbf{H}}[0] & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \vdots & \bar{\mathbf{H}}[0] & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \bar{\mathbf{H}}[D-1] & \vdots & \ddots & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \bar{\mathbf{H}}[D] & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \ddots & \ddots & \vdots & \ddots & \ddots & \mathbf{0} \\ \vdots & \cdots & \ddots & \ddots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{H}}[D] & \bar{\mathbf{H}}[D-1] & \cdots & \bar{\mathbf{H}}[0] \end{bmatrix}, \tag{10}$$

$$\bar{\mathbf{H}}_1 = \begin{bmatrix} \mathbf{0} & \cdots & \bar{\mathbf{H}}[D] & \bar{\mathbf{H}}[D-1] & \cdots & \bar{\mathbf{H}}[2] & \bar{\mathbf{H}}[1] \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{H}}[D] & \ddots & \cdots & \bar{\mathbf{H}}[2] \\ \vdots & \ddots & \vdots & \mathbf{0} & \ddots & \ddots & \vdots \\ \mathbf{0} & \ddots & \ddots & \vdots & \ddots & \ddots & \bar{\mathbf{H}}[D-1] \\ \mathbf{0} & \cdots & \ddots & \ddots & \vdots & \ddots & \bar{\mathbf{H}}[D] \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{11}$$

The effect of ISI on the MIMO ISI system model shown in Fig. 1 is the second component on the right side of the equation (7), ($\mathbf{GH}_1\mathbf{Fs}[i-1]$). With an assumption of $QM_T = \bar{M} + DM_T$, ($\bar{N} \leq \bar{M}$), two conditional methods (one is named by the trailing zero (TrZero) and the other one is the leading zero (LeZero) [6]) can be used to cancel the ISI.

Actually, although the performance of both TrZero and LeZero technical schemes is almost similar, the TrZero scheme is solely considered. For the TrZero technical scheme, the equalization matrix keeps the same ($\mathbf{G}_{\text{TrZero}} = \mathbf{G}$), whereas the precoding matrix changes the last DM_T rows to be zero. The precoding matrix has the form

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_0[0] & \cdots & \mathbf{F}_{(\bar{N}-2)}[0] & \mathbf{F}_{(\bar{N}-1)}[0] \\ \mathbf{F}_0[1] & \cdots & \mathbf{F}_{(\bar{N}-2)}[1] & \mathbf{F}_{(\bar{N}-1)}[1] \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{F}_0\begin{bmatrix}\bar{M}-2\end{bmatrix} & \cdots & \mathbf{F}_{(\bar{N}-2)}\begin{bmatrix}\bar{M}-2\end{bmatrix} & \mathbf{F}_{(\bar{N}-1)}\begin{bmatrix}\bar{M}-2\end{bmatrix} \\ \mathbf{F}_0\begin{bmatrix}\bar{M}-1\end{bmatrix} & \cdots & \mathbf{F}_{(\bar{N}-2)}\begin{bmatrix}\bar{M}-1\end{bmatrix} & \mathbf{F}_{(\bar{N}-1)}\begin{bmatrix}\bar{M}-1\end{bmatrix} \\ \mathbf{0} & \vdots & \mathbf{0} & \mathbf{0} \\ \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{12}$$

where the definition of $\mathbf{F}_{\text{TrZero}} \in \mathbb{C}^{(Q-D)M_T \times \bar{N}}$ is as follows

$$\mathbf{F}_{\text{TrZero}} = \begin{bmatrix} \mathbf{F}_0[0] & \cdots & \mathbf{F}_{(\bar{N}-2)}[0] & \mathbf{F}_{(\bar{N}-1)}[0] \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{F}_0\begin{bmatrix}\bar{M}-2\end{bmatrix} & \cdots & \mathbf{F}_{(\bar{N}-2)}\begin{bmatrix}\bar{M}-2\end{bmatrix} & \mathbf{F}_{(\bar{N}-1)}\begin{bmatrix}\bar{M}-2\end{bmatrix} \\ \mathbf{F}_0\begin{bmatrix}\bar{M}-1\end{bmatrix} & \cdots & \mathbf{F}_{(\bar{N}-2)}\begin{bmatrix}\bar{M}-1\end{bmatrix} & \mathbf{F}_{(\bar{N}-1)}\begin{bmatrix}\bar{M}-1\end{bmatrix} \end{bmatrix}. \tag{13}$$

Subsequently, the optimal criteria is utilized to design the $\mathbf{F}_{\text{TrZero}}$ precoder and the $\mathbf{G}_{\text{TrZero}}$ equalizer jointly to improve the system performance. Thus, when $\mathbf{G}\bar{\mathbf{H}}_1\mathbf{F}\mathbf{x}[j-1]$ is disappeared, it means that the ISI is not completely existent, the equation (7) becomes as follows

$$\hat{\mathbf{x}}[j] = \mathbf{G}_{\text{TrZero}}\bar{\mathbf{H}}\mathbf{F}_{\text{TrZero}}\mathbf{x}[j] + \mathbf{G}_{\text{TrZero}}\mathbf{n}[j]. \tag{14}$$

where the $\bar{\mathbf{H}}$ matrix comprises $(Q-D)M_T$ first columns of the $\bar{\mathbf{H}}_0$ matrix and is given by

$$\bar{\mathbf{H}} = \begin{bmatrix} \bar{\mathbf{H}}[0] & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \bar{\mathbf{H}}[0] & \ddots & \cdots & \mathbf{0} \\ \bar{\mathbf{H}}[D-1] & \ddots & \ddots & \ddots & \vdots \\ \bar{\mathbf{H}}[D] & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \ddots & \bar{\mathbf{H}}[0] \\ \mathbf{0} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \bar{\mathbf{H}}[D-1] \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{H}}[D] \end{bmatrix}. \tag{15}$$

When the transmit power is limited by p_0', the optimal MMSE criterion for joint precoder and equalizer schemes is calculated by

$$\mathbf{F}_{\text{TrZero}} = \mathbf{V}\mathbf{\Phi}\mathbf{U}^H \tag{16}$$

$$\mathbf{G}_{\text{TrZero}} = \mathbf{R}_{xx}\mathbf{F}_{\text{TrZero}}^H\bar{\mathbf{H}}^H(\mathbf{R}_{nn} + \bar{\mathbf{H}}\mathbf{F}_{\text{TrZero}}\mathbf{R}_{xx}\mathbf{F}_{\text{TrZero}}^H\bar{\mathbf{H}}^H)^{-1}, \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices obtained from the eigenvalue decompositions (EVD) algorithm, and $\mathbf{\Phi}$ is a diagonal matrix, whose main diagonal elements $\mathbf{\Lambda}$ matrix can be obtained from a water-filling algorithm,

$$\mathbf{R}_{xx} = \mathbf{U}\mathbf{\Delta}\mathbf{U}^H, \tag{18}$$

$$\bar{\mathbf{H}}^H\mathbf{R}_{nn}^{-1}\bar{\mathbf{H}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H, \tag{19}$$

where $\mathbf{R}_{xx}$ and $\mathbf{R}_{nn}$ are the covariance matrices of the input signal and noise, respectively.

Obviously, the signal to interference noise ratios (SINRs) among the decoupled flat sub-channels are not similar because of the unequal eigenvalues obtained from (19). Hence, the sub-channel with low SINR significantly influences the performance of the system, it means that those sub-channels should be discarded to enhance the system BER [17]. In addition, when the protective interval or the redundancy is utilized, the ISI interference can be cancelled in the frequency selective channel, however the addition of the protective interval or the redundancy lets channel energy be lost.

For solving the above issue, the ideas expressed in [6, 9] were combined and both precoder and equalizer technique schemes relied on the *unweighted* MMSE criterion [10] was proposed. This proposal aimed to share redundancies to both the transmitter and the receiver. Especially, instead of setting the last DM_T rows of the $\mathbf{F}$ precoding matrix, only last $\bar{K}M_T = \left\lfloor \frac{DM_T}{2} \right\rfloor$ rows are set to zero, while the first $(D-\bar{K})M_R$ columns of the $\mathbf{G}$ equalization matrix are also set to zero at the receiver. Herein, the transmitter and the receiver obtain the shared protective interval, it means that the first $(D-\bar{K})M_R$ rows of the $\bar{\mathbf{H}}_0$ channel matrix and the last $\bar{K}M_T$ columns of $\bar{\mathbf{H}}_0$ channel matrix are removed by the $\mathbf{G}$ equalization matrix and the $\mathbf{F}$ precoding matrix, respectively. Consequently, the loss in $\bar{\mathbf{H}}_0$ channel matrix is able to reduced by the proposed design. In the other words, the partial channel energy loss can be reduced. Basing on the above analysis, in our method, the precoder and equalizer are designed as follows (20) and (21).

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_0[0] & \mathbf{F}_1[0] & \cdots & \mathbf{F}_{(\bar{N}-1)}[0] \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{F}_0\left[(Q-\bar{K})M_T\right] & \mathbf{F}_1\left[(Q-\bar{K})M_T\right] & \cdots & \mathbf{F}_{\bar{N}-1}\left[(Q-\bar{K})M_T\right] \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \tag{20}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_0[0] & \cdots & \mathbf{G}_{(Q-D+\bar{K})M_R-1}[0] \\ \vdots & \cdots & \vdots & \mathbf{G}_0[1] & \cdots & \mathbf{G}_{(Q-D+\bar{K})M_R-1}[1] \\ \mathbf{0} & \cdots & \mathbf{0} & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_0\left[\bar{N}-1\right] & \cdots & \mathbf{G}_{(Q-D+\bar{K})M_R-1}\left[\bar{N}-1\right] \end{bmatrix}. \tag{21}$$

Where $\hat{\mathbf{F}} \in \mathbb{C}^{(Q-\bar{K})M_T \times N}$ and $\hat{\mathbf{G}} \in \mathbb{C}^{N \times (Q-D+\bar{K})M_R}$ are given by following equations (22) and (23).

$$\hat{\mathbf{F}} = \begin{bmatrix} \mathbf{F}_0[0] & \mathbf{F}_1[0] & \cdots & \mathbf{F}_{(\bar{N}-1)}[0] \\ \mathbf{F}_0[1] & \mathbf{F}_1[1] & \cdots & \mathbf{F}_{(\bar{N}-1)}[1] \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{F}_0\left[(Q-\bar{K})M_T\right] & \mathbf{F}_1\left[(Q-\bar{K})M_T\right] & \cdots & \mathbf{F}_{\bar{N}-1}\left[(Q-\bar{K})M_T\right] \end{bmatrix}, \tag{22}$$

$$\hat{\mathbf{G}} = \begin{bmatrix} \mathbf{G}_0[0] & \mathbf{G}_1[0] & \cdots & \mathbf{G}_{(Q-D+\bar{K})M_R-1}[0] \\ \mathbf{G}_0[1] & \mathbf{G}_1[1] & \cdots & \mathbf{G}_{(Q-D+\bar{K})M_R-1}[1] \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{G}_0\left[\bar{N}-1\right] & \mathbf{G}_1\left[\bar{N}-1\right] & \cdots & \mathbf{G}_{(Q-D+\bar{K})M_R-1}\left[\bar{N}-1\right] \end{bmatrix}. \tag{23}$$

The *unweighted* MMSE criterion will be used to design the $\hat{\mathbf{F}}$ precoder and the $\hat{\mathbf{G}}$ equalizer [15]. weak eigenmodes are dropped, so they do not affect the BER of system. The transmit power is redistributed for the remaining eigenvalues. Consequently, power distribution to the higher eigenvalues is larger than that of the MIMO ISI channel.

When the ISI cancellation is complete, the equation (7) can be expressed as follows

$$\hat{\mathbf{x}}[j] = \hat{\mathbf{G}}\hat{\mathbf{H}}\hat{\mathbf{F}}\mathbf{x}[j] + \hat{\mathbf{G}}\mathbf{n}'[j] \tag{24}$$

where the $\hat{\mathbf{H}}$ channel matrix is defined as

$$\hat{\mathbf{H}} = \begin{bmatrix} \bar{\mathbf{H}}\left[D-\bar{K}\right] & \cdots & \bar{\mathbf{H}}[0] & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \bar{\mathbf{H}}[D-1] & & & & \ddots & \ddots & \mathbf{0} \\ \bar{\mathbf{H}}[D] & \ddots & & & & \ddots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & & & & \bar{\mathbf{H}}[0] \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{H}}[D] & \bar{\mathbf{H}}[D-1] & \cdots & \bar{\mathbf{H}}\left[\bar{K}\right] \end{bmatrix}, \tag{25}$$

herein, the noise sample block with the length of $(Q-D+\bar{K})M_R$ is denoted as $\mathbf{n}'[j]$.

3.2 In the Case of Imperfect CSI

Factually, the perfect CSI cannot be produced at both the transmitter and the receiver. In this work, we assume that the perfect CSI exists at the transmitter, which means that its feedback is immediate and error-free, whereas the receiver has the imperfect CSI. Herein, the accurate channel matrix and the inaccurate channel matrix with a complex Gaussian distribution are denoted as $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}_{\mathrm{e}}$, respectively, in which $\hat{\mathbf{H}}_{\mathrm{e}} \in CN(0,1)$. Thus, the illustration of the channel matrix is as follows [34, 35]

$$\tilde{\mathbf{H}} = \xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}. \tag{26}$$

The sizes of these matrices are $\left(Q-D+\bar{K}\right)M_R \times \left(Q-\bar{K}\right)M_T$, and the accuracy of channel estimation is illustrated by ξ, where $0 \le \xi \le 1$. Therefore, the equation (24) is rewritten as

$$\hat{\mathbf{x}}[j] = \tilde{\mathbf{G}}\tilde{\mathbf{H}}\tilde{\mathbf{F}}\mathbf{x}[j] + \tilde{\mathbf{G}}\mathbf{n}'[j] \tag{27}$$

The $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{G}}$ matrices of the precoder and the equalizer are designed under the imperfect CSI at the receiver, and optimized in accordance with the *unweighted* MMSE criterion [15]. Therefore, the mathematical description is expressed as

$$\min_{\tilde{\mathbf{G}},\tilde{\mathbf{F}}} : c\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) = E\left\|\mathbf{W}^{1/2}\mathbf{e}\right\|^2,$$
$$\text{Subject to: } tr\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\right) \le p_0', \tag{28}$$

where $\mathbf{e} = \mathbf{x}[j] - (\tilde{\mathbf{G}}\tilde{\mathbf{H}}\tilde{\mathbf{F}}\mathbf{x}[j] + \tilde{\mathbf{G}}\mathbf{n}'[j])$, the weight matrix $\mathbf{W} = \mathbf{I}$ and the transmit power is constrained to p_0'. The expectation (E) relates to the distribution of $\mathbf{x}$ and $\mathbf{n}$.

$$c\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) = E\|\mathbf{e}\|^2 = E\left(tr\left[\mathbf{e}\mathbf{e}^H\right]\right) = tr\left[\mathrm{Re}\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right)\right], \tag{29}$$

where the error covariance matrix is denoted the $\mathrm{Re}\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right)$ and defined as $\mathrm{Re}\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) := E\left(\mathbf{e}\mathbf{e}^H\right)$. Using the expression for $\mathbf{e}$ and the above assumption, we have

$$E\left(\mathbf{x}\mathbf{x}^H\right) = \mathbf{I}, \tag{30}$$

$$E\left(\mathbf{n}'\mathbf{n}'^H\right) = \mathbf{R}_{\mathbf{n}'\mathbf{n}'}, \tag{31}$$

$$E\left(\mathbf{x}\mathbf{n}'^H\right) = 0. \tag{32}$$

Therefore,

$$\begin{aligned} Re\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) = &\left[\tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}} - \mathbf{I}\right]\times \\ &\left[\tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\bar{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}} - \mathbf{I}\right]^H + \tilde{\mathbf{G}}\mathbf{R}_{\mathbf{n}'\mathbf{n}'}\tilde{\mathbf{G}}^H. \end{aligned} \tag{33}$$

For solving the optimization problem in (28), the method of Lagrange duality and the Karush-Kuhn-Tucker (KKT) conditions are utilized, where η is the Lagrange multiplier.

$$L\left(\eta,\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) = c\left(\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) + \eta\left[tr\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\right) - p_0'\right]. \tag{34}$$

Substituting (29) and (33) into (34), we have

$$\begin{aligned} L\left(\eta,\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right) = tr\Bigg\{ &\left[\tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}}+\sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}}-\mathbf{I}\right]\times \\ &\left[\tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}}+\sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}}-\mathbf{I}\right]^H+\tilde{\mathbf{G}}\mathbf{R}_{\mathbf{n}'\mathbf{n}'}\tilde{\mathbf{G}}^H \Bigg\} \\ &+\eta\left[tr\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\right)-p_0'\right] \\ =tr\Bigg\{ &\left[\tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}}+\sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\left(\xi\hat{\mathbf{H}}+\sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)^H\tilde{\mathbf{G}}^H\right]- \\ &\tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}}+\sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}}-\tilde{\mathbf{F}}^H\left(\xi\hat{\mathbf{H}}+\sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)^H\tilde{\mathbf{G}}^H \\ &+\mathbf{I}+\tilde{\mathbf{G}}\mathbf{R}_{\mathbf{n}'\mathbf{n}'}\tilde{\mathbf{G}}^H \\ &+\eta\left[tr\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\right)-p_0'\right] \Bigg\} \end{aligned} \tag{35}$$

By applying the KKT condition, the $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{G}}$ matrices are optimal only if there is an η satisfying the following condition.

$$\nabla_{\tilde{\mathbf{F}}} L\left(\eta,\tilde{\mathbf{F}},\tilde{\mathbf{G}}\right) = 0, \tag{36}$$

$$\nabla_{\tilde{\mathbf{G}}} L\left(\eta,\tilde{\mathbf{F}},\tilde{\mathbf{G}}\right) = 0, \tag{37}$$

$$\eta \geq 0; \mathrm{tr}\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\right) - p_0' \leq 0, \tag{38}$$

$$\eta\left[\mathrm{tr}\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\right) - p_0'\right] = 0. \tag{39}$$

Applying the derivatives of $L\left(\eta,\tilde{\mathbf{G}},\tilde{\mathbf{F}}\right)$ with respect to $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{G}}$ as in [36] and substitute (35) into (36) and (37). As a result, the $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{G}}$ matrices can be calculated as follows

$$\begin{aligned} \left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\bar{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}} = &\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H\times \\ &\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)^H\tilde{\mathbf{G}}^H + \mathbf{R}_{\mathbf{n}'\mathbf{n}'}\tilde{\mathbf{G}}^H. \end{aligned} \tag{40}$$

$$\begin{aligned} \tilde{\mathbf{G}}\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right) = &\tilde{\mathbf{F}}^H\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)^H\tilde{\mathbf{G}}^H\tilde{\mathbf{G}}\times \\ &\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right) + \eta\tilde{\mathbf{F}}^H. \end{aligned} \tag{41}$$

where $\tilde{\mathbf{F}}$, $\tilde{\mathbf{G}}$ are designed by the *unweighted* MMSE criterion. Consequently, the optimal $\tilde{\mathbf{F}}$, $\tilde{\mathbf{G}}$ matrices are derived as follows

$$\tilde{\mathbf{F}} = \tilde{\mathbf{V}}\tilde{\mathbf{\Phi}}_{\mathbf{f}}, \tag{42}$$

$$\tilde{\mathbf{G}} = \tilde{\mathbf{\Phi}}_{\mathbf{g}}\tilde{\mathbf{V}}^H\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)^H\mathbf{R}_{\mathbf{n}'\mathbf{n}'}^{-1}, \tag{43}$$

where the η, $\tilde{\mathbf{\Phi}}_{\mathbf{f}}$ and $\tilde{\mathbf{\Phi}}_{\mathbf{g}}$ are given by

$$\eta^{1/2} = \frac{\sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1/2}\right)}{p_0' + \sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1}\right)}, \tag{44}$$

$$\tilde{\mathbf{\Phi}}_{\mathbf{f}} = \left(\eta^{-1/2}\tilde{\mathbf{\Lambda}}^{-1/2} - \tilde{\mathbf{\Lambda}}^{-1}\right)_+^{1/2}, \tag{45}$$

$$\tilde{\mathbf{\Phi}}_{\mathbf{g}} = \left(\eta^{-1/2}\tilde{\mathbf{\Lambda}}^{-1/2} - \eta\tilde{\mathbf{\Lambda}}^{-1}\right)_+^{1/2}\tilde{\mathbf{\Lambda}}^{-1/2}. \tag{46}$$

Applying (44) in (45) and (46), we can calculate as

$$\left|\tilde{\phi}_{\mathrm{f},jj}\right|^2 = \left[\frac{p_0' + \sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1}\right)}{\sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1/2}\right)}\tilde{\lambda}_{jj}^{-1/2} - \tilde{\lambda}_{jj}^{-1}\right], \tag{47}$$

$$\left|\tilde{\phi}_{\mathrm{g},jj}\right|^2 = \left\{\frac{p_0' + \sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1}\right)}{\sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1/2}\right)}\tilde{\lambda}_{jj}^{-1/2} - \left[\frac{\sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1/2}\right)}{p_0' + \sum_{i=1}^{k}\left(\tilde{\lambda}_{ii}^{-1}\right)}\right]^2\tilde{\lambda}_{jj}^{-1}\right\}\tilde{\lambda}_{jj}^{-1}, \tag{48}$$

where the main diagonal element of $\tilde{\mathbf{\Lambda}}$ is denoted by $\tilde{\lambda}_{jj}$. Moreover, the $\tilde{\mathbf{\Lambda}}$ and $\tilde{\mathbf{V}}$ are matrices obtained from the EVD algorithm

$$\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right)^H\mathbf{R}_{\mathbf{n}'\mathbf{n}'}^{-1}\times\left(\xi\hat{\mathbf{H}} + \sqrt{1-\xi^2}\hat{\mathbf{H}}_{\mathrm{e}}\right) = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^H. \tag{49}$$

The precoder and equalizer matrices are planed as equations (42) and (43); therefore, the calculation of (27) can be expressed as follows

$$\begin{aligned} \hat{\mathbf{x}}[j] &= (\tilde{\mathbf{G}}\xi\hat{\mathbf{H}} + \tilde{\mathbf{G}}\sqrt{1-\xi^2}\hat{\mathbf{H}}_e)\hat{\mathbf{H}}(\tilde{\mathbf{F}}\xi\hat{\mathbf{H}} + \tilde{\mathbf{F}}\sqrt{1-\xi^2}\hat{\mathbf{H}}_e)\mathbf{x}[j] + \tilde{\mathbf{G}}\mathbf{n}'[j] \\ &= \xi^2\tilde{\mathbf{G}}\hat{\mathbf{H}}\hat{\mathbf{H}}\tilde{\mathbf{F}}\hat{\mathbf{H}}\mathbf{x}[j] + \xi\sqrt{1-\xi^2}\tilde{\mathbf{G}}\hat{\mathbf{H}}_e\hat{\mathbf{H}}\tilde{\mathbf{F}}\hat{\mathbf{H}}\mathbf{x}[j] \\ &+ \xi\sqrt{1-\xi^2}\hat{\mathbf{G}}\hat{\mathbf{H}}_e\hat{\mathbf{H}}\tilde{\mathbf{F}}\hat{\mathbf{H}}\mathbf{x}[j] + \left(1-\xi^2\right)\tilde{\mathbf{G}}\hat{\mathbf{H}}_e\hat{\mathbf{H}}\tilde{\mathbf{F}}\hat{\mathbf{H}}_e\mathbf{x}[j] + \tilde{\mathbf{G}}\mathbf{n}'[j]. \end{aligned} \tag{50}$$

Subsequently, the SINR of sub-channels and the capacity of the system (C_{sys}) are able to be given as

$$\mathrm{SINR}_j = \frac{\xi^2\left|\tilde{\phi}_{\mathrm{f},jj}\right|^2\hat{\lambda}_{jj}}{(2\xi\sqrt{1-\xi^2} + (1-\xi^2))\left|\tilde{\phi}_{\mathrm{f},jj}\right|^2\hat{\lambda}_{jj} + 1}, \tag{51}$$

$$C_{sys} = \sum_{j=1}^{j=\bar{N}}\log_2\left(1 + \frac{\xi^2\left|\tilde{\phi}_{\mathrm{f},jj}\right|^2\hat{\lambda}_{jj}}{(2\xi\sqrt{1-\xi^2} + (1-\xi^2))\left|\tilde{\phi}_{\mathrm{f},jj}\right|^2\hat{\lambda}_{jj} + 1}\right), \tag{52}$$

where $\bar{N} \leq rank(\bar{K}M_R, \bar{K}M_T)$, and $\hat{\lambda}_{jj}$ is the main diagonal of $\hat{\mathbf{\Lambda}}$ that is able to be obtained by mathematical calculation as follows

$$\hat{\mathbf{H}}^H\mathbf{R}_{\mathbf{n}'\mathbf{n}'}^{-1}\hat{\mathbf{H}} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^H. \tag{53}$$

4 Simulation Results

To evaluate the performance of the proposed scheme under the condition of imperfect CSI at the receiver, the system model is assumed to have four transmit antennas and four receive antennas. The system performance is evaluated according to the channel estimation accuracy ξ, the order of FIR D and transmit block size Q. The Saleh-Valenzuela indoor channel model is used to generate the CIR as in [33] and the 4-QAM modulation is applied. Additionally, the overall transmit power is standardized through transmit antennas (p_0'=1), and the system performance of the proposed scheme is simulated and calculated according to the Monte Carlo simulation.

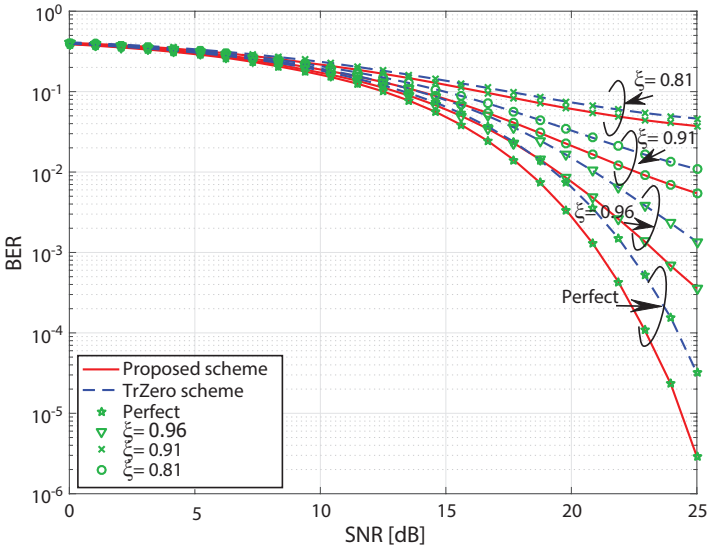


Figure 2: BER performance of the proposed scheme is compared with the TrZero scheme when D = 12 and Q = 26

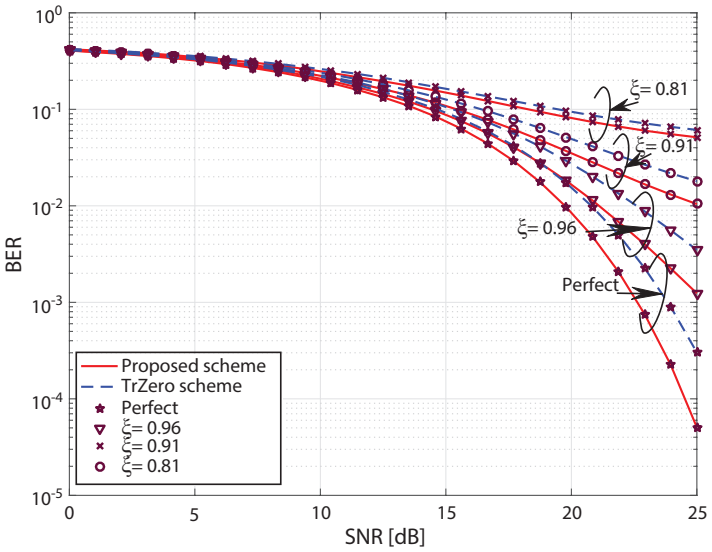


Figure 3: BER performance of the proposed scheme is compared with the TrZero scheme when D = 12 and Q = 30

On the other hand, the BER performance of the proposed and the TrZero schemes is compared in the different situations of the order of FIR and the transmission block dimension as illustrated in Figures 2, 3 and 4, respectively. In general, the BER of the TrZero scheme is significantly higher than that of the proposed one at all considered ξ values, such as ξ = 0.81, 0.91 and 0.96, and 1 (the perfect CSI). Furthermore, it is obvious that the accuracy of channel estimation proportionally affects the BER of the system, which means that the BER and the accuracy of channel estimation increase simultaneously, especially for the high SNR range. This is because of the fact that the residual interference due to imperfect CSI increases when the SNR increases. Consequently, the SINR of every sub-channel extremely decreases compared to the SNR of the perfect one.

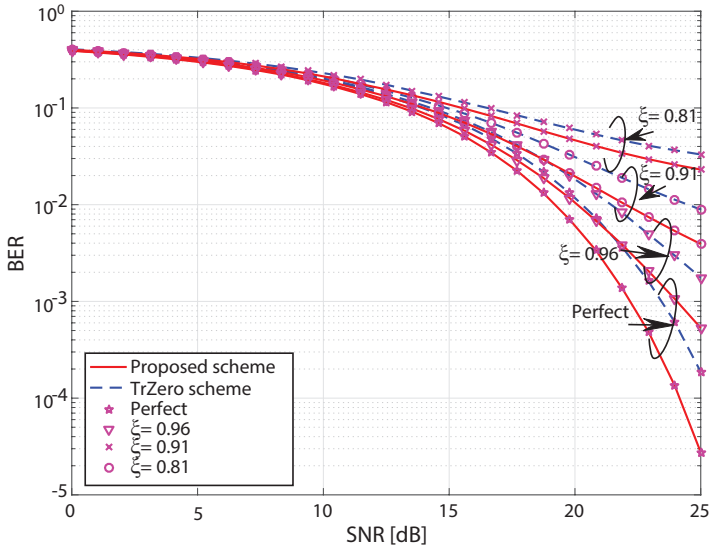


Figure 4: BER performance of the proposed scheme is compared with the TrZero scheme when D = 10 and Q = 26

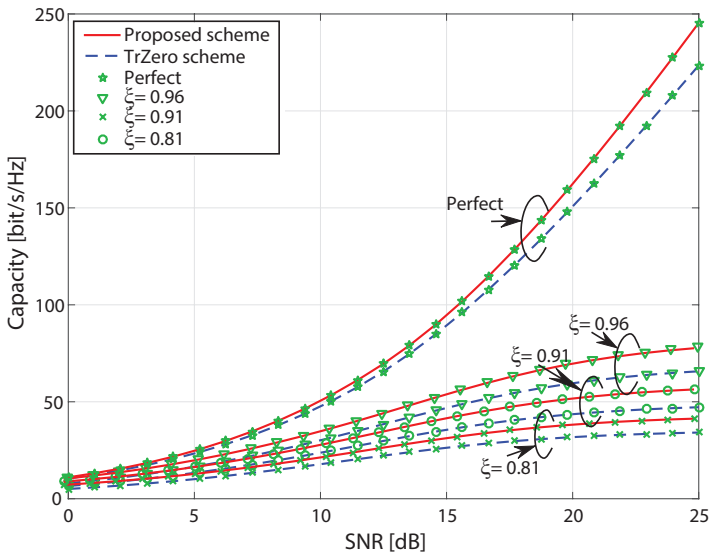


Figure 5: Comparison of system capacity in the proposed and TrZero schemes when D = 12 and Q = 26

As shown in Figures 2 and 3, the BER of the system is deteriorated when the Q increases, or the D decreases (Figures 2 and

4). It can be explained that the increase of Q makes the number of sub-channels increases while the shared redundancy is fixed. Thus, the channel energy for every sub-channel decreases. In contrast, when the D decreases, the shared redundancy decreases while the number of sub-channels is fixed, it leads the decrease in the channel energy for every sub-channel.

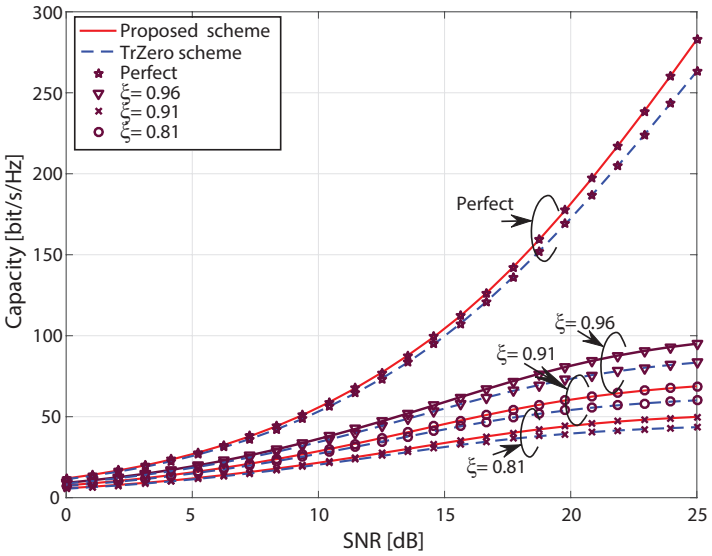


Figure 6: Comparison of system capacity in the proposed and TrZero schemes when $D = 12$ and $Q = 26$

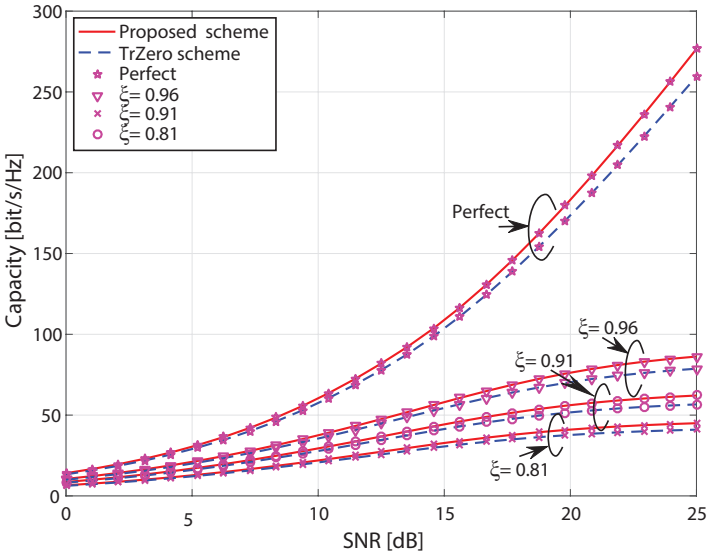


Figure 7: Comparison of system capacity in the proposed and TrZero schemes when $D = 10$ and $Q = 26$

Finally, the capacity of the proposed scheme is slightly higher than that of the TrZero scheme in all analyzed scenarios as in Figures 5, 6 and 7. Moreover, compared to the capacity of system with perfect CSI, the capacity of system with imperfect CSI is decreased dramatically. It is the same for both the TrZero and the proposed technical schemes. For more detail, at the 25 dB of SNR in Figure 5, when the accuracy of channel estimation ξ drops from 0.96 to 0.91, the system capacity witnesses a decrease from 80 to 55 bit/s/Hz. Furthermore, the capacity of system under the imperfect CSI condition at the receiver increases due to the increase of the transmission block size. On the other hand, the capacity of system decreases once the order of FIR increases in both schemes. The reason is that when the transmission block dimension increases and/or the FIR order decreases, the SINR of every sub-channel decreases, however the number of sub-channels in MIMO ISI channels is increased. As a result, the capacity of the MIMO ISI system under the imperfect CSI condition increases.

5 Conclusion

In this paper, the combination scheme of precoder and equalizer relied on the *unweighted* MMSE criterion for the imperfect CSI of MIMO ISI channels has been developed. The proposed scheme lets the loss of channel energy decrease and eliminates some sub-channels with extremely small eigenvalue in order to improve the system performance. The proposed scheme is evaluated by Monte Carlo simulation, and the simulation result shows that the proposed scheme under limited transmit power can achieve the higher system performance comparing to the conventional scheme in the considered scenarios even when the CSI is imperfect. However, the performance of the MIMO ISI system using redundancies is significantly affected by the CSI factor. Furthermore, the relationship between the performance of the system and several parameters, such as the accuracy of channel estimation, the transmission block size and the FIR order, are also discussed. The result in this research can help us to structure a practical MIMO ISI system which is close to the actual situation. The investigation of the impact of imperfect CSI on the advanced future communication system, such as massive MIMO or MIMO filterbank multicarrier systems, with combining designs of precoder and equalizer is left to future works.

References

[1] R. Vannithamby and S. Talwar, *New Physicallayer Waveforms for 5G*. Wiley, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8043580

[2] F. Corno, L. De Russis, and J. Pablo Senz, "On the advanced services that 5g may provide to iot applications," in *2018 IEEE 5G World Forum (5GWF)*, pp. 528–531, July 2018.

[3] B. Kwon, S. Kim, S. Lee, *Scattered Reference Symbol-Based Channel Estimation and Equalization for FBMC-QAM Systems*. *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3522 – 3537, 2017.

[4] J. Wang, Q. Yu ; Z. Li, C. Bi, "Distributed Space Time Block Transmission and QRD Based Diversity Detector in Asynchronous Cooperative Communications Systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 5111 – 5125, 2018.

[5] C. Dong, J. Lin, K. Niu, Z. He, and Z. Bie, "Block-iterative decision feedback equalizer with noise prediction for single-carrier MIMO transmission," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 8, pp. 3772–3776, 2012.

[6] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant filterbank precoders and equalizers. i. unification and optimal designs," *IEEE Transactions on Signal Processing*, vol. 47, no. 7, pp. 1988–2006, 1999.

[7] M.-W. Kwan and C.-W. Kok, "MMSE equalizer for MIMO-ISI channel with shorten guard period," *IEEE transactions on signal processing*, vol. 55, no. 1, pp. 389–395, 2007.

[8] W. A. Martins and P. S. R. Diniz, "Block-based transceivers with minimum redundancy," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1321–1333, 2010.

[9] Y.-P. Lin and S.-M. Phoong, "Minimum redundancy for ISI free FIR filterbank transceivers," *IEEE Transactions on Signal Processing*, vol. 50, no. 4, pp. 842–853, 2002.

[10] B. Q. Doanh, P. T. Hiep, T. C. Hieu *et al.*, "A combining design of precoder and equalizer based on shared redundancy to improve performance of ISI MIMO systems," *Wireless Networks*, pp. 1–10, 2019

[11] Y. Zhai, J. Tong, and J. Xi, "Precoder design for MIMO visible light communications with decision-feedback receivers," *IEEE Photonics Technology Letters*, 2019.

[12] Y. Cheng, L. G. Baltar, M. Haardt, and J. A. Nossek, "Precoder and equalizer design for multi-user MIMO FBMC/)QAM with highly frequency selective channels," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pp. 2429–2433, 2015.

[13] F. Rottenberg, X. Mestre, and J. Louveaux, "Optimal zero forcing precoder and decoder design for multi-user mimo FBMC under strong channel selectivity." in *ICASSP*, pp. 3541–3545, 2016.

[14] F. Rottenberg, X. Mestre, F. Horlin, and J. Louveaux, "Single-tap precoders and decoders for multiuser MIMO FBMC-OQAM under strong channel frequency selectivity," *IEEE transactions on signal processing*, vol. 65, no. 3, pp. 587–600, 2017.

[15] H. Sampath, P. Stoica, and A. Paulraj, "Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion," *IEEE Transactions on Communications*, vol. 49, no. 12, pp. 2198–2206, 2001.

[16] H. Yang, C. Chen, W.-D. Zhong, and A. Alphones, "Joint precoder and equalizer design for multi-user multi-cell mimo vlc systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 354–11 364, 2018.

[17] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1051–1064, 2002.

[18] N. Lee, O. Simeone, and J. Kang, "The effect of imperfect channel knowledge on a MIMO system with interference," *IEEE Transactions on Communications*, vol. 60, no. 8, pp. 2221–2229, August 2012.

[19] D. Le Ruyet, R. Zakaria, and B. Özbek, "On precoding MIMO-FBMC with imperfect channel state information at the transmitter," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*. IEEE, pp. 808–812, 2014.

[20] I. Dvorakova, A. Malyutin, and Y. Nechaev, "MMSE precoder for multipath channels with imperfectly known state information," in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 195–199, July 2015.

[21] M. Huang, S. Zhou, and J. Wang, "Analysis of TomlinsonHarashima precoding in multiuser mimo systems with imperfect channel state information," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 5, pp. 2856–2867, Sep. 2008.

[22] H. K. Bizaki and A. Falahati, "Tomlinson-Harashima precoding with imperfect channel state information," *IET Communications*, vol. 2, no. 1, pp. 151–158, January 2008.

[23] J. Baltersee, G. Fock, and H. Meyr, "Achievable rate of MIMO channels with data-aided channel estimation and perfect interleaving," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 12, pp. 2358–2368, Dec 2001.

[24] A. Maaref and S. Aïssa, "Combined adaptive modulation and truncated arq for packet data transmission in mimo systems," in *Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE*, vol. 6. IEEE, pp. 3818–3822, 2004.

[25] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, April 2003.

[26] A. Goldsmith, "Capacity and power allocation for fading mimo channels with channel estimation error," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[27] B. Q. Doanh, P. T. Hiep, and T. C. Hieu, "Impact of imperfect CSI on capacity of ISI MIMO systems based on joint precoding and equalization designs," in *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, pp. 334–338, 2019.

[28] M. Ding and S. D. Blostein, "MIMO minimum total MSE transceiver design with imperfect csi at both ends," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1141–1150, March 2009.

[29] B. S. Thian, S. Zhou, and A. Goldsmith, "Transceiver design for MIMO systems with imperfect csi at transmitter and receiver," in *2011 IEEE International Conference on Communications (ICC)*, pp. 1–6, June 2011.

[30] Z. Chang, Z. Wang, X. Guo, Z. Han, and T. Ristaniemi, "Energy-efficient resource allocation for wireless powered massive MIMO system with imperfect CSI," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 2, pp. 121–130, June 2017.

[31] P. Aquilina and T. Ratnarajah, "Linear interference alignment in full-duplex MIMO networks with imperfect CSI," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5226–5243, Dec 2017.

[32] J. Cui, Z. Ding, and P. Fan, "Outage probability constrained MIMO-NOMA designs under imperfect CSI," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8239–8255, Dec 2018.

[33] A. A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on selected areas in communications*, vol. 5, no. 2, pp. 128–137, 1987.

[34] S. Serbetli and A. Yener, "MMSE transmitter design for correlated MIMO systems with imperfect channel estimates: power allocation trade-offs," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2295–2304, Aug 2006.

[35] L. Musavian, M. R. Nakhai, M. Dohler, and A. H. Aghvami, "Effect of channel uncertainty on the mutual information of mimo fading channels," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 2798–2806, Sep. 2007.

[36] H. Lütkepohl, "Handbook of matrices. wiley," *New York*, 1996.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 143-147 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Electro-tactile Stimulation for Augmenting Finger Motoric Learning

Daniel Sutopo Pamungkas[*,1], Arjon Turnip[2]

[1]*Mechatronics Department, Politeknik Negeri Batam, 29432, Indonesia*

[2]*Department of Electrical Engineering, Universitas Padjadjaran, 45311, Indonesia*

ARTICLE INFO

Article history:
Received: 28 July, 2020
Accepted: 30 October, 2020
Online: 08 November, 2020

Keywords:
Finger training
Haptic feedback
Electro-tactile feedback
Piano learning

ABSTRACT

Finger skills are achieved through the training process. This training process will be more straightforward if a training system is used. A novel finger motoric training system is introduced in this paper. This system is to facilitate a beginner piano player to learn the basic piano skill. The system is comprised of a tracking finger sensor, a personal computer, and multimedia feedback. Subjects train with passive and active mode. The passive mode, the subject's fingers are doing nothing when the information is given, whereas the subjects are playing a virtual piano under the sensor. While the information which fingers have to tap are provided by visual, audio, and enhanced with electrical haptic sensation. The experimental results demonstrate that the system equipped with electro-tactile feedback makes the subjects more responsive to both modes. Moreover, this system enabled the user to accelerate learning a new skill.

1. Introduction

Motoric skill learning can be trained with repetition, also utilizing stimulants from the surrounding environment [1]. Such motoric skills, for example, are typing, writing, and playing a musical instrument. To acquire dexterous motoric skills, practice with repetitive movements is needed. Moreover, these practices need persistence from the instructors and the learner. On the other hand, a stimulus can be used to augment the learning process.

Visual and audio information are commonly used as stimulants. Learners are showed how to do the skill accurately. In this process, they are played, or listened to, the appropriate sound. A haptic sensation can be applied to the learner using mechanical interaction. Some researchers utilize these stimuli to enhance the method to train the motoric skills, for example, playing the piano [1, 2], writing Japanese characters [3], writing Arabic and Japanese characters [4], and typing the Braille alphabet [5]. However, these systems are cumbersome, and the user finds it difficult to move their fingers.

To overcome these shortcomings, electrical stimulus is used. This information is given to the skin of the user. The electro-tactile feedback system is proven to become feedback used for controlling mobile [6], arm robot [7] also fingers [8], information for haptic sensors [9], enhanced virtual game [10]. To examine the potential usage of electrical signal, a multimedia learning system for a motoric finger is established. This paper is an extension of work initially presented in the 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC) [11]

For the experiment, subjects are divided into three groups. They are trained using different methods. Information in the form of audio and visual means are given to the first group to train the finger movements of each subject. In the other groups, the subjects utilized audiovisual means, enhanced with electrical sensations to practice. These groups consist of subjects which do the experiments with active mode and subjects with passive mode. The results show that the group using an electro-tactile feedback system is capable of playing the music appropriately with both active and passive modes. Moreover, they learn more rapidly than the other group. Besides that, the proposed system is simple and easy to use. The haptic sensation also does not cause pain to the subject.

This paper is organized as follows: Section II delivers a brief overview of related works within this field of study. Section III presents the details of system implementation. Section IV describes the experiment, which demonstrates the effectiveness of

[*]Corresponding Author: Daniel S Pamungkas, Politeknik Negeri Batam, daniel@polibatam.ac.id

the system for finger motoric learning. Finally, Section V delivers the concluding remarks.

2. Background

The process or action to achieve an ability to do, or to gain knowledge, is known as a learning process. The actions done in this process require stimulation from the surroundings. Moreover, repetition is needed in the learning process [12]. In this process, one thing can be connected to others, or known as associate learning. Stimulus or information is given repeatedly, resulting in a reaction, until the reaction can be an automatic response and become habitual. This stimulus can be associated with a reaction. Pavlov shows this condition with his experiments, in which the subject dogs have a new behavior. The dogs are able to associate the sound of a metronome with food [13]. Some physical training can be involved in associate learning, such as finger movements.

The skilled fingers motions are involved for a specific action, such as typing a computer keyboard or playing the piano. These finger skills are challenging to learn using only verbal instructions. Such training will be effective and efficient if the learner does a lot of exercises. Some software have been developed to assist people to learn to type in a keyboard or to play the piano. These software are intended to support the user in typing or playing the piano with all fingers. Visual stimulation is used by these software, for example, colors or lights. Other programs utilize audio information, e.g., the sound of the notes. Moreover, some of the applications apply both audio and visual information. However, another sense has excellent potential to transmit data, namely the sense of touch or tactile taste, which uses the skin to deliver stimulation.

Human skin is sensitive to several sensations, for example, vibration, heat, force, and electric signal (voltage or current). The skin has some sensory receptors, namely mechanoreceptors, thermoreceptors, nociceptors, and proprioceptors. These receptors are to accommodate all the sensations. Moreover, these receptors also enable humans to feel texture, paint, etc. [a]. Tactile sense causes humans to have the ability to detect and classify objects through their skin. This sense is sometimes called haptic sensations.

The term *haptic* comes from the word 'hapto' from Greek. This word means ‘to touch and handle objects’ [14]. Several researchers introduce haptic information utilizing vibration [15], heat [16], force [17], and electric sensations [6]. These information aims to enrich the user to interpret or interact with the application more effectively.

Haptic feedback has been used widely in various applications. A simulator to learn how to perform a surgery [17], remotely controlling a robot’s actions [18], a tactile sensor to sense the roughness of a surface [9] and gaming experience using VR [10] are the examples of the application of this feedback. Also, there are applications for training fingers skills using haptic feedback. For instance, haptic feedback is used for learning to play the piano [1,2], and using a Braille keyboard [5]. Moreover, this feedback is combined with other feedback, namely visual and audio, for writing characters [3] [4].

One of the rehabilitation procedures for a person with spinal cord injury (SCI) is to learn how to play the piano. For this training, subjects are equipped with a vibrating glove. This glove is equipped with a vibration motor on every finger. The vibrations help the subject to know which finger has to strike the keys of the piano [1]. The subjects are required to move their fingers while training. Such a method is called an active mode. Markow et al. report that this system enables the subject to play an easy song.

Others report that a passive mode training using vibration cues has positive results. The passive mode method is the method in which the subject trains without actually doing the activity about which they learn. The learning process is given using their senses. In [2] the subjects are trained to play a song with a piano using a vibration glove. The audio and haptic feedbacks are given to the subjects while they are doing the activity. Subjects are listening to the song, while the vibrations in their fingers give information which finger has to move related to the song. These experiments show that this information, including haptic feedback, enables the user to play the piano reflexively.

In [5], active and passive modes are used to train blind subjects to type in a Braille keyboard. This system has a keyboard that has the capability to train the user to use audio and vibration. The vibration information is given to the user's fingers. The subjects are able to use both hands to learn using this system. In these studies, the experiments are done with active and passive modes. The authors argue that both methods using audio and haptic information can be used to learn to type.

Another haptic information is force feedback. One of the devices able to give kinesthetic feedback to the user is called The Phantom. This device is used to train the learner to write Japanese [3] and Arabic [4] characters. This device helps the hand of the user to write the characters correctly. Both studies show that the force feedback is able to assist the subjects in training to write several newly-learned characters efficiently.

Mechanical feedback, such as vibration and force feedback, is able to transmit information for finger skill training. Haptic feedback has a great potential to be a training method, however it is cumbersome and costly. These devices hinder the movement of users in typing on a keyboard or playing the piano.

One type of haptic feedback is electro-tactile feedback. This applies an electric current to forward information to the skin. This feedback gives sensation to the skin and directly to the nerves. Electrotactile feedback system is simple and comfortable to use. Data from electric sensations can be informed through differences in amplitude or frequency. Thus, electrotactile feedback has various ranges of information that can be transmitted to the user [6]. In the next section, detailed methods to demonstrate the potential of electrotactile feedback are presented.

3. System

3.1. Overview

The proposed system for finger training is a computer with a screen and sound system, in addition to a gesture sensor, also with an electro-tactile feedback system. The electro-tactile feedback has different modes to vary the frequency, amplitude and the width of the electrical pulse. These variations have to be adjust to the subjects wearing this device. This system is comfortable to wear and does not hinder the user’s movements [6-11].

The frequency, amplitude and pulse width have a wide range of options, the frequency was varied between 10Hz to 120Hz. For this experiment stimulus pulses are set at 20Hz frequency (Figure 7). The amplitude of the pulses is adjusted by the user between 40-80V to achieve a comfortable condition. The intensity is controlled by the pulse width. The width of signal is varied, particularly between 10 to 100μs.

This system can be seen in Figure 1. Figure 2 shows the diagram block of the system. The screen of the computer displays the character of the note and which finger(s) to bend, while the corresponding frequency of the note is sounded using the computer's sound system. The electrodes in the fingers of the subject are transmitting the electric sensation. These sensations are given to the fingers appropriate to the note. The gesture sensor is used for tracking the fingers of the subjects. This sensor is placed near the top of the screen facing down to detect the movement of the subject fingers.

The learning is done by a passive and active method. The audio, visual, and haptic information are given to the subject while learning in the passive method, whereas the hand tracking device with the feedback is used when active learning is performed. In the test period, only a hand gesture sensor is used with the computer.

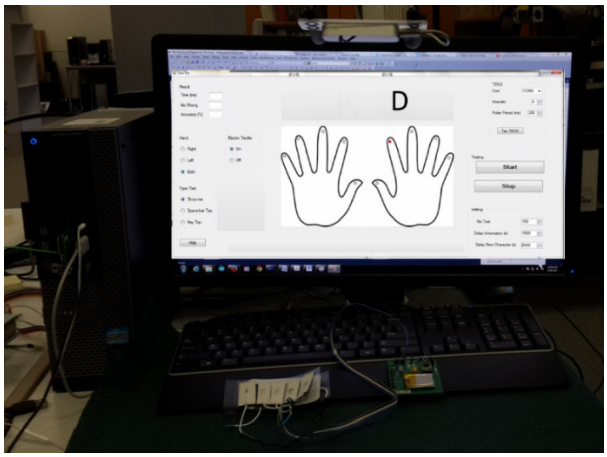


Figure 1: Arrangement of the system

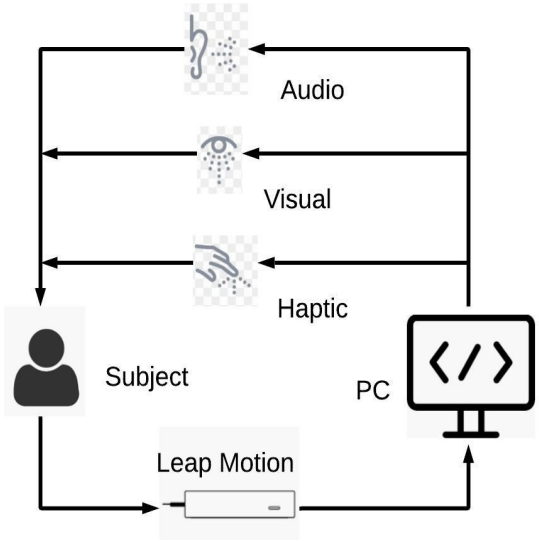


Figure 2: The Diagram block of the system

3.2. Hand tracking and software

One of the main components of the proposed system is the hand and fingers tracking sensor. For these tasks, a leap motion sensor is used. Leap motion is able to detect fingers and hands accurately from around one meter in three dimensions. The accuracy of this sensor is around 0.1 mm. This device is equipped with two infra-red cameras and three infrared LEDs [19]. This sensor has two modes. The first mode faces upward; the hands are above the leap motion sensor, known as a desktop mode. In the second mode, the HMD, the leap motion faces downward. HMD mode is used for this system. A custom bracket is utilized for placing the sensors. In this position, it is possible to have noise from reflective surfaces, so a dark mat is placed to cover the desk. The mat is shown in Figure 1.

The finger training multi senses system has several modes to train the subjects. The first mode is training using three senses; visual, audio, and haptic. The second mode uses visual and audio information, and the third uses audio and haptic information. The first and the second modes will be used to train the subject with the active method, while the third one is used to train with the passive method. The active mode requires subjects to tap the fingers below the gesture sensor.

The software is able to detect the correct or the wrong fingers. The function of this feature is to test the result of the training. The GUI of the system can be seen in figure 3. Moreover, the system is able to train the subject to play piano with only the right hand/notes only, or with both hands (notes and simple chord).

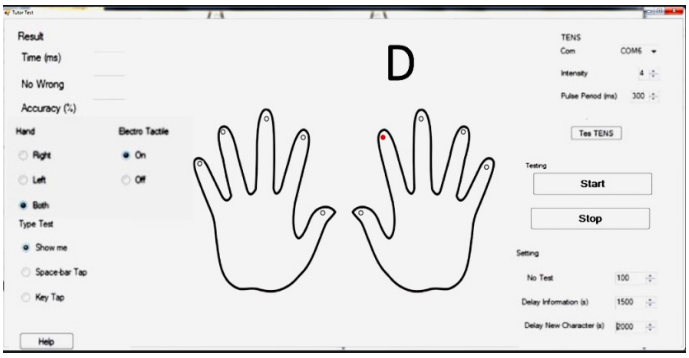


Figure 3: GUI of the system

The software is developed to display the characters of the note. Also, this software shows the appropriate fingers that have to bend. Figure 4 shows the fingers as appropriate to the note in a piano for the left hand (Figure 4.a) and right hand (Figure.4.b). The left-hand fingers of the human are placed in the a lower octave than the right hand. The left hand is trained to press the fundamental note of the chord. For example, if we want to play the C chord (C-E-G) the subjects are trained to press C, and so on such as in Figure 4.a. The left hand and the corresponding notes of the chord are shown in Figure 4.b. The right hand of the subject is trained to play the notes from C to G. Using this arrangement, this software is able to instruct the user to perform simple music. The user can play with only the right hand (melody only) or with both hands (melody with simple chord).

This software also has the capability to produce sound of note. This voice corresponds with the finger as defined in Figure 3. The

output of this information is transmitted to the computer sound system. Moreover, the haptic feedback augments the information for the user. Electric signal is transmitted to the nerve using electrodes patch. Fig 5 shows the illustration of the back of human palms. All the human fingers of both hands are placed the electrode. The software has facility to send electrical sensations appropriate to the visual and audio information. The next section will provide the detailed explanation of the electro-tactile feedback system.

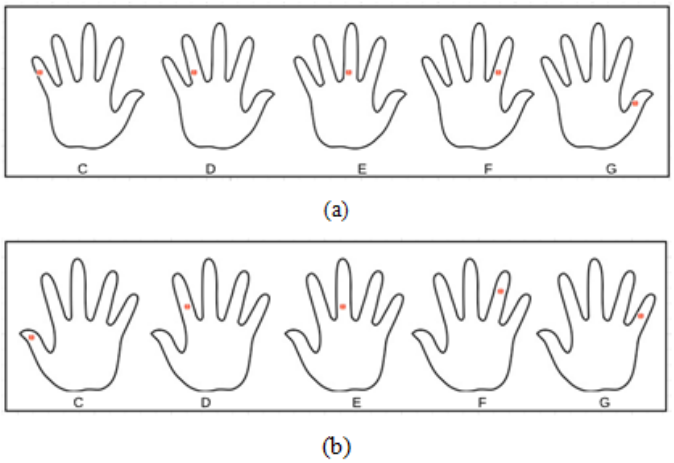


Figure 4: The corresponding fingers with the notes (a) left hand (b) right hand.

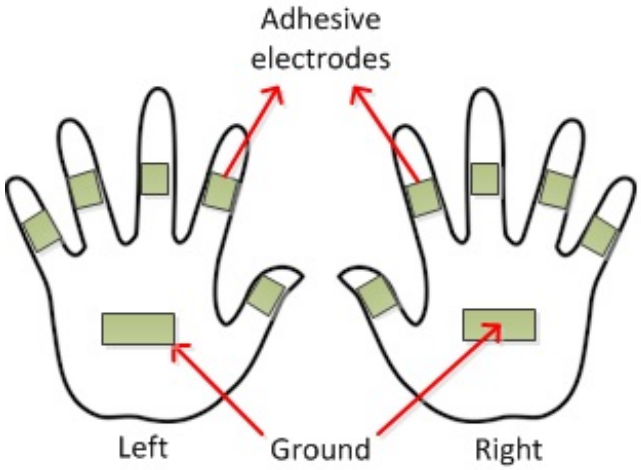


Figure 5: The electrodes in hands of the subject

3.3. Haptic system

The electro-tactile feedback system, which we call TENS (Transcutaneous Electrical Nerve Stimulation) system, comprises of three main components, namely transmitter, receiver and electrodes. The first component is the transmitter; the function of which is to transmit the signal from the computer to the receiver wirelessly. The main element of this component is a Bluetooth transmitter. The receiver receives the information from the transmitter and transformed it into electrical pulses. The five channels receivers with one ground are used. For the experiments, two transmitters are used for both hands. This structure enables the subject to receive ten signal frequencies. The last component is the electrode, placed in the knuckles, and a ground is placed in the middle of each hand. This arrangement allows the nerves of the subjects to obtain the information. The electric signal has frequencies that vary in the range of 10Hz to 120Hz. The width of the pulse varied between 10μs to 100μs. Moreover, the width of the pulse is a factor in controlling the intensity of the signal. The signal has an amplitude range between 40V until 80V. Those features are able to be adjusted. This adjustable function is to attain the subjects' comfort. Every subject has their own preferences. In the following section, the experiment with the results will be shown.

4. Experiment Methods and Results

This learning system is to train the subjects with passive and active mode at improving the basic skills of playing the piano. These trials were directed mainly to confirm the potential of the proposed system using audio and visual information and augmented with electrotactile feedback system. This objective can be done by comparing the experiment with a training system using only audio and visual feedback.

To do the experiment, nine subjects are divided into three groups. All the subjects are never trained to play the piano. They also have never used the electrotactile feedback. The subject's ages vary between 15 until 21 years old. The subjects are comprised of five males and four females. The first group trains using the active mode with audio, visual and haptic information. The second group practices with the active mode, only using audio and visual feedback. The passive mode, with audio and haptic information, is given to the third group. The subjects doing the active mode have to tap their fingers as appropriate as the given information.

Before the trials, all the subjects are given basic music theory, focusing on the use of appropriate fingers and hands for specific note or chords. For the first five minutes, they are trained to use the system. The aim of this training is to enable the subjects play with the virtual piano. Subjects were trained to tap their fingers according to the notes that appear on the screen until they become used to use their fingers under the sensors. This is followed with an initial test. In this state all the subjects tried to play a simple song. The system only shows the letter of the notes that they have to tap, and the errors are counted. Errors include the use of wrong fingers or missed notes.

After the initial test, for the groups where haptic feedback is used, the intensity of the electric signal is adjusted for each subject. This is for the comfort of the subject, also to adapt with the signal. This is then followed with five minutes training of the subjects with the method in their groups. In this time all the subjects are given information about which fingers must be bent in accordance with the letters of the notes shown in the screen. Then, in the last five minutes, all the subjects tried to play music assisted with the feedback and the letters of the notes in the screen. After a short break, the final test is conducted to prove the effectiveness of the haptic feedback. This task is the same with the initial task. The errors are calculated and the percentage of improvement of each subject can be counted.

Based on the experiment, on average, the first group, who learned to bend the fingers utilizing visual and audio information, and enriched with electrical sensation, on average achieved an improvement of around 40%. On the other hand, the second group, who learned through the active method using visual and audio

feedback, but without haptic feedback, achieved 20% improvement. While the last group, which used the passive method, had about 35% improvements. Table 1 shows the number of errors for each subject before and after being trained in the method they use.

Analysis of the outcome of the experiment shows that all the training methods offers positive results. All the methods are able to improve the musical skills of the subjects. However, the methods which employ the haptic feedback show a more significant performance improvement compared to the method not involving haptic feedback. This can be concluded that the systems that involve haptic information are more effective. Moreover, subjects who use electric stimulation feel that the feedback is not painful and easy to use.

Table 1: Error before and after training

Active Electro Tactile Feedback			**Active Without Electro Tactile Feedback**			**Passive With Electrotactile feedback**		
Subject No	**No Errors**		**Subject No**	**No Errors**		**Subject No**	**No Errors**	
	Before	After		Before	After		Before	After
1	25	15	4	28	22	7	30	19
2	27	16	5	27	21	8	23	15
3	22	13	6	26	20	9	24	15

5. Conclusion

This paper introduces a novel system to train fingers to play the piano using multimedia information. The audio, video and haptic feedback is used to teach. A computer's speaker is used to provide audio information. The screen of the computer is used to show visual feedback, while haptic information is transmitted utilizing electrical sensation. This proposed system employs a gesture sensor to become a virtual piano, and a custom software is built to accommodate all the inputs and outputs of the system.

The results of the tests indicate that the system involving electro-tactile feedback using active and passive training methods have positive results. The electro-tactile feedback system enables to train the fingers of the users to move in association with the information reflexively. Moreover, this system augments to train users the new finger skill better than the method only using audio and visual. The improvement of the subjects who use haptic feedback is superior to the subjects who utilize the conventional one. Electro-tactile feedback is inexpensive, convenient to use, simple to wear, and does not hinder the user from moving their fingers.

This system has shown its potential to enhance finger training devices. It is desirable to conduct further research with more subjects, as well as a comparison with existing training systems to measure the performance of this proposed system.

Acknowledgment

This work is financially supported by Penelitian Terapan Unggulan Perguruan tinggi research fund at Indonesian Ministry of Education. Also acknowledgment to Politeknik Negeri Batam for the laboratory facilities.

References

[1] T. Markow, N. Ramakrishnan, K. Huang, T. Starner, M. Eicholtz, S. Garrett, H. Profita, A. Scarlata, C. Schooler, A. Tarun, D. Backus, "Mobile music touch: Vibration stimulus in hand rehabilitation," in 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, Pervasive Health 2010, IEEE Computer Society, 2010, doi:10.4108/pervasivehealth.2010.8.

[2] D. Kohlsdorf, T. Starner, "Mobile music touch: The effect of primary tasks on passively learning piano sequences," in Proceedings - International Symposium on Wearable Computers, ISWC, 2010, doi:10.1109/ISWC.2010.5665877.

[3] M.A. Eid, M. Mansour, A.H. El Saddik, R. Iglesias, "A haptic multimedia handwriting learning system," in Proceedings of the ACM International Multimedia Conference and Exhibition, ACM Press, New York, New York, USA: 103–108, 2007, doi:10.1145/1290144.1290161.

[4] J. Bluteau, S. Coquillart, Y. Payan, E. Gentaz, "Haptic Guidance Improves the Visuo-Manual Tracking of Trajectories," PLoS ONE, **3**(3), e1775, 2008, doi:10.1371/journal.pone.0001775.

[5] C. Seim, D. Quigley, T. Starner, "Passive haptic learning of typing skills facilitated by wearable computers," in Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, New York, New York, USA: 2203–2208, 2014, doi:10.1145/2559206.2581329.

[6] D.S. Pamungkas, K. Ward, "Electro-tactile feedback system for achieving embodiment in a tele-operated robot," in 2014 13th International Conference on Control Automation Robotics and Vision, ICARCV 2014, 2014, doi:10.1109/ICARCV.2014.7064573.

[7] D.S. Pamungkas, K. Ward, "Tele-operation of a robot arm with electro tactile feedback," in 2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics: Mechatronics for Human Wellbeing, AIM 2013, 2013, doi:10.1109/AIM.2013.6584175.

[8] D. Pamungkas, K. Ward, "Electro-tactile feedback system for a prosthetic hand," in 22nd Annual International Conference on Mechatronics and Machine Vision in Practice, M2VIP 2015, 2015.

[9] D. Pamungkas, K. Ward, "Tactile sensing system using electro-tactile feedback," in ICARA 2015 - Proceedings of the 2015 6th International Conference on Automation, Robotics and Applications, 2015, doi:10.1109/ICARA.2015.7081163.

[10] D.S. Pamungkas, K. Ward, "Electro-Tactile Feedback System to Enhance Virtual Reality Experience," International Journal of Computer Theory and Engineering, **8**(6), 465–470, 2016, doi:10.7763/ijcte.2016.v8.1090.

[11] D.S. Pamungkas, A. Turnip, "Electro-Tactile Cues for a Haptic Multimedia Finger Motoric Learning System," in ICSECC 2019 - International Conference on Sustainable Engineering and Creative Computing: New Idea, New Innovation, Proceedings, Institute of Electrical and Electronics Engineers Inc.: 127–132, 2019, doi:10.1109/ICSECC.2019.8906989.

[12] Principles of Neural Science, Fifth Edition (Principles of Neural Science (Kandel)): 9780071390118: Medicine & Health Science Books @ Amazon.com.

[13] I.P. Pavlov, "Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex," Annals of Neurosciences, **17**(3), 2010, doi:10.5214/ans.0972-7531.1017309.

[14] M. Mihelj, D. Novak, S. Beguš, Introduction to Virtual Reality, Kluwer Academic Publishers, 2014, doi:10.1007/978-94-007-6910-6_1.

[15] Y. Yoon, D. Moon, S. Chin, "Fine Tactile Representation of Materials for Virtual Reality," Journal of Sensors, **2020**, 7296204, 2020, doi:10.1155/2020/7296204.

[16] R.L. Peiris, Y.L. Feng, L. Chan, K. Minamizawa, "Thermalbracelet: Exploring thermal haptic feedback around the wrist," in Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, New York, New York, USA: 1–11, 2019, doi:10.1145/3290605.3300400.

[17] K. Rangarajan, H. Davis, P.H. Pucher, "Systematic Review of Virtual Haptics in Surgical Simulation: A Valid Educational Tool?," Journal of Surgical Education, **77**(2), 337–347, 2020, doi:10.1016/j.jsurg.2019.09.006.

[18] A. Abiri, J. Pensa, A. Tao, J. Ma, Y.Y. Juo, S.J. Askari, J. Bisley, J. Rosen, E.P. Dutson, W.S. Grundfest, "Multi-Modal Haptic Feedback for Grip Force Reduction in Robotic Surgery," Scientific Reports, **9**(1), 1–10, 2019, doi:10.1038/s41598-019-40821-1.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 1, 146-152 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Applications of TCAD Simulation Software for Fabrication and study of Process Variation Effects on Threshold Voltage in 180nm Floating-Gate Device

Thinh Dang Cong[1,2], Toi Le Thanh[1,2], Hao Mai Tri[1,2], Phuc Ton That Bao[1,2], Trang Hoang[1,2 *]

[1]*Department of Electronics Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, 72506, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City, 71308, Vietnam*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 23 December, 2020
Online: 15 January, 2021

Keywords:
Floating-gate device
Non-volatile memory
Channel Hot Electron Injection
Fowler-Nordheim Tunnel
CMOS Process
Process variation
TCAD

ABSTRACT

In this work, a study of the process variation effects on the threshold voltage of a floating-gate device is proposed. The study demonstrates the sensitivity of the threshold voltage to five geometrical parameters including gate length, gate width, tunneling gate oxide thickness, bottom oxide-nitride-oxide oxide thickness, and nitride spacer thickness. This paper also proposed a detailed flow to fabricate the floating-gate device for CMOS 180nm process, which is used to design the floating-gate device for the study. This paper used the TCAD tools including Athena, Devedit3D, and Atlas for the simulations.

1 Introduction

The CMOS technology has been developing over the past many years thanks to its benefits like high integration density, low fabrication costs, and high operating speeds. In addition, many novel techniques of circuit design have been shown ranging from memories, analog mixed-signal, RFIC [1] -[7] to mm-Wave IC design [8], [9].

Meanwhile, there has been considerable development in the semiconductor memory field. This field is divided into two main branches which are volatile memories and non-volatile memories, and these types of memories are developing based on the CMOS technology [10], [11]. While the former lose data when the power supply is off, the latter can retain the stored information even after power is removed. Therefore, non-volatile memories are used in a very wide variety of products like cell phones, computers, and communication [12]. Thus, many researchers have focused on the study of this kind of memory. However, to study the memories, the floating gate device needs to be investigated first since it is the core of the memories [13] –[15].

Focusing on designing a simulation model for the floating gate device, many parameters are studied under the process variation. There are two main methods to investigate the process variation effects. The first way is using the SPICE model. This is basically a fast method but the approach of this method lacks many conditions in the fabrication process like doping channel [16]. Therefore, the effects of the doping channel would not be studied. The other method is using the 3D-TCAD simulation for analyzing the geometrical parameters and the doping channel effects [17] -[19].

In this work, a detailed design flow and parameter values to fabricate the complete floating-gate device are proposed to investigate the structure and characteristics of the device. The process variation effects on the threshold voltage, which is a vital parameter of the device, are studied by using the TCAD simulation tools. Authors introduced the floating-gate device and presented ideas about how to study the process variation on threshold voltage in 180nm

*Corresponding Author: Trang Hoang, Department of Electronics Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, 72506, Vietnam, Tel: (+84) 988 071 579 & Email: hoangtrang@hcmut.edu.vn

floating-gate device in the "2019 19th International Symposium on Communications and Information Technologies (ISCIT)" [1] and the following paper is an extension of this work.

The remainder of this paper is organized as follows: Section II introduces an overview structure of the device in fabrication and the operation states also. Following this, the detailed flow and parameter values to fabricate a complete device are shown in section III. Next, section IV shows the analysis and discussion of the process variation results in the threshold voltage. This section also describes the simulation of operation states after fabricating. Finally, section V shows the conclusion.

2 Floating Gate Device

The floating-gate device is the core of almost modern non-volatile memories [20], [21]. With regards to fabrication, the device is recognized through the use of a "floating gate". The gate is completely surrounded by dielectrics [22]. A cross section of a device is given in Figure. 1. Being electrically isolated, the floating gate acts as the storing layer in the device. Therefore, the operation states of the device would be determined based on the status of the gate. More importantly, the insulator around the floating gate must be thick enough to prevent leakage of charges from the floating gate to the substrate when the power is off, and it also must be thin enough to allow the transfer of charges on and off the gate under appropriate bias configuration. Moreover, reducing the thickness of the insulator, especially the tunneling gate oxide thickness has been investigated to reduce the program/erase voltage and enhance the level of integration [13],[23].

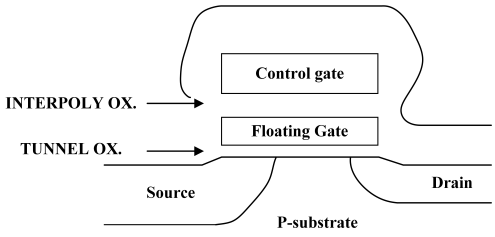


Figure 1: Floating-gate device structure in fabrication

The floating-gate device has three operation states which are programming, erasing, and reading. Programming refers to putting electrons called Channel Hot Electrons on the floating gate. In contrast, removing the electrons from the gate to the substrate would be performed in the erasing process. The removing current is renowned for the Fowler-Nordheim Tunnel current. Reading is used to determine if the device is programmed or erased.

3 Floating Gate Fabrication & Simulation

This flow proposes using an additional Epitaxial grow layer which was not used in traditional CMOS processes. The growth of an epitaxial layer over the P-type substrate offers some advantages including improving the performance of this device as well as floating-gate integrated circuits, minimizing latch-up effects that a CMOS circuit may undergo when powered up, and helping control the doping concentration of this device accurately [24].

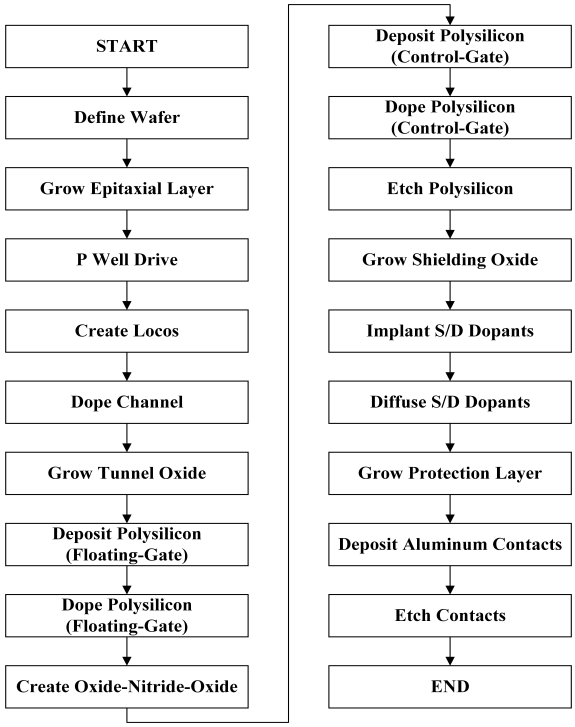
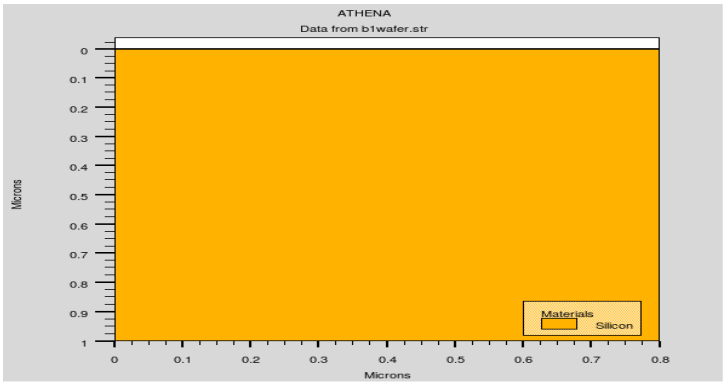


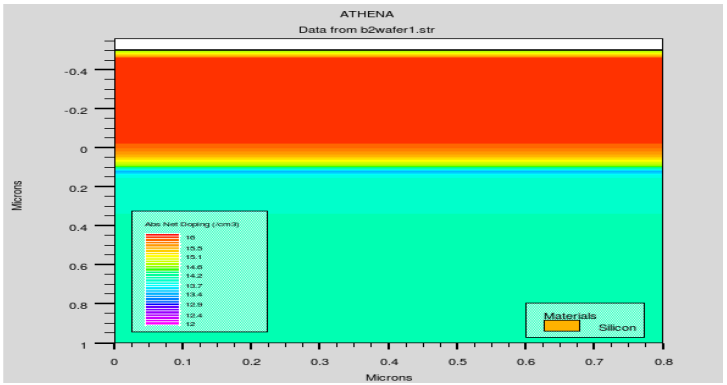
Figure 2: Floating-gate device design flow chart

The main difference in terms of structure between the floating-gate device and CMOS device is that the former has an additional floating gate which is created by two steps (Deposit Polysilicon (Floating gate), Dope Polysilicon (Floating gate) in the flow. These figures below illustrate the main process simulations using TCAD based on the design flow chart presented above (Figure 3a-3h).

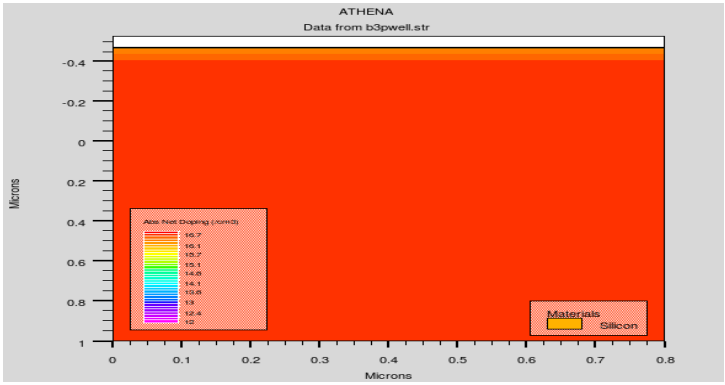
The fabrication process starts from creating P-substrate (initial surface thickness = 1μm) with Boron at a concentration of 1.0e14 cm^{-3} (Figure 3a). Next, a mesh is defined, and the density of the mesh is a trade-off between accuracy and simulation time. After the mesh and wafer definition, a 0.45μm thick Epitaxy layer with Arsenic at a concentration of 1.0e16 cm^{-3} is grown on the top to make device surface thickness increase to 1.45μm (Figure 3b). Then, a P well is implanted using Boron with a dose of 8e12 cm^{-3}. After that, the P well is also diffused with Nitro gas at 1200^0C for about 310 minutes (Figure 3c). Next, Locos and tunnel oxide layer are created (Figure 3d). Creating Locos is an important step in the fabrication of semiconductor devices for the purpose of isolating the operation of two devices on the same wafer, and Oxide is usually used for this isolation. The doping channel is created using Boron at 100KeV and a concentration of 2.5e12 cm^{-3} (Figure 3e). Figure 3f shows the device structure after Polysilicon (Floating gate) is deposited and doped. Figure 3g shows the device structure after Polysilicon (Control gate) is deposited and doped. The Floating gate, Control gate, Tunnel Oxide layer and Oxide-Nitride-Oxide layer which is between Floating gate and Control gate are etched. An oxide layer with a thickness of 0.1μm is deposited and etched on the top to protect the device by the next steps. The next step is to create Source and Drain gates with Arsenic at 50 KeV and a concentration of



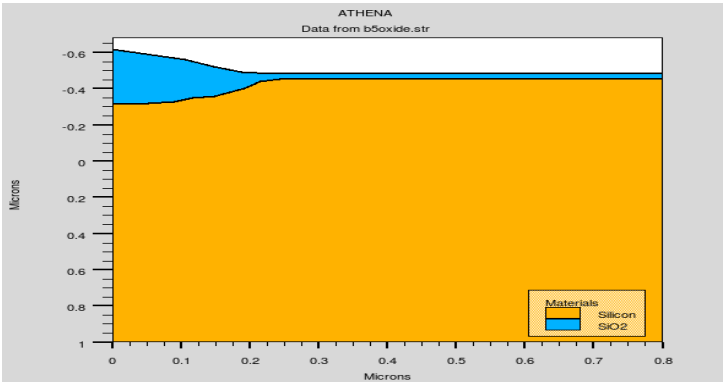
(a)



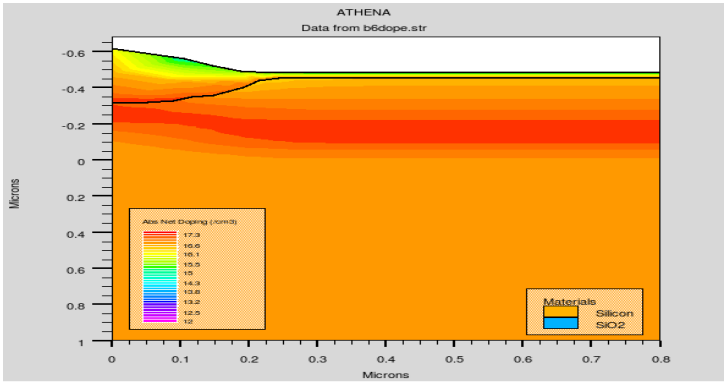
(b)



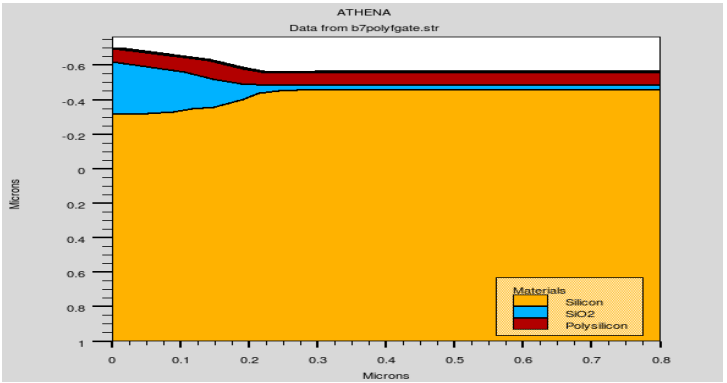
(c)



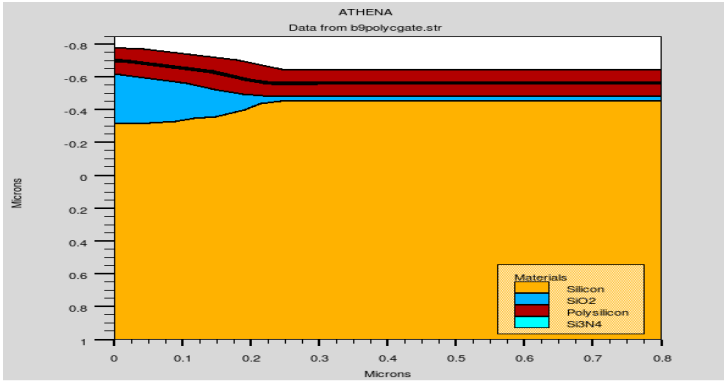
(d)



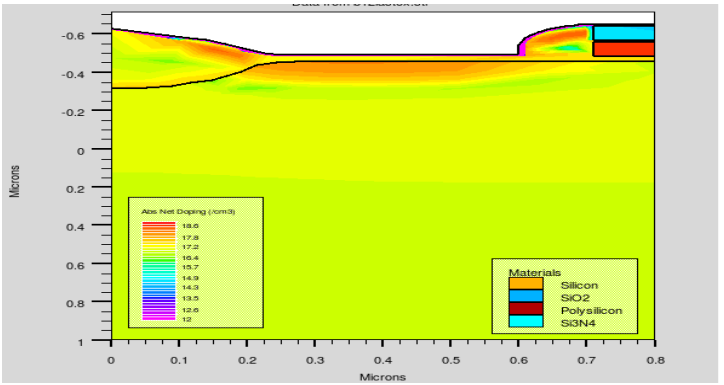
(e)



(f)



(g)



(h)

Figure 3: The main process simulations with TCAD including in a) Definition wafer b) Grow Epitaxial layer c) P well drive d) After creating Locos and grow Gate oxide e) Dope channel f) Create Floating gate and dope polysilicon g) Create Control gate and dope polysilicon h) After implant and diffuse S/D dopants

7e12 cm^{-3} (Figure 3h). After growing the protection layer, the final step is to deposit and etch Aluminum contacts for Source and Drain gates. The detailed process flow to fabricate the device is presented in Figure 2.

In this work, we propose to use the Athena tool to simulate the process of fabrication and the Atlas tool for operation simulations of the device. The following table gives the default parameter values in designing.

Table 1: Description And Default Parameter Values

DESCRIPTION	VALUE [μm]
Thickness of oxide-nitride-oxide nitride	0.013
Thickness of nitride spacer	0.002
Thickness of bottom oxide-nitride-oxide oxide	0.01
Thickness of top oxide-nitride-oxide oxide	0.001
Thickness of tunneling gate oxide	0.02
Thickness of control gate	0.075
Thickness of floating gate	0.075
Length of control gate	0.18
Length of floating gate	0.18

The complete 2D and 3D structures of the device in this work are given in Fig. 4 and Fig. 5, respectively. Regarding the 3D structure, the one is designed by using the 2D structure and the Devedit3D tool.

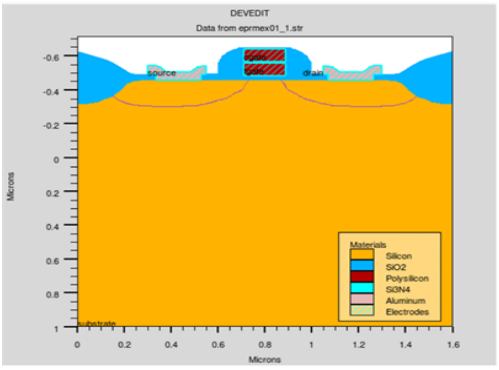


Figure 4: The 2D floating-gate device structure

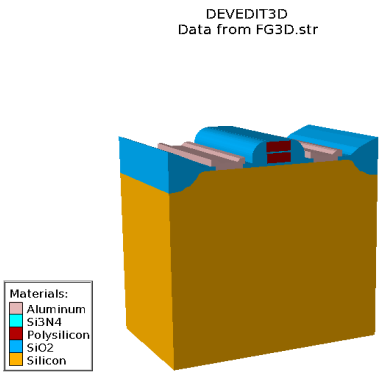


Figure 5: The 3D floating-gate device structure

4 Process Variation Results

This section presents the threshold voltage results of the floating-gate device designed in the previous section before programming and after programming.

Initially, before programming, there are no charges which are stored on the floating gate. The drain current – threshold voltage curve is shown in Fig. 6. The threshold voltage of the device is 0.6V.

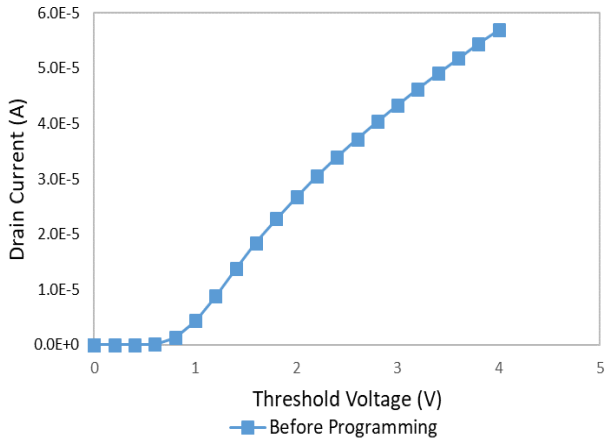


Figure 6: Threshold voltage of the device before programming process

After the programming process, the charges on the gate increase from 0C to around -3.5e-15C since the moving of electrons from the substrate. Meanwhile, the threshold voltage increases from 0.6V to approximately 6V. The charges on the gate and the threshold voltage in the programming process are given in Fig. 7 and Fig. 8, respectively. After that, the device is at the program state.

In contrast, as discussed in section II, the erasing process will remove all the negative charge on the floating gate which is generated from the programming process. Following this, the floating gate will have no charge, and the state of the device is as the initial (before programming state) or erasing state.

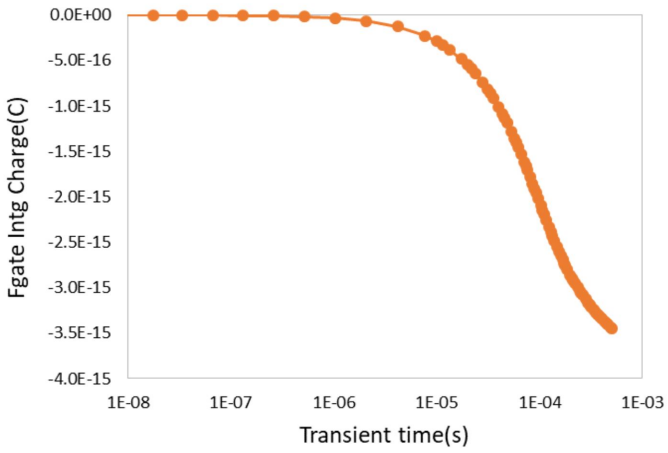


Figure 7: Floating gate charge in the programming process

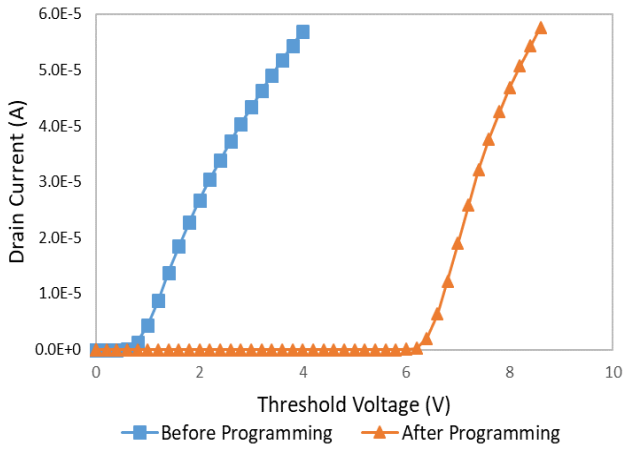


Figure 8: Threshold voltage of the device after programming process

The study of process variation demonstrates the effect of the threshold voltage on different geometrical parameters. In general, what stands out from the study is that whereas the threshold voltage is more sensitive to the gate length, tunneling gate oxide thickness, and bottom oxide-nitride-oxide oxide thickness, the threshold voltage is less sensitive to the gate width and nitride spacer thickness. The details of the study are given as follows.

4.1 Vary in values of the gate length

The figure in this section demonstrates the effect of the gate length parameter on the threshold voltage, it is clear that the increase in this parameter tends to a considerable increase in the threshold voltage, climbing around 18.33%, from 6V to approximately 7.1V. Thus, in the fabrication process, this parameter should be controlled carefully.

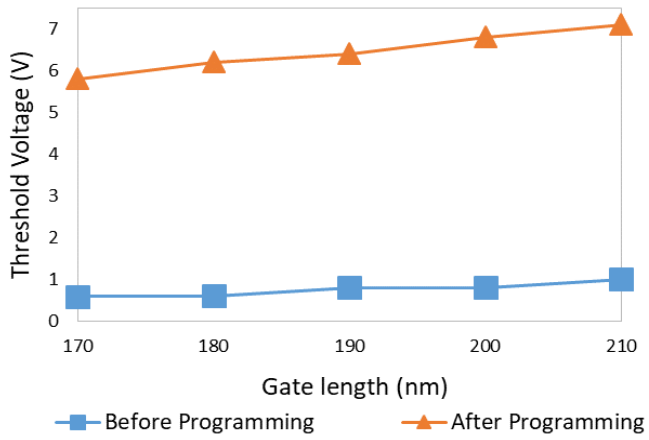


Figure 9: Gate length variation effects

4.2 Vary in values of the gate width

The figure below illustrates the change of the threshold voltage when the width parameter changes from 0.4μm to 1.6μm. This parameter is defined as a variable varied in the z-axis of the Devedit3D tool. In general, what stands out from the graph is that the voltage is not sensitive to this parameter. The threshold voltage remains stable at around 6V.

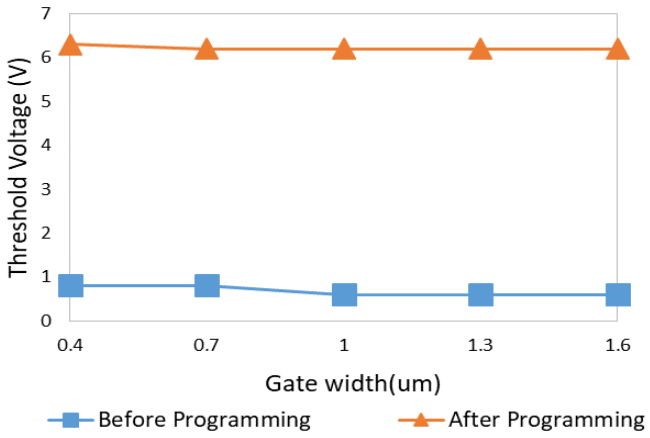


Figure 10: Gate width variation effects

4.3 Vary in values of the tunneling gate oxide thickness

The result is given in the graph in Fig. 11, it is clear that the change in tunneling gate oxide thickness results in a dramatic decrease in the threshold voltage when it varies from 20nm to 40nm. After programming, the voltage witnessed a drastic decrease, falling approximately 50% to about 3.2V. Hence, this is a very important parameter because the variation of the one causes a significant change in the operation of the device. With regard to the fabrication process, this parameter should be controlled carefully.

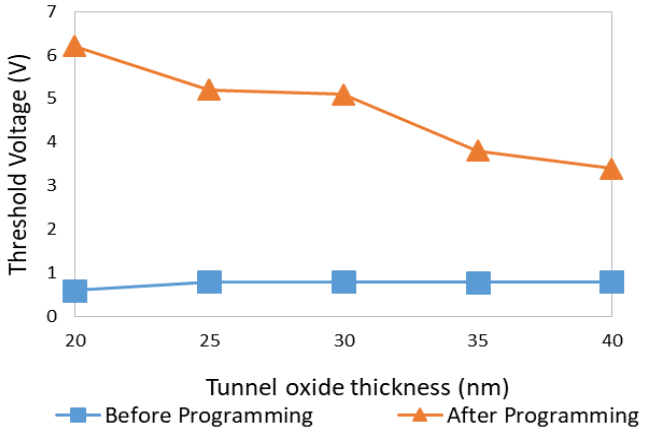


Figure 11: Tunneling gate oxide thickness variation effects

4.4 Vary in values of the bottom oxide-nitride-oxide oxide thickness

The graph in this section shows the results of the threshold voltage when the bottom oxide-nitride-oxide oxide thickness varied from 0.01μm to 0.03μm. Overall, it can be seen that the threshold voltage after programming experienced a considerable decrease of 23.33%, dropping from roughly 6V to approximately 4.6V. Thus, in the fabrication process, similar to the gate length and the tunneling gate

oxide thickness, this parameter should be observed carefully.

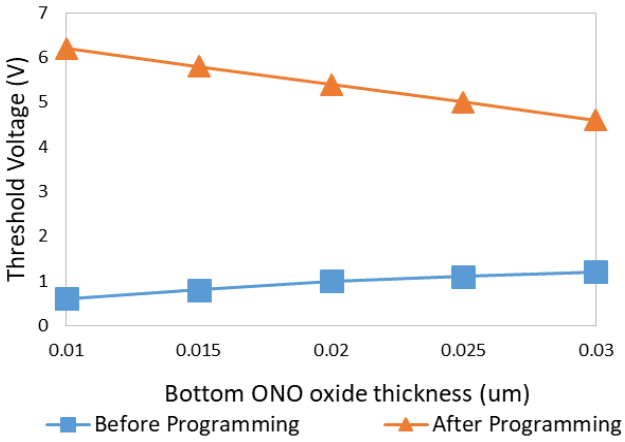


Figure 12: Bottom oxide-nitride-oxide oxide thickness variation effects

4.5 Vary in values of the nitride spacer thickness

The graph in the Fig. 13 presents the impact of the nitride spacer thickness parameter on the threshold voltage. It is clear that the impact of this parameter is very small. The threshold voltage after programming remaining between 5.9V and 6.1V throughout the study. Thus, the parameter would be controlled more easily than the others in the fabrication process.

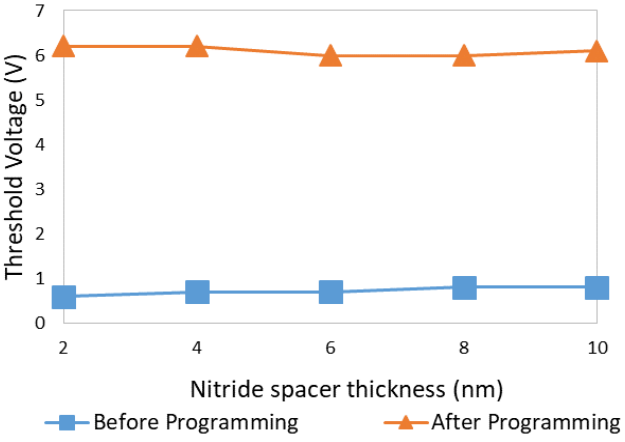


Figure 13: Nitride spacer thickness variation effects

5 Conclusion

This work is successful in the study of the process variation effects on the performance of a floating-gate device. The tunneling gate oxide thickness, bottom oxide-nitride-oxide oxide thickness, and gate length were found out to be sensitive to the threshold voltage, especially the tunneling gate oxide thickness parameter caused 50% change in threshold voltage. The others are 18.33% and 23.33%, respectively. While the gate width, nitride spacer thickness turned out to be less sensitive. This paper also successful in proposing a detailed flow and the parameter values to fabricate the complete floating-gate device based on CMOS 180nm process. The device which is designed by the flow worked as the expectation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is funded by Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, under grant number BK-SDH-2021-2080912.

References

[1] T. D. Cong, T. Le Thanh, H. M. Tri, T. Hoang, "The Simulation Study of Process Variation on Threshold Voltage in 180nm Floating-Gate device," in 2019 19th International Symposium on Communications and Information Technologies (ISCIT), 211–214, 2019, doi:10.1109/ISCIT.2019.8905125.

[2] R. B. Wunderlich, F. Adil, P. Hasler, "Floating Gate-Based Field Programmable Mixed-Signal Array," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, **21**(8), 1496–1505, 2013, doi:10.1109/TVLSI.2012.2211049.

[3] T. Kim, K. Park, T. Jang, M.-H. Baek, Y. S. Song, B.-G. Park, "Input-modulating adaptive neuron circuit employing asymmetric floating-gate MOSFET with two independent control gates," Solid-State Electronics, **163**, 107667, 2020, doi:https://doi.org/10.1016/j.sse.2019.107667.

[4] X. He, X. Zhu, L. Duan, Y. Sun, C. Ma, "A 14-mW PLL-Less Receiver in 0.18- μm CMOS for Chinese Electronic Toll Collection Standard," IEEE Transactions on Circuits and Systems II: Express Briefs, **61**(10), 763–767, 2014, doi:10.1109/TCSII.2014.2345303.

[5] J. Sun, C. C. Boon, X. Zhu, X. Yi, K. Devrishi, F. Meng, "A Low-Power Low-Phase-Noise VCO With Self-Adjusted Active Resistor," IEEE Microwave and Wireless Components Letters, **26**(3), 201–203, 2016, doi:10.1109/LMWC.2016.2521167.

[6] H. Liu, C. C. Boon, X. He, X. Zhu, X. Yi, L. Kong, M. C. Heimlich, "A Wideband Analog-Controlled Variable-Gain Amplifier With dB-Linear Characteristic for High-Frequency Applications," IEEE Transactions on Microwave Theory and Techniques, **64**(2), 533–540, 2016, doi:10.1109/TMTT.2015.2513403.

[7] H. Liu, X. Zhu, C. C. Boon, X. He, "Cell-Based Variable-Gain Amplifiers With Accurate dB-Linear Characteristic in 0.18 μm CMOS Technology," IEEE Journal of Solid-State Circuits, **50**(2), 586–596, 2015, doi:10.1109/JSSC.2014.2368132.

[8] X. Tong, Y. Yang, Y. Zhong, X. Zhu, J. Lin, E. Dutkiewicz, "Design of an On-Chip Highly Sensitive Misalignment Sensor in Silicon Technology," IEEE Sensors Journal, **17**(5), 1211–1212, 2017, doi:10.1109/JSEN.2016.2638438.

[9] S. Chakraborty, Y. Yang, X. Zhu, O. Sevimli, Q. Xue, K. Esselle, M. Heimlich, "A Broadside-Coupled Meander-Line Resonator in 0.13- μm SiGe Technology for Millimeter-Wave Application," IEEE Electron Device Letters, **37**(3), 329–332, 2016, doi:10.1109/LED.2016.2520960.

[10] G. O. Puglia, A. F. Zorzo, C. A. F. De Rose, T. Perez, D. Milojicic, "Non-Volatile Memory File Systems: A Survey," IEEE Access, **7**, 25836–25871, 2019, doi:10.1109/ACCESS.2019.2899463.

[11] E. I. Vatajelu, H. Aziza, C. Zambelli, "Nonvolatile memories: Present and future challenges," in 2014 9th International Design and Test Symposium (IDT), 61–66, 2014, doi:10.1109/IDT.2014.7038588.

[12] R. Bez, E. Camerlenghi, A. Modelli, A. Visconti, "Introduction to flash memory," Proceedings of the IEEE, **91**(4), 489–502, 2003, doi:10.1109/JPROC.2003.811702.

[13] S. Wang, C. He, J. Tang, X. Lu, C. Shen, H. Yu, L. Du, J. Li, R. Yang, D.-X. Shi, G. Zhang, "Nonvolatile Memory: New Floating Gate Memory with Excellent Retention Characteristics (Adv. Electron. Mater. 4/2019)," Advanced Electronic Materials, **5**, 2019, doi:10.1002/aelm.201970018.

[14] E.-J. Park, J.-M. Choi, K.-W. Kwon, "Behavior Modeling for Charge Storage in Single-Poly Floating Gate Device," Journal of nanoscience and nanotechnology, **19**, 6727–6731, 2019, doi:10.1166/jnn.2019.17112.

[15] L. L. M. A. Pavan, Paolo, Floating Gate Devices: Operation and Compact Modeling, Springer US, 2004.

[16] G. Rappitsch, E. Seebacher, M. Kocher, E. Stadlober, "SPICE modeling of process variation using location depth corner models," IEEE Transactions on Semiconductor Manufacturing, **17**(2), 201–213, 2004, doi:10.1109/TSM.2004.826940.

[17] F. Duvivier, E. Guichard, "Worst-case SPICE model generation for a process in development using Athena, Atlas, Utmost and Spayn," in ICM 2001 Proceedings. The 13th International Conference on Microelectronics., 11–18, 2001, doi:10.1109/ICM.2001.997475.

[18] Y. Saad, C. Tavernier, M. Ciappa, W. Fichtner, "TCAD tools for efficient 3D simulations of geometry effects in floating-gate structures," in Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 77–82, 2004, doi:10.1109/NVMT.2004.1380810.

[19] S. Bala, B. Mahendia, "Simulation of Floating Gate MOSFET Using Silvaco TCAD Tools," 2015.

[20] S. Agarwal, D. Garland, J. Niroula, R. B. Jacobs-Gedrim, A. Hsia, M. S. Van Heukelom, E. Fuller, B. Draper, M. J. Marinella, "Using Floating-Gate Memory to Train Ideal Accuracy Neural Networks," IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, **5**(1), 52–57, 2019, doi:10.1109/JXCDC.2019.2902409.

[21] T. Dubey, V. Bhadauria, "A low-voltage highly linear OTA using bulk-driven floating gate MOSFETs," AEU - International Journal of Electronics and Communications, **98**, 2018, doi:10.1016/j.aeue.2018.10.034.

[22] P. Pavan, R. Bez, P. Olivo, E. Zanoni, "Flash memory cells-an overview," Proceedings of the IEEE, **85**(8), 1248–1271, 1997, doi:10.1109/5.622505.

[23] R. Dhavse, F. Muhammed, C. Sinha, V. Mishra, R. M. Patrikar, "Memory characteristics of a 65 nm FGMOS capacitor with Si quantum dots as floating gates," in 2013 Annual IEEE India Conference (INDICON), 1–3, 2013, doi:10.1109/INDCON.2013.6725910.

[24] R. N. T. Stanley Wolf, Silicon Processing for the VLSI Era, Volume1: Process Technology, Lattice Press; 1st edition (January 1, 1986), 1999.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 148-154 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Aviation MRO: Impact of Physical Environment Factors on Job Performance in Aircraft Maintenance Organization

Kamal Jaiswal[1,*], Serdar Dalkilic[1], Evangelos Papageorgiou[1], Balgopal Singh[2]

[1]*Higher Colleges of Technology, Muroor Raod, Abu Dhabi, 25035, United Arab Emirates*

[2]*Banasthali Vidyapith, Banathali, Distt.- Tonk, Rajasthank, 304022, India*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 11 October, 2020
Online: 08 November, 2020

Keywords:
Aircraft Maintenance
Physical environments
Job performance
Safety
Work Efficiency

ABSTRACT

A happy employee with high morale is the best asset of an organization. The morale of an individual working in an organization can be impacted by both positive and negative ways by physical environment factors of the workplace. The physical environment of an organization impacts a lot on the work performance of an individual. It becomes more significant in case of aviation maintenance organization formally called Maintenance Repair and Overhaul (MRO) organization as the work performed by an individual requires full accuracy and directly linked with the safety of the aircraft and its occupants. This can only be achieved by providing a safe and positive working atmosphere to individuals working in MROs. Considering the same, this study targets to deliver a brief status report of physical environment factors disturbing work performance in Maintenance Repair Overhaul (MRO) agencies. This study initially explains the impact of different physical environment factors. The study also discusses the impact of unsafe work due to improper working environment in aircraft maintenance organizations. The individual's working in MROs located in the United Arab Emirates were considered for testing the physical environment factors for this case study. This case study aims to provide the status of different physical environment factor of organization causing problems at work for individuals working in aircraft maintenance organizations. The result of the study indicated that MROs in UAE are providing a conducive working atmosphere to employees, however, the lack of workplace is still an area for improvement.

1. Introduction

The human being is manipulating the environment for his survival and comfort contrary to the environment or nature is human's best friend and nearest surrounding. Unsafe and dangerous conditions may arise if we manipulate nature incorrectly and can make the life of human being very difficult. [1].

The efficiency of individuals and the work morale of individual at the workplace might get affected due to the hazardous environment resulting from the poorly administered surrounding. To determine the performance, morale and efficiency of an individual at a workplace, the workplace environment is a key factor. The workplace environment includes the area and condition in which an employee is executing the job however a positive and effective work environment is the condition when the output of the individual working for an organization is same as expected as per job objective [2]. The physical environment where the job is being carried out by an individual significantly affect the performance of an individual.

The bodily atmosphere as a characteristic of the workplace atmosphere has an uninterrupted influence on human intelligence and can slowly variate interpersonal communications and therefore, efficiency. It happens because the features of space/room or an area have significance on the efficiency and happiness of individuals [3]. Several types of research have concluded that the features of a working area affect the satisfaction of the users. These features of a work area are preferred by the user and extremely important to the user for productivity and work satisfaction. These features include proper lighting, good ventilation, soundproofing, temperature and workspace etc. [4].

*Corresponding Author: Kamal Jaiswal, kamaljswl@gmail.com

The factors like proper illumination of the workplace and suitable work resources such as furniture can have an encouraging effect on the health of an individual and subsequently on the efficiency. Other environment elements of a workplace such as controlled temperature, soundproofing, good ventilation etc. play a significant role in the productivity of an organization [5]. These elements ultimately influence the overall working atmosphere and decide the productivity and satisfaction.

The employees who are bodily and mentally motivated to do their job, the performance is expected to be enhanced, and safety is improved. A well-working environment can significantly reduce the absenteeism and conclusively enhance work efficiency [6]. The aircraft maintenance organization work under a complex system and the productivity of the individual decide the success of the organization. The efficiency of individual working in aircraft maintenance organization largely rely on working condition or physical working environment and it is the most important factor for safety as well [7]. The safety management system used in the aviation sector also address this issue and emphasize on providing a good physical environment to the individuals to perform the job safely. However, it is quite challenging for the management to provide a user-friendly environment to an individual in aircraft maintenance organization as the certain task to be performed on aircrafts requires an individual to work under adverse condition [8].

The output of an organization is directly linked with the satisfaction of employees with the physical environment of the workplace. The administration should regularly monitor and ensure that the physical working environment of the organization is conducive to work for the individual performing the job. The commitment of the administration to deal with work environment influences the motivation of the individual to perform properly and it should be assessed regularly and management should make necessary changes whenever required by employees to improve the efficiency and safety [9]. The changes may include, such as a change in the layout of the workplace or the addition of more resources.

Today's administration has to alter the way of working in order to get real modifications in these facets. The organization has to devote more time on governing these issues of the work situations rather than micromanage. This can be accomplished by enlightening on the physical workplace and amenities, assigning tasks, cumulative liability, and inspiring teamwork. This will encourage faith and devotion between the workers and boosts improved cooperation amongst them.

The capability of an employee to successfully perform a job in a tuff work environment is limited. The studies have revealed that a workspace feature changes the approach of employees and intensify their output. It also states that hazardous and insalubrious workplace atmosphere in terms of poor ventilation, incorrect illumination, disproportionate sound etc. disturb employees' efficiency and healthiness

2. Physical Working Environment

Aircraft maintenance engineers work inside a "complex system". There are plentiful characteristics inside this structure that encroach on the aircraft maintenance engineer, outspreading from his knowledge, skills and capabilities, the atmosphere in which he/she works, to the values of the organization for which he/she works. Even beyond the actual company employee works for, the governing requirements laid down for his/her job evidently impact on his/her comportment. All features of this structure may promote headed for inaccuracies that the engineer might make [10].

To well apprehend efficiency on jobs, it is vital to study the relationship between job performance, individuals, and circumstances elements. Job performance is a very substantial factor manipulating the cost-effectiveness of any organization. Performance is a key element for organizations as employees' performance directs to business success. Similarly, performance is vital for individuals, as accomplishing responsibilities can be a basis of happiness [2].

All the activity and behaviours that are executed by an individual, leading to attain the objectives of an organization can be termed as Job performance. The outcome of a task done by an individual or a group of individual in an organization on a specific time directly indicating that how well the individual or a group of the individual did that job while considering the organization's goal can be described as Performance. The physical environment such as operating conditions, procedures, resources and equipment to perform the job can majority impact the job performance of an employee [11].

In aviation, specifically in MROs, the job performance of an individual is very significant as the job performed by them is safety-critical. High level of job performance from an individual is required in the aviation industry to have a safe operation. The physical work environment and its impact have been enormously considered since the conditions can hamper, interfere with, or set restrictions on the variety of work behaviours that are exhibited which, in turn, possibly influences task performance [3].

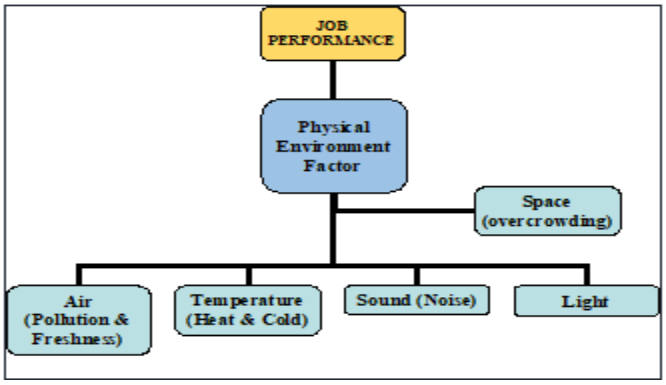


Figure 1: Physical Environment Factor

3. Physical Environment Factor that Influences the Performance

A vigorous interface is discovered between job performance and physical working environment, by the description of what is envisioned by the work environment and literature study findings [12]. The employees' performance, satisfaction, social relations and health is directly linked with physical environment of workplace. Generally, it is presumed that the bodily design of workstations and the ecological condition of workstation are

substantial elements in organizational performance. The study of Badayai (2012) has established earlier studies and categorized five key aspects of an unpleasant working environment as shown in figure 1.

3.1. Air Factor

The behaviour of individuals working in aircraft maintenance organization can significantly alter the job performance due to poor air quality resulting from lack of ventilation or any other factor. The quality of air significantly affect the health of individuals and describe the comfort level of the individual. Sometimes the quality of air can be very poor, specifically individual working with chemicals and in the confined area of aircraft in aircraft maintenance organization [5].

Utmost conceivably unsafe indoor air toxins are radon, asbestos, inorganic material, environmental tobacco smoke, organics, biological and non- ionizing radiation. Supplementary polluting agents such as odours and dust can be the source of serious uneasiness and feelings of reluctance that may lead to a fall in efficiency and job performance. A good ventilation system can overcome this issue and provide a sense of comfort with a combination of controlled temperature. Apart from this, some protective apparatus maybe use to avoid inhalation of pollutants in certain working conditions [10].

3.2. Temperature

The discomfort caused by temperature can be very painful and can significantly reduce job performance. Several types of research have established that the temperature of the workplace has a lot to do with how we interact with others in an organization. The behavior of individuals can significantly change when the temperature of the workplace is not conducive such as too hot or too cold [10]. Working under a condition either hot or cold can cause the lack of attention specifically in those tasks which require a great amount of focus such as crown inspection of rivets on aircraft or Non-destructive testing in workshops.

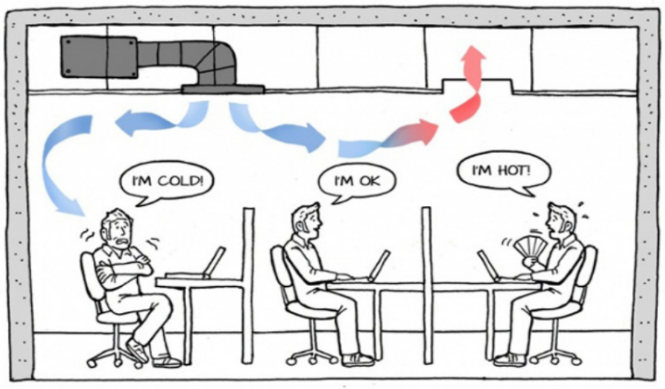


Figure 2: Effect of Temperature

A workplace, which is too hot or too cold, can be the source of stress and can affect the health of individuals. For example, Heat stress can make the body to drop electrolytes and H2O quicker, triggering low psychological act and reduced motor skills. Your musculoskeletal system will be making efforts to save energy, which will absolutely slow your minds, making finishing tasks and escaping errors a hard accomplishment [4].

3.3. Sound (Noise)

Sound is amongst the substantial environmental factors for people's well-being, and it has a vital part in both physical and mental injuries, and it also disturbs individuals' performance and efficiency [6]. The aircraft maintenance engineering is frequently exposed to high noise level such as noise coming from Turbine Engine of the aircraft or from pneumatic equipment used by them. These high level of noise can significantly reduce the ability of an individual to perform efficiently and also have a serious impact on an individual's health both physical and mental [9].



Figure 3: Noise

3.4. Illumination

Poor illumination often gets unnoticed in the workplace, as we talk about mental health and well-being, and the focus is firmly on creating happier and healthier workplaces. But bad lighting is related to a range of ill-health effects, both physical and mental, such as eye strain, headaches, fatigue and also stress and anxiety in more high-pressured work environments [1]. This becomes more significant in terms of aircraft maintenance, as several tasks carried out by aircraft mechanics in aircraft maintenance organization require adequate lighting. The improper lighting can cause serious safety risk as some defects or faults on aircraft may get unnoticed due to lack of lights or bright lights [6].

3.5. Workspace

The individual performing particular task requires an adequate workspace to perform the job properly, however, sometimes it is a bit impossible to have sufficient workspace while performing aircraft maintenance such as C checks performed as a part of major maintenance require several teams of aircraft mechanics to work together [8]. Having a confined space to work, affect the ability of an individual to perform the job and the situation may become worse when it comes with other adverse condition such as noise, time pressure etc.Captions must be written in sentence case (e.g., Macroscopic appearance of the samples.).



Figure 4: Workspace in MRO

4. Research Methodology

The objective of this research paper is to analyze the different physical environment factors disturbing work performance of individuals employed in the Maintenance Repair Overhaul (MRO) organization in the United Arab Emirates (UAE). The profile of intended respondent for this research is individual working in MRO inside the UAE and doing in maintenance activities such as aircraft maintenance technicians, aircraft maintenance engineers and other secondary staff. UAE is a key centre of aviation and patently has the finest airlines in the world. UAE also has the top capability for maintenance of aircraft in the region and unquestionably is the hub of premium MROs in the world. The MROs are functioning under a very multifaceted structure and use innovative technologies for daily operations. The agenda of the research was discussed with the intended audience to have an impartial examination of research elements.

Qualitative and quantitative approaches were engaged in measuring the physical environment factors. The research data was collected through paper-based surveys after a comprehensive conversation with individuals working in aviation MROs. The intended audience considered for this research were the employees of aircraft maintenance organization employed at a diverse level. This research was conducted as a part of the student's project and as project students have to go out and meet industry professionals in different aviation organization. The student involved in the research were studying a four-year Bachelor's Degree program in Aircraft Maintenance Engineering Major.

The structured questionnaire was used and the selection of questions was based on different elements of physical environment factors causing the problem at the workplace. The questionnaire includes the physical environment elements, which can possibly cause problems at the workplace, were related to:

- Inadequate lighting
- Noise
- Excessive Heat
- Excessive Cold
- Overcrowding
- Ventilation
- Fumes
- Maintenance equipment

Cronbach's alpha is used to test the reliability and validity. The Cronbach's alpha value is more than 0.70. The Cronbach's alpha of the questionnaire found to be 0.76. The statistical method used to investigate the data is the frequency and percentage analysis. The data was collected using online survey platform survey monkey. The MS excel software was used to carry out the statistical analysis. The outcomes of the study were arranged after carrying out an in-depth examination of the data collected concerning physical environment elements affecting job performance. Table I shows the statistics of the research survey.

Table 1: Sample Table

STATISTICS OF THE RESEARCH SURVEY		
#	Research Material	
	Type of Material	Quantity
1	Total questions included in questionnaire	22
2	Total questionnaires distributed	200
3	Percentage of responses received	85%

5. Investigation Results

The research outcomes are indicated in the form of the illustration, column and pie tables for the comfort of explanation. The percentage of replies is 85%; it is good and imitates the individual's mindfulness and enthusiasm in relation to the physical environment of the organization.

Figure5 shows that a total of 53% of respondents NEVER had problems due to poor or inadequate lightings at the workplace. Total 38% of respondents choose that SOMETIMES they have problems due to improper lighting. Total 9% respondent OFTEN have problems caused by inadequate lighting at the workplace. In aviation, the adequate lighting is quite critical especially when individuals are working in complex and small aircraft parts during the night and in a dark atmosphere.

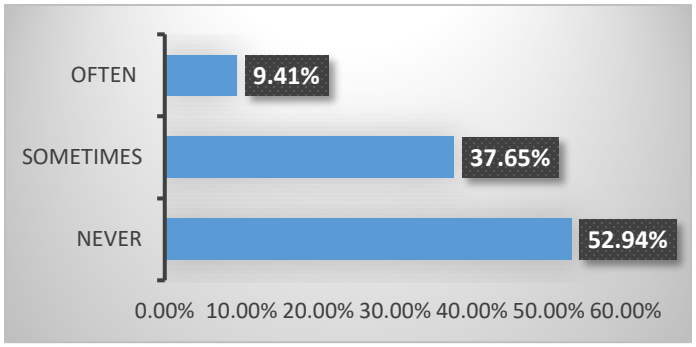


Figure 5: Poor/inadequate Lighting

Figure 6 shows frequency responses to the problem caused by noise at the workplace. As shown, 41% of respondents NEVER have a problem due to noise at the workplace, and 43% of them SOMETIMES have problems due to noise. However, 16% of respondent OFTEN have problems due to noise. Noise can be very annoying and can have a serious impact on health, which can lead to degradation of work performance

Figure 7 shows responses to the effectiveness of individuals exposed to excessive heat. As indicated in the figure, 49%, respondent NEVER have excessive heat working condition, 39% SOMETIMES have excessive heat working condition, and 11%

respondent OFTEN have excessive heat working condition. This is a clear indication that working under excessive heat condition cannot be avoided, but it affects your performance and individuals get less effective when adverse conditions arise.

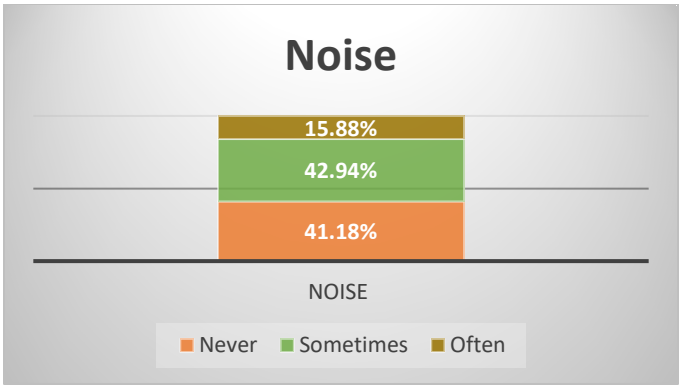


Figure 6: Noise

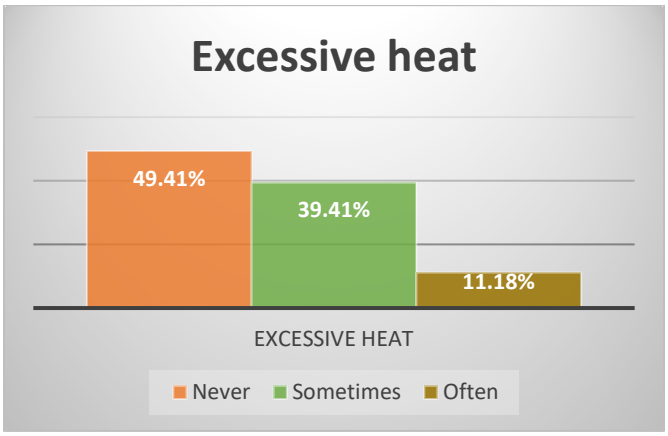


Figure 7: Excessive Heat

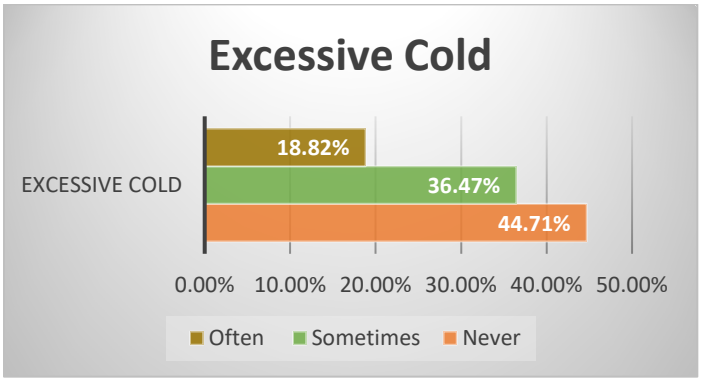


Figure 8: Excessive Cold

Figure 8, shows that 45% of respondents have acknowledged that they NEVER a condition of excessive cold at the workplace while 36% of respondents SOMETIMES have the excessive cold condition at the workplace and almost 19% respondent OFTEN have the excessive cold condition at the workplace. These percentages seem to be quite significant and require action by the organization to control the physical environment condition of the workplace to make it more conducive.

Figure 9 shows responses to the overcrowding of the workplace leading to lack of workspace. About 45%, 40%, and 15% have responded to NEVER, SOMETIMES and OFTEN respectively. This result reflects that the majority of the people have an issue related to overcrowding of the workplace.

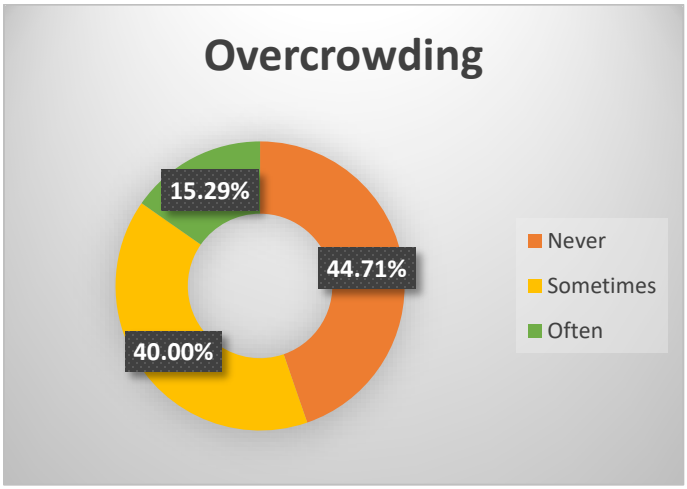


Figure 9: Overcrowding

Figure 10 shows responses to poor ventilation condition at the workplace. About 51%, 38% and 11% have responded to NEVER, SOMETIMES and OFTEN respectively. These results reflect that ventilation condition needs some improvement and it becomes more significant when dealing with hazardous chemicals.

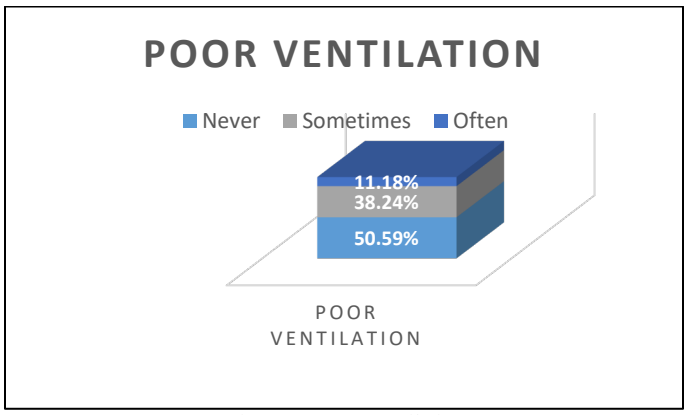


Figure 10: Poor Ventilation

Figure 11, shows responses to the question related to the presence of Dust and Fumes at the workplace. About 54% of respondents NEVER have problems related to dust/fumes, 29% respondent SOMETIMES have problems with dust/fumes and 16% respondents OFTEN have problems with dust/fumes. It indicates that only small no. of the individual having a problem with dust or fumes and thus interfering with their work.

Figure 12, shows the results of poor maintenance condition of equipment used by individuals causing the problem at work. It was found that 52% of respondents NEVER have the problem of poor maintenance of equipment they use, 35% respondent SOMETIMES have the problem of poor maintenance condition of the equipment they use and only 12% respondent OFTEN have the problem of poor maintenance of equipment they use. This

matter is directly related the work performance as equipment used by an individual in the organization provides the basis of good maintenance of aircraft.

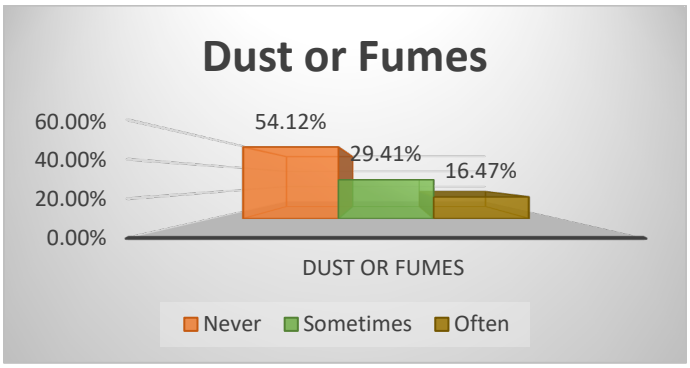


Figure 11: Dust or Fumes

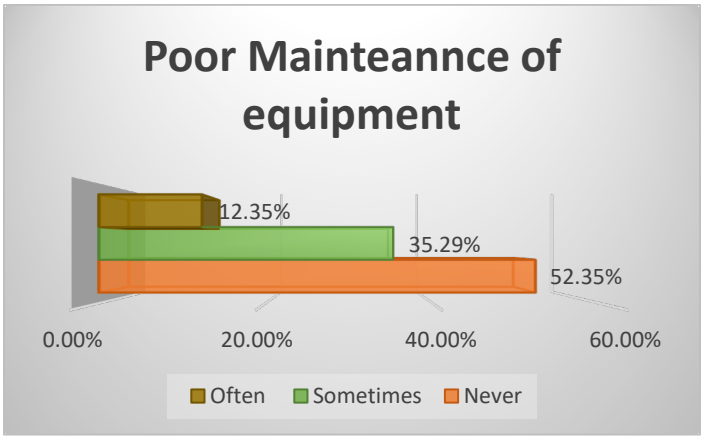


Figure 12: Poor Maintenance of Equipment

6. Discussion and Recommendation

This research specified that (reference to Figure 5) people working in MROs in the UAE typically does not have an issue related to poor illumination, and it does not affect their work. It is a optimistic signal that organizations are aware of this and have the necessary resources in place. However, a substantial number of individuals do have a problem with poor lighting sometimes which further need to be investigated why.

The study presented (a reference to Figure 6) that individuals working in aircraft maintenance organization frequently exposed to Noise, which should be the matter of concern for MROs as noise can seriously degrade the performance of individuals. It is also found (a reference to Figure 9) that a large number of individuals have problems with overcrowding leading to lack of workspace to perform the job specifically for the individuals working in base maintenance.

The analysis from Figure 10, evidently specifies that a great number of persons never have an issue due to poor ventilation which can reduce the air quality of the workplace. However, some respondent reported that they often have an issue with poor ventilation and this need to fix as poor ventilation lead to poor air quality and can result in a poor health condition of individual.

Figure 7 and Figure 8, clearly indicates that the majority the individuals does not have issues related to excessive heat or excessive cold condition. It reflects that temperature control of the workplace is properly maintained to provide a conducive atmosphere to individual working in MRO.

Exploration of figure 12 specifies that the most of the individuals working in MRO's at UAE does not have an issue related to poor maintenance of equipment's they use. However, a significant number of respondents have indicated that they sometimes or often have problems due to poor maintenance of equipment's.

The underneath mention fact specifies the zones of development as per the results of this study, and the succeeding recommendations are identified:

- Increasing attentiveness of the physical work environment and its optimistic effects on the organization and the individuals.
- Management taking proactive action to resolve the problem related to poor working conditions.
- Additional efforts are required to improve working conditions to guarantee enhanced work performance.
- Gathering more often with staff to improve their comfortability to report and to obtain opinion to lift the whole work performance.

Further, a study with hefty no. of the participant and a comprehensive questionnaire is suggested. The study can be very valuable if it is linked to statistics of MRO's worldwide.

7. Discussion and Recommendation

In this exploration, efficiency and significance of physical environment factors affecting work performance in MROs were examined. The MROs in UAE have delivered great backing throughout the course the research and research data was formed with support from MROs. It is concluded that MROs generally have very good physical environment. The MROs are taking all necessary action to provide a conducive atmosphere for employees. However, there are areas for improvement for a better physical environment including the improvement in terms of overcrowding leading to lack of workspace and the introduction of a robust system for proactive actions related to a physical environment.

Acknowledgements

The study was led by endless and unceasing backing from Higher Colleges of Technology (U.A.E.) and different MROs within the United Arab Emirates. The delicate and specialized support by Each Subject Matter Expert has educated me the finest research guidance, and I am appreciative for the generous support delivered by them.

References

[1] A.A. Shikdar, "Identification of ergonomic issues that affect workers in oilrigs in desert environment." International Journal of Occupational Safety and Ergonomics, **10**: 169-177, 2002, doi.org/10.1080/10803548.2004.11076605

[2] F.O. Becker, "Workspace creating environments in organization." New York, N.Y. : Praeger, x, 225 p. : ill, 1981.

[3] F. Becker "Improving organizational performance by exploiting workplace flexibility." Journal of Faculty Management, 1: 154-162, 2002.
[4] C.C. Berger, HH Ehrsson, "Mental Imagery Changes Multisensory Perception." Current Biology, **23**(14), 1367-1372, 2013, doi.org/10.1016/j.cub.2013.06.012.
[5] P.E. Mohr, J.J. Feldman, J.L. Dunbar, et al., "The societal costs of severe to profound hearing loss in the United States." Internationl Journal of Technology Assess. 2000.
[6] R.G. Geen, "Preferred stimulation levels in introverts and extroverts: Effects on arousal and performance." Journal Personality Social Psycholog, 46**,** 1303–1312, 1984.
[7] C.H. Friend, Aircraft maintenance management, Harlow: Longman scientific and technical, 1992.
[8] GCAA-U.A.E., Civil Aviation Requirement-145, 05 revision-01, GCAA-U.A.E., February 2016.
[9] A. Kumar, P.K. Sinha, "Human Error Control in Railways." Jordan Journal of Mechanical and Industrial Engineering, **2**, 183–190, 2008.
[10] S.A. Stansfeld, M.P. Matheson, "Noise pollution: non-auditory effects on health." British Medical Bullettin., 68:243–257, 2003. doi.org/10.1093/bmb/ldg033
[11] K. Persson-Waye, R. Rylander, S. Benton, H.G. Leventhall, "Effects on performance and work quality due to low frequency ventilation noise." Journal of Sound Viberation.;205:467–474, 1997 doi.org/10.1006/jsvi.1997.1013
[12] W.W. Clark, B.A. Bohn, "Effect of noise in hearing." JAMA the Journal of the American Medical Association; **281**(17),1658-1659, 1999. DOI: 10.1001/jama.281.17.1658

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 148-153 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design of Purebred Dog Recommendation System Using MCDM Approach

Phie Chyan*

Department of Informatics, Atma Jaya University of Makassar, Makassar, 90224, Indonesia

ARTICLE INFO

Article history:
Received: 15 May, 2020
Accepted: 18 August, 2020
Online: 09 September, 2020

Keywords:
Dog Breed
Dog Adopter
Purebred Dog
Decision Support System
MCDM Approach

ABSTRACT

The dog is one of the first animals domesticated by human, and for thousands of years, it has been artificially bred into hundreds of types in order to provide certain traits that humans want. Nowadays, the selection of dogs by potential adopters has become a problem due to the availability of different type of breeds with their physical and mental characteristics. This study aims to design a decision support system through an analytical model that uses variety of data on the characteristics of purebred dogs obtained from different sources. Data from official sources is obtained from international purebred dog organizations that set the standards for each breed type, while data from unofficial sources is obtained from the dog lovers community, experts, and kennel owners. The result of the study provides appropriate recommendations to potential adopters in selecting a breed that suits their preferences and needs.

1. Introduction

This paper is an extension of work originally presented in **2018** International Seminar on Research of Information Technology and Intelligent Systems (**ISRITI**) [1]. According to the history recorded since 15 thousand years ago, Dogs were one of the first animals domesticated by humans in East Asia and then spread throughout the world and became involved in the long history of human culture and civilization. Through artificial selection, humans breed dogs into a new kind of breed by mating two species of dogs that have certain characteristics and traits to accentuate the desired combination of their traits [2]. Nowadays dogs have developed into hundreds of breeds with a total of 344 breeds officially recognized by FCI. Among the many types of breeds, the dog has divided into several breed groups with similar physical and psychological characteristics. Each breed group has certain characteristics that can support the role of the dog in social aspects with human's society [3]. Some types of dogs are used by humans to help in herding cattle, others are used as guard dogs to help law enforcement work as sniffer dogs.

The term dog refers to the dog domesticated by Canis lupus familiaris. Dogs were classified as Canis familiaris by Linnaeus in 1758 but in 1993 by the Smithsonian Institute and the American Mammal Experts Association, dogs were designated as subspecies of the gray wolf Canis lupus. Dogs are social animals, but the dog's personality and behavior can vary depending on the breed. In addition to genetic traits, the dog's personality and behavior also depend on the caring environment and treatment received from the dog's owner and people who interact closely with the dog.

Dogs as social animal have a vast variety of personality and are among the most high intelligence animals according to scientific research and field evidence. The level of intelligence of the dog depends on the race and each dog individually. According to Coren (2006), a dog psychologist who conducts dog intelligence studies on more than 100 breeds, the standard of intelligence of a dog is how quickly the dog can follow a new command given and how obedient the dog is to the commands it has learned [4]. Another proof is according to research at the University of Lincoln published, dogs have human-like intelligence who can psychologically recognize emotions and other human psychological states, and this ability is only demonstrated by dogs in interacting with humans, this shows the cognitive social abilities of a dog are more advanced than any other animal that is genetically closer to humans, and this ability is a genetic shift that distinguishes dogs from their wolf ancestors [5].

Based on the physical attributes dogs vary greatly in size, appearance, and behavior compared to other pets. Several of the important physical attribute of dogs highlighted as follows : Dogs have strong muscles, sturdy ankle bones, cardiovascular systems that support physical endurance and running speed, and teeth to catch and tear their prey. Dogs technically walk on tiptoes with toes. In term of sense of sight, research shows dogs can see several colors, although not as much and bright as can be seen by humans with a brightness level of approximately half of the accuracy of

*Corresponding Author: Phie Chyan, Email: Phie_chyan@lecturer.uajm.ac.id

human vision [6]. In sense of hearing aspect dogs can hear low-frequency sound range from 16 to 20Hz (humans only hear frequencies of 20-70 Hz), and high-frequency sounds range from 70 kHz to 100 kHz (humans only hear frequencies 13-20 kHz) and from the sense of smell Dogs have nearly 220 million olfactory cells that are sensitive to odors. According to research conducted by Michell (1999) on the age of various breeds of dogs, dog life expectancy depends on the breed with a median value between 10-13 years [7].

Breed standard is a set of guidelines which is used to ensure that the dog produced by a breeder or breeding facility conforms to the specifics of the standardized breed. Breed standards are devised by breed associations not by individuals and are written to reflect the use or purpose of the species and breed of the dog [8]. Breed standards help define the ideal dog of a breed and provide goals for breeders in improving stock. In essence, a breed standard is a blueprint for an dog fit for the function it was bred - i.e. herding, tracking etc. For that purposes purebred dogs have a non-profit organization/federation that aims to standardize and register purebred dogs that will improve the quality of breeding and use of purebred dogs whose health and physical characteristics meet the standards. Federation Cynologique Internationale(FCI) is the parent organization that oversees almost all organizations/kennels from various countries except the USA, Canada, and the United Kingdom. The American Kennel Club (AKC) is a registry of purebred dog pedigrees in the United States that also has an international reputation. At present, FCI and AKC are the world's largest breed of dog registration institutions that establish the standardization and grouping of purebred dogs, until now various dog organizations in the world use the standards issued by FCI and AKC as guidelines in identifying dog breeds.

Humans are faced with various problems in the selection of dog breed, there are so many breeds available with each of their feature and characteristics and should all of those have aligned with human needs as prospective adopters who also have their respective interests [9]. Selection of the breed that is not compatible can bring problems for both the dog itself and the adopter, a dog may not be well taken care of and even end up in the dog shelter only because of incompatibility with the expectations of the adopter, while the time and cost had wasted for the adopter. More criteria need to be considered by prospective dog adopters apart from physical form characteristics like body size and structure such as the average intelligence level of a breed, ease of care, innate traits, and more. So many criteria that must be considered by prospective adopters in choosing breed types makes decision making a complicated task because of the difficulty in analyzing the influence, interests, and relation between one criteria with another and coupled with many types of breeds available that make this problem requires Multi-Criteria Decision Making (MCDM) approach to handle it. The purpose of this study is to design a Decision Support System (DSS) that can provide analytical recommendations using MCDM approach for the problem of selecting a breed of dog that suit the needs and preferences of the potential adopters.

2. Methodology

2.1. Data Sources

Primary data for this study came from FCI and AKC's online knowledge base sources that contain information on officially recognized breeds, groupings, and their physical and psychological characteristics. Secondary data sources came from interviews conducted with the dog lovers community, dog experts, veterinarians, and kennel owners to find out the experience of the respondents according to their fields and expertise regarding how to choose, care, and handle the breed of dog to be adopted including obstacles encountered in dealing with various problems involving purebred dogs. Data collected from primary and secondary sources were then processed to become a knowledge base of the designed system.

2.2. Analytic Model

To support the developed system model, an MCDM based approach is needed in analyzing data sources to produce an appropriate analytical model [10]. The basic principles of developing an analytical model based on MCDM approach are :

- Decomposition : After defining the problem regarding the selection of dog breeds, then the main problem-solving process was carried out into more specific elements to support structure of the main problem, this process must be carried out until the most specific sub-elements are obtained and formed the hierarchical model.
- Comparative Judgment : Conduct an assessment of the elements relative importance in decision making for the selection of dog breeds at a certain level related to elements at a higher level. The assessment results would be presented in the form of pairwise comparison to get the relative weights among the elements.
- Synthesis of Priority : Each pairwise comparison matrix needs to compute to get local priority value since the pairwise comparison matrix exists at each level, then local priorities must be synthesized to obtain global priorities. Procedures for synthesizing differ according to the shape of the hierarchy. The ordering of elements according to relative importance through a synthesis procedure is called priority setting.

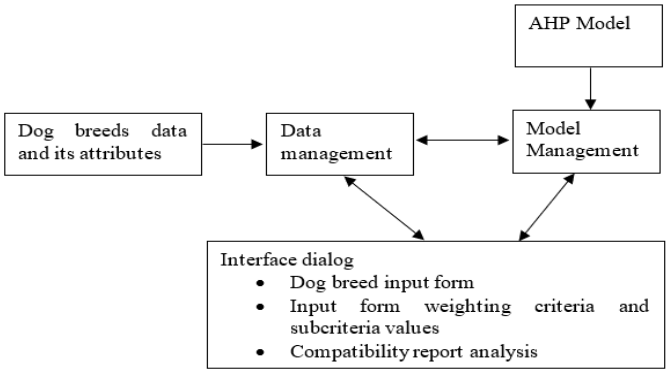


Figure 1: Architecture of DSS for Purebred Dog Selection

2.3. System Modelling

The proposed decision support system model consists of several subsystems, the first is a data management subsystem that utilizes structured data sources as a relational database model, the second is a management subsystem model that utilizes the developed analytic models [11,12]. This analytic model would provide to system an analytical capability by utilizing a

mathematical formula to generate recommendations for the user. Finally, the interface model subsystem is a system component that users used to communicate with the system through an intuitive interface. Figure 1 shows the architectural model of a decision support system for purebred dog selection.

2.4. Data Analysis and Analytic Model

Based on the developed analytic model by processing the data sources, several main criteria and sub-criteria have determined in the selection of dog breeds. These criteria and sub-criteria were obtained based on collected and processed data from a primary data source that was the dog kinology information from the FCI and AKC as the leading world dog registries to obtain standardization information on physical and psychological characteristics of each breed. The collected data was then compared with secondary data sources obtained from dog lovers communities, dog experts, veterinary and kennel owners to know their experience and preferences in selecting and caring for a purebred dog as a result, eight main criteria and seven sub-criteria were formulated with its degree of intensity used in decision making for the selection of purebred dogs as follows:

1. Intelligence: This criterion defines the ability of the dog to obey commands and its tendency to understand new commands. This criterion has two sub-criteria namely obedience and trainability, which defines how good the dog to obey the commands given by its owner at a given time and how good the dog to understand new commands in minimal repetition count, respectively.

a. Obedience

- High: The dog can obey 70 % or more of the command given every time.
- Medium : The dog can obey 30 % or more of the command given every time.
- Low: The dog can obey less than 30% of the command given every time.

b. Trainability

- High: The dog can understand new command with training frequency less than 16 reps.
- Medium: The dog can understand new command with traning frequency 16 to 40 reps.
- Low: The dog can understand new command with training frequency more than 40 reps.

2. Hygiene: This Criterion defines the level of frequency of a dog to produce natural waste that beyond its control like feather loss (shedding) and drooling. This criterion has two sub-criteria namely non-shedding and non-drooling.

a. Non-Shedding

- High: Feather loss rarely occurs.
- Medium: Feather loss occurs periodically/seasonally naturally.
- Low: Feather loss occurs most of the time naturally

b. Non-Drooling

- High: Drooling never or rarely happen..
- Medium: Drooling sometimes occur in certain condition..
- Low: Drooling occurs frequently.

3. Maintenance: This Criterion determines how necessary a specific dog according to the breed type requires special care outside of its daily care, the specific treatment like advanced grooming, specific nail care, and any other special treatment. The intensity degree for this criterion are:

- High: The dog do not need or only require minimal additional care like simple grooming
- Medium : The dog need moderate additional care like full grooming
- Low : The dog needs special care since that affects its hygiene and healthy body. the dog might need advanced grooming, nail trimming, and any other services.

4. Friendliness: This Criterion defines the aggressiveness level of a purebred dog naturally shown against the subject (humans and other animals). Excessive barking toward unfamiliar subject indicates mild aggressive behavior, and severe aggressive behavior has indicated by a tendency to attack the subjects. The intensity degree for this criterion are:

- High: The dog is friendly and generally does not bark if there is no provocative action from the subject.
- Medium: The dog might always bark at any unfamiliar subject.
- Low: The dog has tendency to engage in aggresive behavior including barking and attacking unfamiliar subject.

5. Adaptability: This Criterion defines the level of adaptability of dog breed toward human, other pet, and environment, good adaptability allows the dog to adjust well with the environment and made a good relationship with the adopter and other pets.

a. Adaptation to Human

- High: The dog can adapt well and easy to make new connection with human.
- Medium : The dog can still adapt well but need time to make a new relationship with a human.
- Low : The dog need much time to adapt and build new relationship with human and sometimes the dog may seem uncomfortable until they fully adapt with the human counterparts

b. Adaptation to other pets

- High: The dog highly adaptable and can live along with other pets without problems.
- Medium: The dog can adapt well with a certain pet usually with same species but can behave aggressively to the other one.
- Low: The dog does not like other animal and has a tendency to consider other dog as a rival.

c. Adaptation to environment

- High: The dog highly adaptable to various environmental conditions and has a high tolerance with weather changes.

- Medium: The dog can still adapt well with certain enviromental condition but sometimes need human to help the adaptability process.
- Low: The dog needs a certain environmental condition to live well and very dependant to human to help the adaptability process.

6. Popularity: This criterion defines how popular a breed based on the rankings released by AKC. AKC annually updates the most popular dog ranking data based on genealogy registration, contests, and surveys. For adopters to have a popular dog has benefits such as the ease of joining a large community of dog owners. The intensity degree for this criterion are:

- High: Very popular dog breed, which is a breed that ranks between 1 to 25 according to the latest AKC annual List.
- Medium: moderately popular dog breed, which is a breed that ranks between 26 to 100 according to the latest AKC annual List.
- Low: Less popular dog breed, which is a breed that ranks below 100 according to the latest AKC annual List.

7. Longevity: The life expectancy of a dog breed. According to research, each dog has a lifespan between 5 to15 years depending on the breed [13]. The intensity degree for this criterion are:

- High: The dog breed with an average lifespan above 12 years.
- Medium: The dog breed with an average lifespan between 9 to 12 years
- Low: The dog breed with an average lifespan below 9 years.

8. Impression: Adopters based criteria that express the level of their impression on the certain dog breed subjectively (usually about the dog physically attraction). The intensity degree for this criterion are:

- High: The prospective adopter has a great impression of the breed type.
- Medium: The prospective adopter has a good impression of the breed type.
- Low: The prospective adopter is less interested in breed type.

3. Result and Discussion

3.1. System Design and Evaluation

Based on the designed model, the data of all breeds that have been officially recognized by FCI and AKC were 193 breeds and have been recorded in the database. Table 1 shows some of the breeds available in the system.

Table 1: List of some dog breeds

001	Chihuahua
002	Pekingese
003	Pomeranian
004	Golden Retriever
005	Maltese
006	German Shepherd
007	Border Collie
008	Shih Tzu
009	Labrador Retriever
010	Chow-chow
011	Doberman Pinscher
012	Bulldog
013	Saint Bernard
014	Boxer
015	Bloodhound
016	Shiba Inu
017	Poodle (Toy)
018	Dalmatian

Then from the designed analytical model, each data criteria, sub-criteria, and intensity levels have been structured into the database as shown in table 2 to table 4, there were eight main criteria, which three of them have sub-criteria, finally, there were three degrees of intensity data that will be used for each criterion and sub-criterion.

Table 2: List of Criteria

ID Criteria	Criteria
1	Intelligence
2	Hygiene
3	Maintenance
4	Friendliness
5	Adaptability
6	Popularity
7	Longevity
8	Impression

Table 3: List of SubCriteria

ID SubCriteria	Criteria	Subcriteria
1	Intelligence	Obedience
2	Hygiene	Trainability
1	Maintenance	Non-Shedding
2	Friendliness	Non-Drolling
1	Adaptability	Human
2	Popularity	Animal
3	Longevity	Environment

Table 4: List of Intensity

ID Intensity	Intensity
1	Low
2	Medium
3	High

The system provided several preset (pre-configured set) options that can be used by the user as a prospective adopter for finding the right breed, Besides that users can manually configure the settings tailored to their needs. The preset provided by the system refers to collected data from the FCI and AKC knowledge base in their official website (http://fci.be) and (www.akc.org) plus collaboration with empirical data obtained from the dog lovers community, dog experts, and veterinary. Besides, each preset available will also take the subjective preferences of the user into account by giving the impression criterion certain weight value that will be used in computing to generate recommendations.

There were several pre-configured sets provided by the system that can be used by users in the selection of dog breeds. One

example of a preset provided by the system was namely working-dog preset. This preset was useful for finding dog breed recommendations that have the potential to support humans jobs. This preset has intended to provide recommendations to prospective adopters who need breed types that have great potential to be trained for specific works according to the needs of prospective adopters, for example, dogs that are planned to work as herding dogs for livestock, these dogs must meet the requirement needed as easy to train, independent, easily adaptable, and have good territorial awareness [14]. Another pre-configure set provided by the system wasthe Family-dog preset to help the user to find a dog who can be a great companion for its adopter. The dog expected to be friendly with all family members including children and to other pets such as cats and rabbits. The dogs that are suitable for this role were the dog whose main character is friendly, physically attractive, and have great adaptability to humans and other pets. Both preset certainly have different weighting values and criterion importance ranking, Working-dog preset has higher weighting values for intelligence and adaptability criteria (table 5), while family-dog presets prioritize impression criteria since it is primarily used by an adopter who chooses a dog as a family companion as same as the friendliness criteria which is the second most important criterion for this preset category (table 6). From the criteria weighting matrix, we can get a rank of each criterion based on the eigenvector value. Table 7 and table 8 shows the criteria ranking for working-dog presets and the criteria ranking for family-dog presets respectively.

Table 5: Weighting value matriks for working-dog preset

	Intelligence	Hygiene	Maintenance	Friendliness	Adaptability	Popularity	Longevity	Impression
Intelligence	1	7	4	9	2	5	8	6
Hygiene	0.14	1	0.33	2	0.22	0.5	2	0.5
Maintenance	0.25	3	1	5	0.33	2	4	3
Friendliness	0.11	0.5	0.2	1	0.14	0.25	0.5	0.33
Adaptability	0.5	5	3	7	1	3	6	4
Popularity	0.2	2	0.5	4	0.33	1	3	2
Longevity	0.13	0.5	0.25	2	0.16	0.33	1	0.5
Impression	0.16	2	0.33	3	0.25	0.5	2	1
Total	2.49	21	9.61	33	4.41	12.58	26.5	17.3

Table 6: Weighting value matriks for family-dog preset

	Intelligence	Hygiene	Maintenance	Friendliness	Adaptability	Popularity	Longevity	Impression
Intelligence	1	0.2	4	0.25	3	0.5	2	0.16
Hygiene	5	1	8	2	7	4	6	0.5
Maintenance	0.25	0.125	1	0.14	0.5	0.25	0.33	0.11
Friendliness	4	0.5	7	1	6	3	5	0.33
Adaptability	0.33	0.14	2	0.16	1	0.25	0.5	0.125
Popularity	2	0.25	4	0.33	4	1	3	0.2
Longevity	0.5	0.16	3	0.2	2	0.33	1	0.14
Impression	6	2	9	3	8	5	7	1
Total	14.8	4.38	38	7.08	31.5	14.33	24.83	2.56

Table 7: Criterion importance rank for working-dog preset

Criteria	Eigenvector value	Priority Rank
Intelligence	0.3862	1
Adaptability	0.2407	2
Maintenance	0.1329	3
Popularity	0.0892	4
Impression	0.0623	5
Hygiene	0.0489	6
Longevity	0.0342	7
Friendliness	0.0236	8
	Consistency ratio	2.63%

Table 8: Criterion importance rank for family-dog preset

Criteria	Eigenvector value	Priority Rank
Intelligence	0.3433	1
Adaptability	0.246	2
Maintenance	0.1762	3
Popularity	0.0872	4
Impression	0.0597	5
Hygiene	0.0399	6
Longevity	0.0272	7
Friendliness	0.0204	8
	Consistency ratio	3.69%

The next step after all the criteria and sub-criteria weights have obtained from the preceding process was to perform the computation on all of these weights value to each official breed stored in the database to get the compatibility score of each dog breed. The process of calculating the compatibility value of each alternative has done by weighting the intensity of each criterion and sub-criterion to each dog breed profile based on information stored in the data source. In this example, the subjective preference of prospective adopter was ignored by weighting the intensity of the impression criteria for each breed using the same value. The compatibility result of each breed for Working-dog presets and Family-dog presets were shown in Figure 2 and Figure 3, respectively.

ALTERNATIVE COMPATIBILITY SCORE

Weighting Result | | Summary

Alternative	Intelligence		Hygienity		Maintenance	Friendliness	Adaptability			Popularity	Longevity	Impression	SUM
	Obedience	Trainability	Non-Shedding	Non-Drooling			Human	Pet	Environment				
Pomeranian	0.2455	0.1227	0.0019	0.0368	0.0533	0.0095	0.0618	0.0618	0.0099	0.0892	0.0055	0.0623	**0.7602**
Golden Retriever	0.2455	0.1227	0.0049	0.0148	0.0533	0.0236	0.1541	0.0618	0.0247	0.0892	0.0137	0.0623	**0.8706**
German Shepherd	0.2455	0.1227	0.0049	0.0148	0.1329	0.0236	0.1541	0.0618	0.0247	0.0892	0.0137	0.0623	**0.9502**
Border Collie	0.2455	0.1227	0.0019	0.0059	0.1329	0.0095	0.0618	0.0248	0.0099	0.0358	0.0137	0.0623	**0.7267**
Labrador Retriever	0.2455	0.1227	0.0049	0.0148	0.1329	0.0236	0.1541	0.0618	0.0099	0.0892	0.0137	0.0623	**0.9354**
Doberman Pinscher	0.2455	0.1227	0.0121	0.0059	0.1329	0.0038	0.0618	0.0248	0.0247	0.0892	0.0137	0.0623	**0.8364**
Shiba Inu	0.2455	0.0197	0.0019	0.0148	0.1329	0.0236	0.0618	0.0618	0.0247	0.0358	0.0137	0.0623	**0.6615**
Poodle Toy	0.2455	0.1227	0.0121	0.0368	0.0313	0.0095	0.0618	0.0618	0.0247	0.0892	0.0342	0.0623	**0.7819**
Dalmatian	0.0985	0.1227	0.0019	0.0148	0.1329	0.0095	0.1541	0.0618	0.0247	0.0358	0.0137	0.0623	**0.6531**
Miniatur Schnauzer	0.2455	0.1227	0.0121	0.0368	0.0213	0.0095	0.0618	0.0618	0.0099	0.0892	0.0137	0.0623	**0.7466**
Miniatur Pinscher	0.0985	0.1227	0.0019	0.0368	0.1329	0.0038	0.0618	0.0618	0.0099	0.0358	0.0342	0.0623	**0.6624**
Pappilon	0.2455	0.1227	0.0049	0.0368	0.0213	0.0236	0.1541	0.0248	0.0099	0.0358	0.0342	0.0623	**0.7759**
Rottweiler	0.2455	0.1227	0.0121	0.0148	0.1329	0.0095	0.0618	0.0248	0.0247	0.0892	0.0137	0.0623	**0.814**

Figure 2: Breed compatibility result for working-dog preset

ALTERNATIVE COMPATIBILITY SCORE

Weighting Result | | Summary

Alternative	Intelligence		Hygienity		Maintenance	Friendliness	Adaptability			Popularity	Longevity	Impression	SUM
	Obedience	Trainability	Non-Shedding	Non-Drooling			Human	Pet	Environment				
Chihuahua	0.016	0.0032	0.0245	0.185	0.0204	0.0707	0.007	0.0028	0.0011	0.0872	0.0399	0.3433	**0.8011**
Pekingese	0.0064	0.0032	0.0245	0.185	0.0082	0.0707	0.0174	0.007	0.0011	0.035	0.016	0.3433	**0.7178**
Pomeranian	0.0398	0.0199	0.0098	0.185	0.0082	0.0707	0.007	0.007	0.0011	0.0872	0.0064	0.3433	**0.7854**
Golden Retriever	0.0398	0.0199	0.0245	0.0742	0.0082	0.1762	0.0174	0.007	0.0028	0.0872	0.016	0.3433	**0.8165**
Maltese	0.016	0.0032	0.061	0.185	0.0033	0.0707	0.0174	0.007	0.0004	0.0872	0.016	0.3433	**0.8105**
German Shepherd	0.0398	0.0199	0.0245	0.0742	0.0204	0.1762	0.0174	0.007	0.0028	0.0872	0.016	0.3433	**0.8287**
Border Collie	0.0398	0.0199	0.0098	0.0297	0.0204	0.0707	0.007	0.0028	0.0011	0.035	0.016	0.3433	**0.5955**
Shih Tzu	0.0064	0.0032	0.0245	0.185	0.0082	0.1762	0.0174	0.007	0.0011	0.0872	0.0399	0.3433	**0.8994**
Labrador Retriever	0.0398	0.0199	0.0245	0.0742	0.0204	0.1762	0.0174	0.007	0.0011	0.0872	0.016	0.3433	**0.8287**
Chow Chow	0.0064	0.0032	0.0098	0.0742	0.0033	0.0283	0.007	0.007	0.0011	0.035	0.0399	0.3433	**0.5585**
Doberman Pinscher	0.0398	0.0199	0.061	0.0297	0.0204	0.0283	0.007	0.007	0.0028	0.0872	0.016	0.3433	**0.6624**
Bulldog	0.0064	0.0032	0.0098	0.0742	0.0082	0.1762	0.0174	0.0028	0.0028	0.0872	0.0064	0.3433	**0.7379**
Saint Bernard	0.0064	0.008	0.0098	0.0297	0.0033	0.1762	0.0174	0.007	0.0011	0.035	0.0064	0.3433	**0.6436**
Boxer	0.016	0.008	0.0245	0.0297	0.0204	0.0707	0.0174	0.007	0.0028	0.0872	0.0064	0.3433	**0.6334**
Bloodhound	0.0064	0.0032	0.0245	0.0297	0.0204	0.0707	0.0174	0.007	0.0028	0.035	0.0064	0.3433	**0.5668**

Figure 3: Breed compatibility result for family-dog preset

Based on the results of the compatibility value, the system will display three dog breeds by the highest compatibility value to be used as recommendations for prospective adopters. Figure 4 and Figure 5 shows the three recommended dog breed by the system for working dog presets and family dog presets respectively.

Working Dog Recommendations

Picture	Dog Breed	Total Score
	German shepherd	0.9502
	Labrador retriever	0.9354
	Golden retriever	0.8706

Figure 4: Working-dog recommendation by system

Family Dog Recommendations

Picture	Dog Breed	Total Score
	Shih Tzu	0.8994
	Papillon	0.8882
	Siberian Husky	0.8752

Figure 5: Family-dog recommendation by system

4. Conclusion

Based on research conducted, researchers have designed an analytical model that uses multiple data sources to gather information regarding the physical and psychological characteristics of each dog breed for quantitative analysis without rule out the personal subjectivity of prospective dog adopters. The result was a decision support system that can assist prospective dog adopters in choosing the breed that suits their needs and preferences best by making recommendations based on the dog breed compatibility ranking towards their preference. The final decision remains the prerogative of prospective adopters with all the consequences.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] P. Chyan, "Decision Support System for Selection of Dog Breeds," in 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 343–346, 2018, doi:10.1109/ISRITI.2018.8864327.

[2] S. Barnard, C. Passalacqua, A. Capra, S. Marshall-Pescini, E.P. Previde, P. Valsecchi, "Social behavioral profile of different dog breeds," Journal of Veterinary Behavior, **6**(1), 83–84, 2011, doi:https://doi.org/10.1016/j.jveb.2010.09.004.

[3] M. Nitzschner, A.P. Melis, J. Kaminski, M. Tomasello, "Dogs (Canis familiaris) evaluate humans on the basis of direct experiences only," PloS One, **7**(10), e46880–e46880, 2012, doi:10.1371/journal.pone.0046880.

[4] S. Coren, The Intelligence of Dogs: A Guide to Thoughts, Emotions, and Inner Lives of Our Canine Companions, Atria Books, 2006.

[5] S. Coren, How Dogs Think, First Free Press, Simon & Schuster, New York, 2004.

[6] A. Miklosi, Dog Behaviour, Evolution, and Cognition, 2nd Editio, Oxford University Press, 2016.

[7] A.R. Michell, "Longevit of British breeds of dog and its relationships with-sex, size, cardiovascular variables and disease," Veterinary Record, **145**(22), 625–629, 1999, doi:10.1136/vr.145.22.625.

[8] J. Anderson, M. Krieger, M. Safley, A comparative analysis of alpaca breed type and standards, Northwest Alpaca at Wayback Machine, 2009.

[9] C. Coile, Encyclopedia of Dog Breeds, Barron's Educational series, 2015.

[10] T.. Saaty, Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process, RWS Publication, 2000.

[11] P. Chyan, "Design of intelligent camera-based security system with image enhancement support," Journal of Physics: Conference Series, **1341**, 42009, 2019, doi:10.1088/1742-6596/1341/4/042009.

[12] P. Chyan, "Image Enhancement Based On Bee Colony Algorithm," Journal of Engineering and Applied Sciences, **14**(1), 43–49, 2019.

[13] E. Korec, O. Chalupa, M. Hančl, J. Korcová, M. Bydžovská, "Longevity of Cane Corso Italiano dog breed and its relationship with hair colour," Open Veterinary Journal, **7**(2), 170–173, 2017, doi:10.4314/ovj.v7i2.15.

[14] M.J. Dotson, E.M. Hyatt, "Understanding dog–human companionship," Journal of Business Research, **61**(5), 457–466, 2008, doi:https://doi.org/10.1016/j.jbusres.2007.07.019.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 154-166 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Four-Dimensional Sparse Data Structures for Representing Text Data

Martin Marinov*, Alexander Efremov

Faculty of Automatics, Technical University of Sofia, 1000, Bulgaria

ARTICLE INFO

Article history:
Received: 12 July, 2020
Accepted: 24 August, 2020
Online: 10 September, 2020

Keywords:
Text Classification
Data Cleaning
Data Preparation
Sparse Representations

ABSTRACT

This paper focuses on a string encoding algorithm, which produces sparse distributed representations of text data. A characteristic feature of the algorithm described here, is that it works without tokenizing the text and can avoid other data preparation steps, such as stemming and lemmatization. The text can be of arbitrary size, whether it is a single word or an entire book, it can be processed in the same way. Such approaches to text vectorization are not common in the machine learning literature. This sets the presented encoder apart from conventional text vectorizers. Two versions of the encoding algorithm are described and compared - the initial one and an improved version. The goal is to produce a robust data preparation procedure, capable of handling highly corrupted texts.

1. Introduction

This paper describes a modification of a text encoding algorithm, presented at ICASC 2019. Since the conference proceedings paper [1] was written, there have been further developments and those will be presented here, but first a bit of backstory.

One of the most prolific research teams, who currently experiment with sparse distributed representations (SDR), are Jeff Hawkins, Subutai Ahmad [2] and the people working with them at the private company Numenta. The principle author of this paper found out about the concept of SDR from their work and drew inspiration from their research.

An important thing to note, is that the long term goal of the researchers at Numenta is to distance their developments from conventional machine learning and artificial neural networks. We do not share this goal, in fact we are merely trying to make solutions based on sparse information representation, which are complementary to classic machine learning.

Here we present a type of encoder for text data, which is not described in the research paper on sparse encoders, published by a member of Numenta [3]. Our solution is also different from the Cortical.io SDR text encoder, which is mentioned in point 6 of the Numenta paper (Webber, "Semantic Folding"). We follow the principles described in the conclusions of the Numenta research paper, but the algorithms and implementation are entirely our own. The data structures we use are not binary arrays, we use key-value data structures to implement the representations. Compared to other SDR text encoders, another notable difference in our approach is that we consider characters to be the basic building units of text, not abstract constructs such as words.

This paper focuses on a key enhancement made to our algorithm, presented at ICASC, and the potential benefits it offers. It also covers the development of a benchmark text classification problem, for evaluating the utility of our text encoding algorithms, by comparing them to conventional text vectorizers.

To avoid confusion, our initial algorithm will be referred to as CP (Current-Prior). The modified one, which is the focus of this paper, will be referred to as CPPP (Current-Position-Prior-Position). Their names hint at the way in which they transform strings into sparse distributed representations.

2. Comparing CP and CPPP sparse representations

In brief, CP encoding goes through strings character by character and uses only their sequencing to create a sparse representation of the strings in a two-dimensional encoding space. For a more detailed description of each step, please refer to the ICASC paper. The encoding space is an abstract concept and one of its key features is that the coordinate system uses non-numeric characters, to define the location of points within the space. As the CP encoder moves from character to character, it projects the last

*Corresponding Author, Martin Marinov, e-mail: mu_marinov@abv.bg

character it has read onto the horizontal axis, but it also uses the vertical axis to track prior characters. This is why the algorithm is called current-prior. The result is a specific pattern of points in the encoding space, which is highly dependent on the symbol arrangement of the strings being processed.

	-	a	b	c	d	e	f	g	h
-									
a									
b									
c									
d						de		dg	
e					ed	ee		eg	
f									
g						ge			
h									

Figure 1: edge (CP encoded)

		d						e						g					
		0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
d	0																		
	1																		
	2											de					dg		
	3																		
	4																		
	5																		
e	0																		
	1			ed								ee					eg		
	2																		
	3																		
	4																		
	5																		
g	0																		
	1																		
	2																		
	3											ge							
	4																		
	5																		

Figure 2: edge (CPPP encoded)

The encoding patterns can be illustrated in various ways, with one of the more human-readable ones being tabular representation, as seen on Figure 1. If the representations are being processed by a computer program, however, there are more suitable data structures, which could be used. For example, the CP encoded word shown in Figure 1 can be stored as a set of character pairs: {' e', ' d', ' g', ' ', 'ed', 'eg', 'ee', 'e ', 'dg', 'de', 'd ', 'ge', 'g '}.

CPPP encoding works in a similar way, but in addition to character sequences it also factors in the position of the characters. In other words, the sparse encoding space has two additional sub-axes, which track the positions of the last read symbol and the position of prior symbols. Even though this defines a four-dimensional encoding space, it can still be represented in tabular form (Figure 2). This is the key modification to the CP algorithm and it is noted in the name (Current-Position-Prior-Position).

In theory, the readability of these tables does not diminish, because the encoded strings are sparse patterns by definition. The repetition of the sub-axes below the main axes doesn't make the tabular representation harder to interpret. It still looks like a grid with seemingly random non-empty elements scattered throughout. There is actually nothing random about the distribution of the non-empty elements, however. It is strictly based on the encoded string.

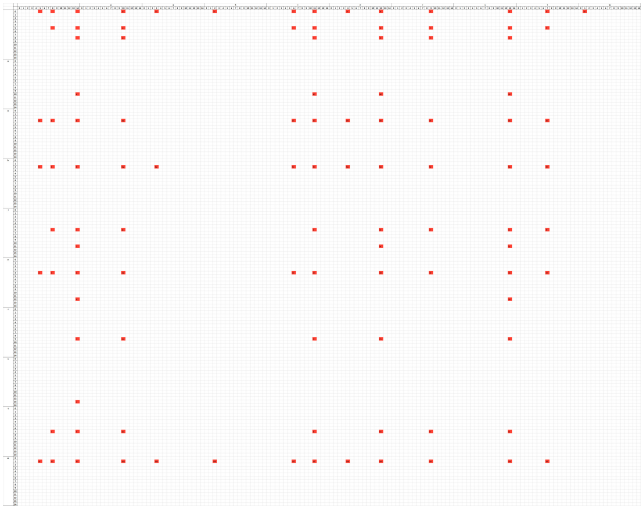


Figure 3: The phrase "when it rains", encoded with CPPP

Difficulties could arise in practice, because of the size of the string being encoded. The sparse representation is still logical and consistent, it is just not possible for a human to comprehend what it means at a glance (Figure 3).

However, these algorithms are meant to be utilized by computers, not people. Machines are far better at processing large amounts of information. The ability to represent the encodings in a human-readable format is useful mainly to the developers, as a debugging tool and a way to illustrate the working principles of these algorithms.

3. Pseudocode implementation

3.1. Input-output

strIn - The string which has to be transformed into a sparse representation. It can be of arbitrary length.

SR - The output from the encoding procedure, i.e. the sparse representation of the input. It is just a key-value data structure. We used python dictionaries, but any other dictionary-like data structures from other languages can be used to implement the algorithms.

winSize - A control parameter, which sets the maximum allowed distance between characters, when forming character pairs.

For example, when encoding an entire book, it would make little sense to form pairs from characters at the start of the book and characters at the very end of the book. That would just create more character pairs (points in the encoding space) than is necessary and significantly increase the size of the sparse representation.

We used winSize = 20, i.e. we told the program not to form points from characters, which were separated by more than 20 characters.

3.2. CP encoding algorithm

Algorithm 1: Current-Prior

```
Result: SR
Receive strIn;
Initialize dictionary object SR;
prior_pos = 0;
while prior_pos < length of strIn do
    current_pos = prior_pos + 1;
    while current_pos < length of strIn do
        chrPair = Concat(strIn[prior_pos], strIn[current_pos]);
        SR[chrPair] = 1;
        current_pos += 1;
    end
    prior_pos += 1;
end
```

3.3. CPPP encoding algorithm

The keys in the CPPP sparse representation key-value data structure are derived in the same way as the ones produced by CP encoding. However, the values are not just a single number. The values are lists of numbers, specifically the positions of the characters, which produced the keys of the dictionary.

Algorithm 2: Current-Position-Prior-Position

```
Result: SR
Receive strIn;
Initialize dictionary object SR;
prior_pos = 0;
while prior_pos < length of strIn do
    current_pos = prior_pos + 1;
    while current_pos < length of strIn do
        chrPair = Concat(strIn[prior_pos], strIn[current_pos])
        if chrPair not in keys of SR then
            SR[chrPair] = list(empty list, empty list);
            append current_pos to list SR[chrPair][0];
            append prior_pos to list SR[chrPair][1];
        end
        current_pos += 1;
    end
    prior_pos += 1;
end
```

4. Advantages and disadvantages of CPPP

When processing repeating syllables and characters, the CP algorithm projects them in the same points of the sparse encoding space. In contrast, CPPP ensures that each repeating symbol will be projected to a different point in the encoding space. A direct consequence of this is that CPPP encoding can handle strings with arbitrary lengths, whereas CP cannot.

It is possible to saturate the two-dimensional encoding space used by the CP algorithm, if a sufficiently long string is processed. CPPP can transform large strings, without decreasing the sparsity of their representations in the four-dimensional encoding space. The significant advantages of this are:

- **Reversible encoding**. It is possible to turn the sparse representations back into strings, without information loss.
- **No need for tokenization**. Text documents (data) can be processed as a single, long, uninterrupted sequence of characters. There is no need for splitting them by delimiters, such as whitespace, commas and such.
- **Increased capacity**. Due to the introduction of the two additional dimensions, for tracking character positions, it is possible to compare not only tokens (words), but entire sentences, phrases, paragraphs and documents.
- **Precise string comparison**. CPPP allows not only the quantification of similarity between strings, but it also makes it possible to determine which specific parts of the strings correspond to each other and which ones differ.

Despite the advantages, the modified algorithm has a notable drawback: the simple methodology for comparison of CP sparse representations is not applicable to CPPP encoded text.

To illustrate the reason, several words encoded by the two algorithms will be examined. The chosen words are **tree**, **trim** and **tree-trimmer**, because these words are short, some of them have double characters and the third one is a compound word, which includes the other two.

CP encoding (Figures 4 to 6) does not represent the following things properly:

- The double **e** in **tree** and the double **m** в **trimmer**. In both cases, the corresponding points for these characters in the encoding space overlap. For the word **tree**, the points are **re** and **te**. For **trimmer**, the points are **im**, **rm** and **tm**.
- The repetition of **tr** in **tree-trimmer** (Figure 6). Again, the representations of the two different instances of this character sequence overlap completely in the two-dimensional encoding space. This means that the initial string cannot be recovered, by using only its sparse representation.

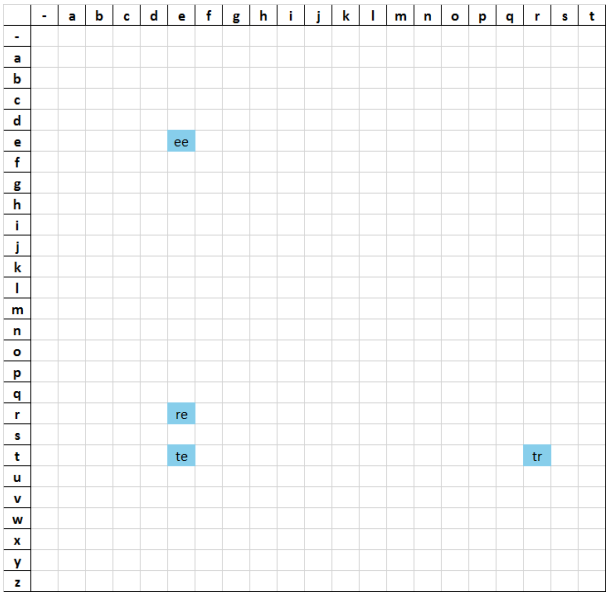


Figure 4: tree (CP encoded)

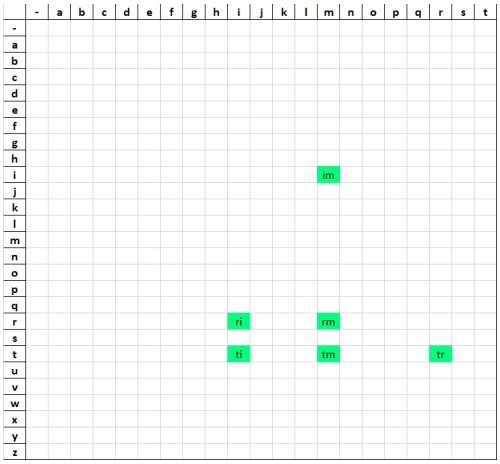


Figure 5: trim (CP encoded)

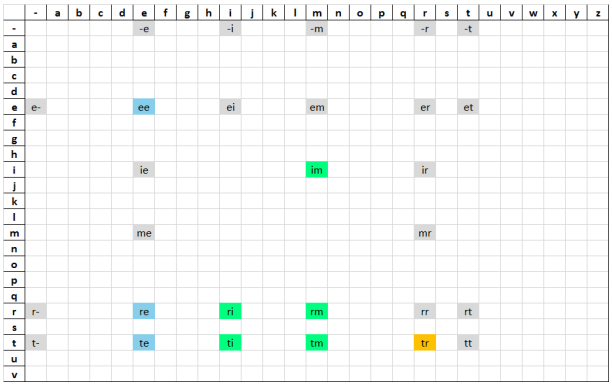


Figure 6: tree-trimmer (CP encoded)

On the other hand, CPPP encoding creates mirror points, when representing repeating character sequences. An example of this can be seen on Figure 7, arrows have been drawn to bring attention to the points in question. They indicate, that the letter **e** is present more than once in the word **tree**. The fact that sectors **re** and **te** have more than one non-empty value is informative, although the algorithmic processing of this information is not as trivial, as the procedure used to compare sparse patterns in CP encoded strings.

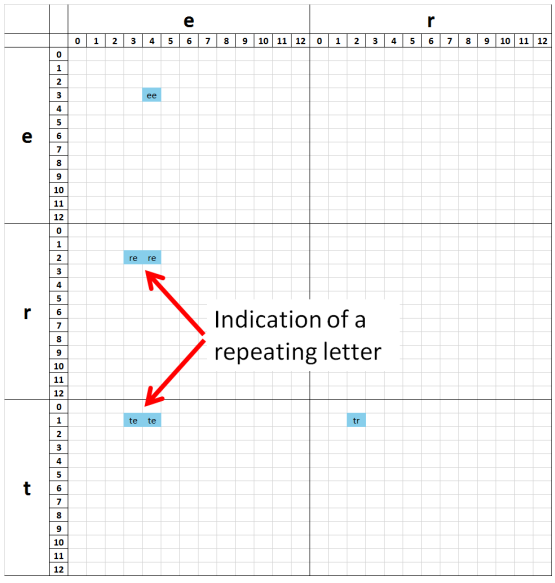


Figure 7: tree (CPPP encoded)

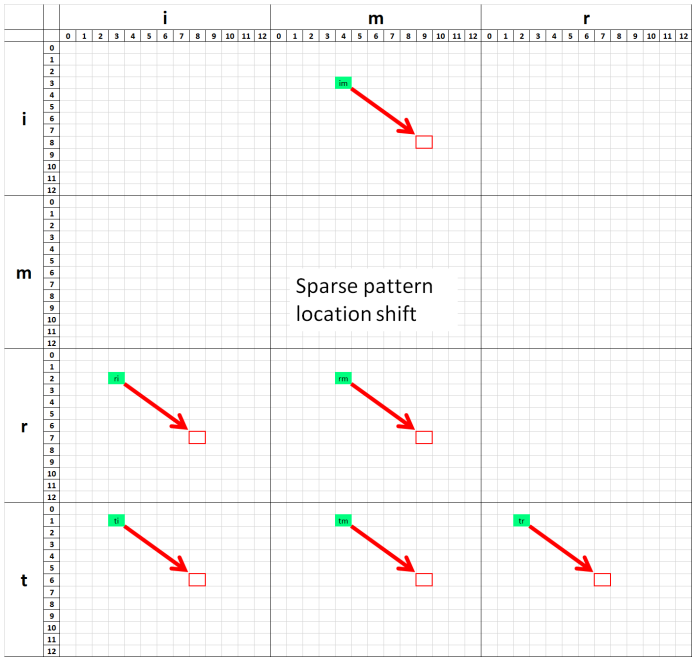


Figure 8: trim (CPPP encoded)

A second complication inherent in CPPP encoding is illustrated on Figures 8 and 9. The points which represent the word **trim** have shifted their positions in the sparse space. This is to be expected, in the third string there is another word before **trim**, which causes a translation of the sparse pattern to the right. This is logical and informative, but again, the algorithmic analysis and quantification of such relations is more complex than the one employed for CP sparse representations of text.



Figure 9: tree-trimmer (CPPP encoded)

These two characteristics of the modified encoding space (mirror points and non-stationary sparse patterns) are the things that give it the aforementioned advantages. However, there is no analogue for them in the two-dimensional encoding space, which means the comparison method developed for CP sparse representations is not applicable to four-dimensional encodings of strings. To utilize the full potential of CPPP encoding, other comparison techniques have to be used.

5. Methods for comparing CPPP sparse representations

In theory, it is possible to feed the encoded strings into artificial neural networks and have them learn to recognize the same types of patterns, which people can notice in small sparse representations. However, in recent years the European Union has begun to implement legal frameworks and guidelines for the use of artificial intelligence, which emphasize the need for algorithmic transparency (see the publications from 2019 by the EPRS, such as "A governance framework for algorithmic accountability and transparency").

It is the opinion of the authors, that the more society delegates decision making to machines, the more the legal regulations and requirements for decision transparency will increase. Conventional ANN are black-box approaches to modelling data and as such, the authors try to avoid relying on them, in favor of more algorithmically transparent models. In addition, it is possible to logically and methodically explain why each element of a CPPP sparse representation exists in the encoding space.

In order to make use of the encoded data, however, specialized procedures have to be developed, for quantifying similarities between encoded strings, as well as processing larger units of written information, such as documents, chapters in books etc. To be more specific, text classifiers based on the bag-of-words approach require accurate word/token counts, in order to build quality TF-IDF representations of textual data. Other machine learning approaches, such as n-gram modelling, require tracking of token sequences as well, in order to take into account the context in which individual words are used in the text.

The methods for distinguishing and locating words in CPPP encoded corpora are currently under development, their implementations could change. However, brief examples will be given, to showcase possible ways of carrying out basic operations with CPPP encoded text.

5.1. Basics of locating words in CPPP text representations

The text is CPPP encoded as a single, uninterrupted string. In addition, the computer has access to a machine-readable dictionary, which consists of separately encoded words. Note the different way of encoding the data and the dictionary. Unlike common practice, tokenization of the dataset is avoided, the text is kept intact and it is CPPP encoded as a whole. This is because it is not known which parts of the text correspond to which words, before some kind of analysis is done. This approach avoids the presupposition, that the text being processed is well formatted and can easily be parsed with a few simple delimiting characters.

The dictionary, however, is a collection of distinct words, by definition, so it is divided into separately encoded strings. It is also possible to curate the dictionary manually and ensure the entries are properly separated, something which usually isn't practicable with the dataset.

5.2. Word search based on maximizing pattern overlaps

For each element (word, token, phrase) in the machine-readable dictionary, the following steps are carried out: finding all potential locations of the search term within the text and filtering the possible results, to leave only the most consistent matches.

5.2.1. Initialization of a character weight vector W.

The purpose of W is to indicate which parts of the text correspond to the dictionary entry, which is under consideration during the entry enumeration. It is initialized as an all-zero vector, with a length equal to the number of characters in the text being analyzed. The first element of W indicates how many times the first character takes part in forming character pairs, which are common to both data and dictionary sparse representations. The second element of W indicates the same for the second character of the text. The third element covers the third character etc.

5.2.2. SPOQ procedure - pseudocode implementation.

The current implementation of the Sparse Pattern Overlap Quantification Procedure, or SPOQ for short, is described here (algorithm 3). It is used to find similar patterns in the encoded data. SR_1 is the sparse representation of the data, SR_2 is a sparse representation of a specific dictionary entry.

Algorithm 3: SPOQ

```
Result: W_SPOQ
Initialize W vector with all zero elements;
commonKeys = Intersection(keys of SR_1, keys of SR_2);
for each com in commonKeys do
    crnt_pos1 = SR_1[com][0];   prev_pos1 = SR_1[com][1];
    crnt_pos2 = SR_2[com][0];   prev_pos2 = SR_2[com][1];
    delta1 = crnt_pos1 - prev_pos1;
    delta2 = crnt_pos2 - prev_pos2;
    for each b in delta2 do
        proximity = abs(delta1 - b);
        W[ crnt_pos1[ index of min(proximity) ] ] += 1;
        W[ prev_pos1[ index of min(proximity) ] ] += 1;
    end
end
```

5.2.3. SPOQ example.

If we look at the sparse representations shown on Figures 8 and 9, their encoding spaces can be divided into sectors (denoted S_T and S_D respectively), by using the primary axes (the ones that track character sequencing). These sectors correspond to the common keys between the two sparse representation data structures. For those sectors, which are not empty and present in both representations, the distances Δp_T and Δp_D are calculated. They indicate how far apart specific pairs of characters are in both representations. For details, see Table 1.

The example shown in Table 1 assumes that the word **trim** is a word from the machine-readable dictionary and **tree-trimmer** is the text. The elements of the weight vector are incremented, based on the closest matches between Δp_T and Δp_D in each sector. This is why indices 6,7,8 and 9 of W will have the highest non-zero values. They correspond neatly to the location of the word **trim**, within the string **tree-trimmer**.

Note that this is a relatively simple procedure and it can produce false positives. The fact that index 1 and 2 also get incremented once is an example of the limitations of SPOQ. In addition, there is no guarantee that all relevant characters will be matched properly.

Table 1: Principle of operation of the SPOQ procedure

S_T	Δp_T	S_D	Δp_D	Incrementations of W
ri	3 – 2 = **1**	ri	8 – 2 = 6	
		ri	8 – 7 = **1**	W_8 += 1 and W_7 += 1
ti	3 – 1 = **2**	ti	8 – 1 = 7	
		ti	8 – 6 = **2**	W_8 += 1 and W_6 += 1
im	4 – 3 = **1**	im	9 – 8 = **1**	W_9 += 1 and W_8 += 1
		im	10 – 8 = 2	
rm	4 – 2 = **2**	rm	9 – 2 = 7	
		rm	9 – 7 = **2**	W_9 += 1 and W_7 += 1
		rm	10 – 2 = 8	
		rm	10 – 7 = 3	
tm	4 – 1 = **3**	tm	9 – 1 = 8	
		tm	9 – 6 = **3**	W_9 += 1 and W_6 += 1
		tm	10 – 1 = 9	
		tm	10 – 6 = 4	
tr	2 – 1 = **1**	tr	2 – 1 = **1**	W_2 += 1 and W_1 += 1
		tr	7 – 1 = 6	
		tr	7 – 6 = **1**	W_7 += 1 and W_6 += 1
		tr	12 – 1 = 11	
		tr	12 – 6 = 6	

5.2.4. *Filtration of weak matches*

The SPOQ search procedure described in points 5.2.2 and 5.2.3 is supposed to generate rough estimations of word positions, within a given text. False positives and overlaps between the weight vectors of multiple dictionary entries are to be expected. Currently, it is assumed that these can be handled with adequate filtration techniques.

Regarding how the filtration of each weight vector is done, the code for the SPOQ procedure was described in point 5.2.2, it is the one which produces the initial W vectors, denoted as W_{SPOQ}. Point 5.2.5 covers the code for the filtration steps, which further modify W_{SPOQ}.

5.2.5. *Pseudocode of the filtration procedures.*

Algorithm 4: Frequency filtration

```
Result: W_F
W = W_SPOQ;  #input
W[W == 1] = 0;
N = length of W – 2; cp = 1;
while cp <= N do
    if W[cp-1] == 0 and W[cp] != 0 and W[cp+1] == 0 do
        W[cp] = 0;
    end
    cp += 1;
end
```

Algorithm 5: Match length filtration

```
Result: W_L
W = W_F;  #input
likely_loc = integer sequence from 0 to length of W;
Keep only the elements of W where W > 0;
W_filtered = zero vector of length W;
i = 0;
j = 1;
while j < length of W do
    if likely_loc[j] - likely_loc[j-1] != 1 then
        substring_size = likely_loc[j-1] - likely_loc[i];
        if substring_size > 2 then
            # This substring is a good match, it is long enough
            W_filtered[ likely_loc[i]: likely_loc[j-1] ] = 1;
        else
            # Bad match, substring is too short
        end
        i = j
    end
    j += 1
end
```

5.3. *Visualization of SPOQ and W filtration.*

To visualize the values of the character weight vector, as it is altered by the filtration steps, we utilize custom plots. In addition to W, they display the actual text under evaluation, so that we can quickly determine if the values of W are as they should be.

The layout of the custom plots is described in Figure 10. For specific example with data, refer to Figures 11 to 14 and point 5.3.1, which describes what those figures illustrate.

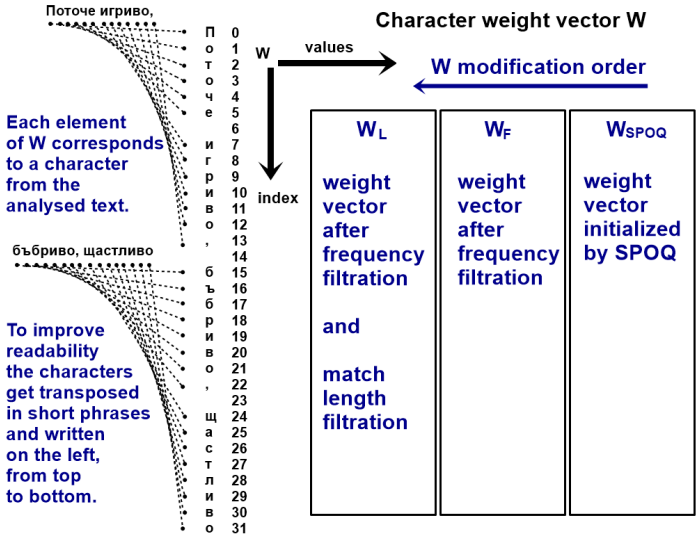


Figure 10: Description of custom plot layout

5.3.1. *Example of SPOQ and W filtration used on text*

To illustrate the filtration process, the Bulgarian poem "Поточе" (Stream), written by Elin Pelin, will be used as an example dataset. When searching for the word песни (songs), the SPOQ procedure is used to obtain an initial weight vector W_{SPOQ}, shown on Figure 11.



Figure 11: search for **песни**



Figure 12: search for **песен**

The largest peak in W_{SPOQ} is roughly equal in size to the length of the search word, about five in this case. As expected, the biggest peak is roughly where the word is in the poem, but there are many other smaller peaks scattered around it. This is why additional filtration steps are necessary.

Many of the individual elements in W_{SPOQ} are equal to 1 or are surrounded by zeros on both sides. Using the frequency filtration procedure, we zero-out all of these and get a clearer indication of where the search-word is, as shown on Figure 11, subplot W_F. Unfortunately, some false positives still remain.

An additional filtration step can be applied, this time targeting the lengths of the matches. Presumably longer sequences of non-zero values indicate better matches. If we zero-out all matches shorter than 2 characters, we obtain the result seen on Figure 11, subplot W_L. This is indeed where the word is in the document.

5.3.2. Necessary refinements to SPOQ and filtration.

There is a caveat to the simple filtrations described here. Their thresholds are static hyper-parameters, meaning they were set by a person at some point and are not necessarily a good fit in all possible word search cases.

For an example of this, refer to Figure 12, which shows the search for the word песен (song). This is the singular form of the word and a more likely entry in a properly curated dictionary. In this case, the match is weaker and a small increase in the threshold of the length filter (from 2 to 4) could zero-out the entire weight vector, thereby suppressing the one relevant match in the dataset.

The filtration thresholds have to be dynamically adjustable by an automated subroutine. Manual adjustment would not be practicable, considering the number of word searches and comparisons that have to be done in an adequately sized corpus.

A single search in the encoded text could produce multiple results, if the search-word is present in the data more than once, as seen on Figure 13. However, multiple results for a single search could also contain false positives, due to the nature of the algorithm.

Figure 14 shows an example of false positives making it through both filtration steps. The reason for this is that the search-word бистро (pure) is contained nearly unchanged in the word сребристо (silvery). This is why the filtered weight vector has 4 peaks, instead of just 2, which is the correct number of occurrences of the search-word. Finding similar character sequences is the intended behavior of the program, but they aren't always a correct match.

Work is still ongoing on the development of dynamic filtration, capable of resolving the issues caused by false positives and ambiguities arising from overlaps between the weight vectors of different dictionary entries. Formulating the problem as an optimization procedure is under consideration, the idea being that this way it will be possible to automatically reach a single interpretation of a given corpus. The aim is to produce a data preparation module, based on CPPP encoding. It should be capable of replacing conventional data preparation steps, such as tokenization, stemming, lemmatization, as well as give sufficiently accurate word counts and position estimations.

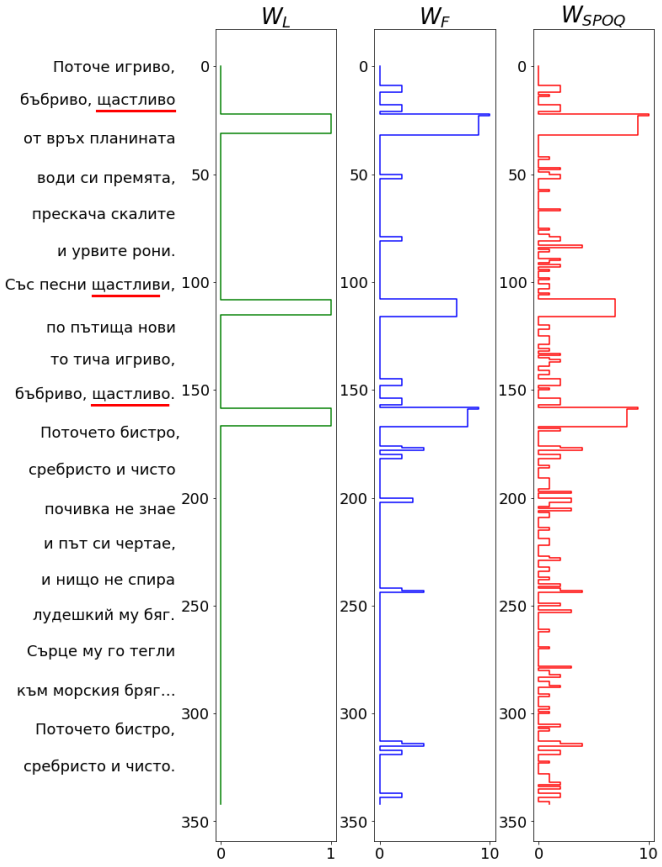


Figure 13: search for **щастливо**



Figure 14: search for **бистро**

6. Text classification benchmark

In order to conduct a more rigorous evaluation, of the performance and quality of CPPP data preparation, a classification problem has been defined, by using a dataset from the LSHTC challenges [4]. Specifically, the small wikipedia dataset from LSHTC3 was used, since it was the only challenge in which the raw text was provided.

The dataset has 456886 documents, each one containing only abstracts of wikipedia articles. The documents are labeled, although the labels are not human readable, they are merely integers representing a specific topic covered in each document. There is also a complex label hierarchy, however for the purposes of our benchmark problem, we do not take it into account. We only use a subset of the entire corpus, which contains documents with a single label. In addition, we alter the labels to reduce everything to a binary classification problem. This way we can train a simple model and focus on the data preparation part of the workflow. The aim is to make changes only in the data preparation stage, the text classifier itself will always be the same, with fixed hyper-parameters.

6.1. Subsetting the dataset and forming binary labels

Looking only at the documents with a single label, the top 10 most frequently occurring label codes are listed in Table 2. Next to the codes are descriptions, based on reading and evaluating a dozen of the documents associated with a given label.

Table 2: Top 10 most used labels

Label	Documents	Label Description
167593	5277	Movies
347803	1470	Businessmen and inventors
324660	1433	Historic buildings
283823	1359	Writers and journalists
284433	1310	Journalists
272741	1006	Film actors
395447	999	Screenwriters
114873	945	Plants
352578	846	Novelists
93043	762	Painters

Label 167593 was chosen as the positive class (the one we want to predict with the model). The decision was based solely on the number of documents tagged with 167593. To create the benchmark corpus, all documents with the chosen label were selected, and reassigned to class 1. An additional 4000 randomly chosen documents, from the remaining 451609 in the dataset, were assigned to class 0, regardless of what their labels were originally. This is how the initial classification of the documents was reduced to a binary classification problem.

In addition, a variable was added to the subset, which randomly separates it into two roughly equal in size parts. This is used to perform a training-validation split, for when the text classification model is trained on the subset.

The only preprocessing of the raw text was the removal of punctuation and the substitution of numbers with the placeholder <NUMBER>. Other preprocessing steps, such as named entity

recognition, have not been implemented, although they are under consideration.

Note that for statistical purposes, this procedure was repeated 30 times, to produce 30 distinct subsets, each with different documents assigned to the negative class 0 and different training-validation splits. The number 47 was used as a seed for generating 30 random numbers, which in turn were used as seeds when the subsets were created. This ensures repeatability of the process. The pseudocode for creating subsets is included (Algorithm 6).

Algorithm 6: Procedure for subsetting the LSHTC3 dataset

Result: Document subsets

```
# Parameters for subset generation
ran_stat = 47;              # keep constant, used as seed
num_of_subsets = 30;        # Can be varied
N_neg = 4000;               # Can be varied
train_test_ratio = 0.5;     # Can be varied

Load dataset (Dw)
Separate docs for positive and negative class (Dn and Dp);

# Generate a list of random integers
rand_int_list = generate_random_integers(
    max value = 1000000,
    length = num_of_subsets,
    seed = ran_stat
);

for each rand_int in rand_int_list do
    Randomly sample (seed = rand_int) N_neg documents
    from Dn;
    Append Dn_sample to Dp to get subset Ds;
    Remove punctuation in Ds;
    Replace numbers with <NUMBER> in Ds;
    Add train-test separation flag to Ds(seed = rand_int);
    Save the subset Ds as "subset_<rand_int>.csv";
end
```

6.2. Corrupting the text

In order to test whether CPPP encoding can provide more robust data preparation than conventional techniques, each of the created subsets had its documents subjected to varying degrees of text corruption. Note that this only affects the documents used for model validation, the training documents were untouched. The types of text corruption used are:

- D – character deletions. Example: **word** becomes **wrd**.
- I – character insertions. Example: **word** becomes **wyozrds**.
- B – character blurring. Repeating one or more character in the string many times. Example: **wworrrd**.

These three basic types of corruption can be combined together in 7 meaningful ways (___, __I, _D_, _DI, B__, B_I, BD_, BDI), if we allow for the possibility that one or more of them can be inactive. The inactive ones are indicated by an underscore.

In addition, it is possible to regulate the number of affected words in a document, as well as the severity of text corruption. Table 3 shows examples of the codes used to denote corruption type and severity and explains how to interpret them.

Table 3: Corruption code examples

Corruption code	Interpretation
___ 000 000 0011528	No corruption, use the subset derived from seed 11528.
__I 001 005 0367540	Insert extra characters, at most 1 per word, for 5% of the words in each document. Use the subset derived from seed 367540.
B_I 009 001 0815602	Blur and insert characters, at most 9 per word, for 1% of the words in each document. Use the subset derived from seed 815602.

Generating the corrupted subsets is done automatically by the benchmark application, which enables the systematic production of thousands of datasets, with varying degrees of corruption. This procedure has a longer code implementation than the algorithms described so far, which is why it has been split into two parts for ease of readability.

The inner word corruption procedure (Algorithm 8) is the one, which actually changes the words/tokens. The outer corruption procedure (Algorithm 7) is the one which tokenizes the documents and selects which words/tokens in which documents to corrupt. The outer corruption procedure also cycles through all test documents and generates different corruption possibilities, depending on three parameters:

- CT – corruption type (see the start of 6.2);
- NCA – number of characters to alter per word;
- NWD – number of words to alter per document.

Note that the user does not simply specify the values of these 3 parameters, the user provides several valid values for each of them. The benchmark application determines all relevant parameter combinations by itself.

Algorithm 7: Outer word corruption procedure

Result: Corrupted document subset

```
Load subset file;
Determine which documents are for training and testing;
Extract rand_int from the subset name, it will be used as a
seed for the random number generator;
ParamCombinations = VaryParameters(
    valid_values_for_CT,
    valid_values_for_NCA,
    valid_values_for_NWD
);
for each parameter combination in ParamCombinations do
    for each docTxt in documents for testing do
        Tokenize docTxt;
        indeces_of_words = randomly select words to alter;
        for each wrd_ind in indeces_of_words do
            Call the inner word corruption procedure;
        end
    end
end
```

Algorithm 8: Inner word corruption procedure

Result: A single corrupted word/token

```
Read parameters from the outer procedure;
Read docTxt and wrd_ind (indexes the target word);

Randomly select only one type of text corruption to use;

# Select positions in the word for corrupting randomly
max_index = length of doc_txt[wrd_ind] – 1;
wrd_seed = rand_int + max_index +
Unicode code point of first character +
Unicode code point of last character;
 if max_index == 0 do
   positions = list with one element (index 0);
 Else
   positions = generate_random_integers(
       from 0 to max_index,
       length = chars_to_alter,
       seed = wrd_seed
   );
 End

 if corruption_type is D do
    if length of positions > 1 do
       Leave only unique values in positions list;
       Sort positions list in descending order;
    End
    for each pos in positions do
       delete character at pos in doc_txt[wrd_ind];
    End
 End

 if corruption_type is I or B do
    if corruption_type is I do
       if length of positions > 1 do
          Sort positions list in descending order;
       End
       seed = wrd_seed;
       chars2ins_length = chars_to_alter;
       chars2ins = random Unicode code points (32 to 126);
    End

    if corruption_type is B do
       chars2ins = empty list
       for each pos in positions do
          Append Unicode code point of
          doc_txt[wrd_ind][pos] to chars2ins;
       End
    End
    i = 0
    while i < length of positions do
       insert chars2ins[i] into positions[i] of doc_txt[wrd_ind];
       i += 1
    End
 End
```

6.3. Predetermining computation requirements

If processing time or memory is an issue, the total number of corrupted datasets Dc can be calculated beforehand with an equation (1).

$$D_c = N_s(1 + CA_cA_w) \quad (1)$$

where:

- N_S – Number of initial subsets.
- C – The number of meaningful combinations of the basic types of text corruption.
- A_C – The number of character alteration levels.
- A_W – The number of word alteration levels.

As mentioned earlier, we prepared an example with 30 distinct subsets from the full corpus. If we want to alter 1, 3 and 9 characters per word for 1, 5, 20, 50, 80 and 95 percent of the words in each document, then:

$A_C = |\{1, 3, 9\}| = 3$

$A_W = |\{1, 5, 20, 50, 80, 95\}| = 6$

Thus, the application would produce 30*(1+7*3*6) = 3810 distinct subsets, all with varying types and severity of text corruption, within the desired ranges.

6.4. Training the text classifiers

The Python module scikit-learn was used, specifically the TfidfVectorizer object for feature extraction and the RandomForestClassifier object for creating the text classification models. A distinct model is trained on each of the subsets. Their hyper-parameters are fixed, so any change in prediction quality is due entirely to the dataset, nothing else. The feature extractor and model were trained only on the training data, the validation data was only used to generate class predictions.

The hyper-parameters of the TfidfVectorizer are as follows:

- stop_words = 'english'
- analyzer = 'word'
- vocabulary = None
- binary = True
- max_df = 1.0
- min_df = 1
- use_idf = False
- smooth_idf = False
- sublinear_tf = False

The hyper-parameters of the random forest classifier were the default ones, the only parameter we explicitly specified was random_state. This was set equal to the final number in the corrupted subset names (see Table 3). This number is the seed used to generate the subset. Using it as the seed for the model ensures that the random forest algorithm will produce consistentl results for each subset file, no matter how many times the benchmark is executed.

6.5. Evaluating the benchmark results

As expected, models evaluated with uncorrupted or lightly corrupted datasets perform very well, models tested with

moderately corrupted documents perform worse and the ones tested with text with severe corruption perform the worst of all.

The full table of classification results is 3810 rows, too large to be included in this paper. Only a snippet is included (Table 4) to show what the format of the output looks like. The full table is summarized with charts in point 6.6.

Table 4: Excerpt from the classification results table

CT	NCC	PWC	Seed	TP	TN	FP	FN
___	0	0	189191	2525	1891	107	95
BDI	3	1	265329	2543	1844	136	96
BDI	3	5	11528	2422	1800	213	169
BD_	1	20	265329	1303	1865	1376	75
BD_	1	50	11528	2387	1247	248	722
B_I	3	95	265329	269	1886	2410	54
B_I	9	1	11528	2520	1866	115	103
_D_	9	95	11528	2380	1436	255	533
_D_	9	5	889991	2513	1905	141	80
_D_	9	50	616169	1321	1880	1327	90
_DI	9	20	837510	1459	1889	1170	82
__I	1	5	998096	2482	1868	146	127
__I	1	20	11528	2348	1826	287	143

Table header descriptions:

- CT – text corruption type.
- NCC – Number of characters altered per word.
- PWC – Percentage of words affected in each test document.
- Seed – The random number generator seed, which was used to create the initial subset and to alter the text.
- T_P – true positives.
- T_N – true negatives.
- F_P – false positives.
- F_N – false negatives.

6.6. Visualizing the benchmark results

In order to visualize the classifier results in a more compact manner, two things were done with the full results table. First, all T_P, T_N, F_P and F_N were averaged and grouped by CT, NCC and PWC. This effectively combines the results for all of the 30 distinct seeds/subsets which were used. The aggregated results table is reduced to only 128 rows as a result of this summarization.

The second step is quantifying the quality of the classification models. Two metrics were calculated, the first one of which is the widely used F_1 score (4), which is just the harmonic mean of the Precision (2) and Recall (3) metrics. The other quality metric is the Matthews correlation coefficient MCC [5], which is specifically designed to give more reliable evaluations for binary classification results (5).

$$P = \frac{T_P}{T_P + F_P} \quad (2), \quad R = \frac{T_P}{T_P + F_N} \quad (3), \quad F_1 = 2\frac{PR}{P + R} \quad (4)$$

$$MCC = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \quad (5)$$

The resulting charts are shown on Figures 15 to 28. Note that the Percent of Words Changed per test document (PWC) is on the x-axis, the classification quality metrics are on the y-axis and there is a separate line for each of the 3 levels of word corruption used (1, 3 and 9 character alterations per word).

From the charts it is evident, that there is a consistent downward trend of model quality versus PWC, regardless of the way in which the datasets were corrupted. An interesting thing to note is that MCC is more sensitive than the F_1 score. This, combined with the fact that it gives reliable results, even when the class sizes are imbalanced, makes it the preferred model evaluation metric by the authors.

A detailed comparison between the many text corruption approaches is beyond the scope of this paper. What matters is that the overall trends are the same, regardless of the exact manner in which the test texts were damaged.

The intent here is to describe the benchmark classification problem, which will serve as an evaluation tool for the quality of the custom data preparation tools, based on CPPP encoding. The experimental setup might change (more granulated levels of PWC, more subsets etc.). Once the custom data preparation solution is more mature, it will be evaluated alongside the conventional feature extraction objects from the nltk package in Python.

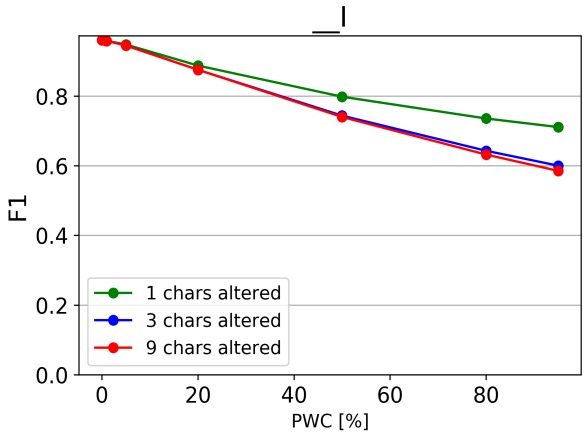


Figure 15: Effect of character insertions on classification quality

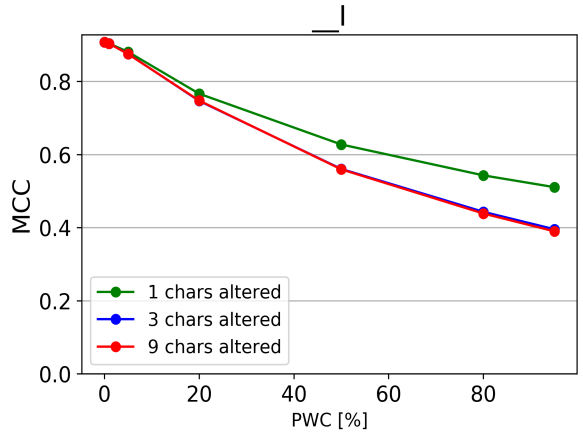


Figure 16: Effect of character insertions (MCC)

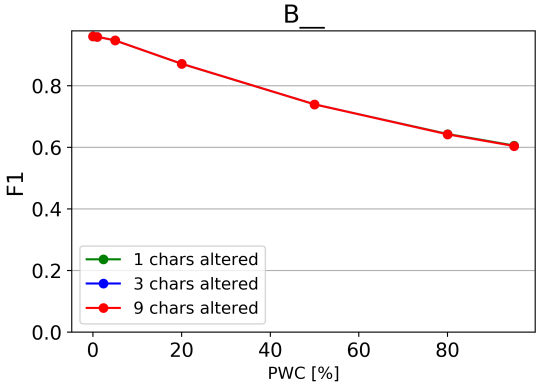


Figure 17: Effect of character repetitions on classification quality

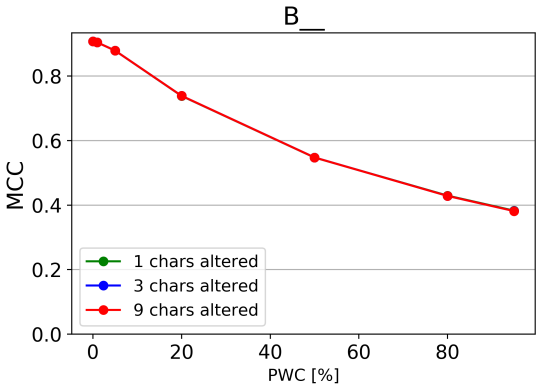


Figure 18: Effect of character repetitions (MCC)

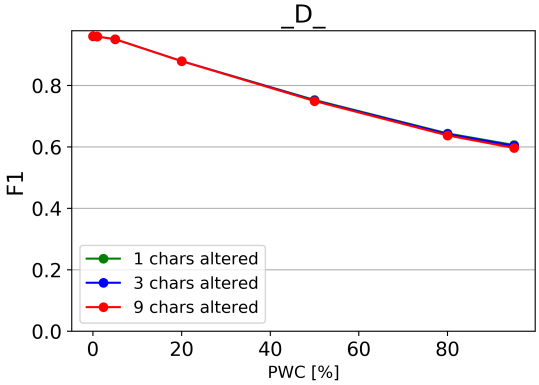


Figure 19: Effect of character deletions on classification quality

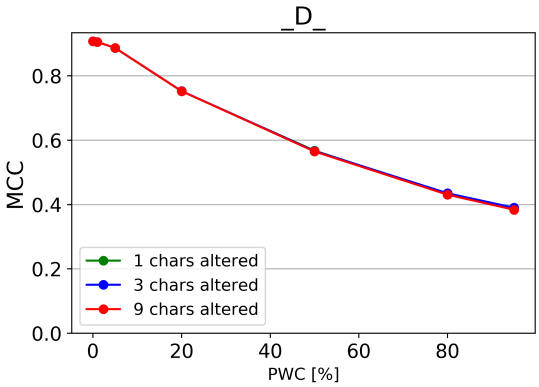


Figure 20: Effect of character deletions (MCC)

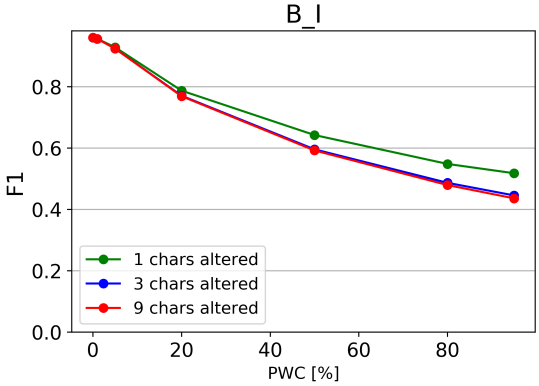


Figure 21: Effect of character repetition and insertion on classification quality

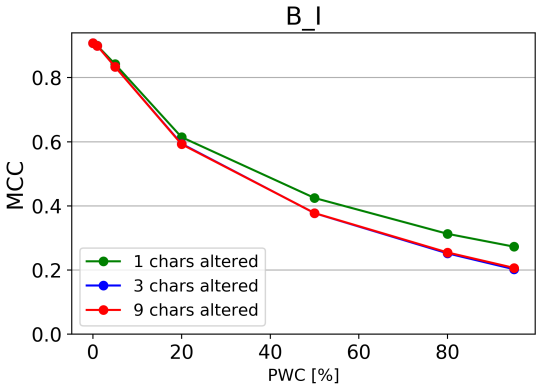


Figure 22: Effect of character repetition and insertion (MCC)

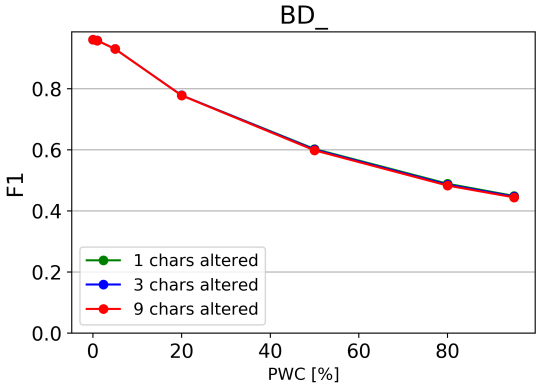


Figure 23: Effect of character repetition and deletion on classification quality

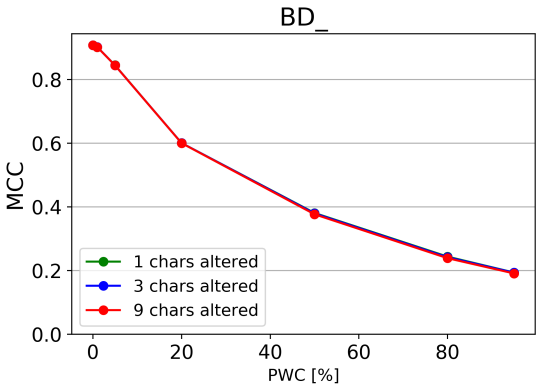


Figure 24: Effect of character repetition and deletion (MCC)

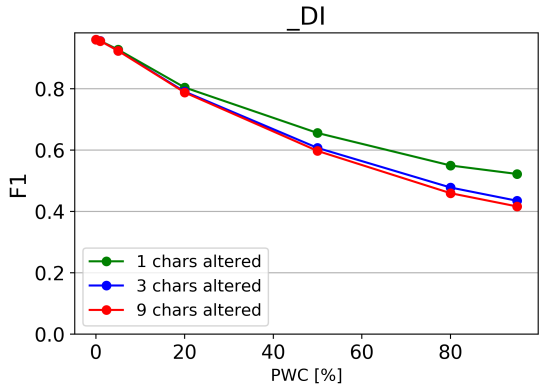


Figure 25: Effect of character deletions and insertions on classification quality

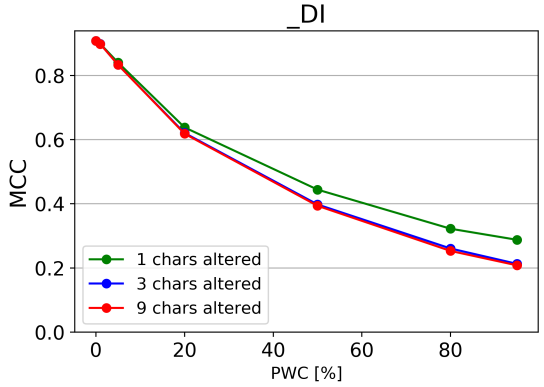


Figure 26: Effect of character deletions and insertions (MCC)

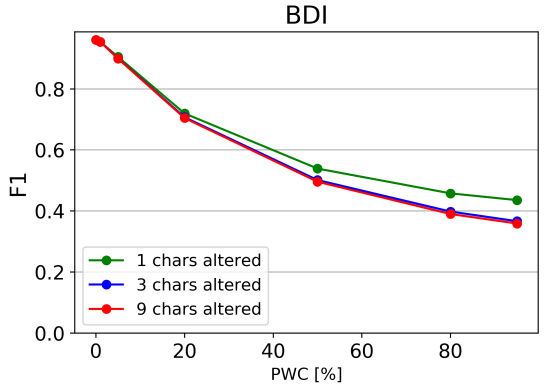


Figure 27: Effect of character repetition, deletion and insertions on quality

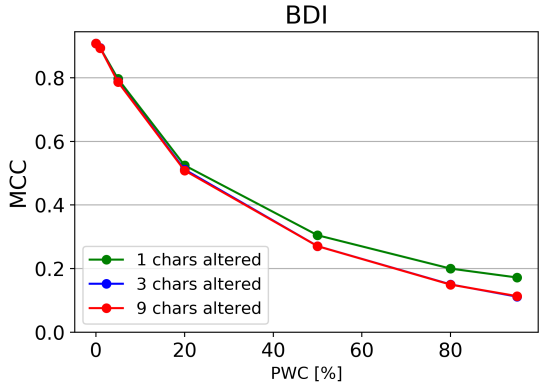


Figure 28: Effect of character repetition, deletion and insertions (MCC)

The CPPP text encoding algorithm has notable advantages over the initial CP algorithm. The significant ones are:

- drastically increased capacity of the encoding space;
- corpora don't have to be tokenized with delimiters;
- the ability to create sparse representations of words, phrases, sentences, paragraphs and even entire documents.

The CPPP algorithm also has advantages over conventional text vectorizers, which rely on tokenization, stemming and other rudimentary techniques of parsing out individual words in a text. The current versions of the string comparison subroutines (SPOQ and its associated filtrations) were presented in the paper. These will be developed further, in order to fully utilize the potential of the sparse representations, produced by the CPPP algorithm.

The benchmark classification problem, described at the end of the paper, will help guide the ongoing development effort. It will also enable comparisons to be made with conventional solutions for data preparation and feature extraction, by using widely known and reliable classification quality evaluation metrics.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank Technical University of Sofia – Research and Development Sector, for their financial support.

References

[1] M. Marinov and A. Efremov, "Representing character sequences as sets : a simple and intuitive string encoding algorithm for NLP data cleaning," in 2019 IEEE International Conference on Advanced Scientific Computing (ICASC), 1-6, 2019, https://doi.org/10.1109/ICASC48083.2019.8946281

[2] Subutai Ahmad, Jeff Hawkins, "Properties of Sparse Distributed Representations and their Application to Hierarchical Temporal Memory", Cornell University, 2015, arXiv:1503.07469

[3] Scott Purdy, "Encoding Data for HTM Systems", Cornell University, 2016, arXiv:1602.05925

[4] Ioannis Partalas, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artieres, George Paliouras, Eric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, Patrick Galinari, "LSHTC: A Benchmark for Large-Scale Text Classification", Cornell University, 2015, arXiv:1503.08581

[5] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, **21**(6), 2020, https://doi.org/10.1186/s12864-019-6413-7

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 160-166 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Nearest Neighbour Search in k-dSLst Tree

Meenakshi Hooda*, Sumeet Gill

Department of Mathematics, Maharshi Dayanand University, 124001, India

ARTICLE INFO

Article history:
Received: 12 May, 2020
Accepted: 26 June, 2020
Online: 18 July, 2020

Keywords:
Nearest Neighbour
Spatial Indexing
k-d tree
Sorted Linked List
Duplicate Keys

ABSTRACT

In the last few years of research and innovations, lots of spatial data in the form of points, lines, polygons and circles have been made available. Traditional indexing methods are not perfect to store spatial data. To search for nearest neighbour is one of the challenges in different fields like spatiotemporal data mining, computer vision, traffic management and machine learning. Many novel data structures are proposed in the past, which use spatial partitioning and recursive breakdown of hyperplane to find the nearest neighbour efficiently. In this paper, we have adopted the same strategy and proposed a nearest neighbour search algorithm for k-dSLst tree. k-dSLst tree is based on k-d tree and sorted linked list to handle spatial data with duplicate keys, which is ignored by most of the spatial indexing structures based on k-d tree. The research work in this paper shows experimentally that where the time taken by brute force nearest neighbour search increases exponentially with increase in number of records with duplicate keys and size of dataset, the proposed algorithm k-dSLstNearestNeighbourSearch based on k-dSLst tree performs far better with approximately linear increase in search time.

1. Introduction

Geospatial datasets are huge and complex in structure and relationships. We need complex spatial operators to fetch the required information from spatial datasets. These complex spatial operators include intersection, overlap, adjacency etc. The traditional indexing structures can't handle queries related to spatial details efficiently. Spatial queries use spatial relationships among different geometries and make use of n-dimensional geometric data such as points, lines and polygons to retrieve required data.

A spatial query retrieves features depending on relationships of spatial data with geometry of queried data. The objective of spatial queries is to find out the spatial relationships in one or more subjects to search for spatial objects. The extracted information help in taking decisions related to various policies or for doing analysis in various fields. Spatial queries permit for the utilization of data types related to geometry like point, line etc. and take into consideration the spatial relationship in these geometries. Spatial indices are used for spatial database to optimize spatial query. The indices provide one of the optimization techniques for improving the quality of services based on location.

1.1 Range Query

The range queries are related with bounded areas. The output of such kind of queries contains some region that might be overlapped. The spatial objects are associated to each other within specific area or distance. These kinds of queries have associated area and require at least two parameters i.e. location and boundary limit.

Example:

Find all the hostels within 12 kilometers of a given university.

Find all towns within 90 kilometers of a given village.

1.2 Spatial Join Query

These queries are the combination of more than one spatial query. We need spatial join operation for retrieving required information. These queries are expensive as join condition(s) involve areas and their proximity to each other.

Example:

Find all the cities near given lake.

*Meenakshi Hooda, Maharshi Dayanand University, Haryana, India & mshthebest@gmail.com

The partitioning and summarization of spatiotemporal data is significant for numerous community based applications like environmental science, public safety and health [1]. The relations among spatial objects can be categorized in different ways such as classifications based on distance, topology and direction relations. These may be further joined using logical operators to depict neighbourhood relation between spatial objects [2].

2. Related Work

2.1 Nearest Neighbour Query

Nearest Neighbour Search (NNS) is also called as Proximity, Similarity or Closest Point Search. It is an optimization problem to find nearest points in metric spaces [3]. In such kind of queries, the data closet to the queried location is retrieved. Here, we search for objects which are nearest to the particular location. We can also query for xNN i.e. x nearest neighbours, in which search is done for x objects nearest to the queried location. The result might be ordered by their respective proximity to queried spatial location. [4] introduce a method to fetch nearest neighbour data in 3-Dimensional space by making use of clustered hierarchical indexing tree structure. Studies show that this approach achieves remarkable improvement in response time analysis as compared to already existing spatial data accessing methods in databases. [5] research work introduces a progressive algorithm for a search of approximate k-nearest neighbor. The most of the KNN algorithms though utilize k-nearest neighbor libraries for many of the data analysis procedures, but the fact is that these algorithms run only after indexing of the whole dataset, which means that the datasets are not online. [6] proposed a method of parallel kd-tree construction for 3-D points on a Graphic Processing Unit. The method consists of a sorting algorithm to maintain high level of parallelism throughout the creation.

Example: Find 5 nearest hostels with respect to given location comprised of n-dimensional coordinate.

In this research paper, we are introducing the algorithm k-dSLstTreeNearest for Nearest Neighbour in k-dSLst tree structure designed in our previous work to index spatial data with duplicate keys.

3. Brute Force Method: Nearest Neighbour Search for Spatial Data with Duplicate Keys

Brute Force exploration process, which is also called as Exhaustive/Blind Search, is extremely generalized technique to solve any problem. It systematically enumerates all possible contenders for the problem's solution and tests to check if the contender fulfills the statement of the problem [7]. In this algorithm, we find the distance of every object in dataset from the position of the queried object. As, multiple objects can be found at the nearest neighbour location, a list is maintained to keep record of all objects with minimum distance. Algorithm 1, **Distance bruteForceNearestNeighbourSearch (SpatialDataset, Location, ResultNN_BF *resultNearestNeighbourBF)**, receives *spDataSet* of *SpatialDataSet* type, k dimensional *queryPosition* of *Location* type to which nearest neighbour is to be found and returns either *INFINITY*, if no record is found in dataset or minimum distance of the nearest found object(s) from *queryPosition* in case record(s) is found. Also, a pointer *resultNearestNeighbourBF* of *ResultNN_BF* type is passed to save the address of result node(s), if found.

If there is no record in dataset and list is empty i.e. *spDataSet* is NULL, **bruteForceNearestNeighbourSearch** will return INFINITY. On the other hand, if records exist in the list, initialize *minDistance* to INFINITY. Now, traverse every node in the *spDataSet* to find the distance between *queryPosition* and *currentRecord*. Also, update the minimum distance *minDistance*, if it is lesser than the *minDistance* calculated up to now. Now, insert object details in *resultNearestNeighbourBF* which satisfy the minimum distance condition.

Algorithm 1 bruteForceNearestNeighbourSearch

```
1:  Begin
2:  if spDataSet is NULL
3:     return INFINITY
4:  end if
5:  Initialize minDistance to INFINITY
6:  Set distanceSqr to 0
7:  Set currentRecord to spDataSet
8:  loop while currentRecord is NOT NULL
9:     Find distanceSqr between currentRecord and
       queryPosition
10:    Insert calculated distaceSqr to currentRecord
       temporarily
11:    if distanceSqr < minDistance
12:    then
13:       minDistance = distanceSqr
14:    end if
15:    Move to next currentRecord
16:    Reset distanceSqr to 0
17: end loop
18: Set currentRecord to spDataSet
19: loop while currentRecord exist
20:    if distanceSqr inserted to currentRecord temporarily ==
       minDistance
21:    then
22:       add currentRecord to resultNearestNeighbourBF
23:    end if
24:    Move to next currentRecord
25: end Loop
26: return minDistance
27: End
```

4. k-dSLst Tree

In our earlier research work, we had introduced an indexing structure k-dSLst tree to index spatial data with duplicate keys. k-dSLst tree is based on k-d tree and sorted linked list. In this paper, we are introducing an algorithm to find nearest neighbour for a given n-dimensional spatial point in a spatial dataset. In k-dSLst tree, first a *kdSLstNode* is created according to n-dimensional spatial location of an object, and then we insert the data related to same object in *dataSNode* in *kdSLstNode* created above. If kdSLstNode related to n-dimensional point already exists, then *dataSNode* is inserted in a sorted way according to *object_Id* of the object in already existing *kdSLstNode*. When we insert the *dataSNode*, we need to find the right location for insertion by traversing the linked list at particular *kdSLstNode*. It will increase the insertion time as compared to k-dLst tree, but when comes to searching for a particular object at particular location, it will outperform k-dLst tree on an average.

Figure 1 shows the structure of 2-dSLst tree for dataset records in 2-d space stored as *kdSLstNodes* in a 2-dSLst tree, which can be generalized for k-d space to create k-dSLst tree. A rectangular space is also represented by the leaf *kdSLstNode* which is further separated into two spaces by the newly inserted point.

The root *kdSLstNode* will have discriminator 0. It will be 1 for two sons at the next level, and will be incremented for every next level until it reaches up to *k-1* on k^{th} level. Then it again starts with 0 for $k+1^{th}$ level and the cycle repeats till the end of k-dSLst tree.

So, **nextDiscriminator (*level_i*) = (*level_i* + 1) mod *k***.

Notations:
K: *Keys of kdSLstNode*
N: *kdSLstNode of k-dSLst tree structure*
$K_0(N), K_1(N) \ldots K_{k-1}(N)$: *k keys of kdSLstNode N*
lSon(N): *Left branch of kdSLstNode N*
rSon(N): *Right branch of kdSLstNode N*
dc(N): *Discriminator of kdSLstNode N*

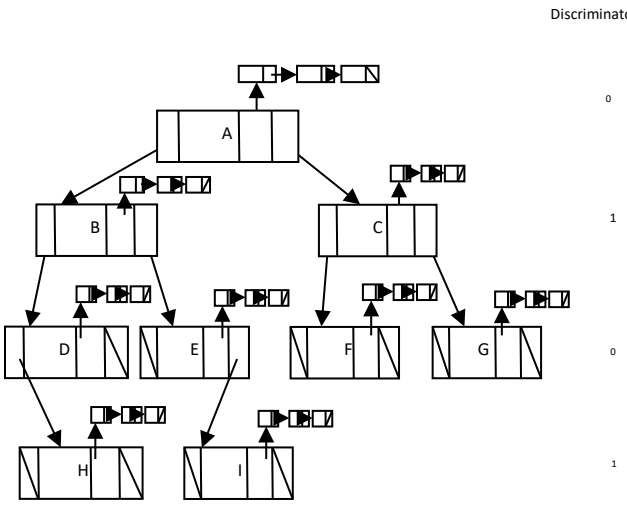


Figure 1: Structure of k-dSLst tree for 2-d keys
(It can be generalized for k-d keys)

Now, whether to insert new *kdSLstNode* as left son or right son depends on the result of comparison of keys. Let *dc* be the discriminator for *kdSLstNode N*. If $K_{dc}(N) != K_{dc}(Q)$ then the successor *kdSLstNode Q* will be inserted either on left or right side of *N i.e. either lSon(N) or rSon(N)*. If $K_{dc}(N) < K_{dc}(Q)$, then *Q* will be inserted on right side of *N i.e. as rSon(N)* else on left side of *N i.e. as lSon(N)*. But, if $K_{dc}(N) = K_{dc}(Q)$, then the keys for remaining dimensions will be compared. A superkey *SK* of *kdSLstNode N* is defined by cyclical concatenation of all keys starting with $K_{dc}(N)$ as

$$SK_{dc}(N) = K_{dc}(N)\, K_{dc+1}(N) \ldots K_{dck-1}(N)\, K_{dc0}(N) \ldots K_{dc-1}(N)$$

Now if $SK_{dc}(Q) < SK_{dc}(N)$ then *Q* will be added as *lSon(N)* else *Q* will be added as *rSon(N)*.

Now, in case of duplicate keys i.e. if $SK_{dc}(Q) = SK_{dc}(N)$, then address of new *dataSNode* will be saved in the same already existing *kdSLstNode* node as a sorted linked list of *dataSNodes*.

5. Nearest Neighbour Search in k-dSLst Tree

Nearest Neighbour Search is a challenge in various domains like computer vision, spatial data mining and machine learning. There is explosive growth of location based data on the Internet and it is becoming a challenge day by day to store and manage this available data in an efficient way. The researchers have designed many indexing data structures using spatial partitions and recursive hyperplane decomposition to index spatial data. These indexing structures also speed up the nearest neighbour search. But, when it comes to spatial data with duplicate keys, many of the indexing structures do not handle them. K-dSLst tree structure is a combination of k-d tree and sorted linked list. The k-d tree having N nodes require O (Log N) inspections to search for nearest neighbour as it requires the traversal to at least one leaf of the tree. Also, as the nearest neighbour search needs to traverse a node at most once, it will not visit more than N nodes [8]. Linked list is maintained to hold all the objects at particular spatial location rather than one only.

To find nearest neighbour for any queried n-dimensional point, the search is started at the root of indexing tree and subtrees are explored recursively using the given rule of pruning spatial subtrees: If the nearest neighbour discovered up to now is nearer than the distance between the queried point and hyperplane coordinates corresponding to the current *kdSLstNode*, we can prune the exploration of this *kdSLstNode* and its subtrees further. A *kdSLstNode* needs to be explored further only if it contains point which is nearer as compared to the best one found so far. The effectiveness of the pruning rule depends on finding a nearby point. To do this, we need to organize the recursive method in a way that if we have two probable subtrees to traverse down further, we must opt for the subtree on the same side of the splitting line as is the queried point. The nearest point searched while exploring the first subtree may enable pruning of second subtree.

In this paper, we are proposing Nearest Neighbour Search **algorithm k-dSLstTreeNearest** for kdSLst tree which was designed and implemented for indexing spatial data with duplicate spatial keys. Algorithms 2 and Algorithm 3 show the proposed work. The **Algorithm 2 k-dSLstTreeNearest** takes two parameters as arguments. First is the root node of the k-dSLst tree and second is the position containing N-dimensional coordinates about which the nearest neighbour is to be searched. Rather than returning a single nearest neighbour object, the algorithm returns the list of all nearest neighbours. If root node of the tree or hyper rectangle points to NULL then INFINITY is returned back to show that there is no nearest neighbour, otherwise root node is considered as nearest neighbour to start with. The variable *distanceSqr* holds the distance in between nearest point up to now and position of point about which nearest neighbour search is queried.

Algorithm 2 k-dSLstTreeNearest

```
1:  Begin
2:  if k-dSLstTreeRoot is NULL
3:      return NIL;
4:  if k-dSLstTreeHRect is NULL
```

```
5:      return NIL;
6:    Initialize resultOfNearest_kdSLst with k-dSLstTreeRoot.
7:    Initialize distanceSqr with 0.
8:    Find distanceSqr between resultOfNearest_kdSLst and
      queryPosition.
9:    CALL k-dSLstTreeNearestIterative(k-dSLstTreeRoot,
      queryPosition,  &resultOfNearest_kdSLst, &distanceSqr, k-
      dSLstTreeHRect).
10:   if (resultOfNearest_kdSLst)
11:   then
12:     Traverse the sorted linked list of dataSNodes to show all
        objects at resultOfNearest_kdSLst
13:     Visualize resultOfNearest_kdSLst
14:   end if
15:   End
```

An iterative **Algorithm 3 k-dSLstTreeNearestIterative** is called with parameters which include root node of the k-dSLst tree, position queried about for nearest neighbour, pointer to save result node, minimum distance up to now and hyper rectangle of the tree.

Algorithm 3 k-dSLstTreeNearestIterative

```
1:    Begin
2:    Initialize currentDim to kdsLstNode's dimension
3:    Set decideLeftRight by difference in queryPosition and kdSLstNode
      coordinate values for currentDim
4:    if decideLeftRight <= 0
5:    then
6:      Set nearerSubTree to lSon(kdSLstNode)
7:      Set fartherSubTree to rSon(kdSLstNode)
8:      Update nearerHyperRectCoordinates and
        fartherHyperRectCoordinates
9:      using k-dSLstTreeHRect
10:   else
11:     Set nearerSubTree to rSon(kdSLstNode)
12:     Set fartherSubTree to lSon(kdSLstNode)
13:     Update nearerHyperRectCoordinates and
        fartherHyperRectCoordinates
14:     using k-dSLstTreeHRect
15:   end if
16:   if nearerSubTree exists
17:   then
18:     Save nearerHyperRectCoordinates to decideLeftRight
        temporarily
19:     Update nearerHyperRectCoordinates with kdSLstNode
        coordinates for  currentDim
20:     CALL k-dSLstTreeNearestIterative (nearerSubTree,
         queryPositions, resultOfNearest_kdSLst,
         resultDistanceSqr, k-dSLstTreeHRect);
21:     Update nearerHyperRectCoordinates with
22:      decideLeftRight;
23:   end if
24:   Reset distanceSqr with 0
25:   loop for every dimension dim
26:     Compute distanceSqr between kdSLstNode coordinates
        [dim] -queryPosition[dim]
27:   end loop
28:   if distanceSqr less than resultDistanceSqr
29:   then
30:     Update resultOfNearest_kdSLst and resultDistSqr
         accordingly
31:   end if
32:   if fartherSubTree exist
33:   then
34:     Save fartherHyperRectCoordinates to decideLeftRight
        temporarily
35:     Update fartherHyperRectCoordinates with kdSLstNode
         coordinates for  currentDim
36:   if closest point of k-dSLstTreeHRect is closer than
37:   resultDistanceSqr
38:   then
39:     CALL k-dSLstTreeNearestIterative (fartherSubTree,
          queryPositions, resultOfNearest_kdSLst,
          resultDistanceSqr, k-dSLstTreeHRect);
40:   end if
41:   Update fartherHyperRectCoordinates with decideLeftRight;
42:   end if
43:   end
```

Depending on the distance of query point from current point for current dimension, we decide nearer and farther sub trees and update the coordinates of nearer and farther hyper rectangles accordingly. If we find any sub tree which is nearer, we call

Algorithm 3 k-dSLstTreeNearestIterative iteratively with updated parameters. Now, check the distance of the queried point from the current point so far and update the value of *resultDistSqr*, if it is nearer. Now, also repeat the search for farther sub tree to test for any other nearer point on other side of the slice. The algorithm saves the nearest neighbour N-dimensional coordinates in *resultOfNearest_kdSLst*. At last, we traverse the complete linked list of *dataSNode* at *resultOfNearest_kdSLst* to show all of the nearest neighbours. Also, all of the nearest neighbours are displayed graphically using QGIS software.

6. Performance Evaluation: bruteForceNearestNeighbourSearch vs k-dSLstNearestNeighbourSearch

We have implemented the algorithms using **language C** and **GNU Compiler Collection (GCC) compiler - version 6.4.3** on **Operating System Ubuntu-10.04.1-Desktop-amd64**. For visualization of spatial datasets and output of **algorithms bruteForceNearestNeighbourSearch** and **k-dSLstNearestNeighbourSearch**, we have used **Quantum Geographic Information System (QGIS) Desktop 2.12.1**.

In our experiments, we are using synthetic datasets which hold spatial details of vehicles. The datasets include vehical-id, spatial coordinates (latitude, longitude) and other non-spatial attributes. We have taken six datasets of different sizes listed in Table 1 given next.

Table 1: Datasets with Duplicate Keys for Performance Evaluation

S. no.	Dataset	Number of records
1	Dataset-01	215
2	Dataset-02	430
3	Dataset-03	860
4	Dataset-04	1720
5	Dataset-05	3440
6	Dataset-06	10320

The six datasets are visualized in Figure 2 through Figure 7 using QGIS Desktop 2.12.1. These figures show the spatial location of different vehicles.

While evaluating performance of searching for nearest neighbour using **Algorithm 1 bruteForceNearestNeighbourSearch** and **Algorithm 2 k-dSLstNearestNeighbourSearch**, we are using the query point Q (43, 20), with Latitude as 43 and Longitude as 20. Both algorithms

give the same results i.e. show the same Vehicles' Id as nearest neighbours for query point Q. But, the time taken by both algorithms is different. As the number of records in datasets and number of nearest neighbours increase, the algorithm **k-dSLstNearestNeighbourSearch** outperforms **bruteForceNearestNeighbourSearch**.

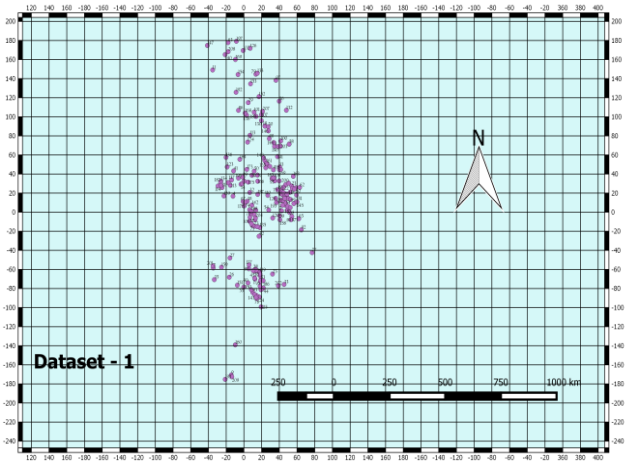


Figure 2: Vehicles' location as per Spatial Dataset – 1
(215 Spatial Records).

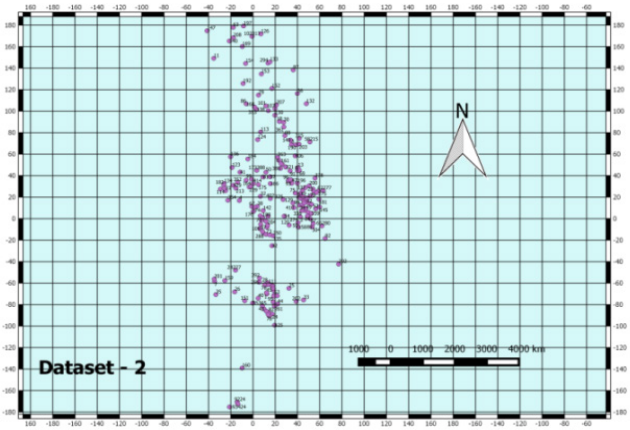


Figure 3: Vehicles' location as per Spatial Dataset – 2
(430 Spatial Records).

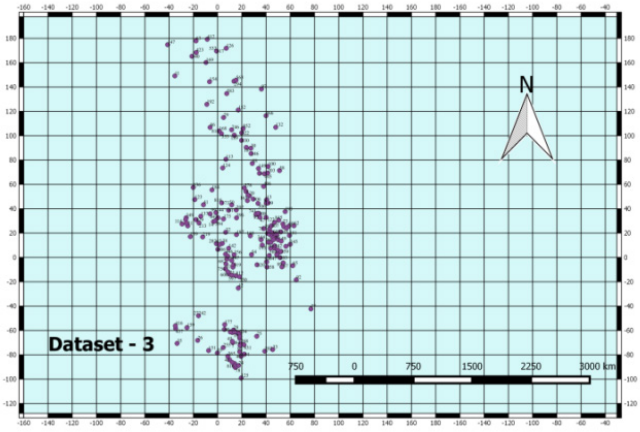


Figure 4: Vehicles' location as per Spatial Dataset – 3
(860 Spatial Records).

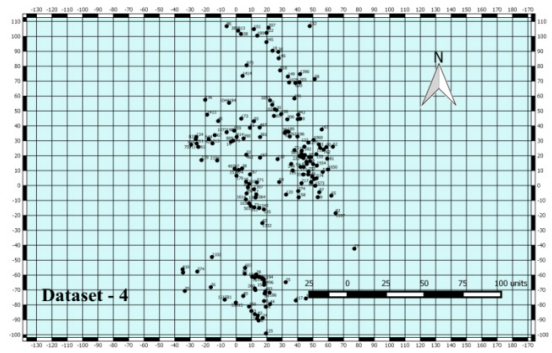


Figure 5: Vehicles' location as per Spatial Dataset – 4
(1720 Spatial Records).

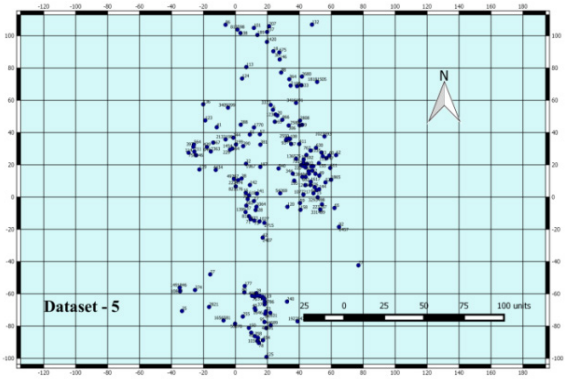


Figure 6: Vehicles' location as per Spatial Dataset – 5
(3440 Spatial Records).

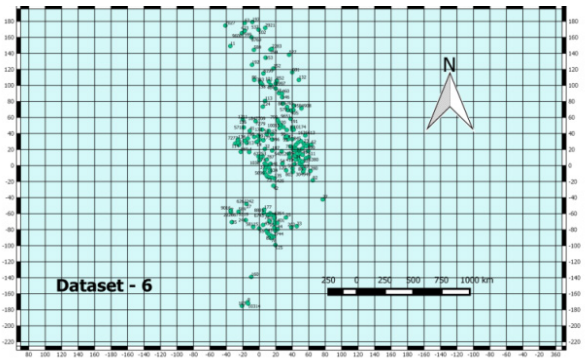


Figure 7: Vehicles' location as per Spatial Dataset – 6
(10320 Spatial Records).

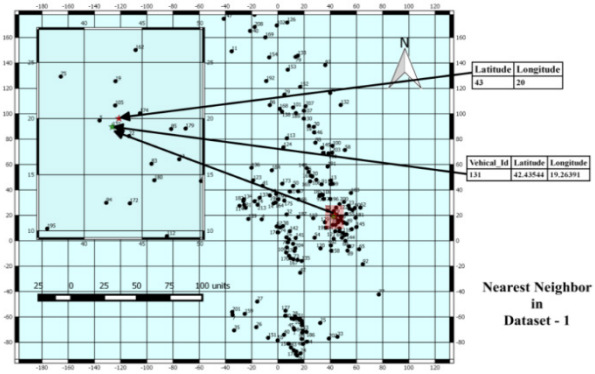


Figure 8: Nearest Neighbors of Q(43, 20) in Dataset – 1.

Figures 8 through 13 visualize the output of nearest neighbour query for query point Q (43, 20). Figures list Ids of all Vehicles found nearest to Q. Figures show that the algorithm **k-dSLstNearestNeighbourSearch** is well efficient to find multiple Vehicle Ids at the same nearest spatial location.

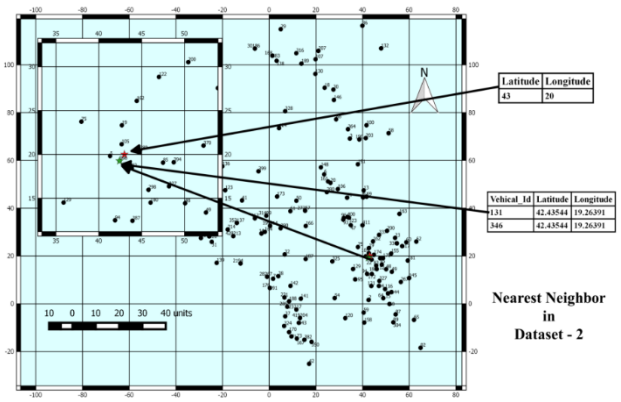


Figure 9: Nearest Neighbors of Q(43, 20) in Dataset – 2.

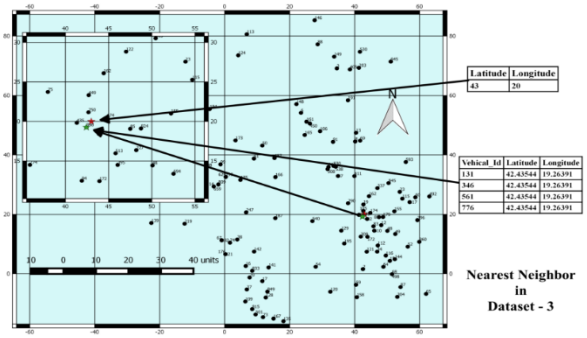


Figure 10: Nearest Neighbors of Q(43, 20) in Dataset – 3.

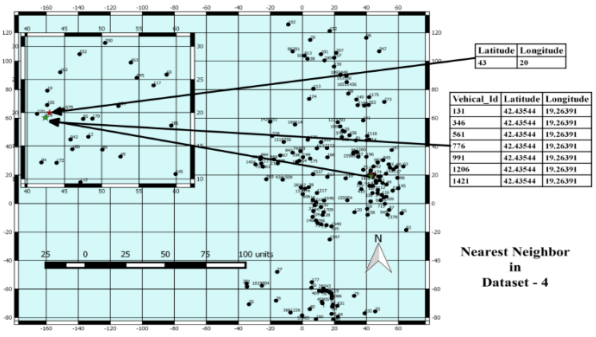


Figure 11: Nearest Neighbors of Q(43, 20) in Dataset – 4.

Table 2 shows the number of records found as nearest neighbours and the time taken by both algorithms in microseconds for searching the same. Results show that algorithm **k-dSLstNearestNeighbourSearch** is more efficient as compared to **bruteForceNearestNeighbourSearch** as the number of spatial records increases and capable of finding all nearest neighbours for spatial datasets with duplicate spatial keys. The search time take by algorithm **bruteForceNearestNeighborSearch** is increasing continuously and there is exponential rise from Dataset-05 to Dataset-06 as number of records increase from 3440 to 10320. Search time of algorithms is also affected by the number of records found at nearest location which are 64 in case of Dataset-05 and 192 in case of Dataset-06. As compared to algorithm **bruteForceNearestNeighborSearch** which takes 407 micro secs. and 1391 micro secs. for nearest neighbor search in Dataset-05 and Datset-06 respectively**,** algorithm **k-dSLstNearestNeighbourSearch** takes much less time i.e. 162 micro secs. and 223 micro secs. for corresponding datasets.

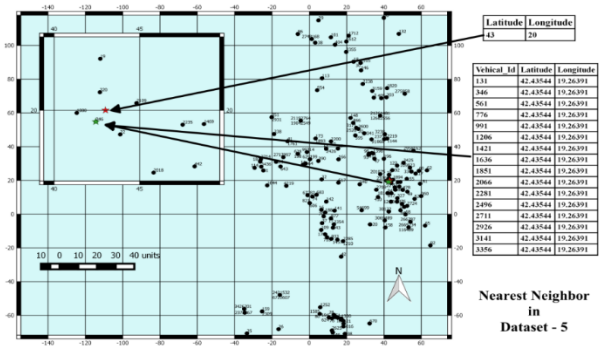


Figure 12: Nearest Neighbors of Q(43, 20) in Dataset – 5.

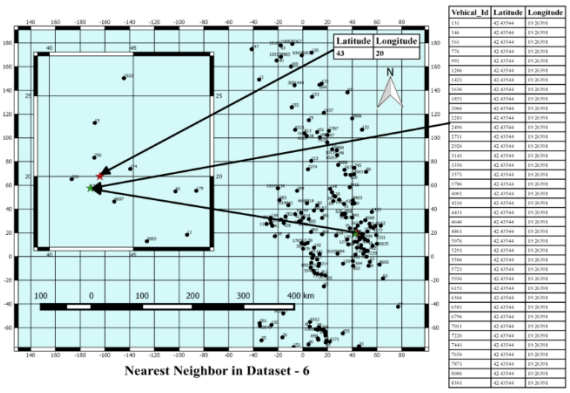


Figure 13: Nearest Neighbors of Q(43, 20) in Dataset – 6.

Table 2: bruteForceNearestNeighbourSearch vs k-dSLstNearestNeighbourSearch

S. no.	Dataset	No. of records found at Nearest location	Time taken for search (in microseconds)	
			bruteForceNearestNeighborSearch Algorithm	k-dSLstNearestNeighbourSearch Algorithm
1	Dataset-01	4	29	106
2	Dataset-02	8	46	89
3	Dataset-03	16	90	96
4	Dataset-04	32	212	110
5	Dataset-05	64	407	162
6	Dataset-06	192	1391	223

Figure 14 shows the results of **bruteForceNearestNeighbourSearch vs k-dSLstNearestNeighbourSearch graphically.**

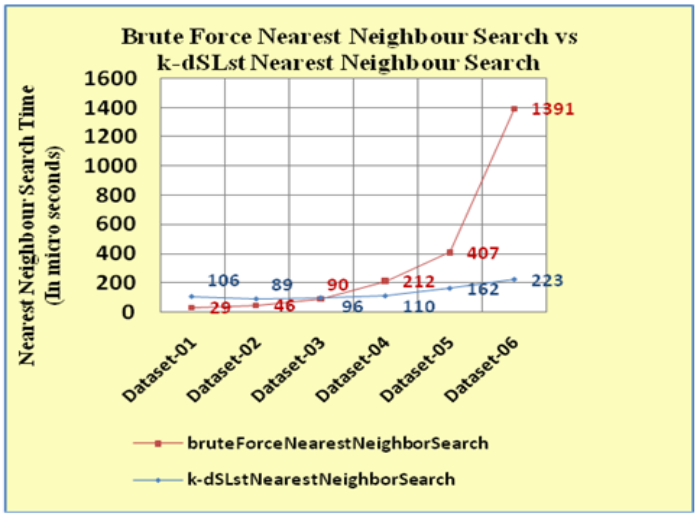


Figure 14: Search time: bruteForceNearestNeighbourSearch vs k-dSLstNearestNeighbourSearch

7. Conclusion and Future Scope

The **k-dSLstNearestNeighbourSearch algorithm** for searching nearest neighbour in spatial datasets with duplicate keys is based on **k-dSLst indexing tree structure** which is further based on k-d tree and sorted single linked list data structures. We have compared the proposed algorithm with brute-force approach. The performance evaluation shows that the search time taken by algorithm **bruteForceNearestNeighborSearch** increases exponentially as numbers on data records and nearest neighbours increases. But in case of algorithm **k-dSLstNearestNeighbourSearch** for the same datasets, the time take for search is much less and doesn't rise exponentially. The performance evaluation of both algorithms shows that the proposed work in this paper revealed better performance when compared to the conventional brute-force approach.

The work can be further expanded to search for N nearest neighbours and objects within a given range. Also, the spatial data structure k-dSLst tree can be extended to accommodate temporal data also to index spatio-temporal datasets.

References

[1] S. Shekhar et al., Spatiotemporal Data Mining: A Computational Perspective, International Journal of Geo-Information ISSN 2220-9964, 04, 2306-2338, 2015.
[2] S. Geetha, S. Velavan, "Optimization of Location Based Queries using Spatial Indexing, International Journal of Soft Computing, 4, 2014.
[3] Verma et al., "Comparison of Brute-Force and K-D Tree Algorithm, International Journal of Advanced Research in Computer and Communication Engineering, January, 03, 2014.
[4] A. Suhaibaha, et al., "3D Nearest Neighbour Search Using a Clustered Hierarchical Tree, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXIII ISPRS Congress, XLI-B2, 2016.
[5] J. Jaemin, et al., "A progressive k-d tree for approximate k-nearest neighbors, IEEE Workshop on Data Systems for Interactive Analysis (DSIA), 2017.
[6] W. David, R. Rafael, "Parallel kd-Tree Construction on the GPU with an Adaptive Split and Sort Strategy, International Journal of Parallel Programming, 46, 1139–1156, 2018.
[7] Stoimen. Computer Algorithms: Brute Force String Matching. Stoemen's web log. [Online] March 2012.
[8] M.W. Andrew, "Efficient Memory-based Learning for Robot Control. Technical Report No. 209, Ph. D Thesis, Computer Laboratory, University of Cambridge, 1991.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 160-168 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Research of the Effect of Rotation and Low-Frequency Vibration on the Robotic Assembly Process

Mikhail Vladimirovich Vartanov, Trung Ta Tran*, Van Dung Nguyen

Faculty of Mechanical Engineering, Moscow Polytech University, Moscow, 107023, Russia

ARTICLEINFO

Article history:
Received: 28 August, 2020
Accepted: 23 October, 2020
Online: 08 November, 2020

Keywords:
Robotic Assembly
A Force/Torque Sensor
The Effect of Rotation
Vibrational Device

ABSTRACT

The prospects for the introduction of assembly automation in the industry are largely determined by the reliability of the assembly process. One of the reasons for reducing its reliability is parts jamming. In order to increase the technological reliability of the assembly process, various physical and technical effects are applied. The article presents the results of an experimental study using the rotational movement effect and low-frequency vibrations of the base part in order to reduce assembly efforts and the probability of part jamming.

1. Introduction

Assembly tasks using industrial robots are constantly becoming more complicated over the years. Their main problem is jamming even with slight mismatching. In the case the assembly gap is comparable to the form error, assembly based on rigid basing becomes impossible. The assembly becomes especially problematic when the assembly gap in the joints is comparable to the repeatability of an industrial robot.

The case of the parallel-plane movement of the assembly parts has been considered in known dynamic models of the robotic assembly of cylindrical joints, using passive adaptation [1, 2], as well as control algorithms. In the case of the plane-parallel movement of the installed part, significant sliding friction forces can occur at the contact points of the assembly parts, in which there is a small gap. It leads to jam. Thus, reducing sliding friction forces is one of the ways to increase the reliability of the assembly process. For that purpose, the rotation and the low-frequency vibration can due to provide this effect.

The efficiency of rotational movement for ensuring assembly conditions during the robotic assembly process was considered in works [3-8] and others.

In the work [6] the assembly using the rotational effect with the airflow was experimentally studied. However, an analytical model was not built. In [7, 8], the influence on the assembly process of oscillations along the conjugation axis and along the inclined axis was studied. In work [3], the assembly of parts was studied when the installed part rotates with respect to the non-inertial reference frame and the base part rotates with respect to the inertial reference frame. However, the simultaneous superposition of the effect of rotation and low-frequency vibrations has not previously been considered.

The use of low-frequency vibration remains one of the most promising areas in the field of automatic assembly. This is confirmed by the results of theoretical and experimental works [9-11].

The modern approach to the theory of experiment assumes the need to achieve the required research quality. For this purpose, it is necessary to build a model of the experiment [12].

The effectiveness of technological operation modeling in robotic assembly and the process as a whole based on statistical experimental design allow studying the influence of many process characteristics. It is necessary to establish a connection among the n independent non -random variables $x_1, x_2 \dots x_n$ and the random variable y that depends on them. When processing the results of a multi-factor experiment, the mathematical model of the process, namely the function $\mathrm{M(y)} = f\,(\mathrm{x_1}, \mathrm{x_2} \dots \mathrm{x_n})$ is usually represented as a polynomial of a specific degree. The expression $\mathrm{M(y)} = f\,(\mathrm{x_1}, \mathrm{x_2} \dots \mathrm{x_n})$ is called the regression equation of the mathematical expectation of a random variable y for non- random variables $x_1, x_2 \dots x_n$, and the type of function can be linear or curved [13].

In this article, the authors study the rotation effect and low-frequency vibration to increase reliability in a robotic assembly process. A mathematical dynamics model of the robotic assembly process was presented. A multifactor experiment of this assembly process was carried out, the results and discussion were presented.

*Corresponding Author: Trung Ta, Tran, Rizhskij proezd, 15k2, Moscow, 129626, Russian Federation, +7 905 531 29 68, trungta82@gmail.com

2. Mathematical Model

2.1. Choosing a kinematic assembly scheme

The shaft *3* is held by the gripper *2* of the robot *1*. The working movements of the robot manipulator are the progressive movement of the gripper vertically downwards to the specified position. The vibrational device *5* is a three-link manipulator, each of its links is driven by a separate drive. The base bushing *4* is kept in the center of the vibrational disk, attached to the last link of the vibrational device (Figure 1).

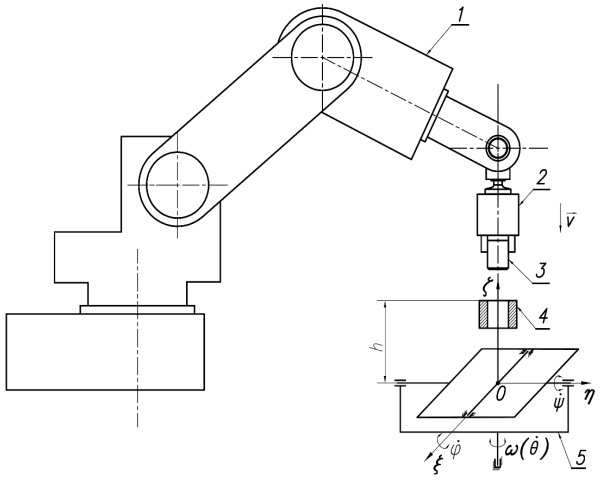


Figure 1: Experimental installation (principle scheme): 1. Industrial robot; 2. The gripper with a force/torque sensor; 3. The shaft, 4. The bushing; 5. The vibrational device;

Four reference frames were defined with a common origin at point O: $Oxyz$ is a fixed reference frame; $Ox_1y_1z_1$ and $Ox_2y_2z_2$ are reference frames, attached to the first and second links of the vibrational device respectively, $O\xi\eta\zeta$ is the reference frame, attached to the third link of the vibrational device. The reference frame $Cx_3y_3z_3$ is attached to the shaft, its origin coincides with the shaft mass center (Figure 2).

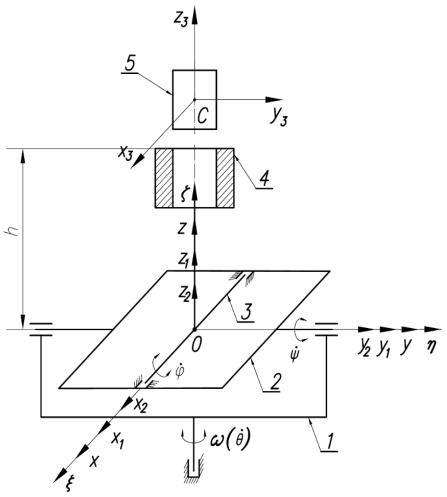


Figure 2: The scheme for determining coordinate systems: 1 - the first link of the vibrational device; 2 - the second link of the vibrational device; 3 - the third link of the vibrational device; 4 - the bushing; 5 - the shaft.

The first link performs rotational motion with a constant angular velocity ω around the fixed axis Oz. In this case, the reference frame $Ox_1y_1z_1$ attached to the first link, at the time t, will rotate with respect to the fixed reference frame $Oxyz$ by an angle θ:

$$\theta = \omega t \tag{1}$$

The second and third links oscillate around mutually perpendicular axes with a phase shift equal to the angle $\pi/2$ as follows:

$$\psi = A\sin kt\,;\, \varphi = A\cos kt \tag{2}$$

where A is the amplitude; k – the angular frequency.

2.2. The mathematical model of the motion dynamics of the shaft mass center with respect to the non-inertial reference frame attached to the base part

Previously, the authors published a mathematical model of the assembly process of a cylindrical joint using rotation and low-frequency vibrations [14]. Therefore, this article presents only a brief mathematical statement.

Formulation of the mathematical model.

On the basis of the theorem on the mass center movement of the system, the movement of the shaft in Figure 3 can be described by a relation of the form:

$$m\vec{\mathbf{a}}_c = m\vec{\mathbf{g}} + \vec{\mathbf{N}} + \vec{\mathbf{F}}_f + \vec{\mathbf{P}} \tag{3}$$

where: m is the mass of the part;

$\vec{\mathbf{a}}_c$ – the absolute acceleration of the cylindrical part mass center;

$\vec{\mathbf{g}}$ – acceleration of gravity;

$\vec{\mathbf{N}}$ – the normal reaction of the vibrational disk plane;

$\vec{\mathbf{F}}_f$ – the sliding friction force;

$\vec{\mathbf{P}}$ – the assembly force.

The relative movement of the mass center of the part with respect to the non-inertial reference frame attached to the bushing can be represented by the equation:

$$m\vec{\mathbf{a}}_{\mathbf{cr}} = m\vec{\mathbf{g}} + \vec{\mathbf{N}} + \vec{\mathbf{F}}_{\mathbf{f}} + \vec{\mathbf{F}}_{\mathbf{ce}}^{\mathbf{fict}} + \vec{\mathbf{F}}_{\mathbf{cC}}^{\mathbf{fict}} + \vec{\mathbf{P}} \tag{4}$$

where $\vec{\mathbf{a}}_{\mathbf{cr}}$ – the relative acceleration of the shaft mass center;

$\vec{\mathbf{F}}_{\mathbf{ce}}^{\mathbf{fict}}$, $\vec{\mathbf{F}}_{\mathbf{cC}}^{\mathbf{fict}}$ – the force of moving space and Coriolis inertial forces of the shaft mass center;

Equation (4) in projections on the axis of the non-inertial reference frame $O\xi\eta\zeta$ has the form:

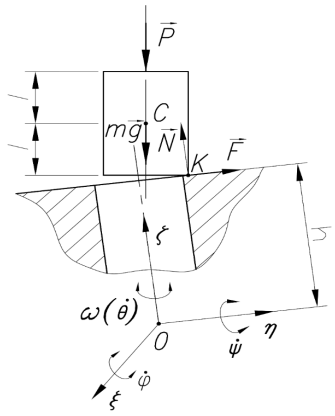


Figure 3: Scheme of forces at single-point contact of the assembled parts

$$\begin{aligned} m\ddot{\xi}_c &= (mg)_\xi + F_{f\xi} + F_{ce\xi}^{fict} + F_{cC\xi}^{fict} + P_\xi, \\ m\ddot{\eta}_c &= (mg)_\eta + F_{f\eta} + F_{ce\eta}^{fict} + F_{cC\eta}^{fict} + P_\eta, \\ m\ddot{\zeta}_c &= (mg)_\zeta + F_{ce\zeta}^{fict} + F_{cC\zeta}^{fict} + N + P_\zeta. \end{aligned} \tag{5}$$

where ξ_c, η_c, ζ_c are the coordinates of shaft mass center (point C) in the reference frame $O\xi\eta\zeta$.

To determine the projections of all forces on the axis of the non-inertial reference frame $O\xi\eta\zeta$, the apparatus of homogeneous coordinate transformation matrices was used [5].

Differential equations of motion of the shaft mass center with respect to the non-inertial reference frame attached to the base part

In [14] the components of (5) were determined, a mathematical model of the dynamics of the relative motion of the shaft mass center in its one-point contact with the plane of the bushing can be as follows:

$$\begin{aligned} m\ddot{\xi}_c &= p\sin\psi - fN\frac{\dot{\xi}_k}{\sqrt{\dot{\xi}_k^2+\dot{\eta}_k^2}} + m(\xi_c\dot{\psi}^2 - \eta_c\ddot{\psi}\sin\varphi \\ &\quad - 2\eta_c\dot{\varphi}\dot{\psi}\cos\varphi - 2\eta_c\omega\dot{\varphi}\sin\varphi - \zeta_c\ddot{\psi}\cos\varphi \\ &\quad + 2\zeta_c\dot{\varphi}\dot{\psi}\sin\varphi - 2\zeta_c\omega\dot{\varphi}\cos\varphi\cos\psi \\ &\quad + \xi_c\omega^2\cos^2\varphi + \eta_c\omega^2\sin\psi\sin\varphi\cos\psi \\ &\quad + \zeta_c\omega^2\sin\psi\cos\varphi\cos\psi) - 2m\dot{\zeta}_c(\dot{\psi}\cos\varphi + \omega\cos\psi\sin\varphi) \\ &\quad + 2m\dot{\eta}_c(-\dot{\psi}\sin\varphi + \omega\cos\psi\cos\varphi); \\ m\ddot{\eta}_c &= -p\cos\psi\sin\varphi - fN\frac{\dot{\eta}_k}{\sqrt{\dot{\xi}_k^2+\dot{\eta}_k^2}} \\ &\quad + m(\xi_c\ddot{\psi}\sin\varphi + \eta_c\dot{\psi}^2\sin^2\varphi + \eta_c\varphi^2 - 2\eta_c\omega\dot{\varphi}\sin\psi \\ &\quad + \zeta_c\dot{\psi}^2\cos\varphi\sin\varphi + \xi_c\omega^2\sin\psi\cos\psi\sin\varphi \\ &\quad + \zeta_c\dot{\varphi} + \eta_c\omega^2\sin^2\dot{\varphi}\sin^2\psi + \zeta_c\omega^2\sin^2\psi\sin\dot{\varphi}\cos\varphi \\ &\quad + 2\xi_c\omega\dot{\psi}\sin\psi\cos\varphi - 2\eta_c\omega\dot{\psi}\cos\psi\sin\varphi\cos\varphi \\ &\quad - 2\zeta_c\omega\dot{\psi}\cos\psi\cos^2\varphi + \eta_c\omega^2\cos^2\varphi - \zeta_c\omega^2\sin\varphi\cos\varphi) \\ &\quad - 2m\dot{\xi}_c(-\dot{\psi}\sin\varphi + \omega\cos\psi\cos\varphi) + 2m\dot{\zeta}_c(\dot{\varphi} - \omega\sin\psi); \end{aligned} \tag{6}$$

$$\begin{aligned} N &= m\ddot{\zeta}_c - m(\xi_c\omega^2\cos\psi\sin\psi\cos\varphi \\ &\quad + \eta_c\omega^2\sin^2\psi\sin\varphi\cos\varphi - \eta_c\omega^2\sin\varphi\cos\varphi \\ &\quad + \zeta_c\omega^2\sin^2\psi\cos^2\varphi + \zeta_c\omega^2\sin^2\varphi - 2\zeta_c\omega\dot{\varphi}\sin\psi \\ &\quad + \xi_c\ddot{\psi}\cos\varphi + \eta_c\dot{\psi}^2\sin\varphi\cos\varphi - \eta_c\ddot{\varphi} + \zeta_c\dot{\psi}^2\cos^2\varphi \\ &\quad + \zeta_c\dot{\varphi}^2 - 2\xi_c\omega\dot{\psi}\sin\psi\sin\varphi + 2\eta_c\omega\dot{\psi}\cos\psi\sin^2\varphi \\ &\quad + 2\zeta_c\omega\dot{\psi}\cos\psi\sin\varphi\cos\varphi) + 2m\dot{\eta}_c(\dot{\varphi} - \omega\sin\psi) \\ &\quad - 2m\dot{\xi}_c(\dot{\psi}\cos\varphi + \omega\cos\psi\sin\varphi) + p\cos\psi\cos\varphi. \end{aligned}$$

where N – the normal reaction of the plane of the bushing. f – the term coefficient of friction; $\dot{\xi}_k$ and $\dot{\eta}_k$ – the projections of the velocity of the contact point $"K"$ on the axis $O\xi, O\eta$ of the reference frame $O\xi\eta\zeta$.

Derivatives $\dot{\xi}_k$ and $\dot{\eta}_k$ are determined as follows [14]:

$$\begin{aligned} \dot{\xi}_k &= r\omega\cos\theta\cos\psi - r\sin\theta\sin\psi\cdot\dot{\psi} \\ &\quad + l\cos\psi\cdot\dot{\psi} + \dot{\xi}_c; \\ \dot{\eta}_k &= r(\omega\cos\theta\sin\psi\sin\varphi + \sin\theta\cos\psi\cdot\dot{\psi}\sin\varphi \\ &\quad + \sin\theta\sin\psi\cos\varphi\cdot\dot{\varphi} - \omega\sin\theta\cos\dot{\varphi} - \cos\theta\sin\varphi\cdot\dot{\varphi}) \\ &\quad - l(-\sin\psi\cdot\dot{\psi}\sin\varphi + \cos\psi\cos\varphi\cdot\dot{\varphi}) + \dot{\eta}_c \end{aligned} \tag{7}$$

and the coordinate ζ_c is determined as follows [14]:

$$\begin{aligned} \zeta_c &= h + r\left(\sin\psi\cos\varphi\sin\theta + \sin\varphi\cos\theta\right) \\ &\quad + l\cos\psi\cos\varphi \end{aligned} \tag{8}$$

The obtained mathematical models (6) and (7), (8) established the relationship between the components of the forces with factors of rotation movement and low-frequency vibrations.

3. Experimental Installation

3.1. Determining the Factors

A multifactor experiment of the process was carried out as follows. Upon reaching the operating mode of vibration, the shaft was fed into the assembly zone in the vertical direction with the contact of the faces of the assembled parts (Figure 1). Under the influence of vibrational and rotational movements of the experimental installation, the shaft was combined with the bushing and the assembly process was carried out. The contact moment of the assembled part faces was taken as the starting point. The completion of the assembly was recorded when the shaft was inserted to a depth of 50 mm. The resulting value of the assembly force was considered as the output optimized parameter (response) of the system when studying the process.

The following parameters were taken as input parameters (factors) in the study of the system: the angular frequency (rad/s); the linear amplitude of the last link of the vibrational device (mm); the linear velocity of the last link of the robot (mm/s); the angular velocity of the first link of the vibrational device (rad/s) and the gap in the joint (mm) (Table 1). All of them meet the appropriate requirements of controllability and integrity used to factors, as well

as the requirements of compatibility and lack of linear correlation applied to a set of factors.

Table 1: Input Factors

Factors	Symbol
The angular frequency of vibrational device (rad/s);	X_1
The linear amplitude of the vibrational device (mm);	X_2
The linear velocity of the last link of the robot (mm/s);	X_3
The angular velocity of the first link of the vibrational device (rad/s)	X_4
The gap in the joint (mm)	X_5

3.2. Selecting the Type of Model

A mathematical description of the assembly force components can be obtained using statistical experimental design [16].

The magnitude of the components of force and torque in the contact zone is measured by a force/torque sensor installed on the last link of the robot. Since the experimental sample was a shaft, then in the experiment 4 components of force and torque were investigated: axial force – F_z, torque around the z-axis – T_z, axial force – F_y and torque around the x-axis – T_x. These force and torque components are dependent on the input process factors. The mathematical model of the research object was understood as the form of the response function connecting the optimization parameter with the process factors. In general, this function can be written as follows:

$$F = f(X_i), \ (i = 1..5) \tag{9}$$

Selecting a type of model means choosing the type of function and its mathematical form. Suppose that the dependencies can be approximated with sufficient accuracy by equations of power type. So, for example, the relationship (9) can be expressed by the equation:

$$F = cX_1^{\alpha} X_2^{\beta} X_3^{\gamma} X_4^{\theta} X_5^{\mu} \tag{10}$$

where $X_i, i = 1..5$, process factors; $c, \alpha, \beta, \gamma, \theta, \mu$ – constant values.

After logarithm, equation (10) is linearized:

$$\ln F = \ln c + \alpha \ln X_1 + \beta \ln X_2 + \gamma \ln X_3 + \theta \ln X_4 + \mu \ln X_5 \tag{11}$$

The possibility of approximating dependences (10) by equations in the form (11) is established by verifying the hypothesis of the linear model adequacy when expressing the results of the experiment with a polynomial in the form:

$$y = b_0 + \sum_{1 \le i \le 5} b_i x_i + \sum_{1 \le i < j \le 5} b_{ij} x_i x_j \tag{12}$$

where $y = ln\, F$, $x_i -$ coded value of X_i.

Coded value x_i of factor i:

$$x_i = \frac{2(\ln X_i - \ln X_{i\max})}{\ln X_{i\max} - \ln X_{i\min}} + 1, \ (i = 1..5) \tag{13}$$

where X_i – natural value of factor i; $X_{i_{max}}$, $X_{i_{min}}$ – natural values of the upper and lower levels, respectively.

3.3. Design of Multifactor Experiments in the Search for Optimal Technological Conditions and Design Parameters of Equipment.

According to the verified experiment results, it is assumed that on the considered small interval, the response function can be described by a polynomial of the first degree. Due to the linearity of the model, the factors varied at two levels. At the same time, the zero levels were taken on the basis of previous studies [2] (Table 2).

Table 2: Levels of Factors and Variable Intervals

Factors	Levels			Range of variation	Dimension
	-1	*0*	*+1*		
X_1	14	16	18	2	rad/s
X_2	1,8	2,0	2,2	0,2	mm
X_3	10	20	30	10	mm/s
X_4	0	0,25	0,5	0,25	rad/s
X_5	0,01	0,03	0,05	0,02	mm

When conducting a full factorial experiment with varying factors at two levels, one replicate of it would involve $2^5 = 32$ experiments. It would make the experiment cumbersome enough if at least three replicates were performed. Therefore, the fractional factorial design of the experiment $N = 2^{5-1}$ was chosen. It reduces the number of experiments of one replicate to 16 due to the system of mixing linear effects of factor x_5 with the least significant interaction effects, but at the same time the number of replicates is increased to three. All experiments, according to the principle of randomization, were carried out in random order. Three replicates for all selected conditions were implemented (Table 3).

Table 3: Experiment Planning Matrix

№	x_0	x_1	x_2	x_3	x_4	x_5	x_1x_2	x_1x_3	x_1x_4	x_1x_5	x_2x_3	x_2x_4	x_2x_5	x_3x_4	x_3x_5	x_4x_5
1	+	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+
2	+	+	-	-	-	+	-	-	-	+	+	+	-	+	-	-
3	+	-	+	-	-	+	-	+	+	-	-	-	+	+	-	-
4	+	+	+	-	-	-	+	-	-	-	-	-	-	+	+	+
5	+	-	-	+	-	+	+	-	+	-	-	+	-	-	+	-
6	+	+	-	+	-	-	-	+	-	-	-	+	+	-	-	+
7	+	-	+	+	-	-	-	-	+	+	+	-	-	-	-	+
8	+	+	+	+	-	+	+	+	-	+	+	-	+	-	+	-
9	+	-	-	-	+	+	+	+	-	-	+	-	-	-	-	+
10	+	+	-	-	+	-	-	-	+	-	+	-	+	-	+	-
11	+	-	+	-	+	-	-	+	-	+	-	+	-	-	+	-
12	+	+	+	-	+	+	+	-	+	+	-	+	+	-	-	+
13	+	-	-	+	+	-	+	-	-	+	-	-	+	+	-	-
14	+	+	-	+	+	+	-	+	+	+	-	-	-	+	+	+
15	+	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+
16	+	+	+	+	+	-	+	+	+	-	+	+	-	+	-	-

3.4. Experimental Installation

Experimental Installation has been established (Figure 4). A force-torque sensor *2* (Schunk FT AXIA 80 Ethernet) was installed on the last link of the robot *1*. The base bushing *6* is kept in the vibrational device *7*, described in [6] and [17].

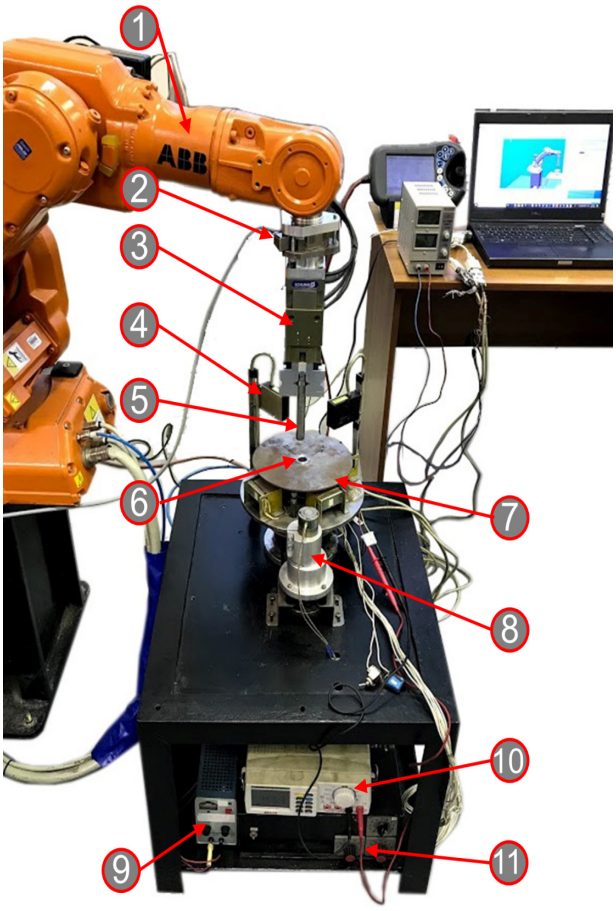


Figure 4: Experimental installation: 1. Industrial robot ABB IRB 140; 2. The force/torque sensor Schunk FT AXIA 80 Ethernet; 3. Gripper; 4. Triangulation Laser Sensors Riftek RF603.2; 5. The shaft; 6. The bushing; 7. The vibrational device; 8. DC motor; 9. Controlled converter; 10. Electronic frequency meter (Mastech M9803R); 11. The generator of low-frequency oscillations.

The experimental procedure was as follows. The shaft *5* was held by gripper *3* of the robot *1*. Vibrations were provided by the original generator of low-frequency oscillations *11*. The generator of low-frequency oscillations was controlled by the electronic frequency meter *10*, and the actual vibration amplitude of the vibrational device was monitored by contactless laser sensors *4*. The first link of the vibrational device performed a rotational motion with a constant angular velocity ω around a vertical fixed axis by a DC motor *8*. The DC motor was controlled by a controlled converter *9*. All sensor signals were collected and processed by the computer. Thus, during the experiments, continuous monitoring of the current values of technological factors was ensured.

The choice of force/torque sensor Schunk FT AXIA 80 Ethernet was determined by the required process sensitivity. The default value of sensitivity is 0.005 (Where: 0 = no damping - robot maintains speed; 1 = full damping - robot decelerates quickly) and its technical data is presented in Table 4.

Table 4: Technical Data of the Sensor Schunk FT AXIA 80 Ethernet

Storage and Operating Conditions			
Storage Temperature [°C]		-20 to +85	
Operating Temperature [°C]		0 to +65	
Electrical Specifications			
Power Source		***DC Power***	
Voltage [V] min. max.		 12 30	
Maximum Power Consumption [W]		1.5	
Measurement Ranges			
Parameter	***F_{xy} [N]***	***F_z [N]***	***T_{xyz} [Nm]***
Measurement Range 0	500	900	20
Measurement Range 1	200	360	8

4. Results and Discussion

4.1. Screening of Coarse Observations in Parallel Experiments

For each row of the experiment planning matrix, according to the results of *n* parallel experiments, $\bar{y}_j$ – arithmetic mean value of optimization parameter was found:

$$\bar{y}_j = \frac{1}{n}\sum_{u=1}^{n} y_{ju} \tag{14}$$

Where u is the number of parallel experiment; y_{ju} is the value of the response in the u parallel experiment of the j row of the experiment planning matrix.

In order to estimate the deviations of the response from its mean value for each row of the experiment planning matrix, the dispersion s_j^2 of the experiment was calculated according to n parallel experiments (Table 5). Statistical dispersion was estimated by an expression of the form:

$$s_j^2 = \frac{1}{n-1}\sum_{u=1}^{n}(y_{ju} - \bar{y}_j)^2 \tag{15}$$

Table 5: Experimental Results and Dispersion of Forces F_y, F_z and Torques T_x, T_z

№	F_y		F_z		T_x		T_z	
	$\bar{y}$	s_j^2	$\bar{y}$	s_j^2	$\bar{y}$	s_j^2	$\bar{y}$	s_j^2
1	2.442	0.015	1.260	0.007	-1.715	0.00759	4.262	0.002
2	2.297	0.016	1.196	0.009	-1.556	0.00983	4.252	0.002
3	2.722	0.006	1.496	0.004	-1.368	0.00656	4.330	0.006
4	2.735	0.037	1.524	0.004	-1.135	0.00272	4.380	0.001
5	2.400	0.062	1.291	0.002	-1.504	0.01142	4.176	0.004
6	2.373	0.012	1.236	0.012	-1.736	0.00990	4.459	0.005
7	2.745	0.020	1.542	0.019	-1.363	0.01318	4.557	0.003
8	2.913	0.017	1.578	0.017	-1.241	0.00276	4.415	0.003

№	F_y		F_z		T_x		T_z	
	$\bar{y}$	s_j^2	$\bar{y}$	s_j^2	$\bar{y}$	s_j^2	$\bar{y}$	s_j^2
9	3.341	0.012	1.858	0.002	-0.742	0.00113	3.942	0.004
10	3.075	0.029	1.814	0.001	-0.613	0.01398	4.306	0.002
11	3.204	0.002	1.873	0.003	-0.597	0.00232	4.368	0.001
12	3.557	0.012	1.919	0.002	-0.589	0.00109	4.180	0.001
13	3.321	0.017	1.930	0.001	-0.892	0.00393	4.169	0.006
14	3.240	0.006	1.854	0.004	-0.771	0.00157	4.214	0.003
15	3.684	0.008	2.050	0.009	-0.688	0.00201	4.297	0.011
16	3.543	0.004	2.067	0.009	-0.621	0.00306	4.448	0.001

4.2. Verification of the Uniformity of Dispersions

The dispersion homogeneity was tested by the Fisher F criterion, which is the ratio of the largest to the smallest variance:

$$F = \frac{s_1^2}{s_2^2} \tag{16}$$

where s_1^2 и s_2^2 - respectively, the largest and smallest values of the response variance in different experiments.

If the observed value is $F < F_t$ (F_t is the table value of the criterion F for the corresponding freedom degrees $f_1 = n_{\max} - 1 = 2$, $f_2 = n_{min} - 1 = 2$ and the accepted significance level of 0.05 [16]) then the variances are uniform (Table 6).

Table 6: Testing the Homogeneity of Dispersion by the Fisher Criterion

$F < F_t = 19.2$							
F_y		F_z		T_x		T_z	
14.62876	Yes	13.23305	Yes	18.30958	Yes	12.82161	Yes

With uniform duplication of experiments, the homogeneity of a number of dispersions was checked using the Cochrane G criterion, which is the ratio of the maximum dispersion to the sum of all dispersions:

$$G = \frac{s_{\max}^2}{s_1^2 + s_2^2 + \ldots + s_N^2} = \frac{s_{\max}^2}{\sum_{j=1}^{N} s_j^2} \tag{17}$$

The dispersions are homogeneous if $G < G_t$ (G_t is the table value of the criterion G for the conditions: N is the number of compared variances $N = 16$, and n is the number of parallel experiments $n = 3$ [16]). If $G > G_t$, then the variances are inhomogeneous, and this indicates that the investigated value y does not obey the normal law (Table 7).

Table 7: Testing the Homogeneity of Dispersion by the Cochrane Criterion

$G < G_t = 0.3346$							
F_y		F_z		T_x		T_z	
0.222947	Yes	0.211286	Yes	0.182417	Yes	0.150301	Yes

4.3. Determination of Regression Coefficients

According to the results of the experiment, the coefficients of the model were calculated. The regression coefficients are determined by the formulas:

$$b_0 = \frac{1}{N}\sum_{j=1}^{N} \bar{y}_j; \; b_j = \frac{1}{N}\sum_{j=1}^{N} x_{ij}\bar{y}_j; \qquad b_{jl} = \frac{1}{N}\sum_{j=1}^{N} x_{ij}x_{lj}\bar{y}_j. \tag{18}$$

where i, l – numbers of factors; x_{ij}, x_{lj} – coded values of factors $i$and l in the j-th experiment. These formulas are obtained using the least squares method.

The significance of each coefficient of the regression equation was checked by using Student t-criterion, comparing the obtained values with the table values in the corresponding degrees of freedom and significance level:

$$t_j = \frac{|b_j|}{s_{\{b\}}} \tag{19}$$

where $s_{\{b\}}$ – the dispersion of the regression coefficients, determined by formula:

$$s_{\{b\}}^2 = \frac{s_y^2}{N \cdot n} \tag{20}$$

If the dispersion s_j^2 of the experiments are homogeneous, the dispersion s_y^2 of the reproducibility of the experiment is calculated by the expression:

$$s_y^2 = \frac{1}{N}\sum_{j=1}^{N} s_j^2 \tag{21}$$

Table 8: Significance of each Coefficient of the Regression Equation

b_i	F_y	F_z	T_x	T_z
b_0	2.975	4.297	1.655	-1.071
b_1	–	0.035	–	0.038
b_2	0.163	0.075	0.101	0.12
b_3	0.053	0.045	0.038	-0.031
b_4	0.396	-0.057	0.265	0.382
b_5	0.045	-0.071	–	–
b_{12}	0.057	-0.051	–	–
b_{13}	–	–	–	-0.028
b_{14}	–	–	–	–
b_{15}	–	–	–	–
b_{23}	–	–	–	–
b_{24}	–	–	-0.044	-0.055
b_{25}	–	–	–	-0.034
b_{34}	–	–	–	–
b_{35}	–	–	–	0.038
b_{45}	0.04	–	–	–

Student criterion t is calculated for each regression coefficient. Statistically insignificant coefficients can be excluded from (16). The verification results are summarized in the Table 8. The regression equations for the considered response functions $y_{F_y}, y_{F_z}, y_{T_x}, y_{T_z}$ can be written based on the data given in Table 8.

4.4. Verification of Model Adequacy

After calculating the coefficients of the model and checking their significance, the dispersion of adequacy was determined. The residual dispersion or the adequacy dispersion characterizes the scattering of empirical values of y with respect to the calculated y determined by the found regression equation. The adequacy dispersion was determined by the formula:

$$s_{ad}^2 = \frac{n\sum_{j=1}^{N}(\bar{y}_j - \hat{y}_j)^2}{f} = \frac{n\sum_{j=1}^{N}(\bar{y}_j - \hat{y}_j)^2}{N-(k+1)} \quad (22)$$

where $\bar{y}_j$ – arithmetic mean value of the optimization parameter in the j-th experiment; $\widehat{y}_j$ – value of the optimization parameter calculated by the model for the conditions of the j-th experiment; f – the number of freedom degrees; k – the number of factors.

The final step in processing the experimental results was to verify the hypothesis of the adequacy of the found model. Hypothesis verification was carry out according to the Fisher criterion:

$$F = \frac{s_{ad}^2}{s_y^2} \quad (23)$$

If the value $F < F_t$ (F_t – table value of the criterion, $F_t = 2.1$ for the accepted significance level of 0.05 and the corresponding numbers of freedom degrees $f_1 = N - (k + 1) = 16 - (5 + 1) = 10, f_2 = N(n - 1) = 32$ [16]) then the model is considered adequate. For $F > F_t$, the adequacy of hypothesis is rejected (Table 9).

Table 9: Result of Model Adequacy Checking

	F_y	F_z	T_x	T_z
s_{ad}^2	0.021959	0.003026	0.005974	0.008727
s_y^2	0.017306	0.003367	0.00647	0.005815
F	1.269	0.899	0.923	1.501
$F < F_t$	Yes	Yes	Yes	Yes

In this case, we can conclude that the resulting model satisfies the adequacy condition.

For the transition from the coded values to natural values of the factors in the regression equation, values x_i factors on expressions (12) were substituted by (13) and regression coefficients values b_i were referenced in Table 8, obtained:

$$\begin{aligned}\ln F_y = {} & 11.111 - 3.111\ln X_1 - 10.875\ln X_2 \\ & +0.096\ln X_3 + 0.188\ln X_4 + 0.117\ln X_5 \\ & +4.521\ln X_1 \ln X_2 + 0.016\ln X_4 \ln X_5\end{aligned} \quad (24)$$

$$\begin{aligned}\ln F_z = {} & -5.322 + 3.062\ln X_1 + 11.931\ln X_2 \\ & +0.082\ln X_3 - 0.188\ln X_4 \\ & -0.088\,\mathrm{n}\, X_5 - 4.405\ln X_1 \ln X_2\end{aligned} \quad (25)$$

$$\begin{aligned}\ln T_x = {} & 1.458 + 0.470\ln X_2 + 0.069\ln X_3 \\ & +0.182\ln X_4 - 0.141\ln X_2 \ln X_4\end{aligned} \quad (26)$$

$$\begin{aligned}\ln T_z = {} & -4.670 + 1.459\ln X_1 - 1.075\ln X_2 \\ & +1.392\ln X_3 + 0.244\ln X_4 - 0.406\ln X_1 \ln X_3 \\ & -0.176\ln X_2 \ln X_4 - 0.421\ln X_2 \ln X_5 \\ & +0.045\ln X_5 + 0.086\ln X_3 \ln X_5\end{aligned} \quad (27)$$

Potentiating expressions (24 – 27), the dependence of forces and torques on the studied factors of the robotic assembly process was found:

$$F_y = \frac{66893.374 X_3^{0.096} X_4^{(0.188+0.016\ln X_5)} X_5^{0.117} X_1^{4.521\ln X_2}}{X_1^{3.111} X_2^{10.875}} \quad (28)$$

$$F_z = \frac{0.005 X_1^{(3.062-4.045\ln X_2)} X_2^{11.931} X_3^{0.082}}{X_4^{0.018} X_5^{0.088}} \quad (29)$$

$$T_x = 4.298 X_2^{(0.470-0.141\ln X_4)} X_3^{0.069} X_4^{0.182} \quad (30)$$

$$\begin{aligned}T_z = {} & 0.009 X_1^{(1.459-0.406\ln X_3)} X_2^{(-1.075-0.176\ln X_4-0.421\ln X_5)} \\ & X_3^{(1.392+0.086\ln X_5)} X_4^{0.244} X_5^{0.045}\end{aligned} \quad (31)$$

According to (28 - 31), nomograms can be constructed, which allow, for practical purposes, to determine the components of the force and torque acting on the installed part when determining factors for the robotic assembly process.

A suitable response for the above regression model F_z (29) was constructed and nomograms (Figure 5) were constructed with paired combinations of these factors while the other factors were at the zero level.

Figure 5 illustrates the dependence of the force F_z on variable factors in the assembly process. It is clear that the angular velocity and the gap in the joint are inversely proportional to the force F_z, while the vibration frequency, the vibration amplitude and the linear velocity are proportional to the force F_z. Also, there is an interaction effect of two factors of the angular frequency and the vibration amplitude. The maximum value of the force F_z was 87.17 N (Figure 5.f), the minimum value of F_z was 57.86 N (Figure 5. g).

Force F_z decreases significantly with increasing the angular velocity (Figure 5. c, f, h, j). Force F_z decreases rapidly at the initial stage, and then more slowly decreases in the investigated values range of the angular velocity. The maximum value of the force F_z in the absence of the rotational movement was 87.17 N (Figure 5. f), the minimum value of F_z with the rotational movement was 65 N (Figure 5. h). Also, F_z decreases with an increase in the gap in the joint. With the small gap, the maximum value of the force F_z was 84.7 N (Figure 5. j). With the large gap, the minimum value of the force F_z was 57.86 N (Figure 5. g).

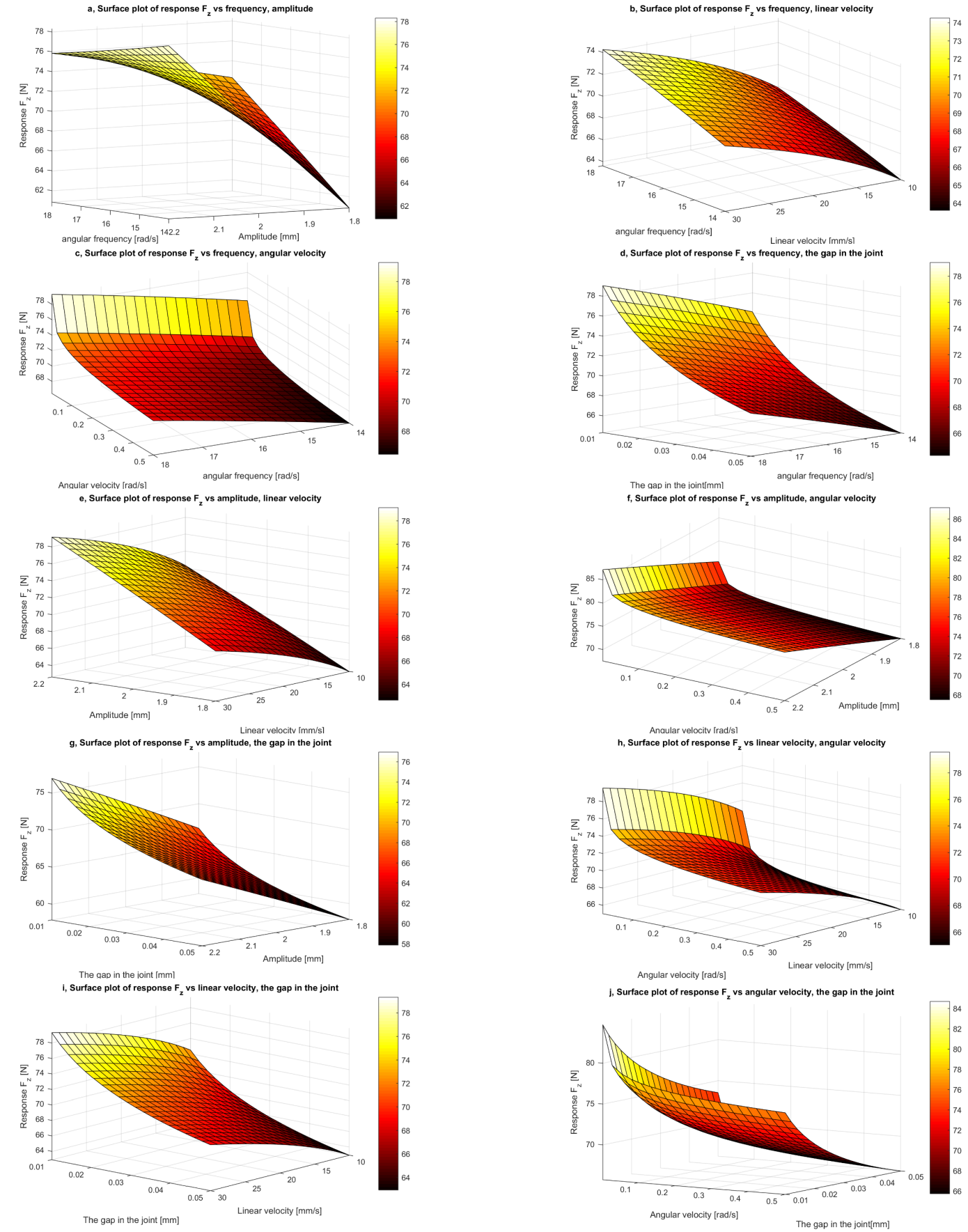


Figure 5: Three-dimensional nomograms showing the behavior of force F_z affected by factors of the robotic assembly process.

The force F_z is proportional to the angular frequency and vibration amplitude. The force F_z decreases with a simultaneous increase or decrease of both these factors (Figure 5. a – Figure 5. g). The angular frequency and vibration amplitude interact with each other, affecting the magnitude of the force F_z. When both factors changed in the investigated values range of them, the change in the value of the force F_z was small (60.8 – 78.3 N) (Figure 5. a). The effect of linear velocity on force F_z is negative (Figure 5.b, e, h, i), however, decreasing the linear velocity the assembly time will increase.

5. Conclusion

The conducted experimental studies revealed the influence of variable factors on the components of the assembly force and torque. Mathematical models that establish the relationship between the components of the friction forces and torques with disturbing factors are formed. The adequacy of the mathematical model of the process is proved. The effects of the interaction of two factors are identified.

The effect of increasing the assembly speed is negative, which correlates with the theory of friction. At the same time, the greatest effect of reducing the assembly force was achieved in the areas corresponding to the lower boundaries of the variation intervals. In various experiments, the decrease in assembly force ranged from 14 to 33%.

The conducted experiments proved the presence of the effect of vibrations and rotation on increasing the technological reliability of the assembly process. Further research will be associated with a change in the zero level of variation of the input factors. It is supposed to check the adequacy of the mathematical model and the results of physical experiments.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors express their gratitude to the Department “Technologies and Equipment of Mechanical Engineering”; Ph.D., Assoc. Prof., Petukhov Sergey Leonidovich, Senior lecturer Mishin Vladimir Nikolaevich and Ph.D., Assoc. Prof., Arkhipov Maksim Viktorovich, Faculty of Mechanical Engineering, Moscow Polytechnic University for the opportunity given to them to work in the laboratory, for significant support and useful suggestions for improving the research.

References

[1] M.V. Vartanov, L.V. Bozhkova, Z.K. Bakena Mbua, “Robot assembly of complex shafts on the basis of passive adaptation and low-frequency vibration”, Russian Engineering Research, **30**, 1092–1094, 2010, https://doi.org/10.3103/S1068798X10110055

[2] A.L. Simakov, Обоснование методов и средств адаптации соединяемых деталей на базе принципов автоматического управления и выявленных взаимосвязей при автоматизированной сборке, Sc.D Thesis, Kovrov State Technological Academy, 2003.

[3] M.G. Kristal, Производительность и надежность сборочных автоматов: монография – Volgograd: VolSTU, 2011.

[4] L.B. Chernyakhovskaya, “Influence of rotational movement of a peg on the automatic peg-in-hole assembly of cylindrical details,” Assembly in mechanical engineering, instrument making, **6**, 7-13, 2016 (article in Russian with an abstract in English).

[5] L.B. Chernyakhovskaya, Кинематический и динамический анализы автоматической сборки цилиндрических деталей: Монография - Samara: Samara State Technical University, 2011.

[6] D.M. Levchuk, Исследование и разработка методов относительного ориентирования сборочных единиц соединения во вращающемся потоке газов при автоматической сборке, Ph. D Thesis, Moscow Automotive Institute, 1974.

[7] B. Bakšys, J. Baskutienė, A. Chadarovičius, “Simulation of vibratory alignment of the parts to be assembled under passive compliance,” Mechanika, **19**(1), 33-39, 2013, doi: https://doi.org/10.5755/j01.mech.19.1.3625.

[8] B. Baksys, J. Baskutiene, “The directional motion of the compliant body under vibratory excitation,” International Journal of Non-Linear Mechanics, **47**(3), 129–136, 2012, doi: https://doi.org/10.1016/j.ijnonlinmec.2012.01.008.

[9] A.A. Ivanov, “Vibration assembly systems," Assembly in mechanical engineering, instrument making, **5**, 7-10, 2013 (article in Russian with an abstract in English).

[10] M.G. Kristal, I.A. Chuvilin, “Research of dynamics of internal member lower flank cradle vibrating meshing,” Assembly in mechanical engineering, instrument making, **4**, 13-17, 2008 (article in Russian with an abstract in English).

[11] V.G. Shuvaev, V.A. Papshev, “Compound oscillatory motion forming device for space orientation of automatically assembling details,” Assembly in mechanical engineering, instrument making, **11**, 15-18, 2009 (article in Russian with an abstract in English).

[12] A.I. Khuri and S. Mukhopadhyay, “Response surface methodology,” Wiley Interdiscip. Rev. Comput. Stat., **2**(2), 128–149, 2010, doi: http://dx.doi.org/10.1002/wics.73

[13] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar, and L. A. Escaleira, “Response surface methodology (RSM) as a tool for optimization in analytical chemistry,” Talanta, **76**(5), 965–977, 2008, doi: 10.1016/j.talanta.2008.05.019.

[14] T. T. Tran, M. V. Vartanov and M. V. Arkhipov, "Application of the Part Rotation Effect for Reliability of the Robotic Assembly Process," in 2019 International Conference on System Science and Engineering (ICSSE), 25-30, 2019, doi: 10.1109/ICSSE.2019.8823286.

[15] Paul, Richard. Modelling, Trajectory Calculation and Servoing of a Computer Controlled Arm [Text] / R. Paul; Translation from English. A. F. Vereshchagin, V. L. Generozov / Ed. E.P. Popova. - Moscow: Nauka, 1976.

[16] M.G. Kristal, Обработка результатов планирования экстремального эксперимента, Volgograd, 2019

[17] L.V. Bozhkova, M.V. Vartanov, E.I. Kolchugin “Experimental setup for robotic assembly on the basis of passive adaptation and low-frequency oscillations,” Assembly in mechanical engineering, instrument making, 1, 5-7, 2009 (article in Russian with an abstract in English).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 167-176 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Modelling and Simulation of Aerodynamic Parameters of an Airship

Anoop Sasidharan Pillai, Venkata Ramana Murthy Oruganti*

Department of Electrical and Electronics Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, 641112, India

ARTICLE INFO

Article history:
Received: 21 May, 2020
Accepted: 17 July, 2020
Online: 22 July, 2020

Keywords:
Airship
Analytical method
Dynamic model
Aerodynamic coefficients

ABSTRACT

The dynamic modelling of an airship is the primary requirement in designing and developing a control system for a particular application. Extracting/predicting/modelling the aerodynamic coefficients is a crucial step towards the modelling of an airship. There is a huge amount of literature on the aerodynamic modelling of airships which presents experimental as well as analytical methods. All these techniques require some experimental data such as the geometrical data, control derivatives, etc. In this work, we are investigating an analytical technique which can calculate the aerodynamic parameters for a high altitude airship. The complete airship model is implemented using the derived aerodynamic coefficients. A MATLAB® Simulink based simulation is used for the investigation.

Nomenclature

α	Angle of attack, *deg*
β	Side slip angle, *deg*
δ_a	Aileron deflection angle, *deg*
δ_e	Elevator deflection angle, *deg*
δ_r	Rudder deflection angle, *deg*
η	Throttle ratio
γ, μ	Angles along the wind axes, *deg*
ϕ, θ, ψ	Airship's Euler angles - roll, pitch and yaw, *deg*
ρ	Density of air, kg/m^3
b	Wing span, m
b_z	Point of action of buoyant force in z direction, m
c	mean aerodynamic chord, m
C_D, C_L, C_Y	Coefficient of drag, lift and side force
C_l, C_m, C_n	Coefficient of rolling moment, pitching moment and yawing moment
d_z	Point of action of thrust in z direction, m
I_x, I_y, I_z	Moment of inertia components about x, y and z axes, kg/m^2
I_{xz}	Product of inertia about y axis, kg/m^2
m	Airship mass, kg
m_x, m_y, m_z	Apparent mass components in the x,y and z axes, kg
p, q, r	Airship angular velocities along the roll, pitch and yaw axes, deg/s
S	Wing planform area, m^2
t, T	Current simulation time and gust time period, s
T_m	Maximum engine thrust, N
u, v, w	Airship linear velocities along the x,y and z body axes, m/s
V	Airship forward velocity, m/s
V_m, V_g	Mean wind and gust velocity, m/s
x_E, y_E, z_E	Position coordinate in the earth frame, m
CoG	Centre of Gravity
CoV	Centre of Volume
EOG	Extreme Operating Gust
LTA	Lighter-Than Air

*Venkata Ramana Murthy Oruganti, Amrita Vishwa Vidyapeetham, ovr_murthy@cb.amrita.edu

1 Introduction

Man's dream of flying has been fulfilled by hot air balloons by the end of the 18th century and the realisation of a controlled flight was done by the design proposal of Jean-Baptiste Meusnier. Airship technology has a great history starting from the hot air balloons used in the medieval ages and has been widely used in the 19th century. Unfortunately, the interest in airships has been lost due to the high maintenance cost, drastic developments in air crafts and helicopters and some catastrophic tragedies such as the Hindenburg disaster [1]. By the end of the 20th century, a resurgence in the interest in airships by researchers and military agencies has occurred (as shown in Figure 1) due to the low speed, long endurance, economic, noiseless and eco-friendly operation of airships. Major applications [2] of airships include transportation of war equipment, intelligence, surveillance, reconnaissance, advertisement, telecommunication, monitoring and inspection of climate and environmental effects, inspection of man-made structures such as buildings, pipelines, railway network, and power lines, planning for urban development, research, exploration and so on.

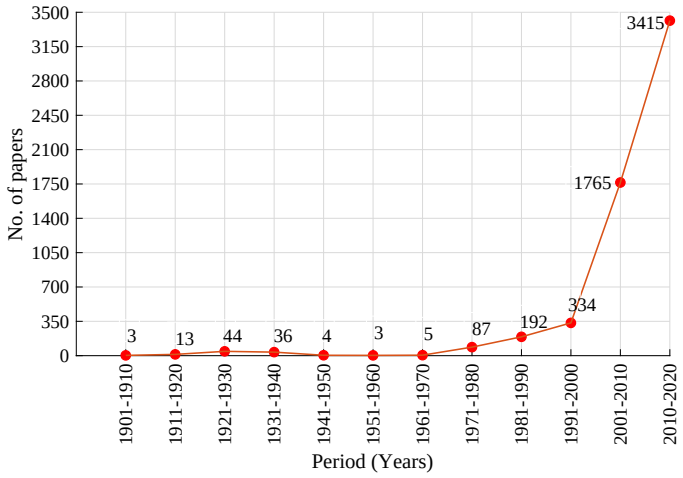


Figure 1: Number of papers published in the field of LTA vehicles (Source: Scopus database)

The airship based application has to be modelled to reflect the expected behaviour. Like any other complex system, dynamic modelling of an airship is one of the challenging areas for researchers in this field. Dynamic modelling of Lighter-Than-Air (LTA) vehicles differ from other members of aerial vehicles such as helicopters and air crafts. There is a close similarity between the dynamics of an airship and an underwater vehicle [3], which researchers exploited for modelling the airship in the early stage of research [4]. A detailed survey of the dynamic modelling of the airship is presented in [1] along with the details of wind tunnel tests; which are being utilized by researchers for the dynamic modelling and structural design of airships. A system level mathematical modelling of a high altitude airship is presented in [5]. The presented model includes the airship dynamic model, solar power system model and atmosphere models. The aerodynamic modeling of the airship was done using the geometrical data available from the literature. A linearized longitudinal and lateral dynamic models of an airship is presented in [6]. An analysis of different stability modes for the two linear models was presented. A dynamic model is presented for an airship in [7] which utilizes the geometrical aerodynamic parameters for the model. A simulation methodology is also presented to demonstrate the effectiveness of the developed model using MATLAB®.

Most of the above cited works on the dynamic modelling of airships have utilised the available database or the method of practical extraction from the actual system for obtaining the model parameters such as the aerodynamic coefficients. There are various methods by which we can estimate the parameters included in the aerodynamic model such as CFD based methods [8], experimental methods using wind tunnel [4, 9] and analytical methods [10]. Computational Fluid Dynamics (CFD) tools can be used to extract the aerodynamic coefficients for airship geometries [11, 12]. A method of extracting the force and moment coefficients of an airship using CFD and Fourier analysis is presented in [11]. Forced sinusoidal oscillations are used in the longitudinal direction to extract the force and moment coefficients. A specialised computational tool used for the aerodynamic coefficient extraction is presented in [12]. The major drawbacks of CFD based analysis are the inability to model the geometry accurately, computational cost and time taken for simulation. The traditional way is to use wind tunnel testing for the extraction of aerodynamic coefficients [4, 9]. An analytical method for calculating the aerodynamic coefficients and other control derivatives is presented in [9]. The method is heavily dependent on the wind tunnel test results. A detailed investigation of LOTTE airship using wind tunnel is presented in [4] in which the force and moment coefficients, pressure distributions and the flow visualization were elaborately discussed. There are practical difficulties in this method such as the size limitation of the test section of the wind tunnel, difficulties in the installation and inaccurate measurements, inaccuracies due to the blockages in the tunnel, etc. Most accurate and realistic data for the calculation of those coefficients can be obtained through actual test flights [10]. But it is not always feasible to conduct flight tests because of fuel consumption, unsteady atmosphere conditions, etc. There are a lot of works carried out in this context which are trying to reduce the efforts in the calculation. A method which requires no experimental data for calculating the aerodynamic coefficients using geometrical parameters is presented in [13]. The problem with this method is the inability to generalise the method to all class of airships. The parameters specified in the study are particularly available for a certain class of airships only. The digital DATCOM software based on Fortran is reported to be useful in the calculation of the aerodynamic coefficients. Such a work is presented in [14, 15] which used the DATCOM to predict the aerodynamic coefficients for a high altitude airship to manoeuvre at an altitude of 21000 m. This method is also limited to the available datasets in the DATCOM database.

There is no doubt that the above cited works are providing enough light to the field of aerodynamic modelling of airships. As per the difficulties posed by those methods, it is evident that there is still scope for research in this field using methods which need not rely much on experimental data, complex geometry development for CFD, etc. The aim of this work is to present a study on the aerodynamic modelling of a high altitude airship. The current study is motivated by the work presented in [13] which describes a method of aerodynamic modeling for the aircraft with a little ex-

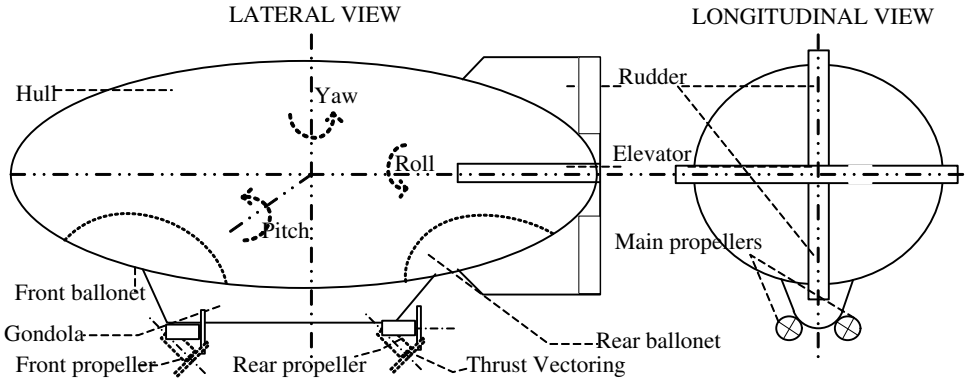


Figure 2: Basic components, actuators, and degrees of freedom of a double ellipsoid airship [16]

perimental data. We are investigating the same technique for a high altitude airship which is modelled using a modified set of aircraft equations of motion. The main advantage of this method is that the experimental data required for the modelling is considerably less. The remaining paper is presented in five sections. After the introduction and literature survey presented in Section 1, a brief description of the principle of operation of airships is given in Section 2 as a preliminary note. Different components of the airship, application of each component, principle of lift and hovering are also presented in Section 2. Section 3 discusses the complete dynamic modelling of a high altitude airship with a detailed study on aerodynamic modelling. Section 4 discusses the simulation results and their analysis. Simulation of the aerodynamic coefficients for different flight conditions and an open loop simulation of the airship are presented along with the results. Finally the conclusions and future works are presented in Section 5.

2 Airship Principle of Operation

The basic principle behind the operation of airships is the natural buoyancy of LTA gases such as hot air, helium and hydrogen. The concept of airship came from the basic hot air balloon, which uses the buoyancy property for lift generation. But to use it as a vehicle, there should be some modifications made on the balloon such as the streamlined body to reduce the air resistance and to enhance the aerodynamic properties. Figure 2 shows the major parts of a double ellipsoid airship [16]. The main body, which is known as the hull, occupies the large volume of LTA gas. Based on its structure, airships are classified as non-rigid, semi rigid and rigid airships. Non-rigid airships, also known as blimps, do not have any rigid structure to maintain its streamlined shape. It is maintained by the pressure inside the envelope. While rigid and semi rigid airships have a fully supported or partial mechanical structure on which the hull envelope is attached. Hull houses the pressure controlling ballonets, which can even control the altitude and pitch of airships. The pressure and thereby volume of the LTA gas may vary with altitude, which results in the expansion and contraction of the gas accordingly. As altitude increases, the helium expands which pushes the air in ballonets out. This will continue till the pressure height which is the designed maximum altitude of operation. Beyond the pressure height, the safety valve will release helium to avoid bursting.

Thrusters or propellers are the aerostatic actuators used for longitudinal and lateral movement of airships. A pair of propellers are placed on both sides of the gondola which provides the forward thrust. Thrust vectoring can be included to have a longitudinal freedom when the aerodynamic actuators are not active due to low wind. Apart from these main propellers, stern propellers may be used for providing a lateral degree of freedom when the aerodynamic surfaces cannot provide much actuation. Aerodynamic control surfaces are provided at the stern end of airships which can control the yaw, roll, and pitch degrees of motion. These control surfaces can be either designed as a '+' configuration or '*X*' configuration. Elevators produce a pitch degree when they act together in the same direction and a roll degree can be generated by differentially actuating the elevators. Yaw is provided by rudders which are the vertical control surfaces. Another part of the airship is the ballast mechanism to control the pitch. By using solid or liquid as a movable mass, the pitch can be controlled by manipulating the location of the centre of gravity of the airship, and sometimes the gondola completely can be the ballast.

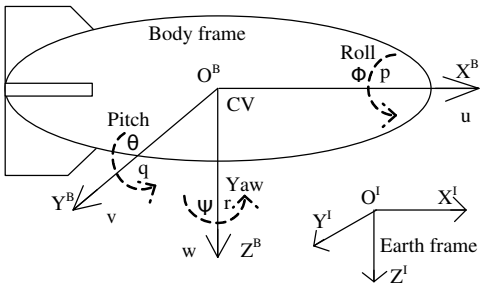


Figure 3: The reference frames used for the dynamic model along with the linear and angular position and velocities [16]

The airship considered in this study is a high altitude airship which is intended to operate at 21000*m*. The hull is of Gertler shape with four tail fins (as shown in Figure 2) and two thrusters. There are some assumptions based on which the following analysis is carried out. They are,

- Hull volume is constant throughout the simulation
- Structure of the airship is rigid
- The airship geometry is symmetric in the *XZ* plane

- Airship is at neutral buoyant condition
- A flat earth approximation is followed

3 Dynamic Model

As mentioned in Section 1, the dynamic modelling of an airship differs largely from that of an aircraft. The displaced Centre of Gravity (CoG) from the Centre of Volume (CoV) and the effect of the added mass of the surrounding air are the main contrasting features of LTA vehicle models. The displacement of the CoG vertically downwards of the CoV is to provide unconditional stability for the airship. The tendency of rotation will be counteracted by the downward displaced CoG. Even though the mathematical model of an airship differs from that of an aircraft model, it is possible to include the added mass, buoyancy and the aerodynamic properties of the airship to the aircraft model. The advantages of this methodology are the availability of well-established flight dynamic equations, availability of experimental data for different flight and airship configurations and more insight into the dynamic behaviour of the airship in the light of aircraft dynamics.

3.1 A mathematical model for an airship

For deriving the dynamic equations of a rigid body there should be a coordinate system which can properly behold the complexity, degrees of freedom and application requirements. There are mainly three reference frames earth fixed inertial frame, wind frame and body frame. The position and orientation of an object in the atmosphere with respect to the earth can be obtained with the help of earth fixed inertial reference frame. In the same way, the forces and moments of a moving body are not only depending on the velocity with respect to earth but also to the surrounding air itself. This leads to the consideration of an atmosphere fixed reference frame i.e., wind reference frame which is a moving frame with respect to earth but not with respect to the vehicle. Another set of a reference frame is the body fixed frame which is fixed to the CoV and is moving with the vehicle. The earth fixed inertia and the body fixed reference frame are shown in Figure 3 (wind frame is not included since it is considered temporarily during the derivation). The inertia frame $O_I X_I Y_I Z_I$ is having the orientation as X_I pointing towards the north, Y_I towards the east and Z_I towards the centre of the earth. Similarly, the body frame $O_B X_B Y_B Z_B$ has its centre at the CoV of the airship and X_B pointing towards the nose of the vehicle, Y_B towards the right side wing and Z_B pointing downwards perpendicular to X_B and Y_B. The wind frame is used only during the derivation of equations and the final equations will be expressed in terms of the earth and body frames only. In a general way, the dynamic equations for an airship can be represented as follows,

$$\begin{aligned}
\dot{V} &= \frac{1}{m_x}[T_m\eta\cos\alpha\cos\beta - \frac{1}{2}\rho V^2 S C_D - (mg - B)\sin\gamma] \\
\dot{\alpha} &= q - \tfrac{1}{\cos\beta}[(p\cos\alpha + r\sin\alpha)\sin\beta + \tfrac{1}{m_z V}(T_m\eta\sin\alpha + \\
&\quad \tfrac{1}{2}\rho V^2 S C_L - (mg - B)\cos\mu\cos\gamma)] \\
\dot{\beta} &= \tfrac{1}{m_y V}[-T_m\eta\cos\alpha\sin\beta + \tfrac{1}{2}\rho V^2 S C_Y + (mg - B)\sin\mu \\
&\quad \cos\gamma] + (p\sin\alpha - r\cos\alpha)
\end{aligned} \tag{1}$$

$$\begin{aligned}
\dot{p} &= \frac{1}{I_x}[(I_y - I_z)qr + I_{xz}pq + \frac{1}{2}\rho V^2 S\, bC_l - Bb_z\sin\mu\cos\gamma] \\
\dot{q} &= \tfrac{1}{I_y}[(I_z - I_x)pr + I_{xz}(r^2 - p^2) + \tfrac{1}{2}\rho V^2 S\, cC_m + T_m\eta d_z \\
&\quad \cos\alpha\cos\beta - Bb_z\sin\theta] \\
\dot{r} &= \tfrac{1}{I_z}[(I_x - I_y)pq - I_{xz}qr + \tfrac{1}{2}\rho V^2 S\, bC_n] \\
\dot{\phi} &= p + q\sin\phi\tan\theta + r\cos\phi\tan\theta \\
\dot{\theta} &= q\cos\phi - r\sin\phi \\
\dot{\psi} &= \sec\theta(q\sin\phi + r|cos\phi) \\
\dot{x}_E &= V\cos\alpha\cos\beta(\cos\psi\cos\theta) + V\sin\beta(\cos\psi\sin\theta\sin\phi \\
&\quad - \sin\psi\cos\phi) + V\sin\alpha\cos\beta(\cos\psi\sin\theta\cos\phi + \sin\psi \\
&\quad \sin\phi) \\
\dot{y}_E &= V\cos\alpha\cos\beta(\sin\psi\cos\theta) + V\sin\beta(\sin\psi\sin\theta\sin\phi \\
&\quad + \cos\psi\cos\phi) + V\sin\alpha\cos\beta(\sin\psi\sin\theta\cos\phi - \cos\psi \\
&\quad \sin\phi) \\
\dot{z}_E &= V\cos\alpha\cos\beta(-\sin\theta) + V\sin\beta(\cos\theta\sin\phi) + V\sin\alpha \\
&\quad \cos\beta(\cos\theta\cos\phi)
\end{aligned} \tag{2}$$

where

$$\begin{aligned}
\sin\gamma &= \cos\alpha\cos\beta\sin\theta - \sin\beta\sin\phi\cos\theta - \sin\alpha\cos\beta \\
&\quad \cos\phi\cos\theta \\
\sin\mu\cos\gamma &= \cos\alpha\sin\beta\sin\theta + \cos\beta\sin\phi\cos\theta - \sin\alpha \\
&\quad \sin\beta\cos\phi\cos\theta \\
\cos\mu\cos\gamma &= \sin\theta\sin\alpha + \cos\alpha\cos\phi\cos\theta
\end{aligned}$$

where the states $[V\ \alpha\ \beta\ p\ q\ r\ \phi\ \theta\ \psi\ x_E\ y_E\ z_E]^T$ are the forward velocity, angle of attack, side slip angle, body axis role, pitch and yaw rates, roll, pitch and yaw angles and position coordinate in the earth frame respectively, m_x, m_y and m_z are the components of apparent mass, T_m is the maximum engine thrust, η is the throttle ratio, ρ is the density, S is the wing planform area, c is the chord, b is the wingspan, b_z is the point of action of buoyant force along the Z_B direction, d_z is the point of action of thrust along the Z_B direction, B is the buoyant force, m is the airship mass, I_x, I_y, I_z, I_{xz} are the inertia components, C_D, C_L, C_Y are the drag, lift and side force coefficients, C_l, C_m, C_n are the rolling, yawing and pitching moment coefficients.

3.2 Aerodynamic coefficients

The set of equations described in Section 3.1 are modified aircraft equations to demonstrate the dynamics of a high altitude airship. The aerodynamic parameters in the equation such as C_D, C_L, C_Y, C_l, C_m and C_n are the major design parameters of the airship. There is an analytical method for the calculation of the aerodynamic coefficients for aircraft using the actuator values [13]. Since the airship model used in this paper is modified from the same aircraft model, this method is investigating for the airship in this paper. The method calculates the aerodynamic coefficients as a function of $\alpha, \beta, p, q, r, \delta_e, \delta_a$ and δ_r. The set of equations for the calculation of aerodynamic coefficients are given by equations (3) to (8).

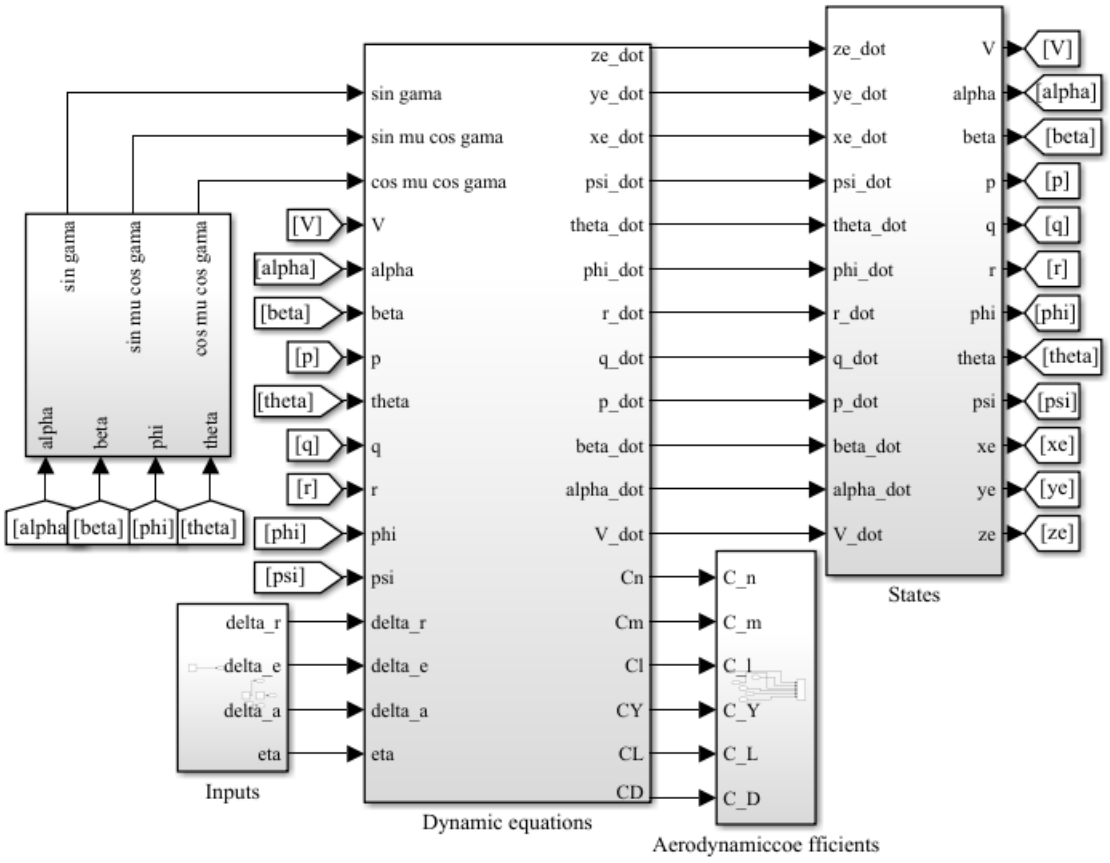


Figure 4: Simulink block diagram of the simulation of airship

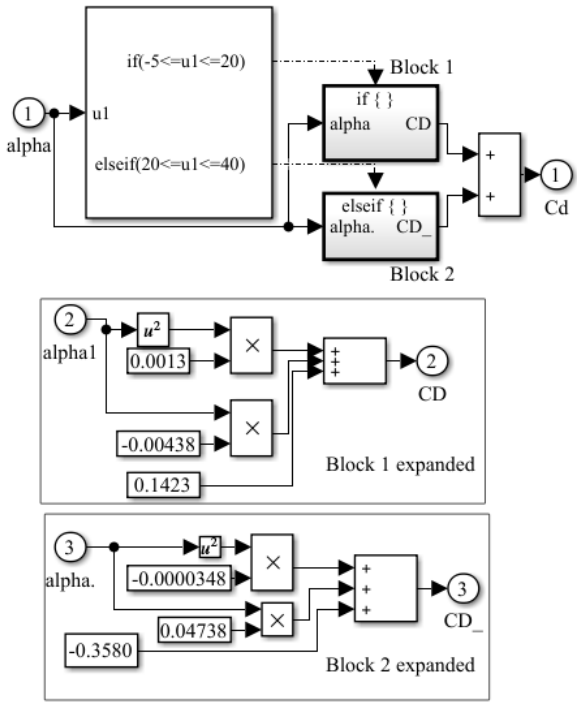


Figure 5: Simulink block diagram of the simulation of coefficient of drag

$$C_D = \begin{cases} 0.0013\alpha^2 - 0.00438\alpha + 0.1423 \\ \text{if } -5 \le \alpha \le 20 \\ -0.0000348\alpha^2 + 0.0473\alpha - 0.3580 \\ \text{if } 20 \le \alpha \le 40 \end{cases} \tag{3}$$

$$C_Y = -0.0186\beta + \frac{\delta_a}{25}(-0.00227\alpha + 0.039) + \frac{\delta_r}{30}(-0.00265\alpha + 0.141) \tag{4}$$

$$C_L = \begin{cases} 0.0751\alpha + 0.0144\delta_e + 0.732 \\ \text{if } -5 \le \alpha \le 10 \\ -0.00148\alpha^2 + 0.106\alpha + 0.0144\delta_e + 0.569 \\ \text{if } 10 \le \alpha \le 40 \end{cases} \tag{5}$$

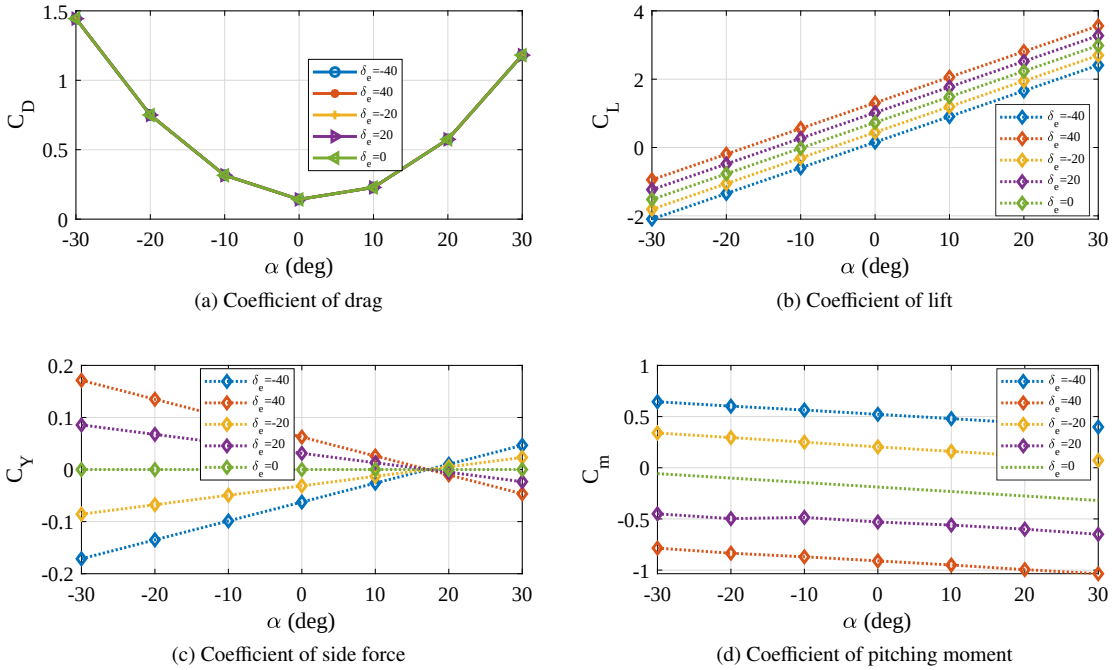
$$C_l = C_l(\alpha,\beta) - 0.0315p + 0.0126r + \frac{\delta_a}{25}(0.00121\alpha + 0.0628) - \frac{\delta_r}{30}(0.000351\alpha - 0.0124) \tag{6}$$

where,

$$C_l(\alpha,\beta) = \begin{cases} (-0.00012\alpha - 0.00092)\beta & \text{if } -5 \le \alpha \le 15 \\ (0.00022\alpha - 0.006)\beta & \text{if } 15 \le \alpha \le 25 \end{cases}$$

$$C_m = -0.00437\alpha - 0.0196\delta_e - 0.123q - 0.1885 \tag{7}$$

$$C_n = C_n(\alpha,\beta) - 0.0142r + 0.0126r + \frac{\delta_a}{25}(0.000213\alpha + 0.00128) + \frac{\delta_r}{30}(0.000804\alpha - 0.0474) \tag{8}$$



(a) Coefficient of drag

(b) Coefficient of lift

(c) Coefficient of side force

(d) Coefficient of pitching moment

Figure 6: Simulation results of the aerodynamic coefficients given by (3)-(8) for various values of δ_e

where,

$$C_n(\alpha,\beta) = \begin{cases} 0.00125\beta & \text{if } -5 \le \alpha \le 10 \\ (-0.00022\alpha + 0.00342)\beta & \text{if } 10 \le \alpha \le 25 \\ -0.00201\beta & \text{if } 25 \le \alpha \le 35 \end{cases}$$

where δ_e, δ_a and δ_r are the deflection of the elevator, aileron and rudder respectively.

4 Results and Discussion

4.1 Simulation

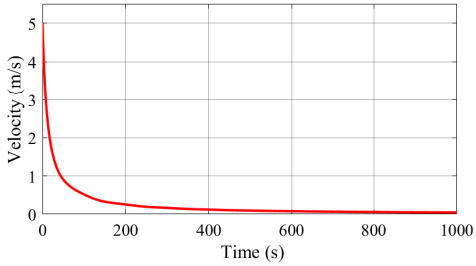
The model of the airship is implemented using MATLAB® Simulink 2018a. Simulink block diagram of the simulation is as shown in Figure 4. The 'Dynamic equations' block shown in Figure 4 holds the realisation of airship dynamic equations given in (1) and (2). The 'Aerodynamic coefficients' block in Figure 4 consists of the realisation of coefficients given in (3) - (8). As a representative figure, the simulink block diagram of the coefficient of drag expressed in (3) is shown in Figure 5. The parameters used for the simulation are given in Table 1. ODE45 solver with variable time step execution is used for the simulation. Five sets of elevator deflection angles are used for the simulation.

4.2 Simulation results

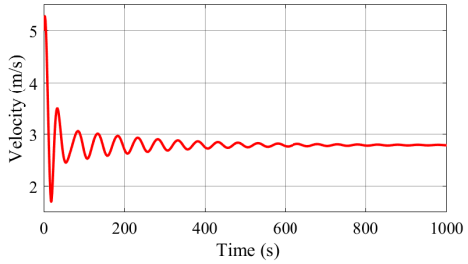
The method of obtaining aerodynamic coefficients explained in (3) - (8) is directly applicable to aircraft. It can be extended to airships also. Since there is no aileron for the airship, the value of δ_a is taken as same as δ_e. The aerodynamic coefficients represented using equations (3)-(8) are obtained for different values of α and δ_e are shown in Figure 6. C_D, C_L, C_Y and C_m are shown in Figure 6 and C_l and C_n are not shown since their values are zero for the simulated conditions. Even though we have shown the coefficients for a wide range of angle of attack starting from −30 to 30 degrees, the airship simulation later in this section shows that the angle of attack remains within 10 degrees. It is evident that the angle of attack obtained in our simulation is far below the stall angle for a stream lined airship [17]. So the problem of stalling is not in the scope of this work.

Table 1: Model parameters used for the simulation

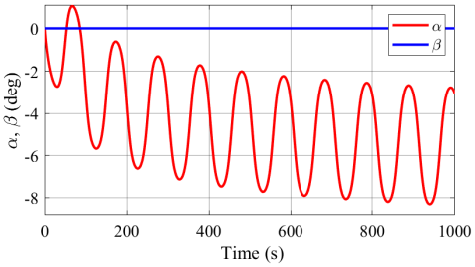
No.	Parameter	Value	Unit
1	m	23145	kg
2	I_x	17400000	kgm^2
3	I_y	245264282	kgm^2
4	I_z	245264282	kgm^2
5	I_{xz}	1920000	kgm^2
6	T_m	6000	N
7	η	23145	kg
8	m_x	25032	kg
9	m_y	43044	kg
10	m_z	43044	kg
11	T_m	6000	N
12	ρ	0.0767	kgm^3
13	S	4748	m^2
14	b_z	16.43	m
15	d_z	14.67	m
16	c	68.9	m
17	b	68.9	m



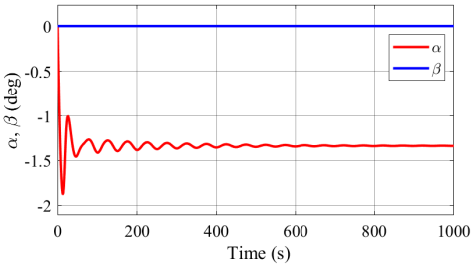
(a) Velocity, V (m/s)



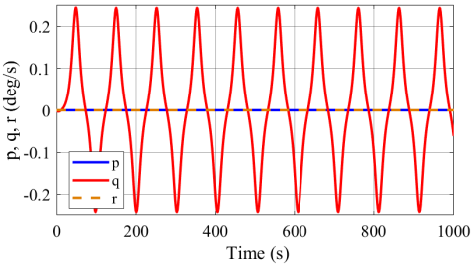
(b) Velocity, V (m/s)



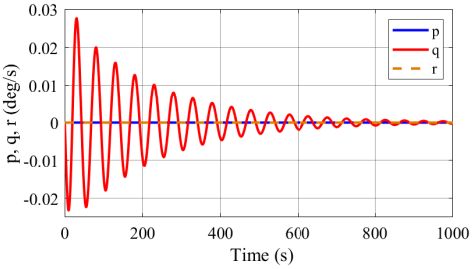
(c) Angle of attack and sideslip angle



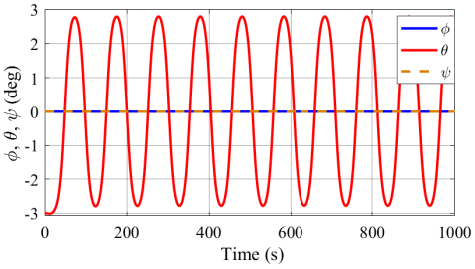
(d) Angle of attack and sideslip angle



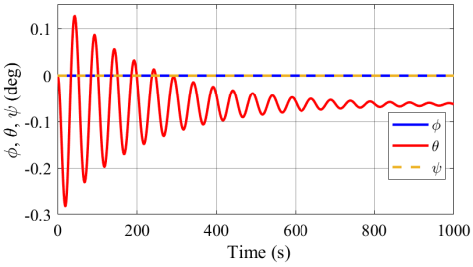
(e) Roll, pitch and yaw rates



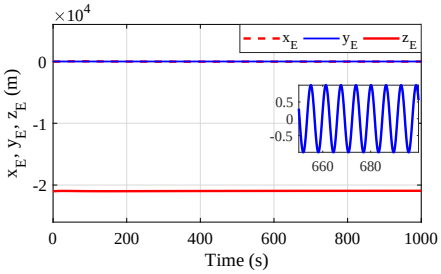
(f) Roll, pitch and yaw rates



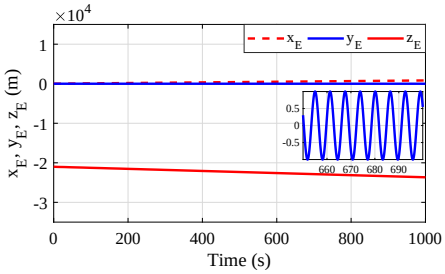
(g) Attitude



(h) Attitude

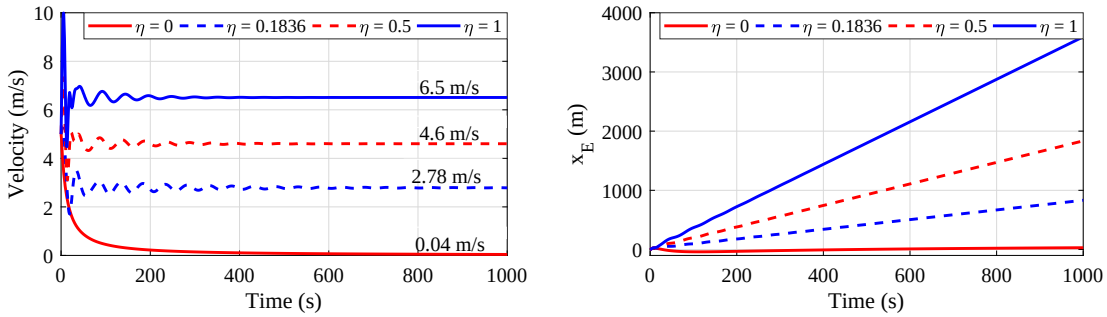


(i) Position coordinates



(j) Position coordinates

Figure 7: Simulation results for the airship model for zero thrust shown in the first column((a), (c), (e), (g), (i) and 0.1836 throttle ratio shown in the second column ((b), (d), (f), (h), (j))



(a) Forward velocity (m/s) (b) X_E (m)

Figure 8: Simulation results of the forward velocity and position along x axis for different η values

Coefficient of drag, C_D is given in Figure 6 (a). The higher values of drag at the higher angle of attacks are as expected and it is clear from the results that there is no effect of δ_e on the drag coefficient of the airship. Even though the coefficients are given for a wide range of angle of attacks, the normal operation (except the take off and landing) of airships will be in a small angle of attack. The drag coefficient at zero angle of attack is 0.1432.

Coefficient of lift, C_L, seems to have a wide range of deviation for the simulated angle of attacks as shown in Figure 6 (b). The model shows a positive lift at zero angle of attack and the above, which is a characteristic peculiarity of airships. Coefficient of side force, C_Y, and the coefficient of pitching moment, C_m are shown in Figure 6 (c) and (d) respectively.

These extracted aerodynamic coefficients are used for the model given in (1) and (2) and are simulated to see the response for different inputs. Some of them are given in the following paragraphs.

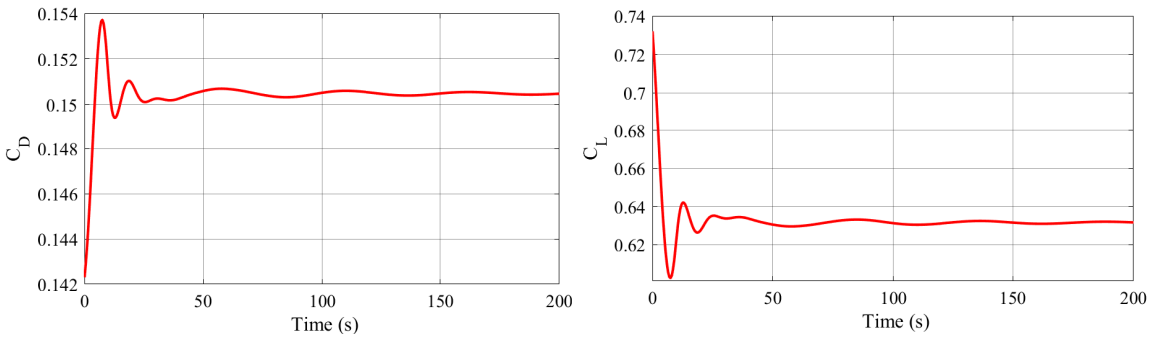
The initial condition for the 9 state variables used in (1) and (2) is as given as follows;

$$x_0 = \begin{bmatrix} V_0 & \alpha_0 & \beta_0 & p_0 & q_0 & r_0 & \phi_0 & \theta_0 & \psi_0 & x_{E0} \\ y_{E0} & z_{E0} & & & & & & & & \end{bmatrix}^T \quad (9)$$

$$= \begin{bmatrix} 5 & 3 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & -21000 & & & & & & & & \end{bmatrix}^T$$

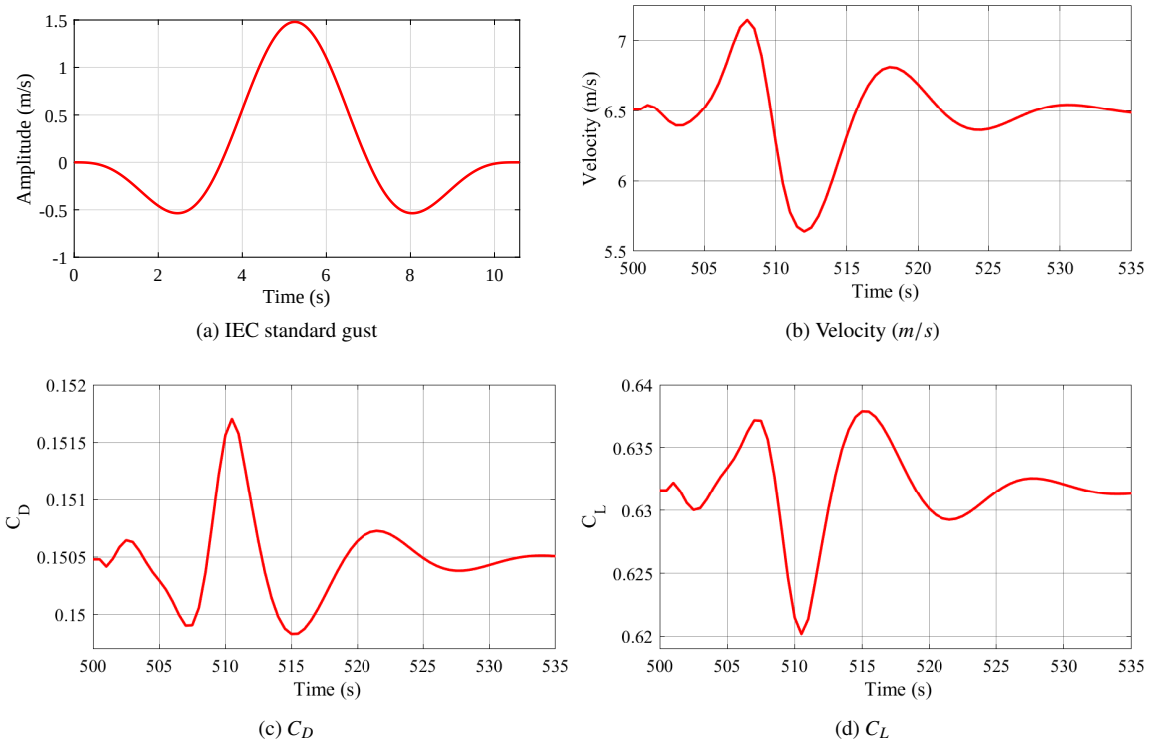
A zero input simulation of the airship is shown in the first column of Figure 7. The simulation was run for 1000 seconds. The velocity of the airship is asymptotically settling to zero as shown in Figure 7 (a). There is an oscillating negative angle of attack around -7 degree which causes a slight decrease in the altitude of the airship from the initial altitude. The side slip angle remains zero throughout the simulation as shown in Figure 7 (c). Figure 7 (e) shows the roll, pitch and yaw rates of the airship. The roll and yaw rates remained at zero throughout the simulation, but the pitch rate shows an oscillation as shown in Figure 7 (e). These oscillations are reflected in the attitude of the airship also. Figure 7 (g) shows the oscillatory pitch angle and the zero roll and yaw angles. The position of the airship with respect to the inertial frame are shown in Figure 7 (i). It is clear from the figure that the airship is staying at the initial position through out the simulation. There are slight deviations in the y direction but it is negligibly small. There is a slight oscillation in the y_E of the order of 1 m/s which is shown in Figure 7 (i).

Another set of results are shown in the second column of Figure 7 with a minimum throttle ratio of 0.1836 by keeping all other inputs zero. The velocity is settling at 2.8 m/s after a prominent overshoot and oscillations as shown in Figure 7 (b). The angle of attack and side slip angle are shown in Figure 7 (d). There is a large improvement in the magnitude of the negative angle of attack. Now the angle is -1.3 degree which can be expected due to the thrust generation off to the centre of gravity. The side slip angle remains zero. The pitch rate oscillation dies out asymptotically as shown in Figure 7 (f). The roll and yaw rates are remaining the same as in the first



(a) C_D (b) C_L

Figure 9: Transient response of the coefficient of drag and lift



(a) IEC standard gust

(b) Velocity (m/s)

(c) C_D

(d) C_L

Figure 10: Gust response of the velocity, coefficient of drag and lift

case. The attitude of the airship is shown in Figure 7 (h) in which the pitch angle θ has a negative angle of -0.6 degree. This indicates the inherent dynamics of the airship due to the displaced CoG from the CoV. An active elevator control is necessary to overcome this. Otherwise, the station keeping of the airship at the desired altitude will be difficult. The roll and yaw angles remain zero. The position of the airship is shown in Figure 7 (j). The position along the x direction shows a linear motion which gives $831m$ at the end of the simulation. There is a deviation of 3660 m in the z direction also, which shows again the necessity of an active elevator control. The y direction motion which is expected to be steady at zero value shows a sustained oscillation throughout the simulation. An enlarged view of a portion of the response (650 seconds - 700 seconds) is shown in the insight of Figure 7 (j).

Figure 8 shows a comparison of the forward velocities and position along x direction for different throttle ratios. Four different cases were considered; zero throttle ratio, initial condition specified in (9), half throttle and a full throttle. It is shown that the maximum velocity this airship can attain is 6.5 m/s. There are considerable overshoots and undershoots present in the velocity as the throttle ratio approaches its maximum. The position of airship also increases with the increase in input.

The transient response of the coefficient of drag and lift for a maximum thrust is shown in Figure 9. The drag shows a considerable overshoot from the initial value which causes a slight damped oscillations lasts for 50 seconds. This increase in drag is due to the motion initiated at the starting of simulation from rest. Similarly for the lift curve, there is a dip in the lift value mainly due to the induced drag.

Even though we haven't considered any closed loop control strategy for the airship, the response to a gust is analyzed. A standard wind gust model, Extreme Operating Gust (EOG) described in IEC standard [18], is used for the analysis. The velocity profile of the EOG gust is given in (10).

$$V_p = \begin{cases} V_m - 0.37V_g \sin(3\pi t/T)(1-\cos(2\pi t/T)) & : 0 \le t \le T \\ V_m & : otherwise \end{cases} \quad (10)$$

A representative gust profile with 10.5 seconds period, zero mean wind velocity and 2 m/s gust amplitude is shown in Figure 10 (a). Since this is an initial stage of this work, we haven't considered real wind scenario. The ability of the model to retain the steady state condition even after the occurrence of a gust is analysed. The gust is applied at 500 second which is sufficiently far from the transient region. The plot of velocity is shown in Figure 10 (b). The direction of application of gust is in the free stream direction of the airship. It is shown that after a surge the velocity falls down quickly at the maximum of the gust and retains to the steady state . Drag increases with velocity and is clearly shown in Figure 10 (c). Since a part of the lift is contributed by the aerodynamic flight, there will be a proportional increase in lift with velocity as shown in 10 (d). With this analysis the open loop stability of the airship in the presence of gust is demonstrated.

5 Conclusion

A study on the modelling and simulation of airships lead us to this work where an analytical methodology to model the aerodynamic

coefficients is investigated for an airship. The implementation of the aerodynamic coefficients given in (3)-(8) was successfully done on the mathematical model of the airship given in (1) and (2). All the six coefficients were simulated for five different values of the elevator deflection angles. Even though the drag coefficient at zero angle of attack is 0.1432, the large drag value (> 1) at a higher angle of attacks demands an optimal sizing of the airship envelope. The inability to be at the initial altitude indicates the necessity of an active position control system for the airship. A gust analysis of the airship is also carried out using IEC standard gust. The open loop stability of the airship in the presence of gust is demonstrated.

The presented method of calculating aerodynamic coefficients is originally applicable to aircraft where the effect of added mass and buoyancy are negligible. Nevertheless, a general outline on the aerodynamic behaviour of the airship can be obtained from the presented analysis. The model behaves as expected using the coefficients extracted using this method.

This study can be considered as an initial investigation of the mathematical modelling of the aerodynamic parameters of airships. As a future work, an investigation on the validation of the proposed method will be done using CFD based analysis on the selected airship configuration.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The project is funded in part by the Amrita University Fellowship.

References

[1] Y. Li, M. Nahon, I. Sharf, "Airship dynamics modelling: A literature review", *Progress in Aerospace Sciences*, **47**(3) pp. 217–239, 2011. doi: 10.1016/j.paerosci.2010.10.001.

[2] F.A. d'Oliveira, F.C.L.d. Melo, T. C. Devezas. "High-altitude platforms - Present situation and technology trends", *Journal of Aerospace Technology and Management*, **8**(3), pp. 249-262, 2006. doi: 10.5028/jatm.v8i3.699.

[3] H. Gopinath, V. Indu and M.M. Dharmana, "Development of autonomous underwater inspection robot under disturbances", in *Int. conf. on Technological Advancements in Power and Energy (TAP Energy)*, Kollam, India, pp. 1–5, 2017. doi: 10.1109/TAPENERGY.2017.8397219.

[4] S P Jones and De Laurier J D. 'Aerodynamic estimation techniques for aerostats and airships', *Journal of Aircraft*, **20**(2), pp. 120-126, 1983. doi: 10.2514/3.44840.

[5] R. P. Kukillaya, A. Pashilkar, "Simulink model development, validation and analysis of high altitude airship", Report. National Aerospace Laboratories (NAL), No. PDFMC/2017/1000, 2017. doi: 10.13140/RG.2.2.11844.22400.

[6] M. V. Cook, J. M. Lipscombe and F. Goineau. "Analysis of the stability modes of the non-rigid airship", *The Aeronautical Journal*, **104**(1036), pp. 279–290, 2000. doi: 10.1017/S0001924000091612.

[7] M Z Ashraf and M A Choudhry. "Dynamic modelling of the airship with Matlab using geometrical aerodynamic parameters", *Aerospace Science and Technology*, **25**(1), pp. 56-64, 2013. doi: 10.1016/j.ast.2011.08.014.

[8] X. Wu, Y. Wang, C. Huang, Y. Liu, L. Lu, "Experiment and numerical simulation on the characteristics of fluidstructure interactions of non-rigid airships", *Theoretical and Applied Mechanics Letters*, **5**(6), pp. 258-261, 2015. doi: 10.1016/j.taml.2015.11.001.

[9] P. Funk, T. Lutz, Siegfried. "Experimental investigations on hull-fin interferences of the LOTTE airship", *Aerospace Science and Technology*, **7**(8), pp. 603–610, 2003. doi: 10.1016/S1270-9638(03)00058-0.

[10] Y. Li, M. Nahon, "Modelling and simulation of airship dynamics", *Journal of Guidance, Control and Dynamics*, **30**(6), pp. 1691-1700, 2007. doi: 10.2514/1.29061.

[11] X. Wang. "Computational Fluid Dynamics Predictions of Stability Derivatives for Airship", *Journal of Aircraft*, **49**(3), pp. 933–940, 2012. doi: 10.2514/1.C031634.

[12] J. L. M. Junior, J. S. Santos, M. A. V. Morales, L. C. S. Ges, S. Stevanovic, R. Santana. "Airship Aerodynamic Coefficients Estimation Based on Computational Method for Preliminary Design", *AIAA Aviation Forum*, Dallas, Texas, 2019. doi: 10.2514/6.2019-2982.

[13] Y. Fan, F. H. Lutze, E. M. Cliff. "Time optimal lateral maneuvers of an aircraft", *Journal of Guidance, Control and Dynamics*, **18**(5), pp. 1106-1112, 1995. doi: 10.2514/3.21511.

[14] D. Gobiha and N. K. Sinha. "Autonomous maneuvering of a stratospheric airship", in *Indian Control Conference (ICC)*, IIT Kanpur, India, pp. 318-323, January 4-6, 2018. doi: 10.1109/INDIANCC.2018.8307998.

[15] S. Agrawal, D. Gobiha, N.K. Sinha. 'Nonlinear parameter estimation of airship using modular neural network', The Aeronautical Journal, pp. 1-20, 2019. doi: 10.1017/aer.2019.125.

[16] S. Anoop, O. V. Ramana Murthy and K Rahul Sharma. 'Analysis of airship dynamics using linear quadratic regultor controller', in 15th *IEEE India Council Int. conf. (INDICON)*, Coimbatore, India, 2018. doi: 10.1109/INDICON45594.2018.8987186.

[17] Casey Marcel Lambert. 'Dynamics Modelling and Conceptual Design of a Multi-tethered Aerostat System', Master's Thesis, Dept. of Mechanical Engineering, University of Calgary, 1999. http://www2.eng.cam.ac.uk/~hemh1/SPICE/papers.

[18] IEC 61400-2:2013. International Standard, Wind turbines-Part 2: Small wind turbines, 2006. https://webstore.iec.ch/publication/5433#additionalinfo.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 177-183 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Industry 4.0 Operators: Core Knowledge and Skills

Olayan Alharbi*

Department of Computer Science, College of Science and Humanities in Rumah, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

ARTICLE INFO

Article history:
Received: 23 April, 2020
Accepted: 07 July, 2020
Online: 22 July, 2020

Keywords:
Industry 4.0
Internet of Things (IoT)
Human Resources

ABSTRACT

One of the most important technological changes due to the arrival of Industry 4.0, an initial, gradual, and complex process of technology transfer is taking place, which strongly relies on the integration of universities, industries, and governments. In this context, to make the Industry 4.0 approach a reality, several requirements need to be met. One of them is the need to qualify people to work in industries. This research paper aims to clarify the required knowledge and learning for a person to operate the manufacturing processes associated with some of the capabilities of Industry 4.0. Interviews were conducted with individuals who are Industry 4.0 employees, including experts of technology, education vendors, and employers who are eager to develop and improve their projects. This study provides several results, the most important of which is the focus on the rehabilitation of operators using modern technologies in alignment with the Fourth Industrial Revolution.

1. Introduction

The era we are living in these days is that of the digital revolution, which first took place the last quarter of the last century. This era is characterized by the fusion of all the techniques that have been developed, so that the physical, digital, and biological fields have overlapped, and the lines between them have blurred. Therefore, there was a need to move from the third revolution to a new one which is considered something unique. There are three reasons why these transformations are unique and not merely an extension of the previous revolution.

These reasons are: the speed of change, its scope, and the impact of these transformations on the prevailing systems. The speed of the current scientific breakthroughs has no parallel in human history, since there is an amazing speed in the development of digital technologies, compared to that of the previous industrial revolutions [1]. Hence, the Fourth Industrial Revolution is the impact of technology, the Internet, and computers on various sectors of development and labor, including material science, robots, nanotechnology, three-dimensional (3D) printing, unmanned aerial vehicles, digital computing, and globalization [2,3].

The Fourth Industrial Revolution is based on the digital revolution, in which technology is an integral part of the society and a link between the digital, physical, and biological worlds. It is characterized by the use of advanced technology in various fields to improve efficiency and promote development and growth [4].

The new mechanisms in the industry focus on technical education and the development of young people skills.

in different sectors of industry, which works to support the direction of political leadership in human development and devotes attention to human capital because this is the basis of any industrial revolution and development. Investing in the human workforce has become more important than ever before, as the world faces economic and social-environmental challenge [5].

The Fourth Industrial Revolution has added new pressures on the labor markets. Therefore, lifelong learning, learning reform, and new skill retraining initiatives will be the key to ensure that individuals have an effective economic opportunity to compete in the new world of work. Companies will also have the opportunity to attract the talented employees they need for future jobs. According to the World Economic Forum, at least 54% of all employees will need to be taught additional skills by 2022 [6]. Therefore, public and private sectors are required to hire new talented employees or train their employees to gain all the required knowledge to implement the latest technologies of Industry 4.0 Revolution.

The manufacturing industry is one of the sectors that are influenced by the fourth industrial revolution. Similarly, manufacturing is moved from Industry 3.0 phase into the Industry 4.0 phase. In industry 4.0 the manufacturing machines are connected to a virtual environment for each physical machine that presents the manufacturing process in real-time [7]. Therefore, factories are one of the entities that are encouraged to retrain and

*Corresponding Author: Olayan Alharbi, Email: O.alharbi@mu.edu.sa

develop their operators to be able to gain all benefits of industry 4.0.

This research paper aims to clarify the required knowledge and learning for a worker to operate the manufacturing processes associated with some of the capabilities of Industry 4.0. This research focuses on operators whose factory is moving from Industry 3.0 into Industry 4.0. Interviews were conducted with individuals who work for Industry 4.0, including vendors of technology, education vendors, employers, and operators who work in factories capable of adopting Industry 4.0 technologies. This study provides several results, the most important of which is the focus on the rehabilitation of operators using modern technologies in alignment with the Fourth Industrial Revolution.

2. Literature Review

The influence of the educational system was obviously noticed through all the three industrial revolutions. Bryan Edward Penprase assures this point and adds that there were positive reactions in providing new educational tracks, majors, institutes for research in addition to new curriculums with the aim of preparing individuals [8]. The first industrial revolution was based on elective courses, enormous discoveries and inventions, centered on the use and adaptation of water and steam energy to the mechanization of productivity, including the gradual transformation of societies from their agricultural nature to a new identity. This new identity was characterized by industries that relied on steam engines, both for production, transport, and communication [9].

The second industrial revolution, from 1860 till 1900, was supported by educational institutes such as Pomona College (1887), University of South California (1880) and others to implement the new technologies to close the gap in operators and labor skills. The revolution was characterized by the use of electrical energy instead of steam and coal, which enabled countries with diverse natural resources to devote them to be exploited in the industry. This gave rise to industrial societies based on the principle of huge productivity. The emergence of electrical machines and their exploitation in peace and war was marked by very important inventions for the development of human civilization [10].

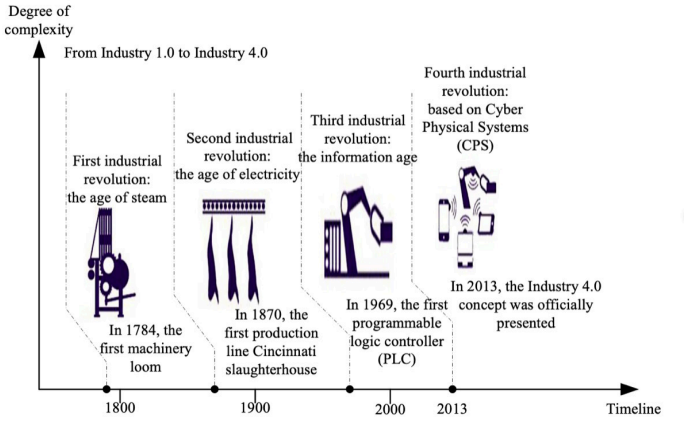


Figure 1. The four stages of the industrial revolution[11]

The third industrial revolution, during 1980s and 1990s, began with the launch of digital technology, such as computers and web-based sites on the Internet. The influence of this revolution was clear in two dimensions. These were the online based education and new majors such as software development, engineering and automated manufacturing. Numerology was used in programming machines to function without the need to human supervision, causing machines to replace the labor force more extensively than the previous industrial revolutions, see figure 1 [11].

2.1. Industry 4.0 in manufacturing

The Fourth Industrial Revolution came with digital capabilities such as the digital-twin that were used in factories to read and analyze data of machines in real time. The revolution beginning was based on the remarkable achievements of the Third Industrial Revolution, particularly, the Internet and its uses, the enormous processing power, the ability to store and recall information whenever it is needed in addition to the unlimited potential for finding the required knowledge. Those achievements have paved the way to unlimited possibilities via the breakthroughs of emerging technologies in various fields. Those fields included many useful applications, inventions and theories such as: artificial intelligence (AI), robots, Internet objects, self-driving vehicles, 3D printing, nanotechnology, biotechnology, materials science, quantum computing, Blockchain, etc [12]. The Fourth Industrial Revolution, or the so-called Industry 4.0, simply means the intensive use of technology and advanced mechanization in the manufacturing processes, the activation of Internet objects, cloud computing, and robotization in the so-called Smart Factory.

The Fourth Industrial Revolution requires new skills such as data analysis and basic information technologies [13]. By these skills, the Revolution can have many impacts on business: Influencing customer expectations, product improvement, collaborative innovation, and organizational forms [11]. Regarding consumers' expectations or businesses, customers are always the centerpiece of the economy, which revolves around how to improve the process of customer service and satisfy them. Moreover, physical products and services can be improved through digital facilities that help increase the value. New technologies provide assets with better qualities and flexibility, while data and analysis change the way these assets are maintained. At the same time, the world of customers' feedback and the data-based services, asset the evaluation process through analysis, new forms of co-operation, especially the provided speed with which innovation occurs, the emergence of global platforms and other new business models which ultimately means that with talent, new culture and organizing methods will emerge [14].

The drastic shift from the simplicity of the Third Industrial Revolution to the innovation based on a combination of technologies (the Fourth Industrial Revolution) is creating a kind of pressure on companies to reconsider the way their businesses are progressing. One of the most experienced aspects of the recent Industrial Revolution (the 4th) is the concept of service-based design. This design is different from the factory settings regarding customers products and those products of companies for specific individuals, according to their specifications. On the other hand, this means that the new products will continue to be automatically and permanently produced by manufacturers [11].

The Fourth Industrial Revolution will affect people in many ways [15,16]. It will change what we do, what we are, and will also affect our identity and everything associated with it. For instance, our sense of privacy, property concepts, consumption patterns, and the time devoted to work and leisure, as well. Furthermore, it will also improve our skills and enhance our relationships. The feedback process for the manufactured products

will be self-explanatory, which will automatically help the manufacturer develop his manufacturing method.

2.2. Benefits and challenges of Industry 4.0

The mechanisms and tools of the Fourth Industrial Revolution reduce the operational costs of factories, improving the levels of corporate profits and the good employment of machines, as well as working to improve and raise the efficiency of human capital to cope with the flourishing era of advanced industry and manufacturing technology. The Fourth Industrial Revolution offers many opportunities for human societies to achieve higher rates of economic, social, and human development in general by minimizing the production costs and consequently providing services, transport, and communication, which combine high efficiency with lower prices. Other positive factors of Industry 4.0 include a contribution to better human healthcare as it shortens the time needed to the process of development, and easily disseminates its achievements throughout the world.

According to [17], the Fourth Revolution has a number of advantages. Three of these are:

1. The speed of its development at the rate of exponential growth, as modern a technology, always drives the development and the emergence of newer and stronger technology rather than the rate of a linear growth.
2. The increase in the volume of benefits per capita in this revolution is different. In the digital age, companies need a few number of employees and materials to manufacture products with great benefits and high qualities. For digital companies, storage, transport costs and product reproduction are reduced to zero level. Some technology-based companies are developing without large capital, e.g. Instagram, WhatsApp.
3. The coordination and integration of different discoveries is widely obvious in this Revolution. The profits of digital manufacturing technology and biotechnology have been shared. Recruiting designers and architects for digital design, implementing modern materials science, and industrial biology have increased the innovation and the production quality of modern products.

At the same time, Industry 4.0 imposes unprecedented challenges on human societies. Among these we find the following:

1. The dominance of large companies in industrial production and the threat of the disappearance of approximately 50% of jobs, which poses a challenge to the spread of unemployment, especially in countries that are not ready for the transformation process.
2. The possibility of employing the capabilities and advantages of the Fourth Industrial Revolution techniques in committing illegal or immoral acts that could harm society, values, or individuals, such as the growth of cybercrime, violation of privacy, spreading hatred, extremism and false news, and preventing the development of further global views.
3. Employees of factories who are moving from Industry 3.0 into industry 4.0 need to reskill themselves to utilize the benefits. This challenge relies on the effort of several stakeholders such as employers, vendors of technologies, and educational institutes.

2.3. Technologies of Industry 4.0

The Fourth Industrial Revolution works in harmony with three technologies: A) Physical technologies which are easily recognized, because they are concrete, such as self-driving cars, production line machines in factories, 3D printing machines, and sophisticated robots [14]. B) Digital technologies which create radically new devices, software and platforms that will revolutionize the way people and businesses conduct their dealings, namely, the IoT which can create the relationship between products, services, places, and people through related technologies and platforms [18]. C) Biological technologies: Industry 4.0 includes technologies to link humans to other technologies such as physical technologies. Currently, a machine can allow only operators who are eligible to work with it based on the operator's record in the factory human resources system [19].

2.4. Human Capital in Industry 4.0

Two of the most important possibilities of human capital are knowledge and education, where they support the Fourth Industrial Revolution to achieve the desired goals, which benefit institutions. The human capital theory believes that knowledge provides greater cognitive skills to individuals, thus it is necessary to impel their productivity and efficiency in a potential way to develop the related activities.

Certain education, experience, knowledge, and skillset have to be possessed by employees. As a result, these are used to create values regarding the success of an organization. Thus, we can see how experience, knowledge, skills, and education are critical for the human capital in organizations, which, in essence, underscore the importance and the role of the human capital in the Industry 4.0 revolution.

The requirements of Industry 4.0 are not only a workforce, but also human capital, nurtured in competitive education systems, which must be well prepared for creative work environments [20]. Nowadays organizations do not require physical and tangible humans as the present and future expectations seem to offer plenty of challenges to organizations and humanity. Therefore, as humans embrace the ushering in Industry 4.0, it has become imperative for nations, as well as organizations, to establish education systems that are more focused on knowledge which is beyond what the world currently addresses. This requires creative teaching particularly to children at early ages, right up to the university levels [21].

As modes of passing examinations never produce thinkers, creators, or ingenuity, there is a need to a new move away from the traditional education systems of writing, reading, and memorization. Therefore, nations need to revolutionize their education systems to produce super-humans capable of implementing the Industry 4.0 revolution. Educational revolutions require a national culture that is supportive of such initiatives of the government, where the citizens feel they have something to contribute towards achieving the goals of the Industry 4.0 revolution. Consequently, the result is the production of human capital that is capable of benefiting the Industry 4.0 revolution [22].

2.5. Education in Industry 4.0

Education based on Industry 4.0 is defined as technology literacy, information literacy, media creativity, social competence and responsibility, workplace skills, and civic engagement. This is

because the available information dramatically increases, therefore people are required to have new skills to critically access and process content to ensure the best social communication and interaction. The Industry 4.0 revolution presents an opportunity, as well as challenges, to nations' education systems, and only those nations whose education systems are anchored in inclusiveness and technological imperatives will remain competitive. To meet all of these demands for the Industry 4.0 revolution, lifelong learning is necessary to ensure that everyone can stay well-trained. Universities have to lead the research efforts not only to identify the relevant skills, but also to produce a high-caliber workforce that own the skills necessary for the Industry 4.0 revolution [23].

Barro [24] pointed out the importance of education and the accumulation of human capital, as well as their positive and moral impact on economic growth. In his study of 1991, he examined the determinants of the economic growth of a sample of 98 countries for the period 1960-1985 and concluded that the human capital has a significant direct effect on the economic growth. In his study of the impact of education on economic growth in 2013, he pointed out that applied studies were based on the neo-classical models of growth that were developed in the 1950s and 1960s. The variables of these models were often expanded to include government policies, institutions, and the accumulation of the human capital. In his 2013 study, he tried to emphasize the importance of his model of growth from to understand the possibility of maintaining long-term growth.

Besides, [25] pointed out that the 2030 Sustainable Development Plan and Goals emphasize the importance of education in promoting change in knowledge, skills, values, and behavior to achieve greater growth sustainability. The emergence of this concept was associated with the emergence of a need to a good educational system to deal with the increasing challenges.

Since the Fourth Industrial Revolution began to appear in the labor market, graduates have to prepared for it, and quality has to be achieved within the educational and research institutions in addition to applying an expansion of modern study disciplines. Curriculum development has become necessary to suit the Fourth Industrial Revolution, as well as enhancing the abilities and skills of students and faculty members in the field of scientific research and innovation.

3. Method

The chosen method was selected by the author based on his approach which was influenced by the phenomenological approach to sociology as it has been used in similar studies [26], [27] [28]. The method was chosen as it helps to understand all the required knowledge areas related to the 4th Revolution. To understand the required knowledge and skills for operators in the early adoption of Industry 4.0, interviews with individuals who work for Industry 4.0, including vendors of technology, education vendors, employers and employees (operators and managers) were conducted. It is believed that individuals who experienced the transaction from Industry 3.0 to Industry 4.0 can identify the required skills and knowledge for the operators who are going to work in Industry 4.0.

3.1. Data collection

The interviews were planned, typed up, and arranged in a way that allowed for enough freedom in discussing related issues. An introductory question was developed to warm up the discussion and to let respondents provide important information. Additional questions were prepared for cases where the respondent was not sure what was meant by the questions or if they fell short in answering a particular question. Permission to record and take notes was always obtained before an interview.

3.2. Data Analysis

The collected data from the semi-structured interviews were then tested thematically. Thematic analysis refers to the identification and analysis of patterns of meaning found in a specific set of data [29]. Specifically, it is a method used to identify, analyze and report patterns in the collected data, which helps to provide an organized, descriptive, and rich interpretation of those data. To correctly identify themes, the analyst needs to understand what counts as a theme. Themes grasp important aspects of the dataset that have a connection to the research question, representing a level of pattern or concept that is found in the analyzed data [30].

This research employed thematic analysis primarily for identifying, interpreting, and explaining the different aspects of the research topic. The adoption of this method in the research was due to its suitability for analyzing the collected data, concerning the research objectives. In addition to other advantages, thematic analysis was also an appropriate match to the research paradigm. To ensure a thorough analysis, the six steps proposed by [30] were followed. These steps are familiarization with the data, generation of initial codes, search for themes, revision of themes, definition, and naming of themes, and, finally, production of the report. The report shows the connections of the qualitative data in answering the research question.

The layered approach was used to strategically recruit participants that could provide us with insights into the required knowledge and skills for operators to work in an Industry 4.0 factory. The researcher interviewed technology vendors first because they would have had the experience of the new technology and the required skills needed for Industry 4.0.

Technology vendors provided comments about Industry 4.0 in general. While technology vendors explained their opinion about the general requirement of building Industry 4.0 solutions, the discussion was led to focus on the new skills required for operators to work with such technology. Moreover, technology vendors shed light on the essential features of Industry 4.0 and the skills required to gain the benefits of transitioning to Industry 4.0 or the quick wins of the transaction. Thus, the interview was led to a discussion on the required skills and knowledge for operators to work in factories that apply Industry 4.0. Furthermore, technology vendors provided stories about how technology has changed over time.

After clarifying the theme of the new technologies of Industry 4.0, participants who had been involved in working in factories that apply some of the Industry 4.0 technologies were recruited. In the second round, people of different positions on the employer side, starting with a chief Executive Officer (CEO) of a rigid plastic factory were interviewed. Lastly, participants from education vendors that have programs for preparing operators to find the missing knowledge in traditional operator programs and the required knowledge and skills for Industry 4.0 operators were recruited.

The first round consisted of eight interviews of technology vendors and was conducted in the summer of 2017 to find the required knowledge and skills for being an operator in Industry 4.0.

The second round consisted of 17 interviews that aimed identify the required knowledge and skills for Industry 4.0, based on the feedback of users and employers of Industry 4.0. The users and employers applied the essential basics of Industry 4.0, such as connecting the machine and starting it, to automatically retrieving production data from the machine immediately. These users are those who, for instance, use manufacturing execution systems to show dashboards for data that are retrieved from the machines directly.

The last round consisted of interviewing educational vendors that teach and prepare operators to work in manufacturing. Educational vendors, who are designing new courses to bridge the gap between Industry 3.0 and the early features of Industry 4.0 (e.g., such as data visualization and connecting production machines), were chosen. This round consisted of interviewing eight participants. In total, 27 individuals were interviewed in person and six over the phone, all of whom had been involved in Industry 4.0 in many different capacities (see Table 1 for a partial list).

Table 1. Selected interviewee roles.

Title	Number of interviews
Factory CEO	1
Factory manager	1
Production manager	1
Operator	3
IoT manager and leader	3
Industry 4.0 pre-sale architectures	3
Program coordinator	1
Instructor	2

3.3. Themes

The collected data were carefully reviewed, looking for the barriers, challenges, and difficulties addressed by respondents, either implicitly or explicitly. The findings were coded and clustered into main and sub-categories. The issues, concepts, and patterns were re-categorized, and refined.

Among the several steps taken to analyze the collected data, the interviews were revisited to find and list all of the mentioned challenges, which were then organized into a list. The list was then refined, excluding repeated issues and combining similar challenges. Similar basic issues were clustered to form organizing themes.

This step reduced the list into organizing themes, which were then combined under wider classifications. The themes are discussed underneath.

The first round of interviews targeted the Industry 4.0 solution vendors. Firstly, the focus of the interviewees was shifted to the new IT capabilities to be added to factories as required competencies for Industry 4.0. These capabilities include edge connectivity, cloud solutions, and microservices. Additionally, there are certain prerequisites, such as the ability to connect a production machine and the availability of process and production sensors.

However, the focus of this research was on the operators. Therefore, the interviewer led the conversation to focus on the new set of skills required for Industry 4.0 operators. After clarifying these concerns, the technology vendors started by stating the tools and solutions, focusing on Industry 4.0 operators, as follows:

One of the essential features of Industry 4.0 is its digital twin. A digital twin is a digital replica of a living or non-living physical entity. By bridging the physical and the virtual world, data are transmitted seamlessly, allowing the virtual entity to exist simultaneously with the physical entity. Operators will be able to obtain different sources of manufacturing data, beyond what their old friend, the human-machine interface (HMI), could provide them with. The digital twin in Industry 4.0 can provide real-time data for the targeted machine and can compare it with similar machines or the process and production data of the same machine historically. Operators need to have essential knowledge of data science activities. These activities include data analysis, descriptive analysis, and problem-solving [7].

Technology providers have industrial platforms that empower users to use and rebuild micro services with minimum programming knowledge. Predix by General Electric (GE) and MindSphere by Siemens are examples of these platforms. Operators will gain access to such technology and could develop basic services for their common issues. These microservices do not require advanced programming skills to be utilized by an operator; however, basic knowledge of programming, web development, and problem-solving are required for operators who want to run the extra mile and develop their tools.

On the other hand, operators work closely with the operational technology that is provided on the machine side. Industry 4.0 will allow operators to access real and historical manufacturing data from different resources. Therefore, operators need to have the ability to work with IT solutions, such as manufacturing execution systems.

The second-round targeted employers and business owners that recently empowered some of their production lines with Industry 4.0 tools and solutions.

The first concern of an employer (a production manager) was that the operators will remain essentially the same. The operators will not be developers for Industry 4.0, but they will need to know how to work and use Industry 4.0 tools. The operators will still need to have the essential instrumentation and process knowledge. However, they will have access to new tools that will help them to improve the productivity of their machines. An example of a new tool is historian trends. This is an analytical tool that allows users to present different process data in one dashboard to analyze a certain issue, such as major downtime.

Surprisingly, a CEO and one of the production managers stated that the English language is important for empowering operators to get the most from Industry 4.0 tools and solutions. The available knowledge and most of the provided solutions are available only in English.

On the other hand, IT literacy is another important prerequisite for operators to utilize Industry 4.0 tools and solutions. IT literacy is different from digital literacy or informatics literacy/computer science literacy, as it puts more emphasis on computer programming, algorithms, and other important mathematical and computational concepts. IT literacy includes an essential area, such as information management, integrated system technology, platform technology, and system paradigms.

The last round consisted of interviews with educational institutions. All of the interviewed educational institutions had an ongoing contract involving the training of operators for local employers. One institution had contracts with food and beverage packaging manufactures and two had contracts with petrochemical, oil, and gas industries. The first theme introduced by the institution members was the IoT. New operators will work on machines that are connected to edge or cloud computing services. The existing sensors in the machine or the new ones will have the same concept of the IoT. Operators will be introduced to the concept of IoT and how platform technology can help. This new theme works smoothly with most of the industrial platform microservices, which are developed to interact with the physical assets to provide better information for operators and the rest of the factory team.

Similar to technology vendors, the interviewees from the educational institutions listed the following concepts to be added to the required skills of Industry 4.0 operators: (1) Basic programming and problem-solving, (2) web development and advanced analytical skills, and (3) basic data science activities, such as anomaly detection. Additionally, IT literacy is a topic that is required by all three different stakeholders. Table 2.0 lists all of the common required knowledge areas, as indicated by all of the stakeholders.

The educational institution interviewees agreed that operators are required to know about programmable logic controllers (PLC), instrumentation, process control, and automation. However, these are not new, and there are diplomas for operators.

In summary, the results from all of the interviews could be organized into two main categories:

1. Operation technology categories which include PLC microcontrollers, pneumatics and hydraulics, and industrial instrumentation.
2. Information technology (IT), which is the goal of this study and identifies the missing knowledge for Industry 4.0 operators as follows: digital industry platform; computer skills and information technology (IT literacy); data analytics and visualization; basic programming and problem-solving skills; introduction to web development; and industrial IoT.

4. Results and Discussion

According to the conducted interviews, the study provided several results, including the following:

It is necessary to have certain knowledge for operators working in factories adopting industry 4.0. Moreover, it is not a must that the required knowledge should be of a high level. This knowledge can be provided in vocational and technical institutes that offer a two-year diploma after finishing secondary school. This was clear in the interview with one of the interviewees who mentioned that a few number of courses can be added to the current ones at vocational institutes. The new added courses, which match the conditions of the work fields, are attractive factors to students at the vocational institutes. Those students chose to study there as their GPA at secondary schools did not enable them to join universities but on the other hand, gave the chance to get a good job immediately after finishing their diplomas.

Table 2. Selected interviewee roles.

Stakeholder Title	Recommended Course/Knowledge
Technology Provider	1. Programming and web development foundation 2. Industrial platform technologies 3. Analytical tools and problem-solving skills
Educational Institution	1. Information and communication technology 2. Programming and web development foundation
Employer	1. Information and communication technology 2. Human-machine interface (HMI) 3. Programmable logic controllers (PLC) 4. Instrumentation 5. Process control 6. Automation 7. Familiarity with technology 8. Programming and web development foundation 9. Data science

Furthermore, the results shed light on the role of the private sector firms in supporting the 4th revolution. One of the interviewees has mentioned that the institute was established by several companies, and added that the courses to be studies are related to the future needs of those companies.

Many modern technologies must be available to suit the Fourth Industrial Revolution, including edge connectivity, cloud solutions, and microservices. Additionally, certain prerequisites must be met, such as the ability of a production machine to be connected and the availability of process and production sensors. The operator is one of the most important factors supporting the Fourth Industrial Revolution due to the fact that he/she, in most cases, is the first user of industry 4.0 products rather than developers or designers. Consequently, he/she must be rehabilitated and trained for that purpose.

The required knowledge can be achieved through:

1. Basic knowledge of programming, web development, and problem-solving, which are required for operators who want to run the extra mile and develop their tools.
2. The ability to work with IT solutions, such as manufacturing execution systems.
3. Mastering the English language, which is important to empower operators to get the most of Industry 4.0 tools and solutions.
4. IT literacy, which is an important prerequisite for operators to utilize Industry 4.0 tools and solutions.
5. Knowledge of PLC, instrumentation, process control, and automation.

6. Knowledge of digital twins (one of the essential features of Industry 4.0). A digital twin is a digital replica of a living or non-living physical entity. By bridging the physical and the virtual world, data are transmitted seamlessly, allowing the virtual entity to exist simultaneously with the physical entity. Operators will be able to retrieve different sources of manufacturing data from those which their old friend, the human-machine interface (HMI), provided them with. The digital twin in Industry 4.0 can provide real-time data for the targeted machine and can compare it with similar machines or the process and production data of the same machine historically.
7. Essential knowledge of data science activities. These activities include data analysis, descriptive analysis, and problem-solving.
8. Knowledge of how to work and use Industry 4.0 tools, without being a developer for Industry 4.0.

5. Conclusion

In light of the rapid changes in financial, economic, political, and social systems, our world is on the threshold of a Fourth Industrial Revolution, which will not only change the form of industries and the methods of production but will impose new requirements on humans. This research paper provides strong evidence of the importance of the role human capital plays in the Industry 4.0 revolution in which the success or failure of most organizations largely depends on how their human capital is managed. This is because the Industry 4.0 revolution provides a space where employee-machine interactions are the order of the day. There is an interconnectedness among the various players and actors. The interfaces created to become the connecting points between workers and machines. The features of the Industry 4.0 revolution require creative and inventive workers i.e. those who are not just creative but are also knowledgeable and have the technical expertise required to work in such environments. Such workers are nurtured through an education system, where creativity, inventiveness, knowledge, and technology flourish and are entrenched in the national culture.

References

[1] M. Peruzzini, F. Grandi, M. Pellicciari, "Benchmarking of tools for user experience analysis in industry 4.0. Procedia Manufacturing", 11, 806–813, 2017. https://doi.org/10.1016/j.promfg.2017.07.182.

[2] J. Bloem, M.V. Doorn, S. Duivestein, D. Excoffier, R. Maas, E.V. Ommeren, "The fourth industrial revolution", Things Tighten 2014, 8, 2014.

[3] G. Li, Y. Hou, A. Wu, "Fourth Industrial Revolution: technological drivers, impacts and coping methods", Chinese Geographical Science 2017(27), 626–637, 2017. doi: 10.1007/s11769-017-0890-x

[4] M. Mindas, S. Bednar, "Mass customization in the context of industry 4.0: implications of variety-induced complexity". Advanced Industrial Engineering 2016, 21–38, 2016.

[5] N.W. Gleason, "Higher education in the era of the fourth industrial revolution; Springer, 2018. https://doi.org/10.1007/978-981-13-0194-0_6;

[6] R. Samans, "Globalization 4.0 shaping a new global architecture in the age of the Fourth Industrial Revolution: A call for engagement. In Proceedings of the World Economic Forum Report; 2019.

[7] T.H.J. Uhlemann, C. Lehmann, R. Steinhilper, "The digital twin: Realizing the cyber-physical production system for industry 4.0. Procedia Cirp 2017, 61, 335–340, 2017. doi: 10.1016/j.procir.2016.11.152.

[8] B.E. Penprase, "The Fourth Industrial Revolution and Higher Education. In Higher Education in the Era of the Fourth Industrial Revolution; Gleason, N.W., Ed.; Springer: Singapore, 2018; 207–229, 2018. doi: https://doi.org/10.1007/978-981-13-0194-0_9.

[9] D. Plinta, "New information technologies in production enterprises, Advanced -industrial engineering, Industry 4.0, Bielsko-Bia, 7-20, 2016.

[10] F. Zezulka, P. Marcon, I. Vesely, O. Sajdl, "Industry 4.0–An Introduction in the phenomenon. IFAC-PapersOnLine 2016, 49, 8–12, 2016. doi: https://doi.org/10.1016/j.ifacol.2016.12.002.

[11] K. Zhou, T. Liu, L. Zhou, "Industry 4.0: Towards future industrial opportunities and challenges. In Proceedings of the 2015 12th International conference on fuzzy systems and knowledge discovery (FSKD); IEEE, 2015; 2147–2152, 2015. doi: https://doi.org/10.1109/FSKD.2015.7382284.

[12] N. Benias, A.P. Markopoulos, "A review on the readiness level and cyber-security challenges in Industry 4.0. In Proceedings of the 2017 South Eastern European Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM); IEEE, 1–5, 2017. doi: 10.23919/SEEDA-CECNSM.2017.8088234 ·.

[13] A. Benešová, J. Tupa, "Requirements for education and qualification of people in Industry 4.0" . Procedia Manufacturing, 11, 2195–2202, 2017. doi: https://doi.org/10.1016/j.promfg.2017.07.366.

[14] M. Hermann, T. Pentek, B. Otto, "Design principles for industrie 4.0 scenarios. In Proceedings of the 2016 49th Hawaii international conference on system sciences (HICSS); IEEE, 3928–3937, 2016. doi: https://doi.org/10.1109/HICSS.2016.488.

[15] M. Armstrong, S. Taylor, "Armstrong's Handbook of Human Resource Management Practice: Edition 13; Kogan page, 2014;

[16] L. Gehrke et al., "Discussion of Qualifications and Skills in the Factory of the Future: A German and American Perspective.; 2015.

[17] H. Kohl, "Holistic Approach for Human Resource Management in Industry 4.0, 2016. doi: https://doi.org/10.1016/j.procir.2016.05.102.

[18] T. Bauernhansl et al., "Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung-Technologien-Migration; Springer, 2014. doi: https://doi.org/10.1007/978-3-658-04682-8;

[19] S. Vasin et al., "Emerging Trends and Opportunities for Industry 4.0 Development in Russia. ERSJ 2018, 21, 63–76, 2018. doi: 10.35808/ersj/1044.

[20] M. Baygin, H. Yetis, M. Karakose, E. Akin, "An effect analysis of industry 4.0 to higher education", In Proceedings of the 2016 15th international conference on information technology based higher education and training (ITHET); IEEE, 1–4, 2016. doi: https://doi.org/10.1109/ITHET.2016.7760744.

[21] S. Weyer, M. Schmitt, M. Ohmer, D. Gorecky, "Towards Industry 4.0 - Standardization as the crucial challenge for highly modular, multi-vendor production systems", IFAC-PapersOnLine 2015(48), 579–584, 2015. doi: https://doi.org/10.1016/j.ifacol.2015.06.143.

[22] B. Sivathanu, R. Pillai, "Smart HR 4.0–how industry 4.0 is disrupting HR. Human Resource Management International Digest 2018. doi: 10.1108/HRMID-04-2018-0059 .

[23] C. Burmeister, D. Lüttgens, F.T. Piller, "Business model innovation for Industrie 4.0: Why the "Industrial Internet" mandates a new perspective on innovation. Die Unternehmung 2016, 70, 124–152, 2016. doi: 10.5771/0042-059X-2016-2-124.

[24] R.J. Barro, "Education as a determinant of economic growth. 2002. doi: 10.1016/j.sbspro.2015.07.156.

[25] A.E. Wals, Y. Mochizuki, A. Leicht, "Critical case-studies of non-formal and community learning for sustainable development; Springer, 2017. doi: https://doi.org/10.1007/s11159-017-9691-9;

[26] F. Baena, A. Guarin, J. Mora, J. Sauza, S. Retat, "Learning Factory: The Path to Industry 4.0. Procedia Manufacturing 2017(9), 73–80, 2017. doi:10.1016/j.promfg.2017.04.022.

[27] I. Neaga, "Applying Industry 4.0 And Education 4.0 To Engineering Education", Proceedings of the Canadian Engineering Education Association (CEEA) 2019. doi: 10.24908/pceea.vi0.13859.

[28] P.A. Quezada-Sarmiento et al., "Curricular design based in bodies of knowledge: Engineering education for the innovation and the industry. In Proceedings of the 2016 SAI Computing Conference (SAI); IEEE: London, United Kingdom, 843–849, 2016. doi: https://doi.org/10.1109/SAI.2016.7556077.

[29] H. Joffe, "Thematic analysis. Qualitative research methods in mental health and psychotherapy, 1, 2012. doi: https://doi.org/10.1002/9781119973249.ch15.

[30] V. Braun, V. Clarke, "Using thematic analysis in psychology. Qualitative research in psychology 2006(3), 77–101, 2006. doi: 10.1191/1478088706qp063oa.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 184-191 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Contextualization of the Augmented Reality Quality Model through Social Media Analytics

Jim Scheibmeir[1,*], Yashwant Malaiya[2]

[1]*Department of Systems Engineering, Colorado State University, Fort Collins, 80523, United States*

[2]*Department of Computer Science, Colorado State University, Fort Collins, 80523, United States*

ARTICLE INFO

Article history:
Received: 15 June, 2020
Accepted: 12 July, 2020
Online: 22 July, 2020

Keywords:
Augmented Reality
Software Quality
Social Media Analytics
Quality Model

ABSTRACT

Augmented Reality applications are gaining popularity while maintaining novelty. Many industries are utilizing the user interface type, and use cases are becoming repeat patterns of problem solutions. Despite this rising popularity, quality has not matured nor has the technology become mainstream. Novelty must be approached as risk, and risk must be evaluated for and tested to assure adequate levels of quality. Quality itself can also be vague and have contextual definition. For these reasons, a quality model for augmented reality was created. This work analyzes over two hundred thousand tweets, collected during 2019 and 2020, relating to augmented reality technology, and contextualizes various data points to the established AR Quality Model. The education industry had the highest mentions among the tweets within the scope of this research while the tweets labeled to the transportation industry had the highest sentiment. Furthermore, the tweets were shown to illustrate the needs of testing against the characteristics within the quality model; presence, perspective, interaction, portability and persistence.

1. Introduction

This paper is an extension of research work producing an Augmented Reality (AR) quality model and originally presented in IEEE Ubiquitous Computing, Electronics & Mobile Communication Conference [1]. Augmented reality was first coined as a term by Caudell and Mizell in 1990 [2]. Augmented reality is accomplished through an application's user interface by merging both physical surroundings as well virtual objects, which may be models or scenes [3].

Research of AR applications covers many different use cases and industries. One study has uncovered findings such as children seem to understand how to use the augmented reality application as soon as they began to interact with the technology. Another research effort found learning English utilizing AR can help students in learning material and motivated students to learn the language. Another study proposes AR mobile applications in the household to educate parents and caregivers of potentially dangerous objects in the home that may endanger infants. Easily and quickly comprehended how innovative the space is becoming, and how useful for humans. In fact, Gartner, Inc. has predicted that "by 2021, at least one-third of enterprises will have deployed a multiexperience development platform (MXDP) to support mobile, web, conversational and augmented reality development" [4].

1.1. The Augmented Reality Application Quality Model

The importance of this technology extends into many industries. AR has been utilized in applications for factory design [5]. Military equipment maintenance is another use case where AR has been applied [6]. Yet another example is in the use of AR for teaching new electronics assembly skills [7].

The effectiveness of the applications requires them to have levels of quality. One model that identifies eight characteristics of software quality is the ISO 25010 model: Functional, Performance, Compatibility, Usability, Reliability, Security, Maintainability, Portability [8]. However, AR applications have not matured, and expectations of AR application users are not currently known [9]. Because the AR interface is a novel one, the intent of traditional usability evaluation methods which have matured over time for traditional software quality are inadequate [10]. For these reasons, a quality model was established that encapsulates the characteristics for AR applications broadly as Presence, Perspective, Interaction, Immersion, and Persistence [1].

*Corresponding Author: Jim Scheibmeir, Email: jimscheibmeir@gmail.com

Beyond the characteristics of quality is the necessity to utilize context in testing. Emerging technologies and designs such as digital twins must still begin with system needs analysis through known methods such as context diagrams and objective trees [11]. And the same is true for established technologies. According to Gartner, "your strategy for API quality needs to be built with the business and environment context in mind" [12].

These contextual concerns drive much of this extended research. Utilizing over 200,000 tweets towards AR and VR, this research contextualizes quality regarding AR and various industries using text analysis with mention, time series and sentiment analysis.

2. Methodology

2.1. Collection

Beyond some popular and academic notes, this research methodology focuses on social media analytics. Data acquisition is completed through a small program written in R and executed daily beginning in August of 2019. The program continues to execute, as of the publication of this research, populating a MySQL database hosted in AWS RDS. The program uses the twitteR library as an API to Twitter social media platform. Figure 1 omits various keys and passwords yet implements the basics of importing the twitteR library and establishing connections to both Twitter and the MySQL database prior to searching on '#AR' and storing the search results.

```
1  library(twitteR)
2  library(RCurl)
3  library(ggplot2)
4  library(reshape)
5  library(sentimentr)
6  library(RMySQL)
7
8  consumer_key <- ''
9  consumer_secret<- ''
10 access_token <- ''
11 access_secret <- ''
12
13
14 register_mysql_backend("", "", "", "")
15 setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
16
17 dbartweets <- search_twitter_and_store('#AR', retryOnRateLimit = 20, lang = "en")
```

Figure 1 Screenshot of example usage of twitteR library within R Studio

Many hashtags and twitter profiles names were utilized in the searches. The next table identifies some of the additional search items, extending the code found on line 17 of Figure 1.

Table 1 Search Terms used in the twitteR API Calls

'#AR'	'@AR_Maxst'
'#VR'	'@wikitude'
'#augmentedreality'	'@Vuforia'
'# omnichannel'	'#sparkar'
'#multiexperience'	'@fbplatform'
'#virtualreality'	'#arkit'
'@GoogleARCore'	'#arkitnews'
'@ArcoreGoogle'	'@AR_Maxst'

The tweets are stored with other certain meta data, made available through the twitteR library and API. Fields beyond the text collected and utilized in this research include discrete data fields towards the number of times a tweet has been marked as a favorite by a Twitter user, as well as the number of times a tweet has been re-tweeted. A date field within the table indicates the date the tweet was created in the Twitter platform.

2.2. Data Preparation and Processing

Social media analytics is not without data quality and research concerns. Tweets lack standards, such as in hashtag usage. A tweet that embeds #AR may be about Arkansas, a state within the United States, and not about augmented reality technology. Prior to formulating findings and information from the collected data, much data preparation and cleaning are required. Data preparation activities may account for 80 percent of time invested in data science effort [13]. The amount of data dropped during preparation from the data set can be large. This is to be expected, as with the example of #AR as Arkansas, and not augmented reality. The search algorithm sought out technology vendors, whose tweets may have been towards other products, process, marketing or communication needs. Tweets are also fixed in length, 280 characters [14]. A short message may come from a targeted technology provider, but also fail to cite key search criteria words. Within this research set, counting only the tweets that mention specifically the words "augmented" or "virtual" (not case sensitive), the original 637,384 tweets were reduced by two-thirds to 211,269 tweets. The breakdown of augmented and virtual tweet counts is illustrated in Figure 2.

The intention of data preparation and cleaning is to move the data from chaotic towards precision [13]. In this research, we create useful subsets without disrupting the raw database contents, as new hypothesis and meaning may yet to be executed and found across the raw data.

To create useful datasets, we must also create new fields, such as categorical values. Tweets have some indication to business and industry, so categorizing the tweets by industry lends toward understanding a contextual need for quality, such as HIPAA requirements and testing concerns towards health or pharmaceutical implementations of AR. A small program in R reads the tweet texts from the accumulated corpus and maps to common industries, counting the industry mentions, including if multiple industries are mentioned in a single tweet.

3. Results

The first finding is that among all tweets that mention either "augmented" or "virtual", augmented reality mentions (135,319) are more frequent than virtual reality mentions (80,540), illustrated below.

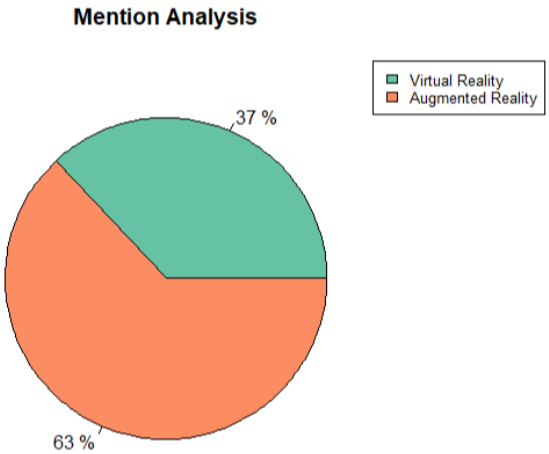


Figure 2 Percentage of "Augmented" versus "Virtual" mention in tweets

This first finding seems to contradict a 2018 survey by Gartner, where survey recipients responded towards virtual-reality apps (20%) more than augmented-reality apps (14%), in terms of apps that will have most impact on business success by year-end 2020 (shown in Figure 3). Although, the Gartner report analysts stated, "It is, however, surprising to see VR apps identified as the second most impactful type of multiexperience app (20%), as AR has more potential use cases and device support. But only 14% of the respondents thought that AR apps would be the most impactful, despite AR app development tools being more widely available. "[15]. The following chart shows where virtual-reality apps were more frequent in survey recipient response compared to augmented-reality apps for business success.

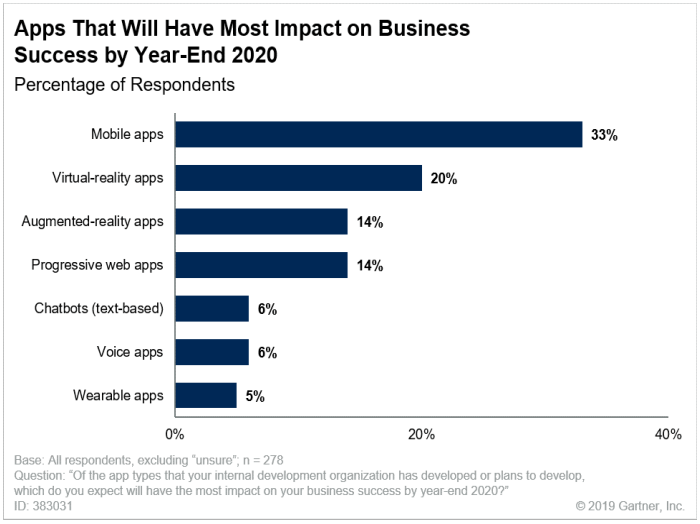


Figure 3 Gartner: Apps expected to have most impact on business success by year-end 2020

3.1. Most Favored and Most Retweeted Tweets about Augmented Reality

Tweets can be liked, or favorited, as well as re-tweeted. These mechanisms, favorite and retweet, are utilized to propagate, share, show appreciation of other user posting, and the number of favorites indicate the count of unique user accounts that like or agree with the content of the tweet [16]. Tweets frequently have meaning, and in this research, the gathered tweets are towards specific technologies.

Within scope of this research, the most favorited tweet is towards AR hardware, wearable glasses, and contains embedded video with a link to a Kickstarter site to raise investment through crowdfunding (tweet image found in Figure 4). The glasses, by TiltFive, are marketed to be the future of tabletop games. A new take on what appears similar to traditional board games. The digital approach gives the characters dimension, allows sharing the digital and augmented experience with other players, as well as the ability to save the game. These AR application quality characteristics are mentioned and defined within the quality model as presence, perspective and persistence. Further evaluation in section four applies this and subsequent text analysis to the previously established quality model.



Figure 4 Most favored tweet within research scope

The following tweet, posted in 2017, had at the time of this research been viewed over 212,100 times and was re-tweeted 4,247 times. The tweet embeds demonstration video to augment the physical environment and makes this statement, "This is ARKit's A-ha moment". ARKit is a development technology for building AR applications. Apple released ARKit with iOS 11 in 2017 [17]. More development tool analysis is found in section 4.5. Within the scope of this research, Figure 5 illustrates the most re-tweeted tweet.



Figure 5 Most re-tweeted tweet within the scope of this research

3.2. Mention, Sentiment and Time Series Analysis

Mention analysis identifies the frequency text occurrence. Research shows augmented reality applications being utilized across many industries; mention analysis from social media provides an additional perspective. Figure 6 indicates that Education, Entertainment and Commerce are the top three mentioned industries. This is determined by first creating a list of known industries from a source [18]. Then the algorithm reads the tweet texts and will count industry mention, including if multiple industries are mentioned in a single tweet.

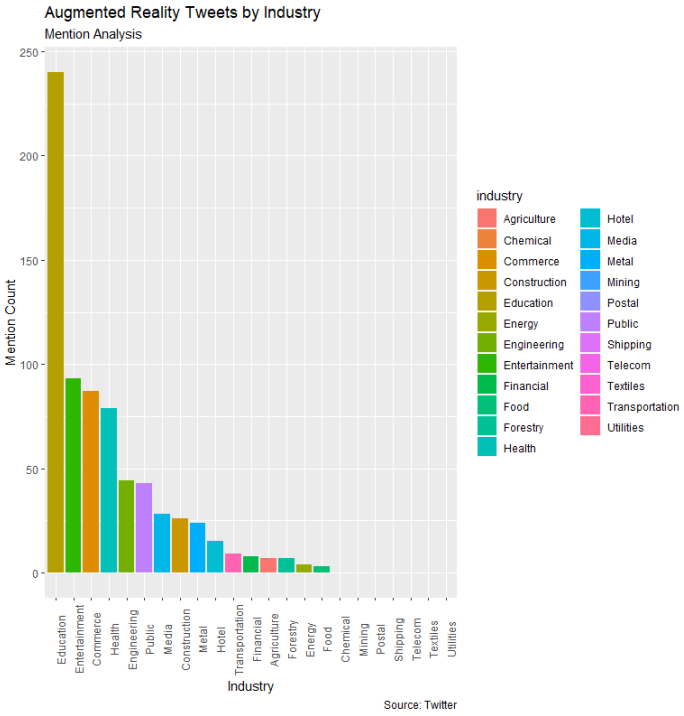


Figure 6 Industry Mention Analysis: Education, Entertainment and Commerce are top three mentioned industries

Further evaluation of the top mentioned industry, education, utilizing a time series chart seen in Figure 7. This chart indicates a slight positive trend in the smoothed line and a peak of tweets in November of 2019.

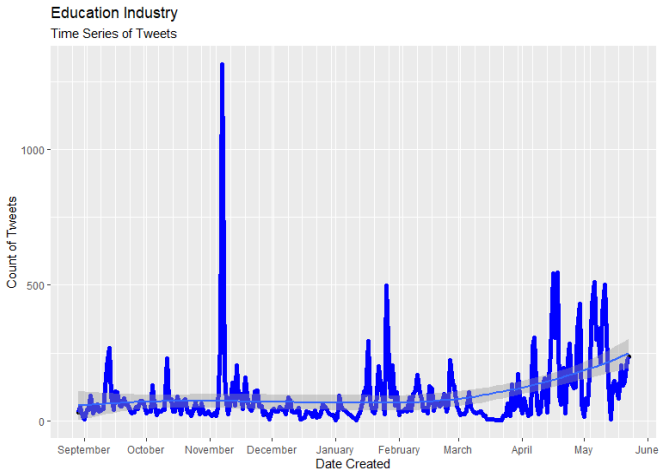


Figure 7 Time series chart of tweets towards Education industry

Figure 8 is a tweet containing embedded AR demo video of a skeletal dinosaur walking on a physical plane as an educational and awareness instrument. This tweet was retweeted over 1,000 times and was the largest contributing factor to the November peak of tweets having mention of education industry.

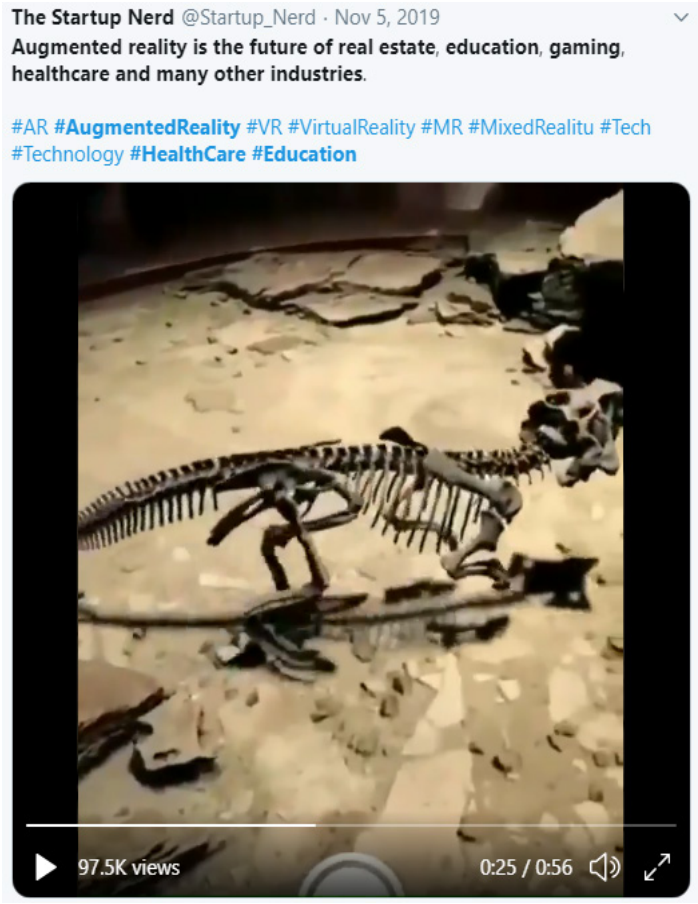


Figure 8 Highly favored tweet was large contribution to the November 2019 peak of Education industry tweets

Tweets are short in character length but offer volume of opinion which can be mined as user-driven data and classified in terms of sentiment. Opinion mining and sentiment analysis techniques can be utilized to support or drive communication plans [19]. Sentiment analysis has also been combined with other analytic methods to produce stock market prediction models with accuracy greater than 60% [20]. Tweets are opinionated text. Opinionated text is the primary resource for shoppers when making purchases [21]. We can utilize user-driven data for measuring system quality. Indeed, the quality of the system is also important in procurement, investment and communication planning.

The highest average sentiment score within the scope of this research is Transportation. Figure 9 illustrates average sentiment scores by the aggregated tweets within industry categories. A tweet which references multiple industries would have a sentiment score towards each referenced industry.

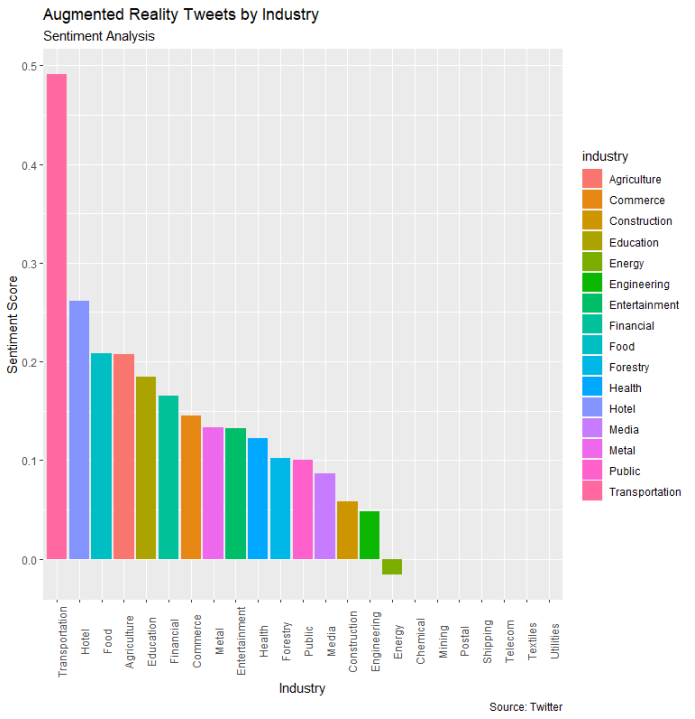


Figure 9 Sentiment Scores by Aggregated Tweet Sentiment Grouped by Industry

Further analysis of Transportation industry tweets indicates a spike in March of 2020. Figure 10 is a chart indicating the slope of the time series data as well as the aforementioned spike.

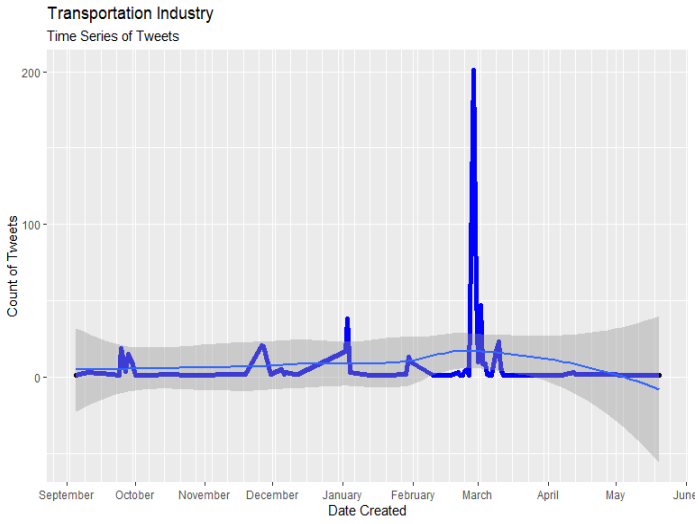


Figure 10 Time series chart of tweets towards the Transportation industry

The tweet found in Figure 11 was created near the end February of 2020 and had hundreds of retweets. This tweet is the main contributor towards the March spike of tweets categorized into the Transportation industry. The tweet references cost optimization of warehouse processing time by helping warehouse employees more quickly identify product location. Beyond cost optimization, the tweet states that AR technology can also help reduce possibility of human error within logistics.



Figure 11 Highly Leveraged Tweet and Main Contributor to the March Spike of Transportation Industry Tweets

The Sentimentr library is utilized within the R code of this research to provide sentiment scores of the collected tweets. The sentimentr library utilizes valence shifters to apply more meaning and accuracy to sentiment, such as negator words (e.g. I do not like it) [22]. The highest sentiment scored tweet, score of 1.540922, is found in Figure 12 (sentimentr library scores range between -2, 2). The tweet references an easier and more fun life due to augmented reality technology and utilizes the hashtag #CX (customer experience). While application users may be customers, they may also be internal employees. This research refers towards user experience as a more generalized approach.



Figure 12 Tweet within research scope having highest sentiment score

4. Application of Analysis to the Established AR Quality Model

Augmented reality applications are novel and require levels of quality to be established to be effective for the users experience. Establishing the quality of an AR application goes beyond traditional testing practices and standards. A new model has been

established for test case focus and categorization, to build system confidence and inform system readiness and release decisions. These are critical decisions when dealing with novel technology.

Gartner's 2020 Digital Friction Survey found that among a survey n size of 4,582 employees surveyed, only 31% experienced High-Quality UX that was productive, empowering and easy [23]. Why might system's need to have this level of quality? The same survey found that "employees with a high-quality UX are 1.8 times more likely to have a high intent to stay as indicated by a lack of interest in leaving their current employer or actively search for a new job." The Gartner research report further explained, "Employees with a high-quality UX are 1.5 times more likely to have high discretionary effort, which includes employees' willingness to help colleagues, take on additional responsibilities, put in extra effort and find better ways of working."

Augmented reality applications are a novel approach to a system's user experience. The user experience has opportunity to benefit users and organizations alike. The technology implementation requires quality for these benefits to be realized.

4.1. Contextual Approaches to Quality

The characteristics of the quality model are quickly identified in the following bulleted text. Further sections then develop the characteristics and apply the findings of the twitter text analysis to the previously established quality model.

- Presence
- Perspective
- Interaction
- Portability
- Persistence

4.2. Twitter Analysis Applied to the Characteristic of Presence

Accurately executed occlusion and collision behavior will make the AR application more realistic [1]. Together they form the characteristic of presence. Presence allows AR application users to experience both the physical and digital worlds accurately. Presence may be exemplified digital models obeying gravity or shadowing as physical objects would which meets the users' expectations. Presence must be tested for. Figure 8, a digital model of a dinosaur overlaid onto a physical landscape (the original tweet is a video of the skeletal dinosaur walking) maintains the existence of shadow beneath the digital model, an example of presence in AR.

According to Gartner, "realism, as well as usability, will be enhanced or undermined by the digital model's ability to have accurate presence within a physical landscape. There should therefore be a test for occlusion which occurs when a far object hides a near object, for example, and which also sometimes occurs when objects are floating above and closer than an original anchor point. Test for occlusion by placing multiple models or scenes and moving throughout the physical environment." [24]. Collision in AR applications exists when a digital object and a physical object appear to occupy the same physical place at the same time. Of the 211,269 tweets related to augmented and virtual reality in this research, only 70 tweets reference occlusion or collision. Average sentiment for an AR tweet is .1176; however, for tweets mentioning collision the sentiment turns negative at -0.0157, and ventures more negative when mentioning occlusion -0.0475.

4.3. Twitter Analysis Applied to the Characteristic of Perspective

Perspective frequently refers to someone's point of view. In an augmented reality, perspective must encompass the physical and digital objects, combined behavior, and how well visual distinction exists among them.

Perspective may be affected by the physical environment, the digital model's attributes, and the interaction between them. This would include the inability to scale a digital model to the correct size when embedded into the physical view. Perspective fails when rendering white font in annotations or digital models given a physical environment containing bright light or predominately white backgrounds. Perspective within AR applications should provide visual distinction and improve the quality of the user experience.

Figure 12, the tweet within the scope of this research having the highest sentiment score, adequately illustrates the characteristic of perspective. This example uses a neon green annotation text color over a mostly grayscale physical scene. Quickly drawing attention to a physical device through digital annotation.

4.4. Twitter Analysis Applied to the Characteristic of Interaction

The characteristic of interaction is directly related to performance. Interaction directly affects the device, the costs, and the execution environment. The tweet text found in Figure 11 refers to AR applications used to model warehouses which may assist in finding specific freight. Warehouses can be quite large, requiring longer application runtime, as well as the rendering of many possible digital directional annotations before arriving at the correct location. If the AR application is utilizing commercial grade devices, we must concern how the interaction of such an application will consume WIFI network, the battery of the device, and whether other applications will run on the device at the same time. The greater the interaction, the more important the performance.

4.5. Twitter Analysis Applied to the Characteristic of Portability

There are many software development kits available to build AR applications. Table 2 references many examples of these technologies, and the runtime platforms they support. Not every AR application will run on your commercial or consumer system. It is not uncommon for some popular mobile AR apps to only run on iOS or Android, as example operating systems [25]. The value and availability of your application will depend upon the portability characteristic, which is a quality concern that must be considered and established during design and development.

Table 1 provides search terms that were utilized to scope the collection of tweets for this research. Many of those search terms are twitter handles or the names of SDK technologies. The top three mentioned development vendors are Apple (966), Unity (713), and Google (601). The popularity found in the mention analysis is not conducive with portability, as Apple is one of the more limited SDKs when concerned with runtime environment support as indicated within Table 2.

Table 2 Augmented reality development tools and supported runtime platform/s

Tool name	Supported Runtime Platform/s
Vuforia Studio [26]	Android, iOS, UWP, Unity
Spark AR [27]	Android, iOS
ARKit [28]	iOS
ARCore [29]	Android, iOS, Unity, Unreal
Wikitude [30]	Android, iOS, Microsoft Tablet
Maxst AR [31]	Android, iOS, Unity

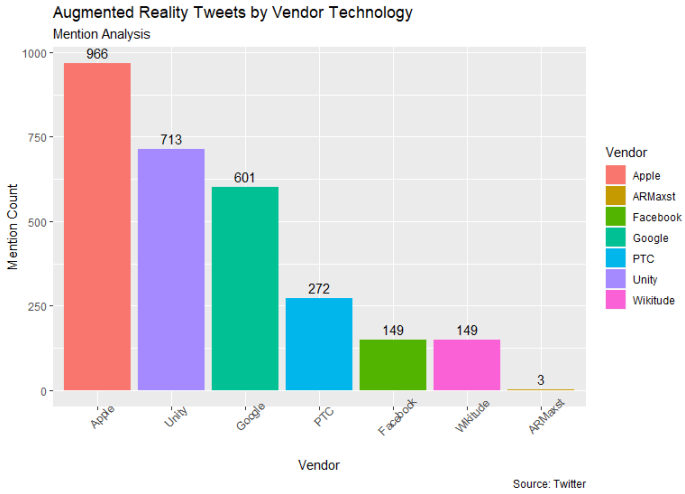


Figure 13 Example augmented reality development technology ordered by mention count

4.6. Twitter Analysis Applied to the Characteristic of Persistence

Scaling augmented reality experiences across multiple users requires centralizing physical maps and virtual object anchors so that multiple users can share immersion [1]. Persistence allows an AR user experience to be shared among other application users from separate devices or application instances during runtime. Additionally, persistence allows the AR experience to be saved and resumed later by the original, or potentially other, application users.

Figure 4 is the most favored tweet within the scope of this research, and it adequately exemplifies many of the characteristics of this quality model. The tweet references the future of tabletop games, a physical board game with digital models rendered above. Persistence of the digital models allows multiplayer action in the game, so that competing players may see and experience the avatars and digital scenes that are being controlled by other's play. It is possible that such a use case could leverage whichever phone device or headset might be handy to a player, which would encompass the portability characteristic. If the game is to be played for many hours, the interaction level will be high, perhaps enough to warrant recharging a device battery, or the requirement to plug an AR hardware device into a power source. Perspective is required so that the digital scenes are relevant to the location, size and position of the physical game board as seen by other's points of view. Presence of the digital models, the prevention of collision, will also provide an experience worth dedicating many hours towards.

While the previously mentioned and exemplified tweet represents the entertainment industry, certainly implementation across healthcare, manufacturing, real estate, education and many other industries would prioritize quality potential during investment in novel technology. Hence the AR quality model, which provides a basis for test case creation and categorization that promotes confidence in the application's readiness and informs decisions towards the application's release.

5. Conclusions and Future Works

5.1. Conclusions

The findings reported in this research study represent essential considerations for the investment and realization of value through augmented reality applications. Augmented reality applications have characteristics that are beyond traditional interfaces. Despite a novel technology, quality still matters. For these reasons, the AR quality model has been established and extended with input from social media and in some cases exemplified through such uses.

Among all tweets that mention either "augmented" or "virtual", augmented reality mentions (135,319) are much more frequent than mentions of virtual reality (80,540). Top mentioned industries within the tweet collection scope were Education, Entertainment and Commerce. Top average tweet sentiment scores by mentioned industries within the study scope were Transportation, Hotel and Food.

AR quality model characteristics of Presence, Perspective, Interaction, Persistence and Portability all had representation among some of the highest retweeted or liked tweets, or among the tweets that drove industry spikes in the time series charts.

Still, limitations exist. Only one social media platform was utilized in this research, Twitter. Facebook, another such social media platform, is also a technology provider of AR software development. What information then may be more prevalent on their social platform? The AR quality model requires further research beyond social media and into specific teams and implementations within industry. Research towards failed AR implementations would also support the model or indicate where omissions of key characteristics exist.

5.2. Future Work

Future work on this quality model for augmented reality applications must extend into cyber security. 475 of the tweets within this research mentioned terms towards privacy or security. That is less than one quarter of one percent of tweets within the scope of this research. While the percentage is small, it is critical to consider augmented reality applications require access to the device camera. The device camera may capture personal or organizational information, potentially confidential or private in nature, that may become vulnerable to threat actors if exploited. Future work should engage in various threat and vulnerability research of AR applications and methods and technology which may help mitigate such risk or exposure.

Future works should also include case studies from many of the named industries within the paper with indication towards

whether the AR quality model adequately supports the users experience of if in fact key characteristics have been omitted, such as current lack of the security characteristic.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to appreciate Tyler Rinker, who has authored the sentimentr package and supported a question regarding sentiment scores during this research.

References

[1] J. Scheibmeir and Y. K. Malaiya, "Quality Model for Testing Augmented Reality Applications," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 0219-0226, 2019. doi: 10.1109/UEMCON47517.2019.8992974.

[2] I. Rabbi, S. Ullah, A survey on augmented reality challenges and tracking, Acta Graphica: J. Print. Sci. Graphic Commun, 24(1-2), 2013, pp. 29-46, 2013.

[3] X. Wang, A. Kotranza, J. Quarles, B. Lok, B.D. Allen, A pipeline for rapidly incorporating real objects into a mixed environment, in The 4th IEEE/ACM International Symposium on Mixed and Augmented Reality, Washington, DC, USA, October 2005, 170-173, 2005.

[4] D. Cearly, et al. "Top 10 Strategic Technology Trends for 2020: Multiexperience." Gartner, 10 Mar. 2020. www.gartner.com/doc/3981950.

[5] S.P. Yun, H.J. Yap, R. Singh, S.W. Chang, K.L.R. Cheong, Augmented reality assisted factory layout planning and analysis for a flexible manufacturing cell, in International Conference on Computer Science and Computational Mathematics, 106-111, 2014.

[6] S. J. Henderson, S. Feiner, Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret, in 8th IEEE International Symposium on Mixed and Augmented Reality, Orlando, FL, USA, 135-144, 2009.

[7] G. Westerfield, A. Mitrovic, M. Billinghurst, Intelligent augmented reality training for motherboard assembly, Int. J. Artif. Intell. Educ. 25(1), 157-172, 2015.

[8] "ISO/IEC 25010." ISO 25000. Accessed May 22, 2020. https://iso25000.com/index.php/en/iso-25000-standards/iso-25010.

[9] A.Dünser and M Billinghurst "Evaluating augmented reality systems," In: Furht B. (eds) Handbook of Augmented Reality. Springer, New York, NY, 289-307, 2011.

[10] A. Sutcliffe and K. Kaur, "Evaluating the usability of virtual reality user interfaces," Behaviour and Information Technology, vol. 19, 2001.

[11] J. Scheibmeir and Y. Malaiya, "An API Development Model for Digital Twins," 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, 2019, pp. 518-519, 2019. doi: 10.1109/QRS-C.2019.00103.

[12] J. Scheibmeir, T. Murphy, and J. Herschmann. "Promote Continuous Quality for APIs to Support Digital Business." Gartner, February 21, 2019. https://www.gartner.com/doc/3902567.

[13] A. Cirillo, RStudio for R Statistical Computing Cookbook. Birmingham (Gran Bretaña): Packt Publishing, 2016.

[14] "How to Tweet." Twitter. Accessed May 23, 2020. https://help.twitter.com/en/using-twitter/how-to-tweet.

[15] J. Wong, A. Leow, Gartner Survey Analysis: Insights to Kick-Start an Enterprise Multiexperience Development Strategy, 12 February 2019. https://www.gartner.com/document/3901479

[16] R.S. Perdana, A. Pinandito, " Combining Likes-Retweet Analysis and Naive Bayes Classifier within Twitter for Sentiment Analysis", Journal of Telecommunication, Electronic and Computer Engineering, 10(1-8), 41-46, 2018.

[17] Apple's ARKit: Cheat sheet, Cory Bohon, TechRepublic, June 7, 2019. Accessed September 28, 2019. https://www.techrepublic.com/article/apples-arkit-everything-the-pros-need-to-know/

[18] International Labor Organization. "Industries and Sectors." Industries and sectors, n.d. https://www.ilo.org/global/industries-and-sectors/lang--en/index.htm.

[19] H. Purohit, A. Hampton, V.L. Shalin, A.P. Sheth, J. Flach, S. Bhatt. What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination Computers in Human Behavior, 29 (6), 2438-2447, 2013. DOI: 10.1016/j.chb.2013.05.007

[20] Y. Tanulia, A.S. Girsang "Sentiment Analysis on Twitter for Predicting Stock Exchange Movement", Advances in Science, Technology and Engineering Systems Journal, 4(3), 244-250 2019.

[21] C. Hung, "Word of Mouth Quality Classification Based on Contextual Sentiment Lexicons." Information Processing & Management 53(4), 51–63, 2013.

[22] T. Rinker, "Sentimentr." README, n.d. https://cran.r-project.org/web/packages/sentimentr/readme/README.html

[23] Application Research Team. "Proven Design Principles to Deliver a High-Value Employee User Experience." Gartner, May 12, 2020. https://www.gartner.com/doc/3985028.

[24] J. Scheibmeir, J. Herschmann, T. Murphy, "Quality Is the Key to Avoiding 'Digital Distortion' With Your Augmented Reality Strategy." Gartner, April 16, 2020. https://www.gartner.com/doc/3983581.

[25] M. Jansen, "Escape Reality with the Best Augmented Reality Apps for Android and IOS." Digital Trends. Digital Trends, June 9, 2019. https://www.digitaltrends.com/mobile/best-augmented-reality-apps/.

[26] Use Vuforia for computer vision and augmented reality apps, 2019, https://www.ptc.com/en/products/augmented-reality/vuforia

[27] Downloads, 2019. https://developers.facebook.com/docs/arstudio/downloads/

[28] Get ready for the latest advances in augmented reality, 2019. https://developer.apple.com/augmented-reality/

[29] ARCore overview, February 28, 2019. https://developers.google.com/ar/discover/

[30] Wikitude Augmented Reality SDK, 2019, https://www.wikitude.com/products/wikitude-sdk/

[31] AR SDK, 2019, Maxst. "Setup Guide." 3.4.x | MAXST Developer Site, n.d. https://developer.maxst.com/MD/doc/unity/setup.

ASTES

Advances in Science, *Technology and Engineering Systems Journal, Vol. 5, No. 6, 185-192 (2020)*

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Toolkit for the Automatic Analysis of Human Behavior in HCI Applications in the Wild

Andrea Generosi[1,*], Silvia Ceccacci[1], Samuele Faggiano[2], Luca Giraldi[2], Maura Mengoni[1]

[1]*Department of Industrial Engineering and Mathematical Sciences, Università Politecnica delle Marche, Ancona, 60131, Italy*

[2]*Emoj s.r.l., Ancona, 60131, Italy*

ARTICLE INFO

Article history:
Received: 03 August, 2020
Accepted: 07 September, 2020
Online: 08 November, 2020

Keywords:
User Experience
Deep Learning
Convolutional Neural Networks
Affective Computing
Gaze detection

ABSTRACT

Nowadays, smartphones and laptops equipped with cameras have become an integral part of our daily lives. The pervasive use of cameras enables the collection of an enormous amount of data, which can be easily extracted through video images processing. This opens up the possibility of using technologies that until now had been restricted to laboratories, such as eye-tracking and emotion analysis systems, to analyze users' behavior in the wild, during the interaction with websites. In this context, this paper introduces a toolkit that takes advantage of deep learning algorithms to monitor user's behavior and emotions, through the acquisition of facial expression and eye gaze from the video captured by the webcam of the device used to navigate the web, in compliance with the EU General data protection regulation (GDPR). Collected data are potentially useful to support user experience assessment of web-based applications in the wild and to improve the effectiveness of e-commerce recommendation systems.

1. Introduction

This paper is an extension of work originally presented in ISCT [1]. Understanding User eXperience (UX) in the wild is of paramount importance for online business (e.g., e-commerce) to gain information about users' opinion and behavior, in order to increase productivity, customize services and better drive strategic decisions. According to the definition provided by the international standard on ergonomics of human-system interaction, ISO 9241-210, UX is "a person's perceptions and responses resulting from the use or intended use of a product, system or service". It follows that the analysis of the sensations and emotions of the user with the product represents a central topic for UX research [2]. Over the last decade, increasingly accessible social media and tools offered by e-commerce sites themselves (e.g. Amazon.com), which encourages users to share their opinions, have made available a wealth of information potentially useful for understanding UX. Alongside, there has been a growing interest in sentiment analysis based on text analysis and user opinion mining. Several automatic approaches have been proposed to understand user emotion and extract UX information from online user online review [3-5]. However, though these tools are becoming increasingly accurate and effective, it is impossible to rely only on online reviews to understand how customers really feel. As reported by Jakob Nielsen in a study about user behavior in online communities [6], a significant number of customers (90%) are lurkers (i.e., they read and observe, but they do not contribute), while the 9% contribute intermittently within the only community and only the 1% account for most contributions. This phenomenon has been named "the 90-9-1 rule". Consequently, while systems able to get direct feedback from all the users (whether they are someone who will leave on online review or not) are still lacking, companies continue to spend millions of dollars on surveys and analysis of their customers to know what really works and what does not.

Moreover, recent years have shown an increasing interest in enhancing possibilities of human-computer interaction with e-commerce. Recommendation systems have started to attract interest in both business and research [7]. In the last few years there was significant improvement especially over the collaborative filtering approaches thanks to the advance in the field of machine learning and deep learning techniques [8]. However, the effectiveness of such a system remains generally low, as the algorithms used by most of the web e-commerce for providing shopping suggestions to customers use data collected from other customers, so that system difficulty meets the

*Corresponding Author: Andrea Generosi, a.generosi@pm.univpm.it

customer's expectation [7]. In this context, emotion detection represents one of the directions that can be taken to enhance the overall UX with these systems. In fact, it has been proved that recommender systems are more precise in providing relevant suggestions, when emotion variables are considered [9].

Nowadays, a variety of technologies exists to automatically recognize human emotions, spanning across facial expression analysis, acoustic speech processing and biological responses interpretation. The extensive deployment of smart devices equipped with sensors and connected to the Internet today opens the possibility of accessing a huge amount of data that until now would have been very difficult to collect. Smartphones and laptops equipped with cameras became an integral part of our daily lives. The use of cameras is so pervasive that we no longer worry about their presence, and we are not fully aware of how much information can be collected through them. For example, using appropriate systems, it is possible to process video images so that demographic (e.g., gender and age) and behavioral (e.g., emotions, eye movements) information can be easily extracted [10, 11]. This opens up the possibility of using technologies that until now had been restricted to laboratories to analyze UX in the wild.

In this context, this paper aims to extend and integrate the contribution of the work presented in ISCT, which introduced and evaluated a new system of emotion detection for mobile applications, in order to define a toolkit for automated collection of data related to users' emotions and behavior useful to:

- Support the collection of data related to the user behavior during the interaction with desktop and laptop web-based applications in the wild;
- Improve the effectiveness of recommendation systems.

Such a tool will take advantage of the system proposed in [1] to monitor user's emotions, through the acquisition of facial expression and implements new system based on deep learning algorithms to track the eye gazes from the video captured by the webcam of the device used to navigate the web, in compliance with the General data protection regulation (GDPR), the European Union regulation on the processing of personal data and privacy.

The remainder of this paper is organized as follows: Section 2 gives an overview of the state of art in the context of systems for the automatic detection of users' emotions and behavior and in particular of facial expression recognition and gaze tracking systems. Section 3 describes the overall system architecture with the Emotion recognition and Gaze detection modules. In Section 4 performance evaluation results of the proposed are reported. The last section summarizes the main paper contributions and highlights future outlook.

2. Research Background

Over the past years producers and the marketing/branding industry demonstrated an increasing interest in emotional aspects of user behavior, as understanding the emotional state of users is crucial for developing successful products and services [12]. Psychophysiological methods may offer data throughout the process of experience, which unfolds new possibilities for UX evaluation. This motivated the demand for automatic emotion recognition techniques as a tool for getting a larger quantity of more objective data. Several attempts have been made to explore the possibility to use sentiments analysis to extract UX information from online user reviews. However, online reviews have proved to be too generic and lacking contextual references to effectively support UX assessment [13]. Several studies, such as [14] and [15], proposed systems that allow us to correlate data from different typologies of data, e.g. eye-tracking fixations, sentiment analysis, body gestures, or facial expressions. However, these systems are designed to support only formal UX evaluation assessment, while are not suitable for studies in the wild. Based on the best of our knowledge, no studies so far have proposed the integrated use of gaze and emotion recognition systems, based on deep learning algorithms, to support the collection of relevant information useful for UX assessment of laptop and desktop web applications nor to power recommending systems. Only [11] proposed a system to support UX assessment based deep learning, but such a solution works only for mobile applications. While only [7] proposed to process data provided by a common pc webcam to enable users' gaze and emotion detection in order to manage an intelligent e-commerce recommendation system: it proposed respectively to use SVM algorithm to enable emotion recognition and a gradient-based method to perform gaze tracking. These algorithms, compared with most of Deep Learning CNNs, are more efficient for what concerns the computational performance, but less accurate [16].

2.1. Emotion recognition

In recent decades, research in the field of Human-Computer Interaction has shown an increasing interest in topics such as Affective computing, emotion analysis and study of human behaviors.

This has encouraged the development of numerous methods and tools for the recognition of human emotions, characterized by different levels of intrusiveness. Most of them relate to three main research areas: facial expression analysis, speech recognition analysis and biofeedback analysis. Instruments based on biofeedback currently available (e.g., ECG or EEG, biometric sensors) are still difficult to adopt for studies in the wild, although the use of smart watches and bracelets is increasingly widespread [17]. Instead, research efforts in the field of facial expression recognition systems have allowed the development of reliable and non-invasive systems. The theoretical model most widely used in the analysis of facial expressions to develop algorithms for the recognition of emotions is certainly represented by the Face Action Coding System (FACS) [18]. It allows the identification of six universal emotions (i.e., joy, surprise, sadness, anger, fear and disgust) by tracking the movements of the face muscles.

The vast majority of facial expression recognition systems are based on Convolutional Neural Networks (CNN): a particular mathematical model of Deep Learning that, unlike classical neural networks, present some layers, in the first part of the entire network, which apply a convolution instead of a general matrix multiplication to the images in input before passing the result to the next layers [10]. Several tools have been proposed in literature, such as [19-21]. Usually these systems refer to trained models using datasets built in controlled environments, where it

is possible to obtain the best accuracy scores, or, trained with data obtained by crawlers on the web, with low accuracy but mostly reflecting real contexts. To the best of our knowledge, approaches able to ensure a good accuracy using data obtained "in the wild" have not yet been evaluated. To ensure good accuracy in the recognition of human emotions in different contexts of use, the system of recognition of emotions proposed in this study adopts a hybrid approach. The resulting CNN, based on Keras and Tensorflow frameworks, has been trained merging three different public datasets.

2.2. Eye-gaze tracking

Tracking one's eyes movement is challenging yet important in HCI and computer vision fields. When it comes to products (as well as website) design, it is readily understandable how much it means to know where a user is looking at. In recent years research has aimed to achieve increasingly accurate results, using less and less invasive and off-the-shelf systems (especially webcams).

According to the results of a recent survey, gaze tracking techniques can be classified into two main categories: feature-based and appearance-based [22]. The first approaches require the acquisition of high-quality images to determine ocular characteristics, such as infrared corneal reflections [23], pupil center [24] and iris contour [25]. To determine the direction of the gaze independently to the head pose, these features are related to 3D eye models. Although such approaches allow high accuracy to be achieved, the accuracy of the geometric calculation is highly dependent on system calibrations that are often difficult or impossible for ordinary users to perform. In addition, as they need to extract small features of the eye, they require the use of special equipment (e.g. IR light sources, special glasses / contact lenses) to acquire special high-resolution infrared images. Such technologies are not normally available in uncontrolled environments. Conversely, appearance-based approaches do not need particular illumination conditions, so that they work well with natural lighting, and require only a single webcam [26]: they take image contents directly in input and try to implicitly extract some relevant features, thus establishing a mapping to screen coordinates. Applicability of this latter kind of methods is remarkably large, although the potential handling of low-resolution images makes their accuracy generally lower [26].

Recently, deep learning (DL) and convolutional neural networks (CNN) have been gaining interest for gaze estimation, so much so that many datasets [27] and network architectures [26, 28, 29] have been proposed over the last years. The gaze-tracking system described in this paper adopts an architecture similar to the one proposed by Krafka et al. [30], which represents one of the more solid CNN actually proposed for gaze tracking. It is based on the AlexNet model described in [31].

3. System Architecture

For the proposed toolkit a centralized architecture has been designed. This kind of architecture has been inherited from the system described in [11] and adapted to be used in synergy with web platforms instead of mobile applications. In Figure 1 are shown the main components: the Web Plugin and the Deep Learning Platform (DLP), respectively the frontend and the backend sides. The Web Plugin is a Javascript library with different functions usable by the developers (APIs) to get info about user's interactions with the web application, i.e. interactions timestamps, clicks and scroll coordinates and above all the webcam handling, so how and when to enable the webcam activation, shooting frequency etc. This feature is the most important because both the Deep Learning engines that compose the DLP make use of the user's face photos to return the information related to his behavior. Photos taken through the pc webcam are encoded in base64 from the Web Plugin and sent through a HTTPS POST to the REST interface exposed through a specific endpoint from the DLP and developed in Python. The received call is then decoded and parsed to get the original data and store it in the server volatile memory, to ensure the compliance of the system with GDPR about privacy. After that, the encoded photo is stored in three different Redis queues: in this way every DL Tracker module can process every photo asynchronously. These queues are used to store and manage the order of arrival of the data that will be given in input to the Gaze Tracker and Emo Tracker for the prediction, respectively, of the gaze coordinates on the screen and the universal Ekman's emotions. Both trackers are Convolutional Neural Networks (CNN) implemented with the Python frameworks Tensorflow and Keras. Whenever the photo processing is concluded for a CNN, the output is saved in a database.

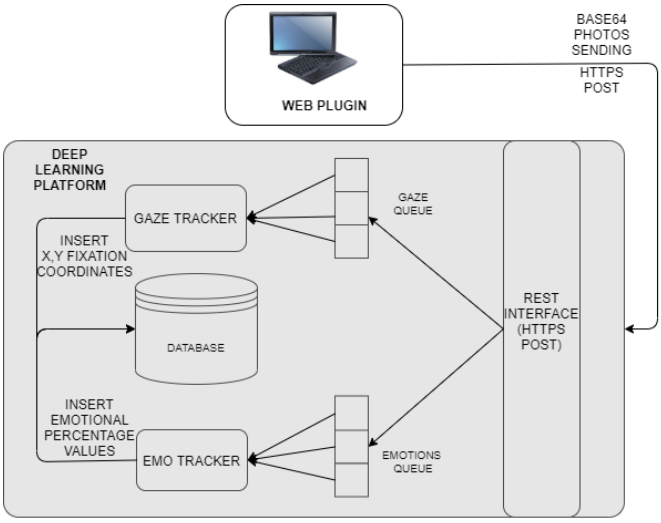


Figure 1: System architecture

3.1. Emo Tracker - the Emotion detection module

Based on [20], a state-of-the-art Deep Learning model that performs well in image recognition should also perform well in a facial expression recognition task: that's because basically the visual discrimination of human emotions is an image classification task. According to different review works, like [19], there exist different DCNNs (Deep Convolutional Neural Networks) that reach various accuracy levels, and those that reach the highest performances are always trained using datasets composed by images retrieved in controlled environments and labeled in labs. Instead, lower accuracy percentages are reached by those datasets with "in the wild" properties, especially retrieved by web crawlers, usually composed by blurred, decentralized or very low-resolution images with misclassified

labels [20]. In return, this kind of datasets are often significantly bigger than the other kind.

To achieve the best emotion recognition accuracy and performances, various tests were conducted. Firstly, by merging the public datasets CK+ (developed in laboratory) [32], FER+ (the FER dataset re-tagged version with crowdsourcing) and AffectNet (manually annotated), a big dataset is constructed. More in detail, FER+ dataset has a label accuracy about 90%, with just 35k photos [20], AffectNet instead has more than 1 million web crawled images, but it also provides more than 400k photos manually labeled by several experts. These datasets have been chosen because they are tagged with almost one of the universal Ekman's emotions labels and because they reach a labeling accuracy relatively high. Moreover, as mentioned, they belong to different dataset categories, so they provide the possibility to improve the resulting model reliability: the assumption is in fact that combining lab generated datasets with those with "in the wild" properties, it is possible to obtain a better accuracy for the in the wild benchmarks. To further improve resulting dataset quality, a Python script to filter all the images with none or more than one face has been developed. The resulting merged dataset at the end counts more than 250k photos, with data distribution shown in Figure 2.

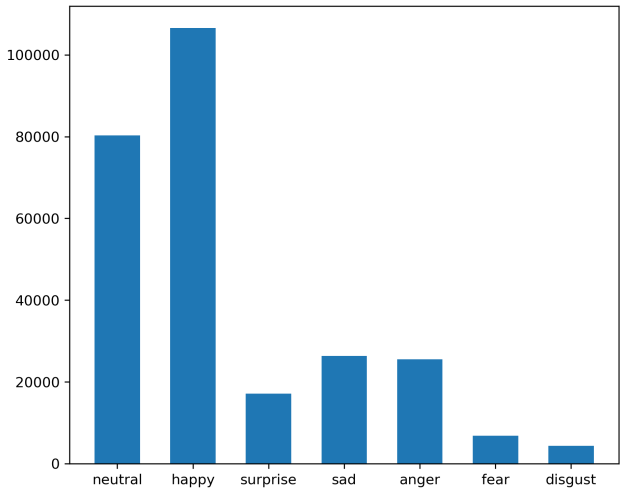


Figure 2: Data distribution

Every single dataset has then splitted into two different parts for training and validation, so to avoid distorting the results during validation phase: for this purpose it was chosen, empirically, to divide each dataset using the ratio 80% for training and 20% for validation, equally for each dataset. After constructing the dataset, preprocessing steps are performed such as facial alignment, centralization of the face respect to the image and face rotation, obtained using Dlib facial landmarks coordinates as reference. All the photos have then been scaled to 64x64 pixels, so to make them homogeneous in size.

For the implementation of the Emo Tracker module, different Python scripts that make use of Dlib, Tensorflow and Keras frameworks have been developed. In particular, Dlib is used to detect one or more human faces in a frame (Face Detection) and provide main face landmarks coordinates, while Keras and Tensorflow provide some of the most used APIs in the Deep Learning field. Trained networks models have been defined to take 64x64 pixels face images in grayscale by the input layer, and to return the Ekman's Emotions (joy, surprise, anger, disgust, sadness, fear and neutral) classification probability by the output layer. Different Keras model architectures such as Inception [33], VGG13, VGG16 and VGG19 [34], were also tested to compare the reached accuracy levels, as listed in Table 1.

Table 1: Performance of different models

Architectures	Accuracy (%)
VGG13	**75.48**
VGG16	74.48
VGG19	73.14
InceptionV2	75.26
InceptionV3	67.20

Network hyperparameters are initialized as it is stated on [20], then the other variations are also experimented such as validation split 0.1, 0.2, number of epochs 30, 50 and 100 and dynamic learning rate defined as

$$lr = lr \times (1 - \frac{epoch}{epoch})$$

Learning rate (lr) is initialized with 0.025 and updated on each epoch accordingly. As shown in table 1, VGG13 model reaches the best accuracy percentage in validation phase, accuracy values over time in relation to the epochs increasing are shown on Figure 3, as a reference of the carried-out experiment. Overfitting occurs during training using other deep neural network architectures, except for the VGG13, when epochs number is bigger than 30.

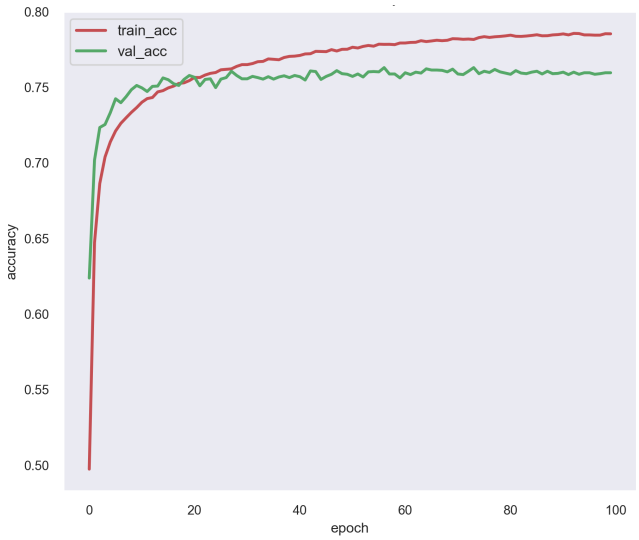


Figure 3: Training and validation accuracy of VGG13

3.2. Gaze Tracker - the Eye-Gaze Detection module

In 2016, Krafka et al. [30] published a paper with a solid CNN proposal for gaze tracking, together with a big dataset of crowdsourcing-collected faces; many researchers are referring to it since then. The Gaze Tracker module's network adopts an architecture similar to the one proposed in [30], which is based on the AlexNet model [31], but respect to that one, a dropout layer

was added to the right after each last convolutional layer. Also in this case, the output is the camera distance, in centimeters. The main challenge using this architecture has been to adapt it to an application context different from the one originally established: iTracker in fact has been trained to work with mobile devices, especially it can reach the best performances on iPhone devices, where it is possible to exactly know, dynamically, the phone camera position. Moreover, practically the whole dataset used to train the model, was built using photos taken by Apple devices. Our purpose, instead, is to propose a gaze detection module optimized to be used on laptop and desktop devices, while it has been decided to not consider a general model that involves also mobile displays in order to not undermine the performance of the model for the specific case taken into consideration by this research.

To achieve this objective, a new training dataset has been built from scratch, collecting pictures from informed volunteers who decided to contribute. For this purpose, a web application has been developed. Each participant was asked to stare at a red dot that was randomly displayed in 30 different screen positions, while the webcam took a picture for each of them; corresponding point coordinates were stored too, to ensure image-screen coordinates association. Concerning screen sizes, this application has been thought to be as general as possible (i.e., capable of dealing with different kinds of displays). However, due to infrastructural reasons, it is not possible to automatically retrieve physical dimensions of a screen from a web page application.

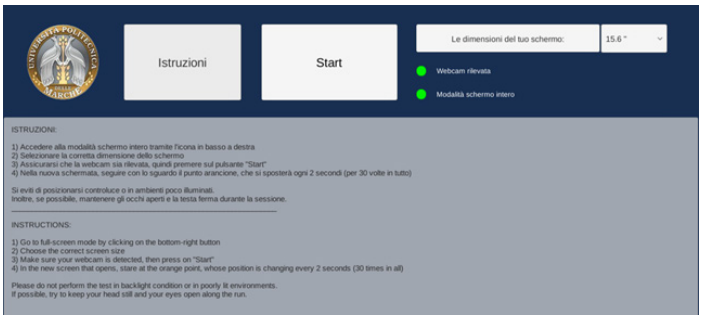


Figure 4: Main screen of the web application for collecting training pictures; by pressing on "Start", a new empty screen appears, in which the dots are shown in succession

For this reason, on the application interface, the user is asked to choose its own screen size, among the ones displayed in a dropdown list, ranging from 13 to 30 inches (i.e., the most common laptop and desktop sizes). By knowing the physical size of the screen and the display resolution (that can be automatically determined instead), it has been possible to express the dots coordinates in cm. At the end, 54 subjects agreed to participate, so the training dataset has been composed of 1620 photos.

The CNN for gaze tracking was implemented and trained using *python* programming language, running *Keras* API upon *TensorFlow* library support and with GPU acceleration.

Inputs of the CNN are the cropped face image (sized 224x224), the two cropped eyes images (sized 224x224) and a 1x4 face grid vector, expressing the portion of the entire image occupied by the face; in order to normalize the dataset, both the face and the eyes images are converted to grayscale, their contrast being increased by means of histogram normalization. The output is a two-dimensional vector containing x and y coordinates of the estimated display point (in centimeters). Faces and eyes are detected and cropped using the dlib [35] frontal face detector and 68 landmarks predictor, respectively. When a face box is found, the screen coordinates of its top left corner and its width and height are stored in the face grid vector; in this way implicit information can be retrieved about the user-screen relative pose. Adadelta [36] has been chosen among the optimizers implemented in *Keras*.

A python script was also realized for visualizing the scan path of the eyes over a displayed image; when launching it, a preset image is shown in full screen, while the webcam repeatedly takes pictures of the user and feeds the network with them. All the coordinates are stored and, when the image is closed, they are automatically processed in a manner that near points are clustered together, thus separating fixations from saccades [37]. Eventually, the colored clusters are depicted over the displayed image, together with straight lines connecting the subsequent ones, to give an idea on the path which the eyes followed overall.

4. Results

The proposed emotion and gaze detection systems were tested separately.

4.1. Evaluation of emotion detection performance

The test accuracy of each emotion category on the confusion matrix plotted as heatmap on Figure 5, shows some interesting results. The images with fear tag misclassified as surprise and disgust tag misclassified as anger has over 20% rate, these are results that are very often found in literature, mostly because these facial expressions are difficult to differentiate even for a human being, so they are often misinterpreted also in the dataset labeling phase. Moreover, as it can be seen, the emotion categories with more training samples resulting in higher accuracy compared to the ones with fewer training samples. Consequently, it is believed that a more evenly distributed dataset may improve the model accuracy further.

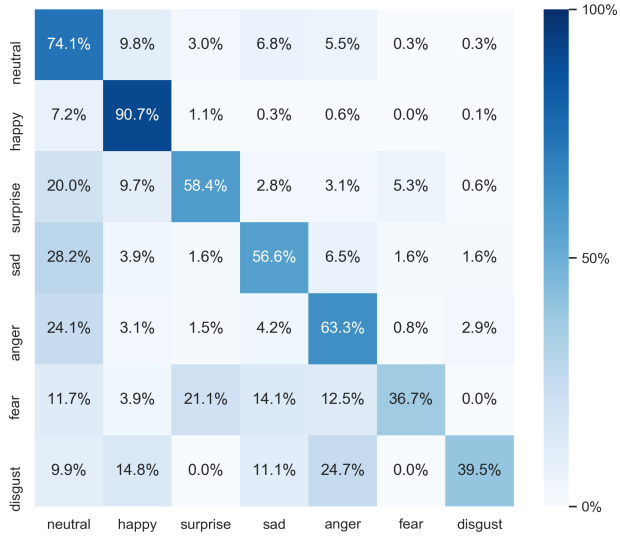


Figure 5: Accuracy of each emotion category

The model VGG13 we trained is also evaluated on EmotioNet [21] 2018 challenge dataset: the Ohio State University, on their website [38], has in fact made available their dataset to give anyone the opportunity to compare their results with those of the challenges of 2017 and 2018. The table shows the evaluation result.

Table 2: Emotionet evaluation results

Categories	Accuracy	F1
happy	**0.9770**	**0.9799**
anger	0.75	0.1198
disgust	0.0128	0.0099
sad	0.5955	0.2888
surprise	0.7059	0.3944

In this case only five emotions have been evaluated: this is because available photos related to Fear emotions were not enough to establish a meaningful comparison. In addition, no photos were made available for the comparison of the Neutral emotion.

4.2. Evaluation of eye-gaze detection performance

To evaluate the performance of the prediction the same software implemented to gather training samples has been used. The human eye detected using dlib [35] landmarks is the input of the python script that crops the region of interests and predicts the coordinates. They are expressed in centimeters and take the top-left screen corner as origin. A total of 20 subjects agreed to take part in the test and for each participant the predictions have been made for 30 different screen positions.

Every time the test is repeated the real and predicted coordinates have been saved in a database. To calculate the error between the predicted point and the real one the following formula has been used:

$$e_m = \frac{\sum_{i=1}^{n} | x_{i\,real} - x_{i\,pred} |}{n}$$

where $x_{i\,real}$ is the real coordinate where the subject *i* is looking at, $x_{i\,pred}$ is the predicted coordinate, *n* represents the total number of participants, and e_m the mean error (in centimeters) that has been calculated for any screen position. The same formula was used to evaluate the error on the y coordinate.

At this stage, 30 values of e_m have been calculated., The mean errors have been aggregated in 12 more meaningful values, as no significant variations have been found in comparison with the high number of screen positions. Each of them represents the error in a specific screen area (see Figure 6 and 7).

Both the heatmaps in Figure 6 and 7 show a higher error value in the top side of the screen (up to 11 cm), but interesting results have been achieved for the other screen positions. In this case the error goes between 0.02cm to 7.2cm. As the heatmap shows, the model still needs some improvement since the accuracy strongly depends on the area the subject stares at, but this experiment proves that it is possible to reach very accurate gaze predictions.

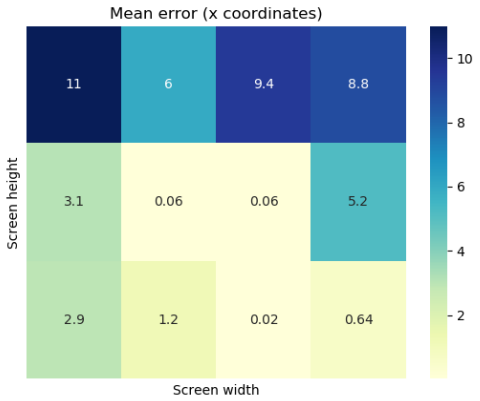


Figure 6: Mean error (cm) for each screen area for the x coordinate.

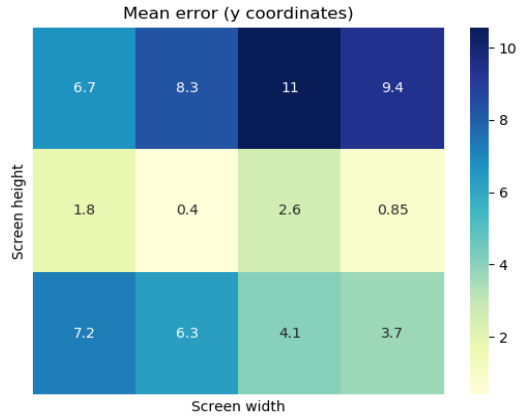


Figure 7: Mean error (cm) for each screen area for y coordinate.

5. Conclusion and discussion

This work proposed a system able to collect, data about user interactions/behavior with web applications potentially useful to support UX analysis. In particular, the proposed toolkit makes use of a deep learning-based platform to allow an "in the wild" data collection during user's interactions with a web application/platform just using a normal webcam. Future studies will be conducted to define an automatic system able to support the analysis of collected data, visualize UX outcomes through the provision of proper KPIs, and suggest useful design guidelines, based on the observed users' behavior, to improve UX design.

Regarding the emotion detection module, experimental results showed that the accuracies of facial expression categories such as fear and disgust are low mainly because of the small dataset of the corresponding category. The other facial expression categories however have reached relatively high classification accuracies. The evaluation results on EmotioNet dataset also supports the results of our experiment. Future work will be focused on the collection of more data on the small dataset categories to improve the classification accuracies of those corresponding categories

and on the exploration of generative adversarial networks to increase recognition accuracy.

About the gaze detection module, has been showed that the average error that can be reached is widely improvable and still does not reach the levels of accuracy of iTracker: this is due both to the small size of the dataset used for training and to the heterogeneity of the environments with which the user may be operating, in particular distance between user and screen, screen size, webcam position, etc. One of the future main works are undoubtedly the expansion of the dataset used to train the gaze network, unfortunately still very limited, adopting crowdsourcing tools.

Finally, data related to users' emotions and behavior, automatically collected through the proposed tool, can be used to improve recommendation systems based on collaborative filtering approaches. Current recommendation systems actually consider only data related to product features, customer preferences, customer demographic data (e.g., age, gender) historical data of purchases, and environmental factors (e.g., time, location). The availability of a huge amount of data related to what users observe the most, and what they feel while browsing e-commerce, opens new possibilities for customization and adaptation of recommendations, based on the real taste and interest of the user. Future studies should go in this direction.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The toolkit is part of the NextPerception project that has received funding from the European Union Horizon 2020, ECSEL-2019-2-RIA Joint Undertaking (Grant Agreement Number 876487). The authors would also thank the research fellow Lorenzo Marchesini for its support in the design and development of the eye gaze detection system.

References

[1] A. Talipu, A. Generosi, M. Mengoni, L Giraldi, "Evaluation of Deep Convolutional Neural Network architectures for Emotion Recognition in the Wild," in 2019 IEEE 23rd International Symposium on Consumer Technologies, 25-27, 2019. IEEE https://doi.org/10.1109/ISCE.2019.8900994

[2] C. Lallemand, G. Gronier, V. Koenig, "User experience: A concept without consensus? exploring practitioners' perspectives through an international survey," Computers in Human Behavior, **43**, 35–48, 2015. https://doi.org/10.1016/j.chb.2014.10.048

[3] J. Jabbar, I. Urooj, W. JunSheng, N. Azeem, "Real-time sentiment analysis on E-commerce application" in 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC) https://doi.org/10.1109/ICNSC.2019.8743331

[4] X. Fang, J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, **2**(1), 5, 2015. https://doi.org/10.1186/s40537-015-0015-2

[5] S. Hedegaard, J. G. Simonsen, "Extracting usability and user experience information from online user reviews," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013. https://doi.org/10.1145/2470654.2481286

[6] J. Nielsen, "The 90-9-1 rule for participation inequality in social media and online communities," 2016. Retrieved 20.07.2020, from https://www.nngroup.com/articles/participation-inequality/

[7] S. Jaiswal, S. Virmani, V. Sethi, K. De, P. P. Roy, "An intelligent recommendation system using gaze and emotion detection," Multimedia Tools and Applications, **78**(11), 14231-14250, 2019. https://doi.org/10.1007/s11042-018-6755-1

[8] I. Portugal, P. Alencar, D. Cowan, "The use of machine learning algorithms in recommender systems: A systematic review," Expert Systems with Applications, **97**, 205-227, 2018. https://doi.org/10.1016/j.eswa.2017.12.020

[9] A. Bielozorov, M. Bezbradica, M. Helfert, "The role of user emotions for content personalization in e-commerce: literature review," in International Conference on Human-Computer Interaction, 2019. https://doi.org/10.1007/978-3-030-22335-9_12

[10] A. Generosi, S. Ceccacci, M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail store," in 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), 2018. https://doi.org/10.1109/ICCE-Berlin.2018.8576169

[11] A. Generosi, A. Altieri, S. Ceccacci, G. Foresi, A. Talipu, G. Turri, ... & L. Giraldi, "MoBeTrack: a toolkit to analyze user experience of mobile apps in the wild," in 2019 IEEE International Conference on Consumer Electronics (ICCE), 2019. https://doi.org/10.1109/ICCE.2019.8662020

[12] E. Ganglbauer, J. Schrammel, S. Deutsch, M. Tscheligi, "Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility," in Workshop on User Experience Evaluation Methods in Product Development, August 25, 2009. Uppsala, Sweden.

[13] S. Hedegaard, J. G. Simonsen, "Extracting usability and user experience information from online user reviews," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013. https://doi.org/10.1145/2470654.2481286

[14] V. Georges, F. Courtemanche, S. Senecal, T. Baccino, M. Fredette, P. M. Leger, "UX heatmaps: mapping user experience on visual interfaces," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016. https://doi.org/10.1145/2858036.2858271

[15] R. Y. da Silva Franco, R. Santos do Amor Divino Lima, M. Paixão, C. G. Resque dos Santos, B. Serique Meiguins, "UXmood—A Sentiment Analysis and Information Visualization Tool to Support the Evaluation of Usability and User Experience," Information, **10**(12), 366, 2019. https://doi.org/10.3390/info10120366

[16] A. Altieri, S. Ceccacci, M. Mengoni, "Emotion-Aware Ambient Intelligence: Changing Smart Environment Interaction Paradigms Through Affective Computing," in International Conference on Human-Computer Interaction, 2019. Springer, Cham. https://doi.org/10.1007/978-3-030-21935-2_20

[17] A. Kartali, M. Roglić, M. Barjaktarović, M. Đurić-Jovičić, M. M. Janković, "Real-time Algorithms for Facial Emotion Recognition: A Comparison of Different Approaches," in 2018 14th Symposium on Neural Networks and Applications(NEUREL), 2018. https://doi.org/10.1109/NEUREL.2018.8587011

[18] P. Ekman, V. F. Wallace. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[19] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," 1–25, 2018. https://doi.org/10.1109/TAFFC.2020.2981446

[20] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution" 2016. https://doi.org/10.1145/2993148.2993165

[21] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, A. M. Martinez, "EmotioNet Challenge: Recognition of facial expressions of emotion in the wild," 2017.

[22] D. W. Hansen, Q. Ji, "In the eye of the beholder: a survey of models for eyes and gaze," IEEE Trans Pattern Anal Mach Intell, **32**(3), 478–500, 2010. https://doi.org/10.1109/TPAMI.2009.30

[23] J. Sigut, S. A. Sidha, "Iris center corneal reflection method for gaze tracking using visible light," IEEE Transactions on Biomedical Engineering, **58**(2), 411-419, 2010. https://doi.org/10.1109/TBME.2010.2087330

[24] R. Valenti, T. Gevers, "Accurate eye center location and tracking using isophote curvature," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008. https://doi.org/10.1109/CVPR.2008.4587529

[25] H. Wu, Y. Kitagawa, T. Wada, T. Kato, Q. Chen, "Tracking iris contour with a 3D eye-model for gaze estimation," in Asian Conference on Computer Vision, 2007. https://doi.org/10.1007/978-3-540-76386-4_65

[26] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, "Appearance-based gaze estimation in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019. https://doi.org/10.1109/CVPR.2015.7299081

[27] S. Winkler, R. Subramanian, "Overview of eye tracking datasets," in 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), 2019. https://doi.org/10.1109/QoMEX.2013.6603239

[28] J. Lemley, A. Kar, A. Drimbarean, P. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," IEEE Transactions on Consumer Electronics, **65**(2), 179-187, 2019. https://doi.org/10.1109/TCE.2019.2899869

[29] X. Zhang, Y Sugano, M. Fritz, A. Bulling, A., "It's written all over your face: Full-face appearance-based gaze estimation," in Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition Workshops, 2017. https://doi.org/10.1109/CVPRW.2017.284
[30] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, A. Torralba, "Eye tracking for everyone," in Proceedings of the IEEE conference on computer vision and pattern recognition 2016. https://doi.org/10.1109/CVPR.2016.239
[31] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012. https://doi.org/10.1145/3065386
[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in 2010 ieee computer society conference on computer vision and pattern recognition-workshops, 2010. https://doi.org/10.1109/CVPRW.2010.5543262
[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition 2016. https://doi.org/10.1109/CVPR.2016.308
[34] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[35] D. E. King, "Dlib-ml: A machine learning toolkit." The Journal of Machine Learning Research, **10**, 1755-1758, 2009.
[36] M. D. Zeiler, "Adadelta: an adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.
[37] L. Chamberlain, "Eye tracking methodology; Theory and practice," Qualitative Market Research: An International Journal, 2007. https://doi.org/10.1108/13522750710740862
[38] EmotioNet Challenge, Retrieved July 30, 2020 from https://cbcsl.ece.ohio-state.edu/EmotionNetChallenge/index.html

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 193-205 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Real-Time Identification and Classification of Driving Maneuvers using Smartphone

Munaf Salim Najim Al-Din*

College of Engineering and Architecture, Electrical and Computer Engineering, University of Nizwa, Nizwa, 616, Oman

ARTICLE INFO

Article history:
Received: 18 August, 2020
Accepted: 01 November, 2020
Online: 08 November, 2020

Keywords:
Driving behavior
Identification and classification
Smartphones

ABSTRACT

The fast-paced development of smart technologies and the prevalence of vehicles, created an urgent demand to study and improve safety issues related to driving. In order to reduce traffic accidents, driving behavior was found to be very important issues to study and investigate. Recently, the advent and widespread of smartphone platforms with advanced computing competence and embedment of a variety sensing elements have greatly contributed to the development of solutions that can detect, and evaluate driving behavior and skills. In this study the development of a real-time smartphone-based identification and classification system for highway driving maneuvers is presented. The proposed system has been designed to detect ten different maneuvers usually performed by drivers when driving on highways. The methodology is based on separating the identification and the classification processes. The identification process is performed by a hybrid pattern matching scheme that combines Dynamic Time Warping (DTW) and neural networks. While a second neural network has been used to classify maneuvers, severity based statistical and time features. The separation of the identification and classification processes simplifies and accelerates the learning processes of the neural networks and greatly improves both system's reliability and accuracy

1. Introduction

During the last century vehicles have prevailed all over the world and have greatly contributed to the development and prosperity of the world economy. According to [1], the number of registered vehicles in the world has increased from 1.015 billion in 2010 to nearly 1.5 billion in 2020. In spite of all their positive impacts, vehicles have negative consequences on environmental pollution, global warming and greenhouse gas emissions, congested roads and noises, fuel consumption, social impacts, road accidents and many others to list. These negative consequences have stringent the environmental and safety regulations and standards for automotive industry. Therefore, to be in compliance with the regulations and standards, innovative solutions and newly emerging technologies have been adopted and utilized to provide the utmost advanced, efficient, safest and comfortable transportation means. In these days, high-end vehicles are equipped with smart systems such as cruise control, advanced emergency braking, safety warning systems, GPS technology for navigation, fuel efficiency, and automated driving system. Furthermore, in the near future, the automotive industry is expected to evolve even further by the production of autonomous vehicles that operate in a fully connected and digitized environment. Such immense leap in transportation technologies is only made possible by intensive studies in the field of vehicular applications especially the studies of driving behaviors.

Although that advanced technologies such as smart safety systems have been integrated within modern, still fully human driven vehicles is and will continue to dominate traffic safety since almost all of vehicles in use at the present time are of conventional type. Conventional vehicles are those fully driven and controlled by human drivers, and have few or none of the new technologies and accessories. The impediment in utilizing the new technologies is due to prohibitive cost that cannot be afforded by most of vehicles' owners especially in developing countries in addition to insulation difficulties [2]. Based on this argument, safety and road accidents due to human errors are still urgent issues that need to be addressed and solved for the majority of conventional vehicles. Based on the World Health Organization reports, vehicle accidents are considered to be the ninth cause of death in the world in addition they are a leading cause of fatal injuries [3]. Previous road safety studies and statistics have showed that the majority of road accidents are attributed to both physiological and behavioral human factors [4]. Therefore, research in the field of driving behavior monitoring systems has gained great momentum and

*Corresponding Author: Munaf Salim Najim Al-Din, munaf@unizwa.edu.om

several research institutes have developed systems to monitor and analyze the driving patterns and activities. Beside their urgent importance these studies serve the research and developments of different applications such as the design of advanced driver assistance system, intelligent transportation systems, and autonomous vehicles [5-6]. Additionally, they helped logistics companies to reduce their operational costs through tracking and monitoring their vehicles [7], and insurance companies to evaluate the driver performance, usage and can be used to resolve insurance tariffs and disputes [8].

The study and analysis of the concepts related to driving processes and problems involve the utilization of Driving Monitoring and Assistance System (DMAS). DMASs are usually designed to at least provide two main tasks, data acquisition and driving maneuvers analysis. The first unit is responsible for automating the collection of driving data, such as driver's attentiveness, vehicle dynamic state, driver's actions and driving environment or surrounding. The second unit is designed to extract patterns and features from the driving data and then utilizes intelligent computational techniques to identify and classify the driving behavior. In general real-time acquisition of driving patterns and parameters can be achieved through various types of sensors, such as GPS, GNSSs, In-Vehicle sensors such as CAN-BUS signals, costumed designed systems, vision systems and smartphones. Among this range of sensing platforms smartphone-based DMAS have growing so rapidly in number and sophistication, due to the prevalence, continuous improvement of the computing power, advancement of integrated sensors and continuous reduction cost in smartphones. Therefore, smartphones are taking an important role in the field of Intelligent Transport Systems such as vehicle monitoring driver behavior analysis [9].

This paper is an extended version of the work published in [10] and it is intended to presents the development of a real-time smartphone-based highway driving maneuvers identification and classification system. The system consists of three layers namely, data acquisition layer, preprocessing layer and the identification and classification layer. It is well known that the estimation driving parameters with high degree of accurately greatly affects the detection and classification of driving maneuvers. Therefore, in this paper a new approach to estimate the speed of a moving vehicle derived from GPS and the IMUs is developed. The identification and classification layer consists of three main units, a pattern matching units, to identify the patterns of estimated driving parameters. The DTW technique is used to identify the pattern of each driving parameter during maneuver execution. The second unit consists of a single feed-forward neural network that will identify the type of the maneuver based on the combination of the identified patterns from the first unit. Finally the third unit is responsible for the manoeuver classification and it is performed by a second neural network.

2. Literature Review

In the past decade several techniques and methodologies have been proposed and used to recognize and classify driving maneuvers and then infer an evaluation for driving behavior. Up to date there is no agreement on the parameters and techniques that could uniquely be used for this purpose. This section presents a short review of the existing technologies and methodologies used for driving maneuvers identification and classification. The literature review illustrates only two major issues related to this field of research namely, the data collection and identification and classification techniques.

Depending on the purpose and functionalities that is required to be achieved, there are various types of data acquisition systems (DAS) developed in the past ten years. The first category of DAS relies on the in-vehicle sensors which broadcast their data through the CAN-BUS which can be captured through the OBD II port. These systems can only provide information that can be used to monitor vehicle' state such as speed, fuel consumption and engine RPM and some limited information related to the driver actions such as throttle pedals positions [11, 12]. Many research centers have developed customized monitoring systems that combine in-vehicle sensors with GPS unit, inertial measurement units, and video systems of the external road and traffic environment and inside the vehicle itself [13]. In general sensors are configured to perceive external and internal environments to provide contextual details about the driving and drivers states. The drawbacks of the simple plug and play monitoring systems are the limited data that can be extracted with. On the other hand, drawbacks of in-vehicle customized systems are their high cost, installation complexity, compatibility and customer acceptance [5]. The developments of smartphones in terms of computational power and integration of sensing devices such as Inertial Measurement Units (IMUs), GPS and many other sensors combined with their ubiquitous presence have led to wide prevalence of smartphones in vehicular applications. Nevertheless, there are shortcomings when employing smartphones as monitoring systems such noisy measurement, coordinates orientation due smartphone placement, battery life and they cannot access vehicles data by their own. Several techniques, as they will be explained later, have been developed to provide the required remedies to overcome these problems [14]. Smartphones have also been utilizes as an intermediate processing hub with the aid of special applications to collect vehicle's sensors data and perform the necessary computation. In general, these applications enable the smartphone to communicate with the OBD-II port, either through Bluetooth or WiFi, and to collect a number of vehicle's sensors data such as the speed, engine RPM and many other diagnosis parameters. The processing of these in-vehicle sensors' data streams can be combined with smartphone sensors to further enrich the analysis [15].

Based on the collected data or the estimated driving parameters and the maneuvers to be detected, different techniques have been suggested to identify the type of a maneuver and classify it. A straightforward approach used to process the driving data is based on specifying a number of thresholds for each maneuver, that are defined merely based on the range of collected driving data and the experience. Once the sensed signals exceed a predefined threshold for a specified period of time then a maneuver is detected [16-19]. The main disadvantage in using threshold-based approaches is related to the absence of both the sensitivity and specificity that are required to obtain accurate identification and classification since improper definition of thresholds can result in poor performance. In a different approach, researchers considered the vehicle driving as a reasoning process, hence rule-base and fuzzy logic systems have been used to model driving behavior in terms of driving maneuvers classification only, i.e. it is not possible to detect the

type of the maneuver by using fuzzy logic schemas [13,20-24]. The implementation of a classical fuzzy logic classifier depends on the definition and setting of fuzzy input variables, the rule-based inference system and the fuzzification and defuzzification processes. Therefore, fuzzy logic classifiers are not constantly precise, since the both the fuzzy rules and conclusions are based on postulation, so it may not be widely accepted. Furthermore, they don't have the capability of machine learning and the setting of the precise inference rules and, membership functions is a difficult task. To overcome the aforementioned limitations in both threshold-based and fuzzy logic classifiers several adaptive methods have been proposed to automatically generate thresholds and fuzzy rules from numerical data [20-24]. Adaptive systems have showed distinctive superiority over classical fuzzy systems due to their learning capabilities, their ability to be modified and extended and finally the ability to extract newly modified rules and fuzzy variables. As a consequence of this combination between the adaptive learning and human experience adaptive vigorous classification systems can be designed with high degree of reliability and accuracy.

Since the data and information used in the identification and classification of driving behaviors and maneuvers are time varying signals, pattern matching techniques, such as DTW, have been employed. In this approach the maneuvers are identified and classified by comparing the similarity of the collected sensors' signals or estimated driving parameters patterns with predefined templates. If the incoming measured sequence has adequate similarity to one of the standard templates, then the system will indicate that a specific driving maneuver has been executed or detected. The major advantage of pattern matching techniques is that the system can sense maneuvers in spite of any differences, in their amplitudes or durations, with predefined templates. Therefore, it would be possible to develop a single set of standard templates that be used for matching maneuvers for different drivers to the same template set [25-28]. The main drawback of pattern matching approach is excessive computational requirements since the algorithms compute the similarity differences for all elements in two signals. This problem will be more significant when the classification analysis is dependent on multidimensional time series. Furthermore, enormous effort is required to extract and select the reference templates and it is also very hard to gather all types of templates due to the exclusive driving styles and behaviors of drivers.

Very recently, machine learning algorithms have been adopted and implemented to identify and classify driving maneuvers. Machine learning algorithms, when are well trained can generate specific rules that will enable them to identify and classify abrupt irregularity on a large set of data. Different approaches have been suggested and implemented to extract features for training these approaches such as simple threshold-based discriminators, statistical values, time domain parameters, frequency domain parameters, and most importantly rules inferred from clustering algorithms. Machine learning algorithms are generally categorized into three different approaches namely the supervised learning, unsupervised learning and semi-supervised learning. In Supervised learning techniques a dataset is first collected and labeled to certain predefined classes. Then a learning model is chosen and trained with some of the labeled dataset to produce a predicted output in different testing samples. Several algorithms under the supervised learning category have been used in the identification and classification of driving maneuvers and behavior such as K-Nearest Neighbor, Naive Bayes, Decision Trees [29], linear regression [30], Support Vector Machines (SVM) [31-32] and Neural Networks [33]. Unsupervised learning methods are used when there are no pre-knowledge about ground truth labels. Therefore, they are used to classify collected unknown data into groups depending on common trends, attributes, patterns, or relationships. There are different unsupervised approaches such as clustering techniques and self-organizing maps. Each approach uses different algorithm for classifying data into groups, and these algorithms could be simple straightforward that can divide collected data based on the common attributes or similar trends in their features. Nevertheless, when performing a pre-processing step of the data, unsupervised learning techniques are usually used with other techniques to enable specific characteristics of the learning model,. The K-means clustering [34] and Principal Component Analysis algorithms [30] are samples of the approaches used in driving behavior analysis.

2.1. Data Collection

As it has been stated modern smartphones are integrated with sensors that can be utilized to collect vehicle driving data for driving maneuver analysis. Previous work [35, 36] showed that precisely preprocessed and calibrated smartphone sensors are becoming a competitive approach to the high price in-vehicle customized DAS. In this study ten frequent maneuvers usually executed by drivers during highway driving are going to be detected and classified. Four of these maneuvers namely, braking, acceleration and left and right lane changes will be analyzed for both straight and curved road segments, in addition to another two maneuvers which are the merging on or exiting a highway. In the preliminary stages of this study an online vehicle driving data recording system has be developed and installed on three different android-based smartphones. The sensors that have been used are the Accelerometer, Gyroscope and the GPS. The developed App has been test by comparing its performance with well-established available Apps like AndroSensor. The sampling rate for the IMUs is in the range between 50 and 100 Hz, from which the optimal sampling rate then will be obtained.

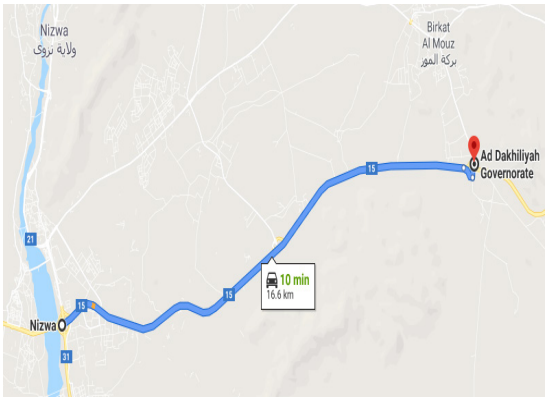


Figure 1: The route used to collected training data

The development of the system proposed in this study passed through two phases, the development and testing phase and the final real-time testing phase. In the first phase adequate data for training and testing the system were first collected. Ten drivers

with different vehicles were enlisted to drive along a highway route of 16km long that covers varying road conditions as shown in Figure 1. In order to have adequate data for training and testing each driver repeat each of the basic divining maneuvers for at least five times with different severity levels from the defensive to the aggressive level. Data collected from inertial sensors in these tests are first pre-processed, then segmented and finally labeled into types and classes.

In addition to the collected data by the smartphone the state of the surrounding vehicles and the state of the driver have been recorded by two video cameras. These two recorded videos are combined with the manually segmented vehicle's driving data to define each recorded maneuver's irregularity levels by three experts. Each expert examined all recorded maneuver and assigned a category or a score for them. Each extracted maneuver will be assigned with an inter-rater agreement score that is calculated from the scores provided by each expert. If the inter-rater agreement score is low then the maneuver is assigned with deviated scores by the experts and there is no agreement on the classification. Therefore, experts are requested to iterate their evaluation to reach a high inter-rater agreement score for most of the extracted maneuvers.

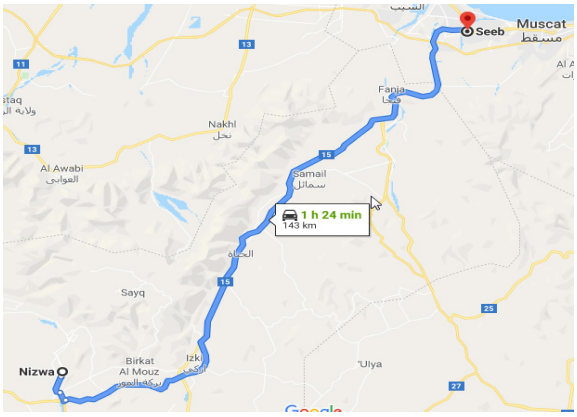


Figure 2: The route used to test data

One of the common problems that obstruct the development of reliable and accurate classification systems is the lack of adequate training and testing data. The lack of sufficient data for any classification problem could results in a noisy and instable class distribution. Therefore, for the sake of improving the system a second dataset was collected by installing a special app, developed by the author, in the smartphones of different 15 drivers, whom they drive on a daily base from the city of Nizwa to the capital Muscat over the route shown in Figure 2. The aim of collecting the second dataset is to enrich the training and testing data by performing a naturalistic driving through a route that is very dynamic and contains almost every different road types.

3. Calibration and Signal Processing

Smartphones' sensors are cheap low-grade sensors that have several deficiencies that may impose restrictions to their use in certain class of applications. In vehicular applications, it is hard to measure and guarantee smartphones accuracy and reliability due to five main prominent problems and limitations namely:

1- Sensor and algorithm complexity

2- Limited battery power.

3- The perturbation of the measurements due to the usage of the device that causes change in its orientation.

4- The low accuracy of smartphone sensors.

3.1. Sensors Calibration

Smartphones' IMUs have poor performance and are prone to error when especially used in measuring dynamic motion applications such as vehicle driving patterns. Therefore, sensors calibrations are used to correct sensors' deterministic errors. The most common types of these sources of errors are biases, scale factors and axes misalignment. In addition to the above mentioned deterministic errors, smartphones' sensors are also exposed to random noise and vibration, non-linearity due to thermal and magnetic effect and many others [37, 38]. To overcome these deficiencies a proper calibration process has been employed to eliminate Smartphones' sensors deterministic errors namely, fixed biases, scale factor errors, and misalignment errors.

The accelerometers are used to measure the instantaneous forces acting on the sensor. The measured acceleration along any axis consists of three components namely; linear acceleration, gravitational acceleration field and noise. On the other hand, the gyroscope measures the rate of change in the device's angular displacement, i.e. angular velocity, along the smartphone's three orthogonal axes. As mentioned errors in an IMU's measurements are contributed to three main categories: biases, scale-factor errors, and misalignment errors. The following equations (1) and (2) are used to model the accelerometer and gyroscope instantaneous measurement:

$$a_m(t) = S_a a_L(t) - g + b_a + n_a \quad (1)$$

$$\omega_m(t) = S_g \omega(t) + b_g + n_g \quad (2)$$

Here a_m, a_L, S_a, g, b_a and n_a represent the instantaneous measured acceleration component, the linear acceleration, scaling factor, gravitational acceleration, bias and noise respectively. While ω_m, ω, S_g, b_g and n_g represent the measured angular velocity along one of the three axes, the true angular velocity, scaling factor, bias and noise respectively. In this paper a new set of calibration procedures were proposed and used to estimate and calibrate smartphone's IMU sensors deterministic errors by considering the effects of vehicle vibration and noise [39-40].

3.2. Filtering

It is well known that data collected using IMU sensors contains substantial noise level, specially the accelerometer and the gyroscope readings. The noise in raw data can be contributed to several sources but mainly it is due to vibration noise and white Gaussian noise. There are different sources of vibration in vehicles such as road roughness, engine induced vibration, whirling of shafts, worn out parts, speed bumps, wheels, etc. In general, the data collected from the sensors contains random white Gaussian noise signals that oscillate with high frequency and typically have zero mean value. A sample for the measured noisy accelerometer and gyroscope signals are shown in Figure 3). In order to attain reliable and accurate recognition, the noise needs to be eliminated as much as possible. There are different approaches to filter

distorted signals; for example statistical filters [41], digital filters [27], Kalman Filter [42] and many others.

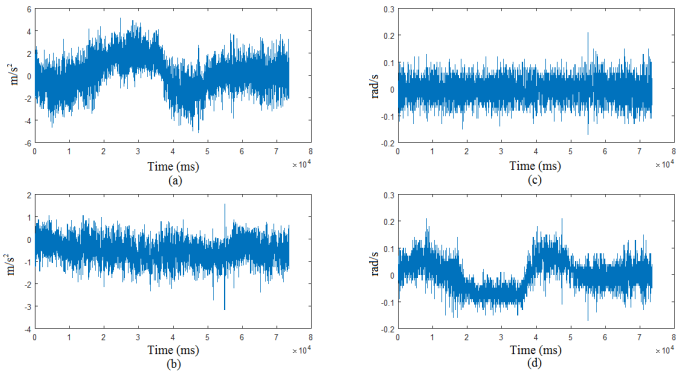


Figure 3: Noisy Smartphone Sensors Signals: (a) Accelerometer Signal x-Direction, (b) Accelerometer Signal y-Direction, (c) Gyroscope Signal x-Direction and (d) Gyroscope Signal z-Direction

In this paper three types of low-pass filters have been investigated namely; one-dimensional Kalman filter, simple moving average filter, and locally weighted running line smoother (LOSS). The performances of the aforementioned filters were compared to select the filter that effectively removes high frequency noise and maintain the real time variation of different driving patterns to the utmost extent.

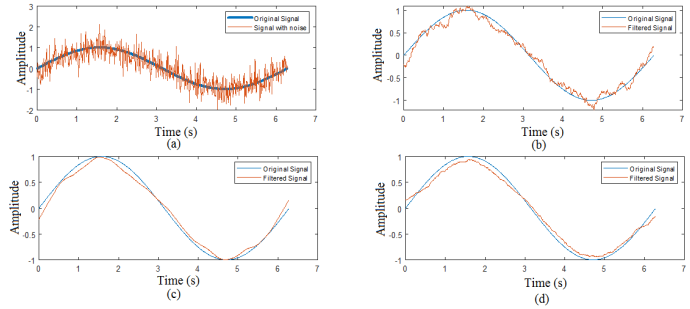


Figure 4: Filtering Results for Test Signal; (a) Original Test Signal; (b) Moving Average Filtering; (c) Kalman Filtering; (d) LOSS Filtering

The simplest filter is the moving average filter which filters the measured sampled signal by replacing each sampled value with the average of the neighboring sampled values defined within a given span or window. Kalman filter is a recursive predictive filter that is based on the use of state space techniques and recursive algorithms. The filter observes over time a series of measured data values, including noise, and then estimates new values that tend to be more accurate than those based on the actual measurement. The LOESS filter estimates the latent function in a point-wise fashion, where for each value of the measured sampled signal a new estimated value is found by using its weighted neighboring sampled (known) values. Detailed theory and implementation of this type of filters is presented in [43]. Due to the random irregular road pavement and vehicle vibration, it is found that the separation between of the real signals through total elimination of noise is not possible. Therefore, a test sinusoidal signal with imposed noise is used to assess the performance of the three filtering approaches. Figure 4 shows the results obtained for the three filters. It is clear that the LOESS technique can achieve very close results to Kalman filter where both effectively remove the nose while preserving the shape of the original signal. Figure 5 shows a sample of the measured accelerometer signals, from which the same conclusion regarding the two filters can be drawn. The LOSS filter has been used in the implementation of noise filtering unit of the complete system because of it is simpler to implement and faster to execute.

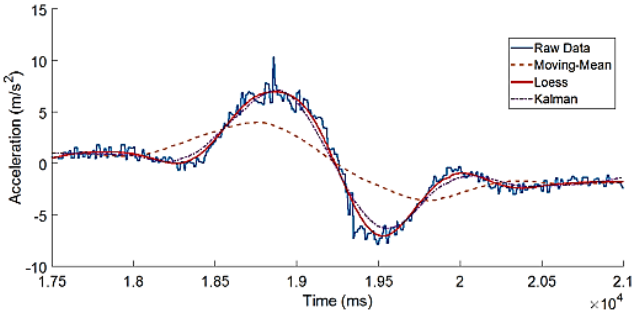


Figure 5: Raw and Filtered Accelerometer Signal

3.3. Smartphone Orientation

When smartphones are used to detect vehicle's driving patterns they could be placed at any location with any arbitrary orientation. Since sensors' measured signals are expressed in a specific frame that is static relative to the device, then one of the most interesting problems of smartphone-based vehicular applications is the separation of the dynamics of the smartphone from the dynamics of the vehicle. In this paper the smartphone is fixed to the windshield by a holder and it could have any orientation. Therefore, a reorientation correction process for the smartphone's coordinate systems is integrated as a pre-processing module to transform the measured sensors data to the vehicle coordinate system. This process is performed by a sequence of geometrical rotations using Euler Angles by assuming that the vehicle is horizontally aligned during the initial calibration period, so that the vehicle roll and pitch angles relative to a tangent frame both are zero. Assuming that the vehicle does not experience any acceleration, the smartphone's roll and pitch angles can be estimated from accelerometer measurements of the gravity vector as shown in Figure 6.

The determination of Euler angles has fully detailed in [44]. Upon the determination of rotational angles, the referenced coordinate system for the smartphone can be transformed to the vehicle-referenced coordinate system by multiplying with the reorientation matrices as follows:

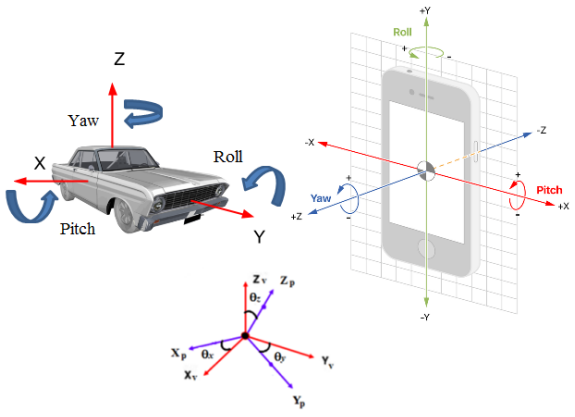


Figure 6: Smartphone and Vehicle Coordinate Systems

$$\begin{bmatrix} f_{xv} \\ f_{yv} \\ f_{zv} \end{bmatrix} = \mathbb{R} * \begin{bmatrix} f_{xp} \\ f_{yp} \\ f_{zp} \end{bmatrix} \quad (3)$$

where f_{iv} is the transformed measured signal to the vehicle reference, f_{ip} the measure signal in the phone reference and R is rotation matrix defined by:

$$\mathbb{R} = R(\theta_x) * R(\theta_y) * R(\theta_z) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{pmatrix} * \begin{pmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{pmatrix} * \begin{pmatrix} \cos\theta_z & \sin\theta_z & 0 \\ -\sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (4)$$

4. System Overview

The general architecture of the proposed system and its design methodology is shown in Figure 7). The system contains five main modules: raw data acquisition, pre-processing, driving parameters estimation, maneuvers detection unit and maneuvers identification and classification unit. The function of each module is briefly described as follows:

a. Android smartphones were used to capture vehicle's raw data through by using its IMUs, i.e. the accelerometer and gyroscope at a rate of 50 samples/second. Furthermore, smartphone's GPS data are also collected and used to provide vehicle's location and to correct speed estimation.
b. The pre-processing unit is designed to perform four functions namely; signals filtering, sensors error correction, transformation of sensors data to vehicle's coordinate system and finally estimates the driving parameters.
c. The manoeuvre detection unit is designed to detect and slice and executed manoeuvre in real-time by implementing the endpoint detection algorithm that identifies the starting and ending time.
d. The identification and classification unit is responsible to identify manoeuvre's type and its irregularity class. In this study the identification and the classification processes were performed separately by using a hybrid pattern matching and neural networks schemes.

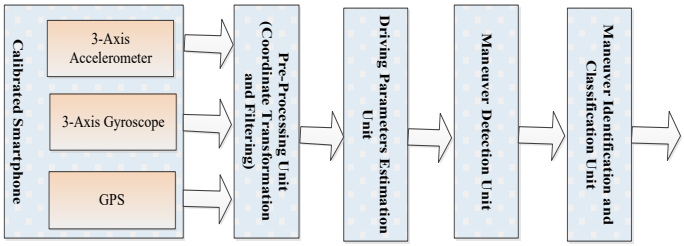


Figure 7: The Proposed System Architecture

5. Driving Parameters Estimation

A successful implementation of a driving maneuvers identification and classification system depends on the set of features and parameters that are used in the this process. In the literature there are two main approaches used to define the set of features and parameters namely, driver actions based approach and vehicle state based approach. In the first approach a number of essential driving signals can be collected from the vehicle's CAN-BUS such as vehicle speed, gas pedal pressure, brake pedal pressure, steering wheel angle, and acceleration. The second approach utilizes different approach by monitoring the vehicle's dynamic state in terms of longitudinal and lateral movements. These longitudinal and lateral movements can be detected and identified by using IMUs, GPS and the fusion of both the IMUs and the GPS. In this study it is believed that the description of vehicle' dynamical state in terms of actual driving parameters is more informative than other approaches and useful for future investigations. Therefore, the vehicle's acceleration, orientation and speed are used and represented as a time-series signals.

5.1. Longitudinal and Lateral Accelerations

The first and the simplest parameters that can be estimated directly are the longitudinal and the lateral acceleration components. The two components can be determined directly by subtracting the gravity the from the filtered accelerometer's Cartesian components, a_{xp}, a_{yp} and a_{zp}. The resulted singles are then transformed from the smartphone's to the vehicle's coordinate system by using Eqn. (4). The two components of the acceleration can be obtained by:

$$a_{lat} = a_{xv}\sin(\gamma) + a_{yv}\sin(\rho) \quad (5\text{-a})$$

$$a_{lon} = a_{xv}\cos(\gamma) + a_{yv}\cos(\rho) \quad (5\text{-b})$$

where, a_{lat} and a_{lon} are the lateral and longitudinal accelerations, a_{xv}, a_{yv} are the x and y components of the acceleration represented in the vehicle coordinate system and the angles γ and ρ are given by:

$$\gamma = \cos^{-1}\frac{a_{xv}}{\sqrt{a_{xv}^2+a_{yv}^2}} \quad \text{and } \rho = \cos^{-1}\frac{a_{yv}}{\sqrt{a_{xv}^2+a_{yv}^2}} \quad (6)$$

5.2. Vehicle's Orientation

The second estimated parameter is vehicle's heading or orientation. The attitude of a vehicle can be described by its angles of rotation along three axes as indicated in Figure 6). The attitude of any object can be modeled by different methods such as Direction Cosine Matrix, Quaternions and Euler angles. The first method has nine parameters to be estimated, hence it is difficult to implement. On the other hand, Quaternions has less parameters and it has advantages over Euler angles since it does not suffer from the gimbal lock problem, but still the Euler angles method is used in this study because it is more comprehensible and easier to decompose the rotations into separate degrees of freedom. Furthermore, since the smartphone is held in a fixed position inside the vehicle then the gimbal lock problem will not appear.

Euler angles can be found directly by using the measured gyroscope's data and integrate them with respect to time [45]. The accelerometer can also be used to estimate Euler angles, where simple trigonometric projections can be utilized to obtain both the Roll and the Pitch angles. Nevertheless, both methods could results in low prediction accuracy due to the stochastic bias variation in gyroscopes and the noisy accelerometer signals that results in an accumulated drift error [45]. To overcome the limitations of the direct approaches, sensor fusion techniques are developed to join all sensors' data to compensate the deficiency of each method. There are two filtering techniques namely; Kalman and complementary filters are widely used for sensor fusion. Because

of its simplicity and reliability, the digital complementary filter is utilized to estimate of the orientation. The filter uses a low-pass filter for the accelerometer's data to remove the high frequency noise and the dynamic bias, and utilizes a high-pass filter to suppress the drift from the gyroscope's data. Figure 8) shows a simple block diagram illustrating the implementation of a digital complementary filter. The filter can determine the angles using the following equation:

$$\theta_i = \alpha\left(\theta_{i-1} + \Delta\theta_{gi}\right) + (1-\alpha)\theta_{a\,i} \quad (7)$$

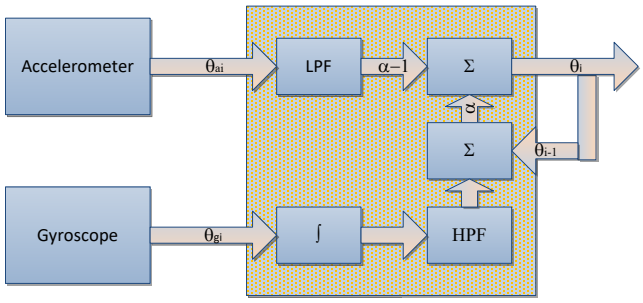


Figure 8: The Implementation of Digital Complementary Filter

In Eqn. (7), θ_i represent any of the estimated angles (Roll, Pitch or Yaw) and θ_{i-1} is the formerly determined angles, while the subscripts *g* and *a* designate the sensor type. The empirical constant α is selected to set the time constant of the error-signal calculations and stabilize the rate-sensor angle calculation.

5.3. Vehicle's Speed

Based on the kinematic equations, vehicle's speed can be found directly from the smartphone accelerometer data by integrating the corresponding acceleration component, longitudinal or lateral, as follows:

$$v_f(t) = v_i(t) + \int_0^t a(t)dt \quad (8)$$

where $v_f(t)$ is final speed, $v_i(t)$ is the initial speed and a(t) is vehicle's acceleration. The system is a discrete one, hence a sequence of the acceleration data is collected and equation (8) can be rewritten as:

$$v_f(n) = v_i(0) + \sum_{i=0}^{nT} \frac{1}{n} a(i) \quad (9)$$

where T is the sampling rate, n is the present nth sample and a(i) is the ith accelerometer value.

As it can be noted the basic idea of estimating the speed is very simple, and it is expected that with rigorous calibration and noise filtering of the raw data will improve the estimation process and a reasonable accuracy would be achieved. Nonetheless, speed estimation by the direct integration of accelerometer data suffer from significant deviations and unrealistic drifts because the accelerometer data is not pure and still contains some noise and other effects. This can be clearly shown in Figure 9), where the variation of the speed obtained from the GPS and those obtained by using the direct integration are shown. The doted green curve shows the estimated speed when the accelerometer data are not calibrated neither filtered while the dashed orange curve shows the trend in the estimated velocity for the preprocessed accelerometer data. As it can be seen, the error is accumulated over time and will obviously cause serious errors in the estimation of the speed. Therefore, a correction method is required to instantaneously correct the speed estimation by eliminating the accumulative error.

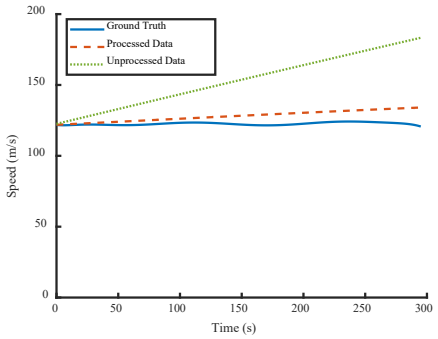


Figure 9: Comparison between Measured and Estimated Speed

In the literature there are mainly three methods that can be used to correct the accumulated error in numerical integration of the accelerometer data namely; high pass correction filter, Kalman filter and sensor fusion techniques. In this study, a simple but effective sensor fusion technique has been adopted and utilized to provide accurate velocity estimation [46].The basic idea of the technique is based on fusing the high frequency estimated speed with a low frequency ground truth speed obtained through the GPS. To illustrate the implementation of the vehicle speed estimator, consider Figure 10) that shows two equally spaced time intervals at which correction process take place at these instants. Suppose that the estimated speed and that measured by the GPS at time instant $t = T_A$ are v_{AE} and v_{AG} respectively, while those at the instant $t = T_B$ are v_{BE} and v_{BG}. The error of the speed at reference points then are given by:

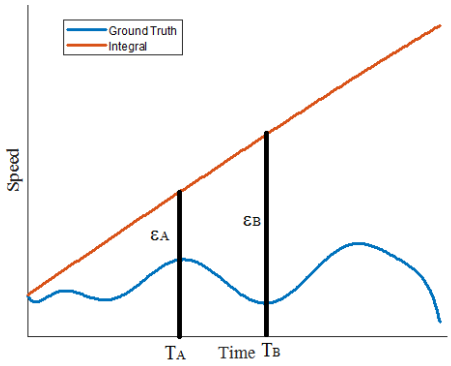


Figure 10: Vehicle Speed Correction Process

$$\varepsilon_A = v_{AG} - v_{AE} \quad \text{and} \quad \varepsilon_B = v_{BG} - v_{BE} \quad (10)$$

If we can consider that the rate of change in the accelerometer error is constant, i.e. that the accelerometer error varies linearly in this short interval, then this rate change can be expressed as follows:

$$\Delta\bar{a} = \frac{\varepsilon_B - \varepsilon_A}{T_B - T_A} \quad (11)$$

From Eqn. (11) the accumulative error at any instant (t) can be given by:

$$\emptyset = \int_{T_A}^{t} \Delta\bar{a}\ dt = \Delta\bar{a}kT_s \quad (12)$$

The constant k is the number of samples passed from the time instant T_A to t, and T_s is the sampling rate of the accelerometer. By knowing the accumulative error at any instant, then the estimated speed at instant (t) can be given by:

$$v_{tE} = v_{AG} + \sum_{i=T_A}^{t} \frac{1}{k} a(i) - \emptyset \quad (13)$$

Further enhancement been has been achieved by using exponential moving average correction to the error rate change as follows:

$$\Delta\bar{a}(i) = \alpha\Delta\bar{a}(i-1) + (1-\alpha)\frac{\varepsilon_B - \varepsilon_A}{T_B - T_A} \quad (14)$$

Figure 11 shows the estimated speed and the speed obtained through the GPS. It can be clearly seen that there is some small error between the estimation and the ground truth, which indicates that the estimator has excellent accuracy.

6. Maneuvers Detection

The main task that is required to be performed by any driving monitoring system is to reveal driving maneuver with reasonable level of reliability and accuracy. Table (1) presents a list of the maneuvers that are detectable by the proposed system. The system should be able to process the measured signals or the estimated parameters and attempts to detect the maneuver. The system detects the variation in the vehicle's longitudinal and lateral states by continuously monitoring the variation in the speed and orientation at a rate of 10Hz. Once a maneuvers is detected, the system can then proceed to recognize and classify it. There are different methods that can be used to detect the starting ending times of a maneuver [47-48]. In this paper the short-term energy endpoint detection algorithm with sliding window has been adopted, where the speed and orientation are continuously separated to event and non-event segments.

Table 1: Maneuvers Classes

1-	Acceleration straight road segment	2-	Acceleration curved road segment
3-	Braking straight road segment	4-	Braking curved road segment
5-	Left lane change straight road segment	6-	Left lane change curved road segment
7-	Right lane change straight road segment	8-	Right lane change curved road segment
9-	Merging into highway	10-	Exit from highway

The detection module consists of three key phases. In the first phase the signal is divided into non-overlapping windows of a 100ms period. The short-term energy is first computed for the data sequence in each window. For an infinite sequence of a discrete signal the energy is defined by:

$$\hat{x}[m] = x[m]\Lambda[n-m], n-N+1 \le m \le n \quad (15)$$

where Λ is a window function given by:

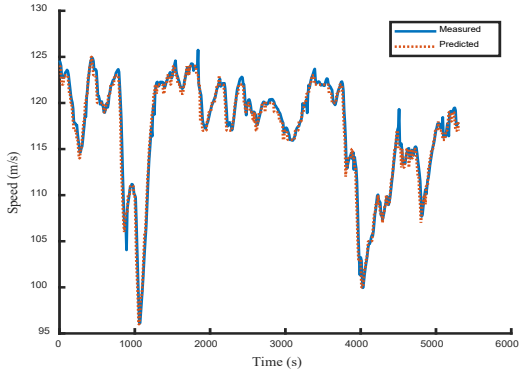
$$\Lambda[n-m] = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{Otherwise} \end{cases} \quad (16)$$

The energy contained in this short interval then can be computed by:

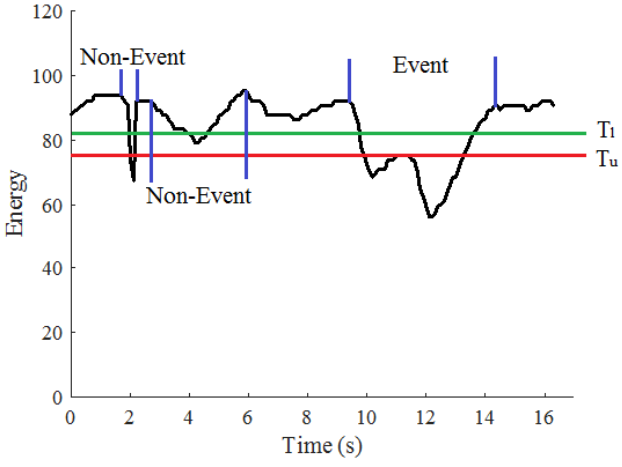
$$E_w = \sum_{m=n-N+1}^{n} (x[m]\Lambda[n-m])^2 \quad (17)$$

The value of the energy computed by Eqn. (17) is then compared to a threshold value, see Figure 12, and if the computed energy is lower than a specific threshold, then this frame is rejected and considered as a non-event segment. Otherwise, if the computed energy is higher than a lower threshold level T_l then the second stage will be executed.

Once the second stage is triggered, the starting time of the detected maneuvers is logged and the short-term energy is computed for a rectangular sliding window function as in Eqn. (17). Depending on the computed short-term energy in the second stage one of the following possibilities could be activated:



(a)



(b)

Figure 11: (a) Estimated and Measured Speed (b) Maneuver Detection using Short-Term Energy

- If the computed short-term energy is maintained to be less than the upper threshold T_u for less than 2 seconds or the energy fall back to a value lower than T_l, then this segment is treated as a non-event segment and the execution will resume from stage 1.

- If the computed short-term energy becomes higher than the upper threshold T_u for more than one second, then this segment will be treated as an event segment. In this case the system will continue computes the energy and compare it with the threshold until it drops back to the lower than T_l again. It is should be noted that if the energy falls down to be lower than T_u before the end of the first second then this segment will be regarded as a non-event segment, and the execution will start again from stage

7. Driving Maneuvers Identification and Classification

The Identification process is used to describe a driving maneuvers by a well-defined unique type. On the other hand the classification process evaluates the driving maneuvers to a certain class according to given criteria. In general there are two main approaches used for maneuver identification and classification, the first is based on the selection of a number of distinctive features and uses a machine learning technique, and the second approach is based on pattern matching algorithms. In this study 60% of the first dataset, described in section three, were used to study the time variation of each estimated driving parameters when a certain maneuvers has been executed. Figure 12) shows a sample of parameters' variations for a brake maneuvers. Intensive off-line observations and investigation were made to study and analyze the behavior of the parameters variation for each maneuvers type and to infer a general trend in their variations. It has been clearly noticed that the variation of the parameters follow common variation patterns and this fact agrees with what been presented in the literature. Figure 13) shows ideal approximated patterns that could be generated during any driving maneuver. As it can be notices that the driving parameters have limited number of patterns, for example the longitudinal acceleration, the lateral acceleration and the speed have three different patterns each, while the heading angle has five different patterns. Based on these facts, the identification and classification processes are performed separately where the identification process is performed by using a hybrid pattern matching neural network unit and the classification processed is performed by using another neural network.

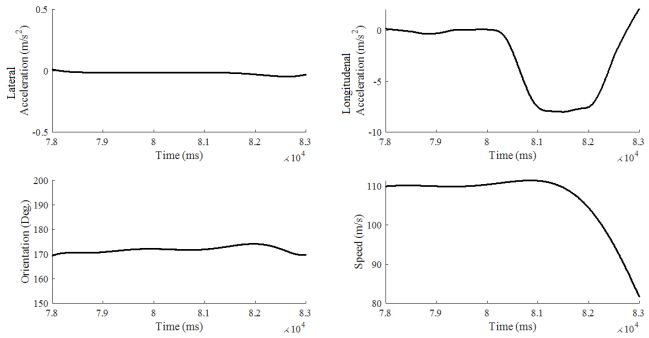


Figure 12: Driving Parameters Variation during a Break Manoeuvre

Figure 14) shows the basic units of the identification and classification module. The module consist of three units, a pattern matching units, to identify the patterns of estimated parameters, the second unit consist of a single feed-forward neural network that will identify the type of the maneuvers based on the combination of the identified patterns. Finally the third unit, which is responsible for the manoeuver classification, is performed by another neural network.

The pattern matching unit utilizes four DTW units, each one is designed to identify the pattern type for one of the parameters. The DTW is a pattern matching technique uses the discrete dynamic programming model to compute the alignment between two discrete time signals or sequences by assessing the similarity between them, regardless of the time/speed synchronization and distortion effects due to dynamic spatiotemporal differences.

Event Type	Longitudinal Acceleration	Lateral Acceleration	Heading Angle	Speed
Acceleration				
Braking				
Left- Lane Change				
Right- Lane Change				
Entry				
Exit				

Figure 13: Driving Parameters Variation for Different Maneuver

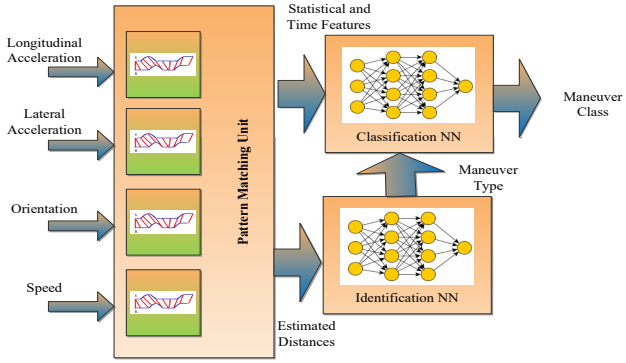


Figure 14: Structure of Identification and Classification Unit

To illustrate how the DTW algorithm works, consider two discrete signals, $R = \{r_1, r_2..., r_m\}$ which is the reference signal and an unknown signal $S = \{s_1, s2..., s_n\}$ that need to be identified, where m and n represent the number of samples in R and S, respectively. In Figure 15) the two time series that are highly similar, are shown. To align these two sequences an $(m \times n)$ local distance matrix is constructed by populating its elements with the Euclidean distances as follows:

$$d(i,j) = (R(i) - S(j))^2 \quad (18)$$

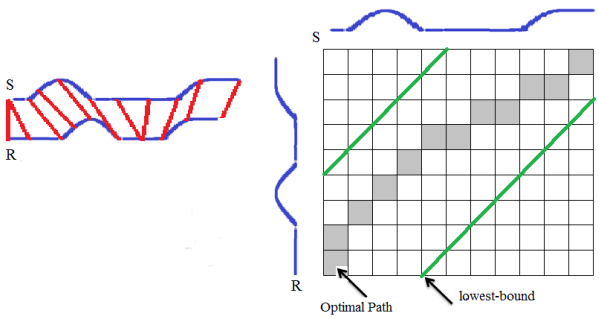


Figure 15: A DWT Matrix with a Warp Path

The correlation between the patterns of these two sequences is expressed by the cumulative distance measured over the optimal path within the distance matrix. For this purpose a warping matrix (D) is constructed and filled by searching for the minimum distance between the two sequences. This minimum distance can be found by computing the total distance D(i, j) for each entry (i, j) as:

$$D(i,j) = \begin{cases} d(1,1), & \text{if } i = j = 1 \\ d(i,j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}, & \text{otherwisw} \end{cases} \quad (19)$$

In the final stage the optimal warping path (P), as shown in Figure 15), and the DTW distance are to be computed. The Warping path is a set of adjacent matrix elements that detect the similarity between R and S and it represents the minimum distance between R and S. The matrix (P) contains K elements which range in value as:

$$\max(|R|, |S|) \leq K < (|R|, |S|) \quad (20)$$

Finally the optimal path Po is the one that minimizes the warping cost given by:

$$DTE(R,S) = \min\left\{\frac{1}{K}\sqrt{\sum_{k=1}^{K} P_k}\right. \quad (21)$$

To reduce the complexity of the DTW technique and thus speeding up the execution time, not all the warping paths are evaluated. To reduce the number of paths considered during the matching process four constraints are usually applied namely; monotonicity, continuity, boundary and windowing. To prevent the warping path from rolling back on itself, the monotonicity constrain is imposed to assure that any two adjacent elements, $p_k = (p_i, p_j)$ and $p_{k-1} = (\hat{p}_i, \hat{p}_j)$ will follow the inequalities $(p_i - \hat{p}_i) \geq 0$ and $(p_j - \hat{p}_j) \geq 0$. On the other hand, to guarantee a one-step advancement in the warping path, the continuity constrain is used to restrict the movement of the point (i, j) to the next point must be (i+1, j+1), (i+1, j), or (i, j+1). Finally, to assure that the warping path contains all points of both sequences, the boundary constrain is used to force the warping path to start from the top left corner and ends at the bottom right corner. Finally the windowing will prevent pathological mappings between two sequences by restricting one point in the first sequence to be mapped to a limited number of points in the other sequence; hence the path is likely to be around the diagonal without deviating much from it [49].

In order to identify the types of the maneuver by the DWT technique, a set of templates for each maneuvers are required as reference signals. To find a template signals for different set of sequences, one can consider that template sequence could be the longest common sequence, or the medoid sequence, or the average sequence. The selection of the reference pattern for each specific maneuvers class is not a simple task because the set of sequences collected for each maneuvers class have different time durations and amplitudes. In this study, the sequence with the minimum average distance from all other sequences for a specific class is selected as the reference template. The details of the averaging strategy are given in [50].

When dealing with single 1-D signal identification, the DTW technique will directly identify the test signal. Nevertheless, when it is required to identify a maneuvers or an action by using multiple measured signals, the DTW won't be able to identify its type directly. This is because each signal will need a special DTW unit to identify the pattern of the signal only. In order to predict the type of the maneuvers, the combination of the DTW distances were used as features to train a neural network that will be used for the final identification of the maneuvers.

Statistical and time metrics have been utilized as features to classify maneuvers' irregularities. In general maneuvers are categorized into one of three main types; Hard, Normal and Light. By studying all the possible patterns for all the classes, see Figure 13), it was found that there are seven distinctive patterns as shown in Figure 16). For each pattern time metrics and statistical values such as max, min, mean, standard of deviation and the variance are used as classification features. The segmented maneuvers collected in the first dataset were represented by a vector that contains the statistical and the time metrics for each driving parameter and the classification label. As it has been mentioned the classification labels are obtained from classification made by the experts. It is should be noted that the percentage of the rejected samples was less than 10%.

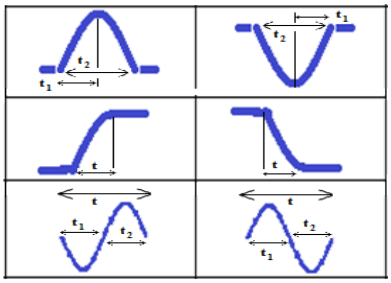


Figure 16: Time features for the common signals

8. System Validation

This section presents an experimental evaluation for the three main components of the proposed system namely, detection, the identification and classification units. To test the performance of the system, two different datasets have been used. In the first phase the first dataset described in section three has been used to train and conduct a preliminary testing of the system. In this phase a single smartphone (Samsung Galaxy S5) have been calibrated and used in all the tests. As been mentioned ten drivers with different car models were asked to conduct each of the basic maneuvers, listed in table (1), at least five times with different severity levels from the Light to the Hard level. From this dataset 1500 maneuvers were collected and segmented then each was combined with two videos, the first used to record the surrounding traffic state and the second used to record the driver's state, i.e. hands and feet. To obtain the ground truth classification of each maneuvers, experts were consulted to evaluate the abnormality level. Each assessor assigned abnormality level scores for the entire recorded maneuvers and the final score for every recorded maneuvers was calculated using the inter-rater reliability method.

For this initial dataset 60% of the collected samples were utilized to extract the information and the features that were used in the training phase. The information and features includes the computation of the DTW reference templates, the DTW's distances and the statistical and time metrics. The DTW reference

templates have been stored and used to measure a vector that includes the distances of the warping for each recorded driving parameter of a specific tested sample. These vectors are used to train the first neural network which is used to identify the type of the maneuvers, while another set of vectors that contain the statistical and time metrics features used to train the classification neural network.

The remaining 40% of the first dataset have been used to validate the performance of the system. Figure 17 shows results obtained when the 40% testing samples fed to the identification module. It can be seen clearly that the identification rate is very high and reached a 100% successful rate for some of the maneuver and the minimum identification rate scored is 85%, and this is obvious since the identification process combines all the similarities for the four parameters. Figure 18, shows the confusion matrix classification for the 40% testing samples. Again it can be noticed that the classification rate is very high nearly 95% for some cases. This high identification and classification rates are expected since the two processes have been separated.

In the second phase of this study, the author developed a complete Android-based APP that detects, identify and classify the highway driving maneuver in real-time. The architecture of the system is described in section five. The first test for the system was to check its ability to detect the starting and end of a driving maneuvers. The complete trips performed by the ten drivers, from which the first dataset was collected, were used just to detect the driving maneuver. Figure 19, shows the detection of maneuvers for a sample trip. It should be noted that the first 1500 maneuver collected in the first data set were used to determine the lower and upper thresholds for short-term energy for both the vehicle's speed and orientation. The results obtained by the system were compared with those already registered manually by the author. The system showed a high detection rate where more than 96% of the manually registered maneuver were detected successfully by the detection unit.

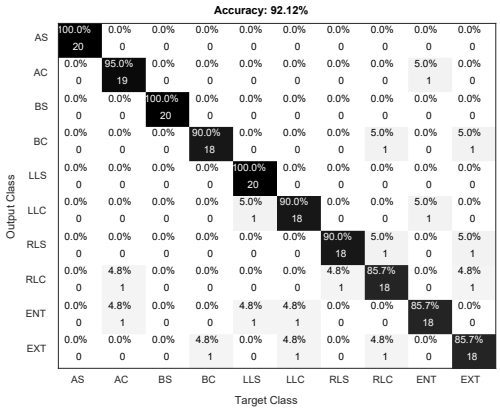


Figure 17: Confusion Matrix for Maneuvers Identification

The developed application allows two types of analysis namely, the real-time analysis and off-line analysis. The graphical interface of the developed application is designed to achieve user friendly interaction with the driver. All the driving maneuver are presented instantly to the driver using both audio and visual indicators. Each maneuvers type is represented by a special icon that it will be lighted up; along with a human voice that utters the name of the maneuvers when it is detected by the application. Furthermore, the icon's color will be changed in accordance to the class of the maneuvers. This attractive and user friendly interface helped in measuring the performance and the reliability of the system in the second stage of the study. As been mentioned in section three a second dataset was collected from a naturalistic driving performed by 15 drivers who are daily travel a distance of 140km back and forth. We have asked driver's companions to check the alerts produced by the application and check them. Data recoded by the companions and those stored in the SD card are then analyzed. Table (2), presents a comparison between the number of maneuver performed by each driver and those detected by the system. Note that the first number in any cell represents the performed maneuver, while the second number represents the detected maneuver. Figure 20 and 21 show the confusion matrix for the maneuvers' identification and classification respectively.

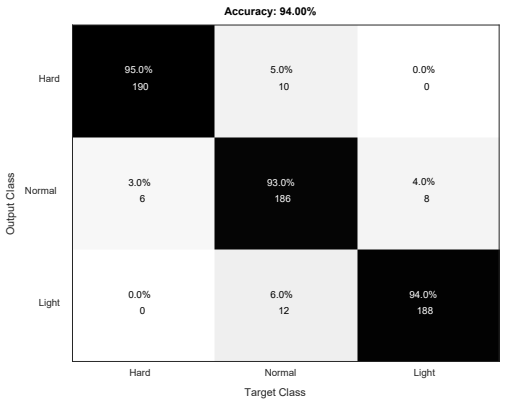


Figure 18: Confusion Matrix for Maneuvers Classification

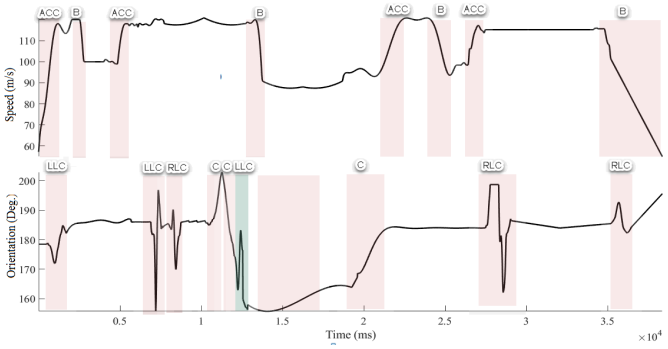


Figure 19: Driving Maneuvers Detection in a Complete Driving Trip

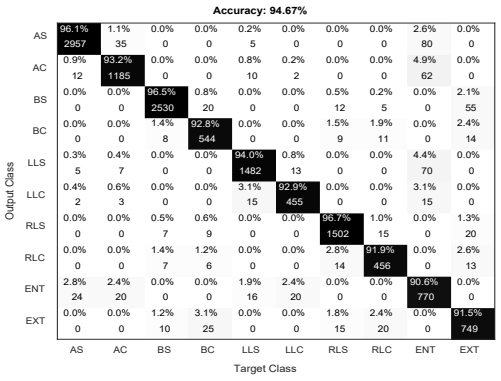


Figure 20: Confusion Matrix for Maneuvers Identification (Second Dataset)

Table 2: Comparison between Performed and Detected Maneuver for each Driver

Driver	ACC/S	ACC/C	B/S	B/C	LLC/S	LLC/C	RLC/S	37 : 34	Ent.	Exit
1	221: 210	95 : 87	184 : 176	48 : 42	112:109	37 : 33	114 : 111	40 : 36	60 : 57	60 : 56
2	211: 207	86 : 81	159 : 156	41 : 36	105:99	30 : 26	104 : 99	35 : 32	58 : 52	59 : 54
3	199: 189	85 : 81	154 : 142	43 : 38	108:100	41 : 36	107 : 102	39 : 36	59 : 57	59 : 57
4	230: 222	87 : 84	200 : 194	44 : 40	110:104	36 : 33	109 : 105	31 : 28	58 : 56	58 : 56
5	219: 213	93 : 86	194 : 186	46 : 42	114:109	40 : 36	111 : 106	34 : 29	60 : 56	60 : 57
6	216: 209	91 : 85	188 : 181	45 : 40	107:104	39 : 34	106 : 101	41 : 35	59 : 56	60 : 56
7	220: 212	89 : 83	166 : 162	41 : 36	109:108	38 : 32	110 : 104	37 : 34	59 : 55	59 : 55
8	223: 214	90 : 85	172 : 168	42 : 37	111:107	34 : 30	113 : 108	40 : 36	58 : 57	59 : 57
9	218: 210	94 : 86	190 : 187	47 : 40	112:110	37 : 34	105 : 100	35 : 32	55 : 51	55 : 52
10	210: 202	93 : 88	185 : 179	48 : 44	108:102	40 : 36	108 : 104	39 : 36	56 : 53	57 : 54
11	200: 193	92 : 84	184 : 180	43 : 39	115: 110	35 : 32	112 : 107	31 : 28	58 : 55	57 : 53
12	207: 201	90 : 84	180 : 173	44 : 39	104:99	39 : 36	110 : 106	34 : 29	60 : 56	60 : 56
13	215: 208	89 : 85	176 : 172	40 : 35	106:102	31 : 28	104 : 99	41 : 35	57 : 55	57 : 54
14	201: 196	88 : 83	185 :180	42 : 37	110:106	34 : 29	101 : 97	37 : 34	59 : 56	60 : 56
15	196: 191	95 : 89	190 : 186	45 : 41	114:108	41 : 35	110 : 104	40 : 36	57 : 52	56 : 51

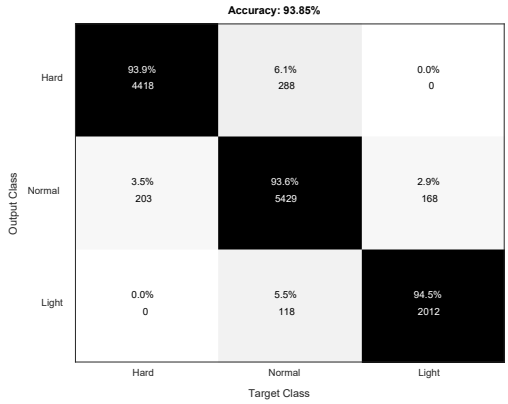


Figure 21: Confusion Matrix for Maneuvers Classification (Second Dataset)

9. Conclusion

This paper proposes a novel real-time smartphone-based highway driving maneuvers identification and classification system. The system collects vehicle's longitudinal and lateral movements using smartphone's IMU sensors and vehicle's location by using the GPS. All the smartphones used in this investigation were calibrated the collected smartphone's raw data were filtered and smoothed by using the statistical LOESS filter. Driving parameters namely vehicle's acceleration, orientation and speed were used in the identification and classification of the driving maneuver. Sensor fusion techniques have been used to estimate vehicle's orientation and speed.

It was noted that driving parameters for each event class have common patterns, thus the DTW technique combined with a feed-forward neural network are used for the identification. In the classification process a second neural network was employed, and it is trained to classify a maneuver classes. Two datasets were collected through a number of experiments conducted on highway routes. Results obtained in this study show an excellent detection, identification and classification rates were achieved.

References

[1] S.C. Davis and Robert G. Boundy, Transportation Energy Data Book, Ed. 38, Oak Ridge National Lab TN-United States, 2020.

[2] D. Searsona , G. Pontea , T Paul Hutchinsona , Robert Andersona and Mary Lydona, "Emerging vehicle safety technologies and their potential benefits: Discussion of expert opinions" in 2015 1st Australasian Road Safety Conference (ARSC2015), 1-11, 2015,

[3] World Health Organization: "Global status report on road safety 2018", WHO: Geneva, Switzerland, 2018.

[4] A. Haghi, D. Ketabi, M. Ghanbari and H. Rajabi, "Assessment of Human Errors in Driving Accidents; Analysis of the Causes Based on Aberrant Behaviors", Life Science Journal, **11**(9), 414-420, 2014.

[5] M.Q. Khan and Sukhan Lee, "A Comprehensive Survey of Driving Monitoring and Assistance Systems", Sensors **19**(11), 2574-2606 ,2019, doi: 10.3390/s19112574

[6] C.M. Martinez ; M. Heucke ; Fei:Yue Wang ; Bo Gao ; Dongpu Cao, "Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey", IEEE Transactions on Intelligent Transportation Systemsm **19** (3) , 666–676, 2018, doi: 10.1109/TITS.2017.2706978.

[7] M. VivaldiniI; Silvio R. I. PiresI; Fernando Bernardi de SouzaII, "Improving logistics services through the technology used in fleet management", Journal of Information Systems and Technology Management, **9**(3), 541-562, 2012, doi: 10.4301/S1807-17752012000300006

[8] M.J. Prasad, S Arundathi, Nayana Anil, Harshikha, B. S. Kariyappa, "Automobile black box system for accident analysis", in 2014 International Conference on Advances in Electronics Computers and Communications (ICAECC), 1-5, 2014, doi: 10.1109/ICAECC.2014.7002430.

[9] T.K. Chan ; Cheng Siong Chin ; Hao Chen ; Xionghu Zhong, "A Comprehensive Review of Driver Behavior Analysis Utilizing Smartphones" IEEE Transactions on Intelligent Transportation Systems, 21(10), 4444 - 4475, 2020, doi: 10.1109/TITS.2019.2940481.

[10] M. S. N. Al-Din, "Highway Driving Events Identification and Classification using Smartphone," in the 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 1-6, 2018, doi: 10.1109/ICCSDET.2018.8821090.

[11] R. Kridalukmana, H. Y. Lu and M. Naderpour, "An object oriented Bayesian network approach for unsafe driving maneuvers prevention system", in the 2017 12th International Conference on Intelligent Systems and Knowledge

Engineering (ISKE), 1-6, 2017, doi: 10.1109/ISKE.2017.8258758.
[12] I. Silva and JoséEugenio Naranjo, "A Systematic Methodology to Evaluate Prediction Models for Driving Style Classification", Sensors, **20**(6), 1692-1713, 2020, doi: /10.3390/s20061692
[13] H. Malik, G. S. Larue, A. Rakotonirainy and F. Maire, "Fuzzy Logic to Evaluate Driving Maneuvers: An Integrated Approach to Improve Training," IEEE Transactions on Intelligent Transportation Systems, **16**(4), 1728-1735, 2015, doi: 10.1109/TITS.2014.2371061.
[14] T. K. Chan, C. S. Chin, H. Chen and X. Zhong, "A Comprehensive Review of Driver Behavior Analysis Utilizing Smartphones," IEEE Transactions on Intelligent Transportation Systems, **21**(10). 1-32, 2019, doi: 10.1109/TITS.2019.2940481.
[15] N. AbuAli, "Advanced vehicular sensing of road artifacts and driver behavior," in the 2015 IEEE Symposium on Computers and Communication (ISCC), 45-49, 2015, doi: 10.1109/ISCC.2015.7405452.
[16] L. M. Bergasa, D. Almería, J. Almazán, J. J. Yebes and R. Arroyo, "DriveSafe: An app for alerting inattentive drivers and scoring driving behaviors," in the 2014 IEEE Intelligent Vehicles Symposium Proceedings, 240-245, 2014, pp. doi: 10.1109/IVS.2014.6856461.
[17] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya and M. C. González, "Safe Driving Using Mobile Phones," IEEE Transactions on Intelligent Transportation Systems, **13**(3), 1462-1468, 2012, 1462:1468, doi: 10.1109/TITS.2012.2187640.
[18] Yang Zheng, A. Sathyanarayana and J. H. L. Hansen, "Threshold based decision-tree for automatic driving maneuver recognition using CAN-Bus signal," in the 2014 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 834-2839, 2014, doi: 10.1109/ITSC.2014.6958144.
[19] T. Pholprasit, W. Choochaiwattana and C. Saiprasert, "A comparison of driving behaviour prediction algorithm using multi-sensory data on a smartphone," in the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 1-6, 2015, doi: 10.1109/SNPD.2015.7176249.
[20] H.R. Eftekhari and Mehdi Ghatee, "Hybrid of discrete wavelet transform and adaptive neuro fuzzy inference system for overall driving behavior recognition", Transportation Research Part F, **58**, 782-796, doi: 10.1016/j.trf.2018.06.044.
[21] H.R. Eftekhari, M. Ghatee,: "A similarity:based neuro:fuzzy modelling for driving behavior recognition applying fusion of smartphone sensors", Journal of Intelligent Transportation Systems, **23**(1), 72-83, 2019, doi: 10.1080/15472450.2018.1506338.
[22] C. Arroyo, L. M. Bergasa and E. Romera, "Adaptive fuzzy classifier to detect driving events from the inertial sensors of a smartphone," in the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 1896-1901, 2016, doi: 10.1109/ITSC.2016.7795863.
[23] A. Aljaafreh, N. Alshabatat and M. S. Najim Al:Din, "Driving style recognition using fuzzy logic," in the 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012), 460-463l, 2012, doi: 10.1109/ITSC.2014.6957822.
[24] M. Najim, Ahmad Aljaafreh, Nashat Albdour and Maen Saleh:, "Driving Styles Recognition Using Decomposed Fuzzy Logic System", International Journal of Electrical, Electronics and Computer Systems, **16**(1) 820-824, 2013.
[25] D.A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), 1609-1615, 2011, doi: 10.1109/ITSC.2011.6083078.
[26] C. Saiprasert, Pholprasit, T. and Thajchayapong, S., "Detection of Driving Events using Sensory Data on Smartphone", International Journal of Intelligent Transportation Systems Research, **15**(1), 17-28, 2015, doi:10.1007/s13177-015-0116-5.
[27] G. Singh, D. Bansal and Sanjeev Sofat, "A smartphone based technique to monitor driving behavior using DTW and crowdsensing", Pervasive and Mobile Computing, **40**, 56-70, 2017, doi: 10.1016/j.pmcj.2017.06.003.
[28] P. Tkachenko, J. Zhou, Davide Gagliardi and Luigidel Re, "On:line Maneuver Identification in Highway Traffic Using Elastic Template Matching", IFAC:PapersOnLine, **51**(15), 557-562, 2018, doi: 10.1016/j.ifacol.2018.09.214.
[29] M. Wu; Zhang, S.; Dong, Y., "A Novel Model-Based Driving Behavior Recognition System Using Motion Sensors", Sensors, **16**(10), 1746-1769, 2016, doi: 10.3390/s16101746.
[30] C. Chen, 1 Xiaohua Zhao, "Driver's Eco-Driving Behavior Evaluation Modeling Based on Driving Events", Journal of Advanced Transportation, **2018**, 1-12, 2018, doi: 10.1155/2018/9530470.
[31] J. Yu, Z. Chen, Y. Zhu, Y. Chen, L. Kong and M. Li, "Fine-Grained Abnormal Driving Behaviors Detection and Identification with Smartphones," IEEE Transactions on Mobile Computing, 1**6**(8), 2198-2212, 2017, doi: 10.1109/TMC.2016.2618873.
[32] J. F. J´unior, E. Carvalho, B. V. Ferreira, C. de Souza, Y. Suhara, A. Pentland, and G. Pessin, "Driver behavior profiling: An investigation with different smartphone sensors and machine learning," PLoS one, 12(4), e0174959, 2017. doi:10.1371/journal.pone.0174959
[33] P. Brombacher, J. Masino, M. Frey and F. Gauterin, "Driving event detection and driving style classification using artificial neural networks," in the 2017 IEEE International Conference on Industrial Technology (ICIT), 997-1002, 2017, doi:10.1109/ICIT.2017.7915497..
[34] U. Fugiglando et al., "Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment," IEEE Transactions on Intelligent Transportation Systems, **20**(2), 737-748, 2019, doi: 10.1109/TITS.2018.2836308.
[35] Manon Kok, Jeroen D. Hol and Thomas B. Schon, "Using Inertial Sensors for Position and Orientation Estimation", Foundations and Trends in Signal Processing: **11**(1-2), 1-153, 2017, doi:10.1561/2000000094.
[36] I. Aicardi , P. Dabove , A. Lingua and M. Piras, "Sensors Integration for Smartphone Navigation: Performance and Future Challenges", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, **XL**(3), 9-16, 2014 .
[37] I. P. Prikhodko, B. Bearss, C. Merritt, J. Bergeron and C. Blackmer, "Towards self-navigating cars using MEMS IMU: Challenges and opportunities," in the 2018 IEEE International Symposium on Inertial Sensors and Systems (INERTIAL), 1-4, 2018, doi:10.1109/ISISS.2018.8358141.
[38] J. Almazán, L. M. Bergasa, J. J. Yebes, R. Barea and R. Arroyo, "Full auto:calibration of a smartphone on board a vehicle using IMU and GPS embedded sensors," in the 2013 IEEE Intelligent Vehicles Symposium (IV), 1374-1380, 2013, doi: 10.1109/IVS.2013.6629658.
[39] M. S. Najim Al:Din, "Calibration and Pre-Processing Techniques for a Smartphone:Based Driving Events Identification and Classification System," in the 2018 IEEE Electron Devices Kolkata Conference (EDKCON), 396-402, 2018, doi: 10.1109/EDKCON.2018.8770487.
[40] Munaf Salim Najim Al:Din; Atef Saleh Al:Mashakbeh, "Development of a highway driving events identification and classification using smartphone", International Journal of Nanoparticles, **12**(1-2), 152-173, 2020, doi;10.1504/IJNP.2020.106006
[41] T. Pan, C. Kuo and M. Hu, "A noise reduction method for IMU and its application on handwriting trajectory reconstruction," in the 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 1-6, 2016, doi: 10.1109/ICMEW.2016.7574685.
[42] M. Wu, S. Zhang and Y. Dong, "A novel model:based driving behavior recognition system using motion sensors", Sensors, **16**(10), 1746-1769, 2016, doi: 10.3390/s16101746 .
[43] Wendy L. Martinez and Angel R. Martinez, "Exploratory Data Analysis with MATLAB", Chapman & Hall/CRC Press, 2005.
[44] F Bruwer and M Booysen, "Vehicle Acceleration Estimation using Smartphone:Based Sensors", in the 2015 34th Annual Southern African Transport Conference (SATC 2015), 118-129, 2015, doi: 10.13140/RG.2.1.3244.4640 .
[45] Y. Huang, Y. Rizal and M. Ho, "Development of attitude and heading reference systems," in the 2015 International Automatic Control Conference (CACS), 13-18, 2015, doi: 10.1109/CACS.2015.7378358.
[46] T. Lei, A. A. Mohamed, and C. Claudel, "An imu-based traffic and road condition monitoring system," HardwareX, **4**, e00045, 2018, doi: 10.1016/j.ohx.2018.e00045
[47] Xu Dawei,Wu Bian, Zhao Jianwei et al., "A real time algorithm for voice activity detection in noisy environment", Computer Engineering and Application, **24**(1), 115-117, 2003.
[48] Liang-Sheng Huang and Chung-Ho Yang, "A novel approach to robust speech endpoint detection in car environments," in the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings 1751-1754, 2000, doi: 10.1109/ICASSP.2000.862091.
[49] B. Johnen and B. Kuhlenkoetter, "A Dynamic Time Warping algorithm for industrial robot motion analysis," in the 2016 Annual Conference on Information Science and Systems (CISS), 18-23, 2016, doi: 10.1109/CISS.2016.7460470.
[50] F. Petitjean, J. Inglada, and P. Gancarski, "A global averaging method for dynamic time warping, with applications to clustering," Pattern Recogn., **44**(3), 678–693, 2011, doi: 10.1016/j.patcog.2010.09.013.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 211-220 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Trace-Driven Simulation of LoRaWAN Air Channel Propagation in an Urban Scenario

Eugen Harinda, Hadi Larijani *, Ryan M. Gibson

Glasgow Caledonian University, Networks and Cybersecurity, School of Computing and Built Environment, Glasgow, G4 0BA, United Kingdom

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 18 October, 2020
Online: 08 November, 2020

Keywords:
LoRaWAN
Trace-Simulation
Received Signal Strength
Deterministic models
Empirical Models
Urban Environment

ABSTRACT

Long-range, Low-Power Wide Area Network (LoRaWAN) is a very scalable solution for the Internet of Things (IoT). Due to the air channel environment's complexity, connectivity is a crucial parameter for successfully planning and deploying the IoT networks. Measurements and simulations have been used to evaluate LoRaWAN propagation models in the Urban environment, but it is a challenging task. While practical propagation evaluation has been prohibitively expensive, the theoretical modeling results have been less accurate. This paper uses real-world measurements and a trace-driven simulation technique to evaluate the RF propagation models' prediction performance for LoRaWAN 868 MHz propagation. First, a novel LoRaWAN trace-driven simulation of Glasgow city centre has been performed. Second, LoRaWAN 868 MHz measurements have been used to perform a critical analysis of LoRaWAN trace-driven Radio Frequency (RF) propagation models and validation. The processed trace dataset is composed of GPS coordinates, and the corresponding LoRaWAN received signal strength. The dataset has been extracted from 5017 datasets of LoRaWAN measurements taken from Glasgow city centre. A trace simulation program built-in ICS-Telecom was used to simulate LoRaWAN propagation in the real-world urban environment. Comparison of LoRaWAN simulation traces and the real-world data was performed to evaluate the prediction performance accuracy of Deygout 94, ITU-R 525/526, and COST-Walfish Ikegami (COST-WI) propagation models. All models over-estimated LoRaWAN trace-simulated RSS levels in comparison to collected measurement samples. While Deygout 94 prediction accuracy was higher with mean absolute error (MAE) at 0.83 dBm and standard deviation (SD) at 4.17 dBm, COST-WI performed poorly with MAE and SD at 2.87 dBm and 10.96 dBm respectively.

1 Introduction

This paper is an extension of work originally presented in The International Conference on UK-China Emerging Technologies, Glasgow, UK, August 2019 [1]. The purpose of this work was the application of a trace-based approach to evaluate the prediction capabilities of the conventional propagation models for LoRaWAN networks. As the Internet of Things (IoT) network is steadily growing, an understanding of conventional radio propagation models' prediction accuracy is key to the successful planning and deployment of Long-range, Low-Power Wide Area Network (LoRaWAN).

Unlicensed IoT technologies allow anyone with an idea to create and deploy IoT applications. Authors in [2] have predicted that a global number of IoT connected devices will rise to 75.44 from 15.4 billion in 2025. LoRaWAN [3] is one of the Low Power Wide Area Network (LPWAN) technologies for the interconnection of physical things. It uses Long-Range (LoRa) modulation technology [4] at the physical layer to achieve long-distance connectivity among the connected things. LoRaWAN end-devices connect to LoRa gateway using Radio Frequency (RF) propagation through the air channel and LoRaWAN network and application servers. Due to the complexity of the air channel environment, connectivity is a crucial parameter for the successful deployment of the IoT networks. To understand whether a LoRaWAN network meets the acceptable connectivity performance, a detailed investigation of the network using radio propagation prediction models is requisite to the network's deployment. However, Received Signal Strength (RSS) prediction in LoRaWAN networks has become a complex task due to prediction

*Corresponding Author: Hadi Larijani, Cowcaddens Road, Glasgow, G4 0BA, U.K, (0044)-141-331-3190, H.Larijani@gcu.ac.uk

performance differences in RF propagation models. These models' ability to accurately predict the RSS depends on how correctly they capture factors affecting RF signal propagation in the air channel.

Propagation models rely on natural and environmental factors to calculate signal power attenuation, path loss in the air channel. In the Line-of-Sight (LOS), the signal power attenuators include RF and the distance between end-devices and LoRa gateways. On the other hand, in cities, where Non-Line-of-Sight (NLOS) is dominant, environmental attenuators of signal power include human-made structures such as dense tall and short buildings, cars, and bridges. The path loss incurred by the propagating signal due to these obstructions is a consolidation of LOS and NLOS path loss due to multipath, scattering, diffraction, and reflection losses. To incorporate sufficient natural and environmental factors in the RF propagation models, numerous empirical propagation models [5] and deterministic propagation models [6] have been designed.

Different empirical and deterministic RF propagation modeling techniques highlight contrasts in prediction performance accuracy and application scenario. Empirical models were designed to predict the RSS in the event of insufficient knowledge of terrain profile data. These models were designed based on actual measurements taken from the propagation environment conditions. They are made of simple mathematical formulae with no reliance on the digital terrain data and are less accurate for a distance range of less than 1 *km* [7]. On the other hand, deterministic models largely depend on the digital terrain models to predict the RSS. These models take into consideration the free space and NLOS path loss. Accordingly, detailed knowledge of digital terrain models is a prerequisite for their prediction performance. Further contrasts exist within empirical and deterministic models in terms of parametric design consideration and prediction performance accuracy. Empirically, measurements and simulation methods are accurately used to investigate these models' ability to predict the RSS in radio networks.

Taking measurements from the real-world networks and simulation of the air channel propagation are vital methods used to evaluate the RF models. Studies [5], [6] and [8] have used measurements and simulations to evaluate the performance of various propagation models, but less has been done to validate their applicability on LoRaWAN networks in an urban environment. The studies in [9]–[13] have used only measurements to understand the behavior of air channel propagation on LoRaWAN network operations, but this is an expensive approach as best practices involve taking measurements at the later stages of network optimization. To validate the applicability of empirical propagation models for LoRaWAN in the city environment, our previous work in [14] showed less satisfactory empirical models' prediction performance against measurements. Trace-driven simulation can bridge the gap between measurement and simulation to evaluate RF propagation models' performance for LoRaWAN networks.

In this paper, we use real-world measurements and trace-driven simulation technique to evaluate the RF propagation models' prediction performance for LoRaWAN networks. Trace-driven simulation method can be used to evaluate computer-based models, but it works well if a trace contains the experience required by the trace simulation program [15]. The RF models are implemented in ICS-Telecom [16], and the simulation required a digital terrain model of Glasgow city centre. The simulation that uses a trace that contains the system's experience makes a more realistic simulation performance [17]. The study utilized a digital terrain model at 25 *m* and a dataset from Glasgow city centre. This dataset contained 5017 measurements with GPS coordinates and their corresponding LoRaWAN RSS. Trace-driven simulation results were compared against the real-world measurements to evaluate the prediction accuracy of Deygout 94, ITU-R 525/526, and COST-WI. All models over-estimated LoRaWAN trace-simulated RSS but showed superior prediction performance over empirical models in [14]. The result shows that the LoRaWAN network's trace-driven simulation with these models provides a more realistic RSS prediction than the conventional simulations.

The main contributions of this work are given in the summary below:

- A novel use of measurement and trace-driven simulation of LoRaWAN 868MHz propagation prediction for Glasgow city centre.
- The trace-driven simulation and extraction of LoRaWAN RSS traces spread over the digital terrain model of Glasgow city centre with the minimum, maximum range of trace signal levels, and the total number of trace measurements.
- Critical analysis of LoRaWAN trace-driven RF propagation models and validation with real-world measurements. Deygout 94 prediction accuracy was higher with *MAE* at 0.83 *dBm* and *SD* at 4.17 *dBm*.
- Visual comparative analysis of trace patterns between the LoRaWAN real-world measurement site in Figure 2. and trace-simulated site in Figure 5.
- Shown the functional relationship (Figure 4.) between trace-driven simulation and the data collection, data processing, and input to the prediction models.

This paper is arranged as follows: Section I briefly introduces LoRa technology, LoRaWAN network, conventional RF propagation models, motivation for performance and evaluation of RF models with trace-driven simulation, and the related work. Section II provides details about measurements and simulation. Section III contains the RF models' performance analysis and discussion. Finally, section IV presents the conclusion and prospective future work.

1.1 Overview of LoRaWAN and LoRa

LoRaWAN is one of the media access control protocol for low-power Wide Area Network (WAN). It is a long-range wireless communication technology that enables low-powered IoT devices and applications to communicate over the Internet [18]. LoRaWAN specification V1.0 [3] published in 2015, provides LoRaWAN details to describe the network communication protocol and architecture. The typical LoRaWAN network architecture is a star topology consisting of the end-devices, applications, gateway, and the server. These devices may be connected through the air or cables. LoRaWAN-enabled devices may be categorized as Class A, B, or C, depending

on the capacity for a trade-off between downlink latency and battery-life saving. Figure 1. describes a typical star-topology of LoRaWAN architecture. While LoRaWAN defines the architecture and system communication protocol, LoRa is a modulation technology defined at the physical layer.

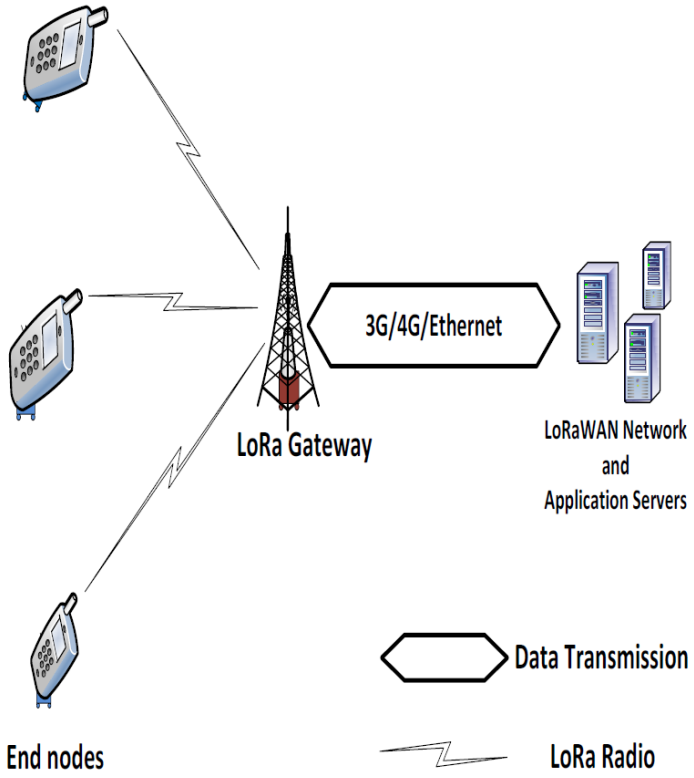


Figure 1: Basic LoRaWAN Network Architecture

LoRa is a spread spectrum technique adapted by Semtech [4] and implemented at the physical layer for digital wireless modulation. It was derived from chirp spread spectrum technology [19] to achieve a long-range radio communication in LoRaWAN networks while retaining the ability to operate at low-power levels. Furthermore, the LoRaWAN network employs different spreading factors (SF), another key technology to allow a trade-off between the coverage and data rates [20]. Higher SF configuration helps the LoRaWAN network to transmit further, but at the cost of low data rates, and vice-versa. Thus, the LoRaWAN network delivers the required connectivity to enable the LoRa end-devices to transmit data over the air channel to the gateway and then to the network and application servers.

1.2 Radio Propagation Models

RF propagation models[21] characterize the propagation of radio waves as function carrier frequency, distance, and other NLOS factors in the air channel environment. These RF models, deterministic, semi-empirical, and empirical, employ several attenuation factors, including LOS, diffraction, multipath, to estimate the RSS. These models perform connectivity analysis in radio networks to fast-track radio coverage information. Radio propagation effects are heavily site-specific and depend on configuration parameters such as terrain cover, operating frequency, distance, transmitter, and receiver antenna height. In this study, commonly used RF models in ICS-Telecom, Deygout 94, ITU-R 525/526, and COST-231 Walfish-Ikegami models are evaluated for RSS prediction accuracy in LoRaWAN network. These models are respectively, deterministic, semi-deterministic, and empirical.

1.3 Deygout 94

Deygout 94 model determines diffraction attenuation for N successive knife-edge or rounded obstacles. The model is based on the idea of the main signal blocking component [22], which exerts considerably high obstruction losses to the link between the transmitting and receiving ends. The overall diffraction loss, L_D, is given as:

$$L_D = \sum_{i=1}^{N} L_1(V_i) \tag{1}$$

where L_1 is the diffraction loss in case of one main obstruction between the transmitter and receiver, and V_i is the diffraction parameter for the ith knife-edge. The total path loss is the sum of Deygout 94 diffraction losses, free space, sub-path attenuation, and 3D losses that are implemented in ICS-Telecom.

1.4 ITU-R 525/526

The ITU-R 525/526 uses ITU-R P 525 [23] to calculate the free space loss and ITU-R P 526 to calculate attenuation due to diffraction [24]. The concept of ITU-R 526 calculation of diffraction losses is similar to Deygout 94. It calculates sub-path and diffraction losses using the Delta Bullington diffraction model [25]. The ITU-R 525 calculates the basic free space transmission loss, P_{A_0} between isotropic antennas, as follows:

$$P_{A_0} = 32.4 + 20log_{10}(d) + 20log_{10}(f) \tag{2}$$

where f is frequency (MHz) and d is the distance (km).

1.5 COST-231 Walfish-Ikegami

COST 231 Walfish-Ikegami (COST-WI) the model [26] is a combination of Walfish and Ikegami models. Unlike the deterministic model that requires digital terrain models, including the buildings layer, COST-WI is an empirical model that requires characteristic values. It improves the path loss prediction by considering more data to characterize medium-sized flat urban environments [27]. Compared to other empirical models, the COST-WI parameters include the heights h_b of the building, the separation between buildings d, the widths of roads w, and the angle θ to the direct radio path. The range of fundamental parameters considered is 800-2000 MHz for frequency, 1-3 m, 4-50 m, 0.02-5 km for distance, and end-device and gateway antenna height, respectively. The model calculates path loss due to the LOS and NLOS. The mathematical formulae for LOS and NLOS are respectively defined in (3) and (4) below.

$$P_{LOS} = 42.64 + 26log_{10}(d) + 20log_{10}(f) \tag{3}$$

If NLOS is the case, then path loss calculation is a combination of attenuation due to free space P_{LOS}, the rooftop to street diffraction, and the scattering L_{rts}, and the multiple screen diffraction loss L_{msd}. The totality of this (NLOS) path loss is described mathematically [27**?**] as follows:

$$P_{NLOS} = P_{LOS} + L_{rts} + L_{msd} \tag{4}$$

The L_{rts}, the diffraction loss from the rooftop to the street is determined as in the following formula:

$$L_{rts} = -16.9 - 10log_{10}(w) + 10log_{10}(f) + 20log_{10}(h_b - h_r) + L_{ori} \tag{5}$$

where, w is width of the roads, h_b and h_m are the height of building and end-device mobile station respectively. The street orientation correction factor, L_{ori}[28] is given as:

$$L_{ori} = \begin{cases} -10 + 0.35\alpha & \text{for } 0° < \alpha < 35° \\ 2.5 + 0.0755(\alpha - 35) & \text{for } 35° < \alpha < 55° \\ 4 - 0.0114(\alpha - 55) & \text{for } 55° < \alpha < 90° \end{cases} \tag{6}$$

where α is the street orientation angle, L_{msd}, the multiscreen loss, represents diffraction loss from multiple obstacles. The following mathematical representation determines it:

$$L_{msd} = L_{bsh} + K_a + K_d log_{10}(d) + K_f log(f) - 9log_{10}(s_b) \tag{7}$$

where the correction factors, L_{bsh} and K_a represent path loss when the gateway is above and below the rooftops, respectively. s_b represents buildings separation. The terms K_d and K_f quantify the diffraction loss as a factor of the distance, d and frequency, f, and are defined in[26] as follows:

$$L_{bsh} = \begin{cases} -181log(1 + h_t - h_b) & h_t > h_b \\ 0 & h_t \leq h_b \end{cases} \tag{8}$$

$$K_a = \begin{cases} 54 & h_t > h_b \\ 54 - 0.8(h_t - h_b) & h_t < h_b \text{ and } d_{km} \geq 0.5km \\ 54 - 1.6(h_t - h_b)d & d_{km} < 0.5km \end{cases} \tag{9}$$

$$K_d = \begin{cases} 18 & h_t > h_b \\ 18 - 15(h_t - h_b)/h_b & h_t \leq h_b \end{cases} \tag{10}$$

$$K_f = -4 + \begin{cases} 0.7(f_{MHz}/925 - 1) & \text{for medium-size city and suburban} \\ 1.5(f_{MHz/925-1}) & \text{for metropolitan centres} \end{cases} \tag{11}$$

1.6 Motivation for Performance and Evaluation of RF Models with Trace-driven Simulation

The evaluation of how the air channel affects the propagation of wireless communication signals is a challenging task. In cities, the physical environment, including buildings, bridges, people's motion, and vehicles, adversely affect propagating RF signals. To accurately evaluate the effects of the physical environment in cities, simulation methods and models have become an integral part of wireless communication systems' planning and deployment. However, it has been challenging to identify which RF propagation models, from their significant number, produce realistic predictions of the RSS. The assumption that simulation models can be applied to modeling propagation effects in different environments, protocols, and applications have often been proved less accurate.

The literature is yet to contain either the RF propagation prediction models or an established understanding of the actual RF propagation prediction models' performance for the LoRaWAN propagation system in cities. Studies that attempted to evaluated LoRAWAN propagation system performance in different urban scenarios used the analysis of the real-world measurements [9]–[13]. However, the use of field measurement to characterize the impact of the air channel on RF propagation is usually costly as it involves real systems. In wireless communication networks, taking real-world measurements is usually the last phase of network planning and deployment, aiming to optimize a wireless network system performance. Instead, the network planning tools use propagation prediction models to estimate path loss due to air channel conditions. In our previous work [14], we evaluated conventional empirical propagation models' performance to determine applicability for predicting LoRaWAN RSS in an urban scenario. Although the study showed that COST-WI had better prediction performance for LoRaWAN RSS than other models considered for the study, this result was over-predicted by 6.48 dBm. Therefore, this indicates that empirical RF models lack accuracy in attempts to capture paths loss over the rooftops, diffraction, and reflections arising from the buildings and streets set up, and these factors are better handled with deterministic models.

In this paper, we intend to evaluate Electromagnetic Wave (EM) based deterministic models to understand the applicability of these models to predict LoRaWAN RSS in urban areas. The EM deterministic models' ability to accurately consider the phenomena of path loss over the rooftops, diffraction, reflections from buildings, dispersion, and wave running are essential features for dense urban outdoor RF propagation [29]. Deterministic models require the use of accurate and high-quality terrain models of the environment. However, in the presence of low-quality terrain models, the use of trace-driven simulation for evaluation of deterministic and semi-deterministic models produced more accurate LoRaWAN RSS prediction than empirical RF models. This study can give insight into the effectiveness of trace-simulation with standard propagation models for evaluation of IoT connectivity with LoRaWAN networks at 868 MHz in the NLOS urban environment.

1.7 Related Work

Trace-driven simulation is a technique that utilizes both measurements and simulation for evaluation and prediction of the performance of the computer-based systems [30]. All the trace-driven studies reviewed in this paper used a software probe to collect measurement trace data except [31], which used device-specific data collection from the testbed. Trace-driven simulation technique has been introduced in [32] as a technique that uses analysis of measurement trace data to redo similar experimental circumstances in the simulations. Trace-driven simulation has been used to perform a realistic simulation.

Trace-driven simulation use analysis of the experimental trace data in an attempt to recreate simulation of the experimental conditions. Authors in [33] stated that sufficient and accurate measurement datasets could directly serve as a basis of propagation performance prediction or fitting the RF models. Trace-driven simulation was used in [34] to model the error behavior of the WaveLAN wireless channel. Chia et al [35] proposed an Automata-Synchronized Replay (EAR) system software to perform trace-based simulation of the measured packets received over the Wireless Local Area Network. The EAR was used to control trace packet, synchronize the environmental effects in trace packet with the packets, and other signals received from the real-world environment. The system implementation results show that the EAR achieved an event reproduction rate of 92.45% under a fading environment. Trace-driven simulation can be used to evaluate wireless link models.

In [36], the authors presented a Wireless Link Simulator (WiLinkSim). This simulator implemented models based on mathematical distribution to correctly represent wireless errors (bit error rate) due to the air channel's effects. WiLinkSim utilizes trace-driven models for realistic simulation of Rayleigh and multipath fading and during the deployment of a real-world network. G. Judd and P. Steenkist [37] indicated that trace-driven simulation is one of the accurate methods to model RF signal propagation. Authors in [38] used measurement error traces to configure stochastic error models to reproduce the signal-to-noise-plus-interference ratio errors of the received wireless signal. The accuracy for Bernoulli, Gilbert-Elliot, and chaotic map error models' ability to reproduce or estimate the measured dataset was performed by comparing the measured traces with models' simulated traces. The results showed that Gilbert-Elliot and chaotic map error models attained enough accurate reproduction of the measured traces.

2 Measurement Setup and Simulation

In the first step, the measurement trace dataset was collected, and then, the RF propagation prediction models were trace simulated using the measurement processed trace data. The measurement trace dataset was collected from an experimental LoRaWAN network installed in Glasgow City using LoRaWAN end-devices and Servers. The processed trace data needed for trace-simulation was selected from the measured data. This trace data served as input data to the trace simulation program. Below is the detailed description of data collection, data processing, and trace-simulation setup.

Table 1: Measurement and Trace-simulation Parameters.

S/N	Parameters	Values
1	Operating frequency Band	868 *MHz*
2	Bandwidth	125 kHz
3	End-device transmit power	14 *dBm*
4	Number of Gateway	3 m
5	Gateway antenna heights	30, 27 and 27 *m*
6	End-device antenna height	1.5 *m*
7	Measurement spreading factor (SF)	7-12
8	Trace-simulation spreading factor (SF)	7, 8 and 12
9	Maximum transmit and receive distance	2275 *m*

2.1 Data Collection Setup

Capturing the field measurements is essential in characterizing air channel impact on LoRaWAN system applications under real-world conditions. We performed an experimental field study to acquire real-world LoRaWAN measurements from Glasgow city, Scotland, the UK. On the map, the city's latitude and longitude are 55.8642° N and 4.2518° W, respectively. The city is built at the bank of River Clyde and constitutes tall and huge buildings and open spaces. Data collection was performed in this city's environment to characterize the effects the NLOS has on the air channel for RF propagation of the LoRaWAN network. The leading equipment used to collect data is the LoRaWAN end-device and three gateways. The former was a Multitech mDot module [39] with LoRa and LoRaWAN protocol, regulated by a single board Raspberry Pi computer, whereas the later constituted Kerlink gateways, enabled with Semtech LoRa SX1301 chip [40]. The Raspberry Pi computer was connected to a 3G/GPS module to keep location records for the received from transmitting end-device operated at 868 *MHz* and 14 *dBm*.

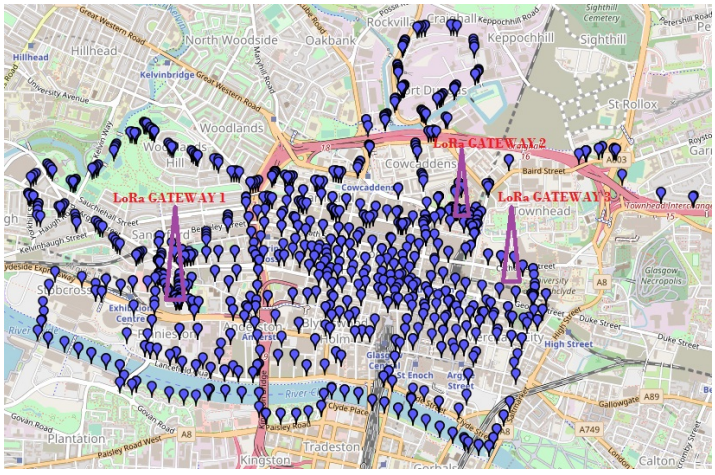


Figure 2: A Google map showing the location of the three LoRa Gateways installed in Glasgow city centre and measurements waypoints. The blue markers on the map represent GPS and their corresponding LoRaWAN RSS waypoints obtained during the measurements

The gateways operated at three locations. The first gateway at Glasgow Caledonian University, 30 *m* on top of George More building. The second gateway at Strathclyde University, 27 *m* on top of James Weir building, and the third gateway placed at 27 *m* on top of

Skypark. The distance between gateway at Skypark and James Weir building is 2005 *m* apart, and both buildings are respectively, 590 *m* and 1900 *m* from George More building. Figure 2. is the display of a 1:300 Google map showing the topography of Glasgow city centre used for the LoRaWAN RSS measurements. It indicates the location of the three LoRa Gateways and measurement waypoints. The three gateways utilize the cable and mobile networks to connect to the network server to record packets received from the transmitting end-device. The mobile end-device transmissions operated the 868 *MHz* and 14 *dBm*. Table 1. shows details of LoRaWAN transmission parameters. The packets were sequentially transmitted, and success or failure to receive an ACK message was recorded. These packets altered between 7-12 the spreading factor. We collected the data in a backpack logger unit at the walking speed in Glasgow city. The data collected in the network server and backpack logger unit were combined and processed in a database. This data shall be known as measurement trace data.

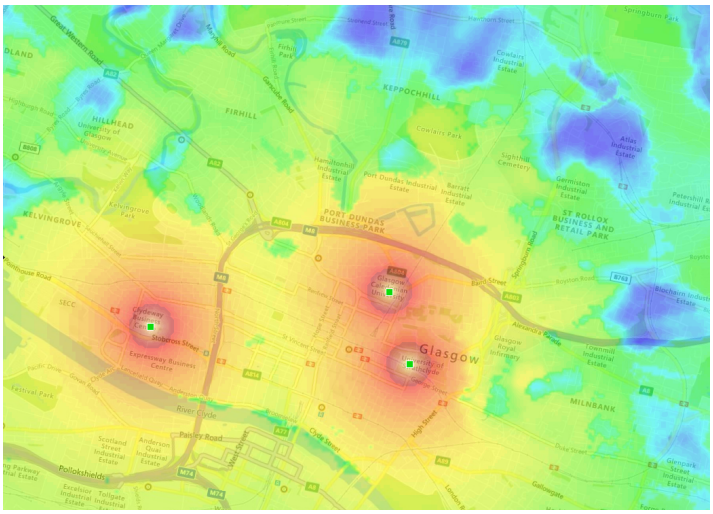


Figure 3: LoRaWAN 868 *MHz* simulation without traces over a Bing digital terrain model of Glasgow city centre. The three green squares in a reddish/brown color represent the simulated LoRa gateways. Colors represent the variation of signal levels, with blue indicating the weakest and reddish the strongest signals.

2.2 Data Processing and Input to Models

To prepare LoRaWAN system measurements for a trace simulation program, we processed measurement trace data. Measurement trace data contained feature data and metadata not required by the trace simulation program. A total of 5008 measurement trace data collected in the backpack logger unit and network server was loaded into the database. We used the command line manipulation commands to pare measurement trace data down to locations coordinates and the RSS suitable trace simulation. Consequently, manipulation of measurement raw data resulted in 5008 datasets for latitude, longitude, and corresponding RSS required for the ICS-Telecom trace simulation function.

The trace simulation program is built from LoRaWAN and RF propagation system models. LoRaWAN system configuration parameters in Table I. and RF propagation models are introduced in section I. The LoRaWAN parameters in this table are based on the LoRaWAN band regulation in the European Union and the physical position of gateway and end-device antenna heights used during the measurements. Similarly, the maximum transmission is the distance between the transmission and reception attained during the measurements. However, due to lack of mobility during the simulation, the SF was heuristically fixed at 7, 8, and 12. The trace simulation program imported the processed trace data through the measurement import function to perform a trace-driven simulation. It executes the LoRaWAN and RF propagation environment using the processed trace data. The trace-driven simulation program's output, which is of interest to this work, was the trace-driven RSS. This output is compared with the RSS in the processed trace data for performance analysis. Figure 4. shows a block diagram of the trace-driven simulation model. This model's main parts are; data collection, data processing, and LoRaWAN and RF propagation conceptual models.

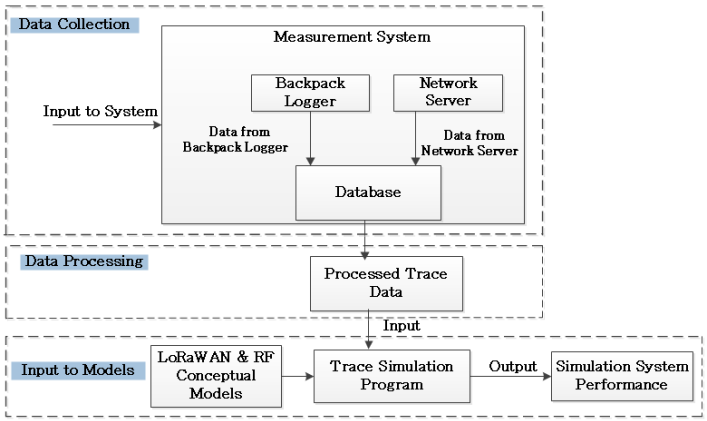


Figure 4: Block Diagram of Trace-driven simulation

2.3 Trace-Driven Simulation

The trace-driven simulation requires a combination of system measurements and simulation [41]. In this paper, trace-simulation set up involved importing processed trace data and configuration of the Glasgow city centre digital terrain model. When the trace simulation program starts, it loads the processed measurement trace data and the LoRaWAN RF propagation system model to reproduce the same experimental conditions in the ICS-Telecom. Authors in [**?**] showed that trace-simulation produces more realistic simulation results than conventional simulation. In [42], the authors provided a detailed account of the advantages of trace-simulation over the conventional simulation. To simulate Glasgow city centre terrain, we imported and configured a 1:300 Bing terrain model of Glasgow city centre at 25 *m* resolution into ICS-Telecom. However, low-resolution terrain maps such as 25 *m* provided a digital elevation model and the city's clutter to the simulator, leaving out the building layer. This input facilitated the simulation of the LoRaWAN system in Glasgow city centre. Over the terrain map, we placed three system gateway models. Each gateway model location configuration on the terrain map matched the latitude and longitude of system gateways used during the LoRaWAN measurements. Like the measurement system, the distance between LoRaWAN gateway models at Strathclyde University and Skypark on the terrain map represents 2005 *m* apart. The respective distance between these

gateway models from Glasgow Caledonian University represents 590 m and 1900 m on the ground. Figure 3. shows the LoRaWAN 868 MHz simulation system set up without traces over a Bing digital terrain model of Glasgow city centre. The next step was to load the processed trace data for trace-driven simulation of the LoRaWAN 868 MHz system.

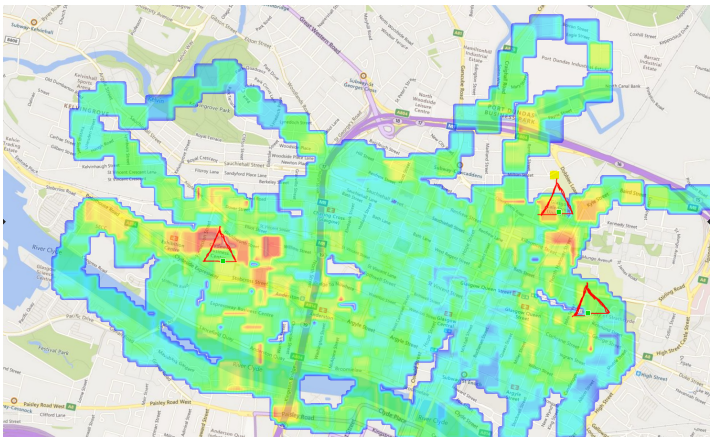


Figure 5: A trace-driven simulation of LoRaWAN 868 MHz over a Bing digital terrain model of Glasgow city centre. The three green squares at the base of red triangles represent the simulated LoRa gateways. Colors represent trace-driven RSS data. The blue color refers to the weakest signals, and reddish/brown refers to the strongest signals. The trace-driven data follows waypoints of the real-world measurements.

Through the measurement import function, the trace simulation program imported the processed trace data. Presented to the digital terrain map are LoRaWAN system measurement locations, a set of latitude and longitude coordinates. Figure 5. shows a trace-driven simulation of LoRaWAN 868 MHz over a 1:300 Bing digital terrain model of Glasgow city centre. It presents the trace-simulated RSS in the previously identified location with the processed trace data, which contains latitude, longitude, and corresponding RSS. At the base of the trace simulation program, there are LoRaWAN and RF propagation system models. These models are the designed base of the trace-driven simulation program. LoRaWAN and RF conceptual system models comprise LoRa modulation, LoRaWAN network, RF propagation models, and the terrain model. The configuration of the LoRaWAN simulation system parameters is similar to LoRaWAN system parameters shown in Table I. On the other hand, the configured RF propagation models are described in Section I. After the trace-simulation, extraction of the predicted trace-driven LoRaWAN RSS followed for each of the RF models.

The extraction of trace-simulated LoRaWAN RSS from the LoRaWAN simulation system provided a trace-simulated dataset required for prediction performance analysis. The minimum trace signal level was set to -140 dBm and the maximum trace signal level set to -61 dBm. The total number of trace data was set to 5008. The trace-simulated LoRaWAN RSS are spread over the digital terrain map of Glasgow city centre in Figure 5. and the color code shows the varying signal strength, which depends on different factors, including clutter obstructions in the environment. The predicted trace-driven dataset for each propagation model was extracted from the LoRaWAN simulation system digital terrain map. Alternatively, the predicted trace-driven dataset can be exported from the simulation system. To validate the obtained trace-simulated dataset, we compared the RF propagation models' trace-simulated LoRaWAN RSS with the LoRaWAN system RSS, measured from Glasgow city centre.

3 Performance Analysis and Discussion

To evaluate models' prediction performance accuracy, we validated LoRaWAN trace-simulation with measurement trace data. Common mathematical error methods measured the prediction accuracy of the models. The error performance metrics used to represent the error between measurement and simulation are the *Mean Square Error (MSE)*, Δ'_{yi}, *Mean Absolute Error (MAE)*, $|\Delta y|$, and *Standard Deviation (SD)*, σ_e. In this paper, N indicates the total number of measured or simulated data samples, whereas Δ_{yi} denotes the difference between predicted and measured datasets. While *MAE* is used as a measure of RF models prediction error, *SD* measures the average distance between measurement and the predicted dataset. C. J. Willmott and K. Matsuura [43] examined the ability of *MAE* and *Root-Mean-Square Error (RMSE)* to represent the average error performance of a model. The authors concluded that *MAE* is a preferred measure of one dimension average error performance compared to *RMSE*, which is suitable for more than one error performance of a model. However, in [44], authors argued that both *MAE* and *RMSE* may be used to evaluate the average error prediction accuracy of the models. In this study, the *MAE* is considered to be an intuitive measure of average error prediction accuracy in RF models.

$$\Delta_{yi} = (RSS_{predicted} - RSS_{measured}) \tag{12}$$

$$\Delta'_{yi} = \frac{1}{N}\sum_{i=1}^{N}(RSS_{predicted} - RSS_{measured})^2 \tag{13}$$

$$|\Delta y| = \frac{1}{N}\sum_{i=1}^{N}|\Delta_{yi}| \tag{14}$$

$$\sigma_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta_{yi} - |\Delta_y|)^2} \tag{15}$$

Table 2: Error Performance Metrics

Parameters	Deygout94	ITU-R525/526	COST-WI
MSE	18.7812	20.978	24.2155
MAE	0.83	1.01	2.87
SD	4.17	5.84	10.96

The comparative analysis of RF models prediction accuracy showed that all the three investigated models over-estimated the RSS measurements. However, Deygout 94 registered higher prediction accuracy with *MAE* at 0.83 dBm. The *MAE* for ITU-R 525/526 and COST-WI was 1.01 dBm and 2.87 dBm, respectively. In terms of prediction error, Deygout 94 prediction had the lowest *Mean Square Error, MSE* at 18.7812 dBm, followed by the ITU-R 525/526 and COST-WI with *MSE* at 20.978 dBm and 24.2155

dBm respectively. Furthermore, Deygout 94 exhibited the lowest *standard deviation,* σ_e at 4.17 *dBm*, followed by the ITU-R 525/526 and COST-WI at 5.84 *dBm* and 10.96 *dBm* respectively. This result implies that Deygout 94, a deterministic model, performs better RSS predictions than the ITU-R 525/526 and COST-WI models. COST-WI poor prediction performance is attributed to its reliance on the buildings' information. Table II. shows models' prediction error performance metrics calculated in equations 12-15 for MSE, MAE and σ_e using measured and trace-driven dataset in the NLOS city environment. Overall, it can be observed that Deygout 94 obtained low prediction error performance.

Density and vertical obstructions presented to the application through a high-resolution DTM is crucial for modeling RF propagation [45]. The DTM at 25*m* resolution that presented the NLOS obstructions over Glasgow city centre terrain landscape into the ICS-Telecom lacked detailed buildings' information. The graphical representation in Figure 6. shows the relationship between the LoRaWAN 868 MHz measurement dataset and models trace-simulation performance in the NLOS conditions in Glasgow city centre. The measurement and trace-driven simulated data are plotted as a RSS function, in *dBm* and distance, d in meters. The visible gap between measurements and trace-driven simulation may be attributed to the low resolution of the DTM, which fails to represent actual NLOS conditions in Glasgow city centre to the Trace Simulation Program. However, trace-driven RSS trend against the distance follows RSS measurements with Walfish-Ikegami keeping more distance and Deygout 94, making a closer drift with measurements. The clustered data visible on the graph indicates many RSS collected at specific locations due to reduced NLOS obstructions. On the other hand, the flattened regions are an indicator of either low variations of LoRaWAN RSS as signal reception reduce or completely blocked due to an increase in density and the height of buildings in the City's air channel.

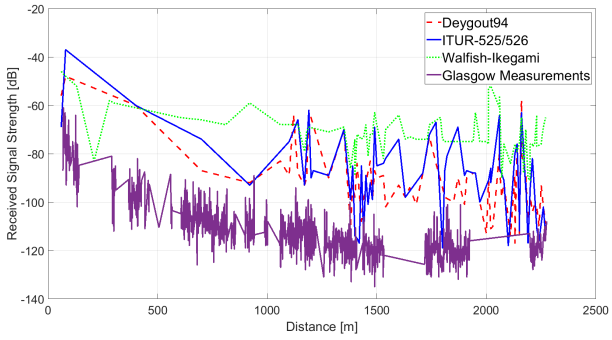


Figure 6: Models trace-simulated performance comparison with LoRaWAN 868 MHz measurements. The vertical axis plots real-world LoRaWAN RSS measurements, *dBm* and models' trace-simulated LoRaWAN RSS data, *dBm*. The horizontal axis represents the distance, d in m.

The trace-driven simulated models improved prediction accuracy performance compared to empirical models in the previous study [14]. Both studies were carried out for LoRaWAN 868 MHz in Glasgow city centre, and the same measurement dataset was used for validation. Table III. shows the comparison of prediction accuracy of empirical models, in cyan color heading and trace-simulated models, in the yellow color heading. The improvement in trace-simulated predictions compared to empirical RF models can be attributed to measurement traces and differences in model design and application. The spreading of LoRaWAN RSS over the Glasgow city centre map shows how trace-driven simulation leads closer to measurements.

Table 3: Comparison of Emprical and Trace-simulated Models Prediction Accuracy

Parameters	*Deygout94*	*ITU-R525/526*	*COST-WI-Traced*	**Okumura-Hata**	**COST231**	**COST-WI**
MSE	18.7812	20.978	24.215	48.84021	57.2056	96.746
MAE	0.83	1.01	2.87	5.564	6.131	7.413
SD	*4.17*	*5.84*	10.96	9.158	11.425	7.454

The analysis of measurements and simulation mapping over Glasgow city centre's map shows significant variation in the RSS spreading. The simulation mapping of RSS in Figure 3. spreads all over the map compared to the mapping of RSS measurement, which follows the data collection path. On the other hand, LoRaWAN trace-driven simulation mapping in Figure 5. attains a close match to the mapping of RSS measurements in Figure 2. when the minimum and maximum trace signal level and the total number of trace dataset is set. The closer pattern between the measured and trace-driven RSS mapping matches the claims that trace-driven simulation performs realistic simulation [32]. This pattern shows that the trace-driven simulation technique can perform a realistic evaluation of RF models for the planning and deployment of LoRaWAN in cities.

4 Conclusion and Future work

We have created a trace-driven simulation of LoRaWAN 868 MHz for Glasgow city centre. Trace-driven simulation techniques using the real-world measurement trace dataset make it possible for RF propagation models to produce realistic RSS predictions of LoRaWAN radio links. This paper has used measurements to perform trace-driven simulation over the DTM of Glasgow city centre. We have used LoRaWAN trace-driven RF propagation models to analyse and evaluate several RF propagation models' prediction performance accuracy for LoRaWAN 868 MHz. The RF models considered in this study are Deygout 94, ITU-R 525/526, and COST-WI. The trace simulation program, built in the ICS-Telecom, imported 5017 processed trace datasets measured from an experimental LoRaWAN 868 MHz network deployed in Glasgow city centre. The processed trace dataset contained the latitude, longitude, and the corresponding measurement RSS. The comparative performance analysis of trace-driven LoRaWAN simulation traces and the real-world measured LoRaWAN processed trace dataset showed that RF propagation models produce realistic LoRaWAN RSS predictions.

All models over-estimated LoRaWAN trace-simulated RSS levels in comparison to real-world collected data samples. However, a comparison of trace-driven simulated models and empirical RF models shows that trace-simulated RF models produce a more accurate prediction. Deygout 94 prediction accuracy was the higher with *MAE* at 0.83 *dBm* and *SD* at 4.17 *dBm*. COST-WI cannot

be used for simulation of LoRaWAN coverage estimation in an urban environment unless high-resolution DTM models are used in ICS-Telecom. Currently, most studies use measurements to evaluate LoRaWAN performance. This approach is expensive, and trace-driven simulation can be a better option. Although it requires DTM, the use of even low-resolution DTM produces results better than conventional models. The random neural network could be used to design a more accurate and inexpensive RF propagation model in the future.

References

[1] E. Harinda, H. Larijani, R. M. Gibson, "Trace-Driven Simulation for LoRaWan868 MHz Propagation in an Urban Scenario," in 2019 UK/ China Emerging Technologies (UCET), 1–5, 2019, doi:10.1109/UCET.2019.8881854.

[2] R. Taylor, D. Baron, D. Schmidt, "The world in 2025 - predictions for the next ten years," in 2015 10th International Microsystems, Packaging, Assembly and Circuits Technology Conference (IMPACT), 192–195, 2015, doi:10.1109/IMPACT.2015.7365193.

[3] T. E. T. K. I. O. N. Sornin (Semtech), M. Luis (Semtech), "LoRaWAN™ Specification," LoRaWAN™ Specification 1.1 — LoRa Alliance™, **2015**, 2015.

[4] LoRa®Alliance, "A technical overview of LoRa® and LoRaWAN™," https://lora-alliance.org/sites/default/files/2018-04/what-is-lorawan.pdf/, 2015, [Online; accessed 25-September-2018].

[5] V. S. Abhayawardhana, I. J. Wassell, D. Crosby, M. P. Sellars, M. G. Brown, "Comparison of empirical propagation path loss models for fixed wireless access systems," in 2005 IEEE 61st Vehicular Technology Conference, **1**, 73–77 2005, doi:10.1109/VETECS.2005.1543252.

[6] E. Greenberg, E. Klodzh, "Comparison of deterministic, empirical and physical propagation models in urban environments," in 2015 IEEE International Conference on Microwaves, Communications, Antennas and Electronic Systems (COMCAS), 1–5, 2015, doi:10.1109/COMCAS.2015.7360394.

[7] C. Braithwaite, M. Scott, UMTS network planning and development: design and implementation of the 3G CDMA infrastructure, Elsevier, 2003.

[8] O. Fratu, A. Martian, R. Craciunescu, A. Vulpe, S. Halunga, Z. Zaharis, P. Lazaridis, S. Kasampalis, "Comparative study of Radio Mobile and ICS Telecom propagation prediction models for DVB-T," in 2015 IEEE Interna-tional Symposium on Broadband Multimedia Systems and Broadcasting, 1–6, 2015, doi:10.1109/BMSB.2015.7177260.

[9] R. Sanchez-Iborra, J. Sanchez-Gomez, J. Ballesta-Viñas, M.-D. Cano, A. Skarmeta, "Performance Evaluation of LoRa Considering Scenario Conditions," Sensors, **18**(3), 772, 2018, doi:10.3390/s18030772.

[10] A. Augustin, J. Yi, T. Clausen, W. Townsley, "A Study of LoRa: Long Range & Low Power Networks for the Internet of Things," Sensors, **16**(9), 2016, doi:10.3390/s16091466.

[11] L. Feltrin, C. Buratti, E. Vinciarelli, R. De Bonis, R. Verdone, "LoRaWAN: Evaluation of Link- and System-Level Performance," IEEE Internet of Things Journal, **5**(3), 2249–2258, 2018, doi:10.1109/JIOT.2018.2828867.

[12] A. J. Wixted, P. Kinnaird, H. Larijani, A. Tait, A. Ahmadinia, N. Strachan, "Evaluation of LoRa and LoRaWAN for wireless sensor networks," in 2016 IEEE SENSORS, 1–3, 2016, doi:10.1109/ICSENS.2016.7808712.

[13] J. Petajajarvi, K. Mikhaylov, A. Roivainen, T. Hanninen, M. Pettissalo, "On the coverage of LPWANs: range evaluation and channel attenuation model for LoRa technology," in 2015 14th International Conference on ITS Telecommunications (ITST), 55–59, 2015, doi:10.1109/ITST.2015.7377400.

[14] E. Harinda, S. Hosseinzadeh, H. Larijani, R. M. Gibson, "Comparative Performance Analysis of Empirical Propagation Models for LoRaWAN 868MHz in an Urban Scenario," in 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), 154–159, 2019, doi:10.1109/WF-IoT.2019.8767245.

[15] J. W. C. Fu, J. H. Patel, "Trace driven simulation using sampled traces," in 1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, **1**, 211–220, 1994, doi:10.1109/HICSS.1994.323170.

[16] I. Telecom, "ATDI," 2014.

[17] S. W. Sherman, J. C. Browne, "Trace Driven Modeling: Review and Overview," in Proceedings of the 1st Symposium on Simulation of Computer Systems, ANSS '73, 200–207, IEEE Press, 1973, doi:10.1109/HICSS.1994.323170.

[18] D. Singh, O. G. Aliu, M. Kretschmer, "LoRa WanEvaluation for IoT Communications," in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 163–171, 2018, doi:10.1109/ICACCI.2018.8554713.

[19] B. Reynders, S. Pollin, "Chirp spread spectrum as a modulation technique for long range communication," in 2016 Symposium on Communications and Vehicular Technologies (SCVT), 1–5, 2016, doi:10.1109/SCVT.2016.7797659.

[20] F. Van den Abeele, J. Haxhibeqiri, I. Moerman, J. Hoebeke, "Scalability Analysis of Large-Scale LoRaWAN Networks in ns-3," IEEE Internet of Things Journal, **4**(6), 2186–2198, 2017, doi:10.1109/JIOT.2017.2768498.

[21] S. Faruque, Radio frequency propagation made easy, Springer, 2014.

[22] J. Deygout, "Multiple knife-edge diffraction of microwaves," IEEE Transactions on Antennas and Propagation, **14**(4), 480–489, 1966, doi:10.1109/TAP.1966.1138719.

[23] R. ITU, "ITU-R P. 525-3," Calculation of Free-Space Attenuation, ITU-R Recommendation, **2016**.

[24] P.-. ITU-R, "A path-specific propagation prediction method for point-to-area terrestrial services in the VHF and UHF bands," 2014.

[25] C. Lee, S. Park, "Diffraction Loss Prediction of Multiple Edges Using Bullington Method with Neural Network in Mountainous Areas," International Journal of Antennas and Propagation, **2018**, 2018, doi:10.1155/2018/8737594.

[26] D. Har, A. M. Watson, A. G. Chadney, "Comment on diffraction loss of rooftop-to-street in COST 231-Walfisch-Ikegami model," IEEE Transactions on Vehicular Technology, **48**(5), 1451–1452, 1999, doi:10.1109/25.790519.

[27] J. Milanović, G. Šimac, K. Mazor, "Performance analysis of empirical propagation model for long term evaluation (LTE) network," in 2016 International Conference on Smart Systems and Technologies (SST), 69–74, 2016, doi:10.1109/SST.2016.7765635.

[28] P. Frangos, D. Jaggard, S. Sautbekov, G. Georgiev, S. Savov, "Propagation of electromagnetic waves in terrestrial environment for applications in wireless telecommunications," 2014, doi:10.1155/2014/297082.

[29] W. J. Krzysztofik, "Radio Network Planning and Propagation Models for Urban and Indoor Wireless Communication Networks," 2018.

[30] H. Lucas Jr, "Performance evaluation and monitoring," ACM Computing Surveys (CSUR), **3**(3), 79–91, 1971, doi:https://doi.org/10.1145/356589.356590.

[31] G. T. Nguyen, R. H. Katz, B. Noble, M. Satyanarayanan, "A trace-based approach for modeling wireless channel behavior," in Proceedings of the 28th conference on Winter simulation, 597–604, 1996, doi:10.1145/256562.256764.

[32] P. Owezarski, N. Larrieu, "A trace based method for realistic simulation," in 2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577), 4, 2236–2239 Vol.4, 2004, doi:10.1109/ICC.2004.1312915.

[33] G. Hampel, D. Abush-Magder, A. Diaz, L. M. Drabeck, M. J. Flanagan, J. M. Graybeal, J. D. Hobby, M. MacDonald, P. A. Polakos, J. Srinivasan, H. Trickey, L. Zhang, G. Rittenhouse, "The new paradigm for wireless network optimization: a synergy of automated processes and human intervention," IEEE Communications Magazine, **43**(3), S14–S21, 2005, doi:10.1109/MCOM.2005.1404593.

[34] G. T. Nguyen, B. Noble, R. H. Katz, M. Satyanarayanan, "A trace-based approach for modeling wireless channel behavior," in Proceedings Winter Simulation Conference, 597–604, 1996, doi:10.1145/256562.256764.

[35] C. Ku, Y. Lin, Y. Lai, P. Li, K. C. Lin, "Real traffic replay over WLAN with envi-ronment emulation," in 2012 IEEE Wireless Communications and Networking Conference (WCNC), 2406–2411, 2012, doi:10.1109/WCNC.2012.6214199.

[36] H. Bai, M. Atiquzzaman, "Error modeling schemes for fading channels in wireless communications: A survey," IEEE Communications Surveys Tutorials, **5**(2), 2–9, 2003, doi:10.1109/COMST.2003.5341334.

[37] G. Judd, P. Steenkiste, "Using Emulation to Understand and Improve Wireless Networks and Applications," NSDI'05, 203–216, USENIX Association, USA, 2005, doi:10.5555/1251203.1251218.

[38] M. Wellens, M. Petrova, J. Riihijarvi, P. Mahonen, "Building a better wireless mousetrap: need for more realism in simulations," in Second Annual Conference on Wireless On-demand Network Systems and Services, 150–157, 2005, doi:10.1109/WONS.2005.6.

[39] MultiTech, "Long Range LoRa® Modules (MTDOT Series)," https://www.multitech.com/brands/multiconnect-mdot, 2019, [Online; accessed 3-November 2020].

[40] LoRa®Alliance, "Wireless and sensing products datasheet," https://www.semtech.com/uploads/documents/sx1301.pdf, 2017, [Online; accessed 25-September-2019].

[41] S. W. Sherman, "Trace driven modeling: An update," ACM SIGSIM Simulation Digest, **7**(4), 87–91, 1976, doi:10.1145/1013610.807302.

[42] S. W. Sherman, J. Browne, "Trace driven modeling: Review and overview," in Proceedings of the 1st symposium on Simulation of computer systems, 200–207, 1973, doi:10.5555/800241.807240.

[43] C. J. Willmott, K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Climate research, **30**(1), 79–82, 2005, doi:0.3354/cr030079.

[44] T. Chai, R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," Geoscientific model development, **7**(3), 1247–1250, 2014, doi:https://doi.org/10.5194/gmd-7-1247-2014.

[45] R. J. Peckham, J. Gyozo, Digital terrain modelling, Springer, 2007.

Acronyms

IoT	Internet of Things
LPWAN	Lower Power Wide Area Network
RF	Radio Frequency
ITU-R	International Telecommunication Union
COST-WI	COST (European Cooperation in Science and Technology - Walfish Ikegami
LoRaWAN	Long-range, Low-Power Wide Area Network
ITU-R	International Telecommunication Union
LOS	Line-of-Sight
NLOS	Non Line-of-Sight
RSS	Received Signal Strength
ICS-Telecom	Image Cartography System - Telecommunication
SF	Spreading Factor
MAE	Mean Absolute Error
SD	Standard Deviation
RMSE	Root Mean Square Error
EM	Electromagnetic Waves
EAR	Automata- Synchnronised Display
WiLinkSim	Wireless Link Simulator
DTM	Digital Terrain Model

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 221-227 (2020)
www.astesj.com *Special Issue on Multidisciplinary Innovation in Engineering Science & Technology*

ASTES Journal
ISSN: 2415-6698

Multi-Directional Light Sensing Using A Rotating Sensor

Hoang Anh Dung[*,1], Nguyen Manh Cuong[2], Nguyen Phan Kien[3]

[1]*Faculty of Electronics Technology and Communication, Hanoi Open University, Hanoi, 100000, Vietnam*

[2]*Military Technical Academy, Hanoi, 100000, Vietnam*

[3]*School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, 100000, Vietnam*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 22 October, 2020
Online: 08 November, 2020:

Keywords:
Light Intensity Sensors
Smart Sensor
The Internet of Things
Intelligent Lighting Control
Control and Automation

ABSTRACT

The performance of indoor illumination control in many applications, such as in an intelligent building, relies on the quality of the light sensors. In many cases, the light level is not uniform and depends on the direction of the illumination source. It usually requires multiple sensors set up in different directions to gather the overall light level. We propose a system that can provide multi-directional light sensing data by rotating a single sensor. This approach overcomes the problem of static sensors network by dynamically changing the measuring angular of the light sensor. We present a sensor system prototype using the ESP8266 controller board, BH1750 light sensor, stepper motor, and 3D printed rotation base mechanism. The system can calculate the sensing angle and transmit sensing data to the monitoring unit or Internet of Things platforms for visualization and analysis. The testing results in normal workrooms show that the rotating sensor can measure the light level in different directions and detect the direction of the main illumination source. Even blocking some directions, the sensor still is able to accurately measure and provide sensing information on the remaining directions. Our sensor system is useful in both whole lighting and local lighting control applications.

1 Introduction

Light sensing information has a vital role in illumination control applications as in industrial lighting or a sensor-actuator (S-A) building. The characteristics and capability of light sensors highly affect lighting control performance. The indoor illuminance depends on both artificial and natural light. The light measuring is not just based on the photodetector in the sensor, but also other factors like sensor direction and occupant activities. Typically, the S-A system uses multiple light sensors or a network of sensors that are located distributively. As mentioned in [1], there are challenges in the use of a network that contained a large number of sensors such as unreliable sensor nodes (fault tolerance), difficult to add a new sensor or to change sensors direction (scalability), and late respond of light changing events (low latency). For a single sensing node, the sensor could be blocked by unaware objects or illuminated by an offensive light beam (not the lighting source), causing an inaccurate measurement. We proposed in [1] a rotating light sensor system that helps to reduce these issues. This paper presents a new design of the system with more details of the evaluation results.

The system proposed in [1] used a single light sensor rather than multiple sensors, and the sensor can rotate freely to measure the light illuminance from numerous directions. The sensor also recognizes the measuring angle as a part of the sensing data. In [1], we introduced a demonstration version of the sensor system that had a relatively large size. We present here a new version with a reduction in the overall size. For validation purposes, we add a static sensor to the system. The significant advantage of our system is that it is able to provide light sensing data at different angles using just a single sensor compared to a system that uses multiple sensors. The rotating sensor can detect the temporary light changing at a specific angle or determine which measuring direction may have been blocked. These features allow the sensor to accurately tracking the light information in real indoor applications. The system can also work on the Internet of Things (IoT) [2] applications since it can connect and transfer data to the popular IoT platforms.

The paper is organized as follows. The following section provides the related works. The methodology, design, and features of the rotating sensor system are presented in Section 3. The performance analysis and experimental results are provided in Section 4. The last section concludes the paper.

[*]Corresponding Author: Hoang Anh Dung, Email:dungha@hou.edu.vn

2 Related Work

Light sensing data is used in the illumination control system is often come from either a single sensor or a network of sensors. The use of light sensors networks has become more popular, especially in medium and larger size applications. Using a single sensor may not provide enough sensing information, while a network of sensors has the challenges of fault tolerance [3], scalability [4], and low latency [5]. We have observations about these challenges based on our work in light sensing and lighting control research. Firstly, the fault tolerance feature in the light sensors network requires the system to work correctly when some sensors may have been failed. It is difficult to satisfy this requirement when the light sensors are distributed at a different location. For the scalability, there is a challenge in the application that requires to increase illumination areas such as in the commercial center or industrial working floors. In such a case, the component cost would increase, and the sensor system should be portable enough to adapt to the extensions. Lastly, the static sensor system may have a low latency issue system when it places the sensor in a fixed direction and orientation. The static system may not quickly sense the changing of illumination level in the zones uncovered by the sensor. In [6], the lighting control system used the wireless sensor networks data to control the illumination to reduce the energy costs. A significant deployment and maintenance costs of the sensors was reported. In addition to the components cost, there was another complexity in the controlling system related to the variation of the light level caused by the sunlight [6]. The study in [7] indicated the effect of other natural factors like sun movement, clouds, and shadows on changing the light levels in the testing applications. The effect of distance and direction of the sensors and light sources on the sensing data was studied in [8]. They studied the controlling of both overall and local illumination. The controlling plan uses the local light sensing data to provide illumination upon the user's needs. A limitation is that the system requires a light sensor to be carried with the user to measure the local light level. These local measurement data were used in the illumination controlling plan to meet the user requirement.

For the location-based illumination control, the whole lighting provides background light, and the local lighting provides concentration light [9]. In this system, the control system uses light sensing data to control both background and concentration illumination. The work in [10] had evaluated the effect of the local light sensing on controlling the whole lighting devices. It also requires to determine the location zone of the user to provide the personalized light control [11, 12].

In [13], the author studied the distributed light control system to accommodate both the user occupancy and the natural daylight level. A large number of light sensors were used in their system to examine the relationship between the occupation bases and the relevant light sensors. The testing result also indicates the dependence of the light control performance on the accuracy of the local and whole light sensing information.

We have not seen such a system using a rotating sensor in light sensing applications. In the field of radio tomography, a study presented in [14] introduced a rotating RF sensor system to sense the radio system from multiple directions and improve the accuracy of tomographic imaging. As we stated in [1], since the light ambient is heterogeneous, it is necessary to create a light sensor system that can measure light from different angles and reduce the problem of multiple sensor networks. Such a system is useful to increase the quality of intelligent lighting control applications.

3 Design of the new Sensor System

The rotating sensor motioned in [1] was a demo prototype. We have designed a new version of the sensor system that could be used as a real measurement tool. An image of the new sensor system is presented in Figure 1. The new design took care of the following factors:

- Improving the rotation system.
- Reduce the distance between the sensor and the rotation module.
- Optimizing the overall diameter of the sensing system.
- Adding a static sensor for referencing.

3.1 Multi-Directional Light Measurement Vector

Consider a visible light source s, the Lumens from this source can be calculated based on the I_s intensity [15] as:

$$\Phi_s = 4\pi I_s. \tag{1}$$

For the isotopic light source, where the intensity is the same in all directions, the measurement of the luminous flux on a unit area, also known as or light level in a specific area or illuminance, is the ratio of the Lumens Φ over the area of the measuring surface. However, in the indoor environment where the source is not isotopic, for a flat radiating surface of a sensor, the calculation of the illuminance need to consider cosine the observation angle (θ) with respect to the surface normal as shown in the Equation (2)

$$E = \frac{\Phi_s \cos\theta}{A}. \tag{2}$$

In the lighting control system, the observed ΔE, changing of illuminance on the light sensor surface, are used to control light bulb brightness:

$$\Delta E = \frac{\Delta\Phi_s \cos\theta}{A}. \tag{3}$$

In our proposed system, the light level is measured in multiple directions. Hence, we use a vector of illuminance as:

$$\mathbf{E} = \begin{bmatrix} E_0 \\ E_1 \\ \vdots \\ E_{N-1} \end{bmatrix}, \tag{4}$$

where N is the number of the measuring directions and E_n is the measured illuminance at the direction $n \in [0, N-1]$.

There are two major advantages of the sensing data provided by a rotating light sensor comparing to a static sensor. First, the light intensity measured by a static sensor depends on a single sensing

direction, the angle θ as in Equations (2) and (3), while the rotating sensor provides a combination of illuminances from multiple directions. In the real application, such as lighting control, the measuring direction could be blocked by an object that makes the sensing value of the static sensor inaccurate. The rotating sensor can be able to void this issue as it measures the light from different angles. Second, the rotating sensor uses a vector of illuminance (4) that can indicate the main light source direction. Furthermore, we can apply the gradient calculation (∇) on the vector of illuminance to measure the rate and direction of change in the light intensity.

3.2 *Features of the Rotating Sensor System*

The new sensor system is shown in Figure 1 has an overall diameter of 100 mm. This design is small enough to use in most indoor applications. Table 1 summarizes the features of the sensor system. The system uses one light sensor to dynamically measure multi-directional light intensity. In this system, there is no need to use many sensors as in a sensor network. By rotating the light sensor, our system may avoid the issues when the static measuring angle has been blocked by an object.

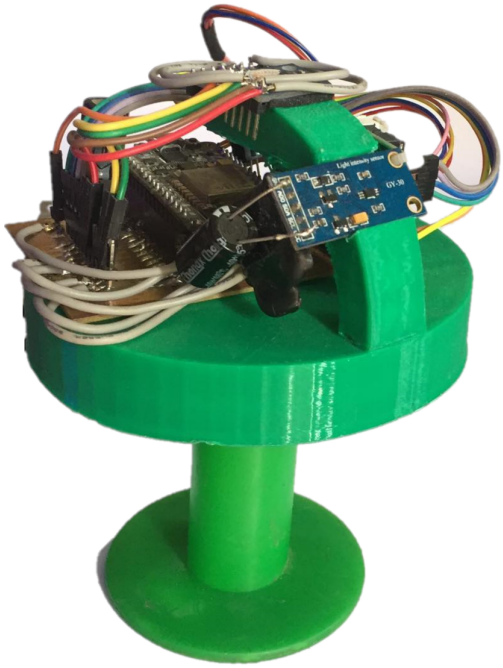


Figure 1: A sample image of the new rotating light sensor system. The light sensor is attached to the side of the curved handle of the rotation base structure.

Table 1: Features of the sensor system

System Features	Value
Light Sensor	BH1750
Stepper Motor	28BYJ-48
Stepper Motor driver board	ULN2003
Microcontroller Module	ESP8266
Overall Dimension (mm)	100
No. of Measuring Directions	8 and 16
Sensor Measuring Resolution (Lux)	1 (high) or 4 (low)

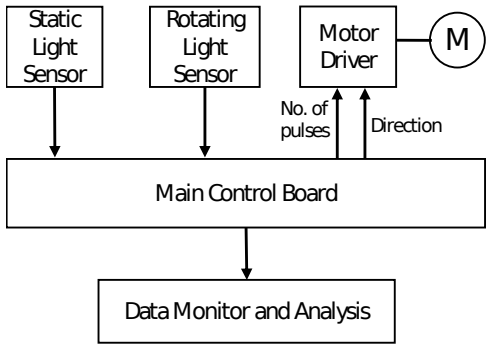


Figure 2: Diagram of the sensor system.

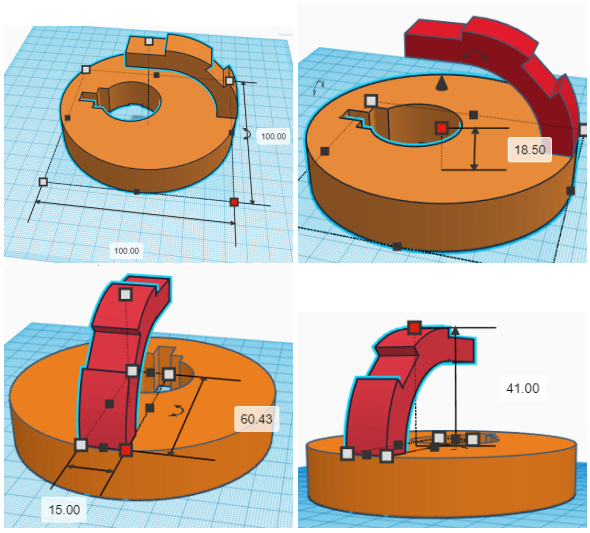


Figure 3: Sample views of the rotation base structure design.

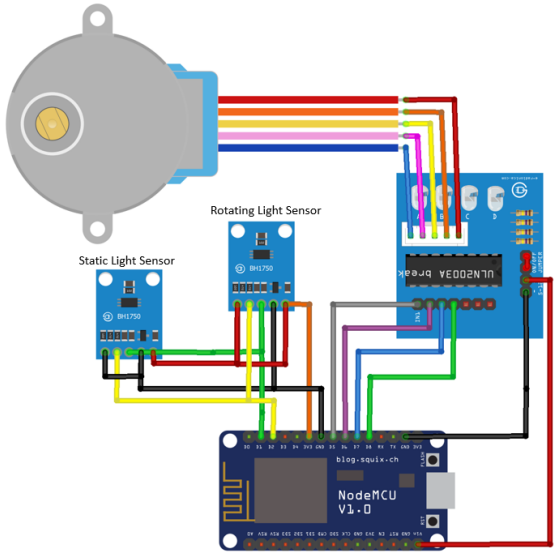


Figure 4: Conceptual logic circuit of the sensor system.

As shown in the system diagram in Figure 2, the major components of the system are the light sensor, stepper motor, motor diver board, and the main control board. The system uses a micro-

controller to control the rotation, sensing the light illuminance, and transfer data to the monitoring platform. In this design, we have improved the rotation system and created a new base structure of the sensor. Figure 3 shows several views our the new base structure. The structure was 3D printed and used in the sensor system.

As a modified version in [1], the new system uses an additional static sensor to provide a reference as requested. In this study, we consider the static sensor is the sensor that has a fixed sensing surface normal vector, thus the observation angle θ for any lighting source stays unchanged (see Equation 2). The rotating sensor is the sensor that its observation angle can be changed. By rotating the sensor in an axis different than the normal vector of its sensing surface, we can change its observation angle.

The logic circuit of the new system is shown in Figure 4. The sensors communicate with the microcontroller using the I^2C protocol. Both static and rotating sensors are the BH1750 [16] digital light intensity measuring module. The sensor module has an internal converter to convert the analog signal to the digital value. The module also contains a preprocessor to process the digital value and return the measuring value in the range of 1 to 65535 Lux.

The circuit uses a ULN2003 stepper motor driver to provide a drive interface between the microcontroller and the stepper motor. The ULN2003 driver pins, IN1, IN2, IN3, and IN4, are connected to the microcontroller digital pins. The power supply pins (VCC and GND) of the ULN2003 driver board are connected to the microcontroller 5V pin and GND pin, respectively. The stepper motor, model 28BYJ-48, runs in 32 steps mode and has a built-in gear with a 64 to 1 reduction ratio to produce 2048 steps per revolution. The step-angle relation is calculated as:

$$\alpha = \frac{i \times 360}{2048}, \tag{5}$$

where α represents the angle, and i is the number of controlling step of the motor. Sample images of the motor and the driver board were presented in [1]. In this new design, the motor is attracted to the center hole of the base structure shown in Figure 3. The motor can rotate the whole structure hence rotate the light sensor. For the microcontroller, we used the ESP8266 [17] board that runs in the Arduino environment. This module has a Wi-Fi microchip with a full TCP/IP stack to transfer the data. We program the system in the Arduino code using the C++ programming language.

Table 2: Relation of the sensing directions, angle interval, and motor steps interval

No. of Sensing Directions	$\Delta\alpha$	Steps Interval
8	45°	256
16	22.5°	128
32	11.25°	64
64	5.625°	32

3.3 Measuring Multi-Directional Light Illuminance

The main procedures of the illuminance measurement are kept similar to in [1] except for the additional signal of gathering the static sensor data. As mention in [1], for the applications that require the data of light intensity in different directions, the users may have to manually set up several sensors at different orientations. We developed an algorithm to automatically rotate the sensors, calculate the angles, and gather the illuminance in multiple directions. Figure 5 presents the major steps of the algorithms. In the beginning, the program set up the initial direction of the sensor. The sensor can rotate in a full 360° revolution, hence any angle can be set as the initial sensing direction. Then the sensor system starts collecting the angle and illuminate data. In each iteration of the major loop shown in Figure 5, the control program rotates the sensor by increasing the motor step, calculates the sensing angle, gets the light level information, and transmits the data to the monitoring software. The counting of the motor steps is used to calculate the angle based on the Equation (5). When the motor reaches a full circle (at 2048 counted steps), the counting is reset to Zero, and the rotating direction is reversed. The sensor system can measure light in any direction or at a group of interval directions. Table 2 summarizes some combinations of the sensing directions, angle interval ($\Delta\alpha$), and the motor steps interval. The visualizations of 8 sensing directions with 45° apart and 16 directions with 22.5° apart are shown in Figure 6.

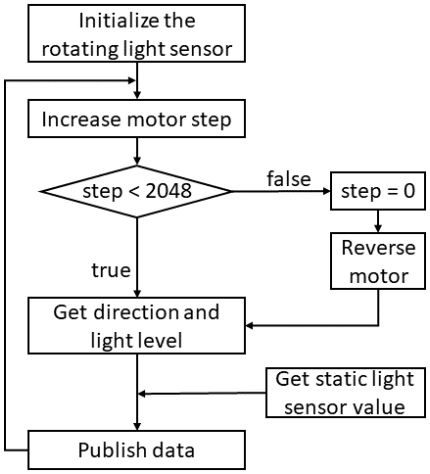


Figure 5: Flowchart of the control algorithm of the light sensor system. The major procedures are to control the rotating light sensor. The command to get static light sensor data is optional.

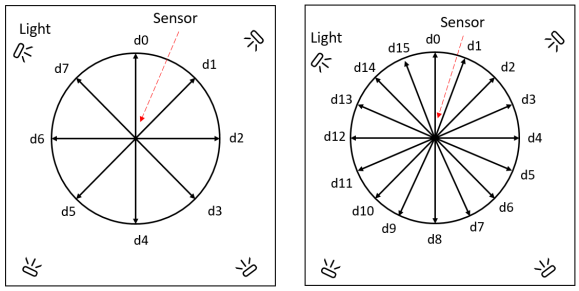


Figure 6: Sample views of the rotating light sensor with different number sensing angles. **Left**: 8 directions from $d0$ to $d7$ with $\Delta\alpha = 45°$. **Right**: 16 directions from $d0$ to $d15$ with $\Delta\alpha = 22.5°$. The lights are for presentation purpose only.

4 Performance Analysis of the Sensor System

Indoor lighting control is the major application of our proposed sensor system. We have tested the system in a workroom with the effect of the popular parameters on the sensor, like the changing of light power and blocking object. We present here the experiments

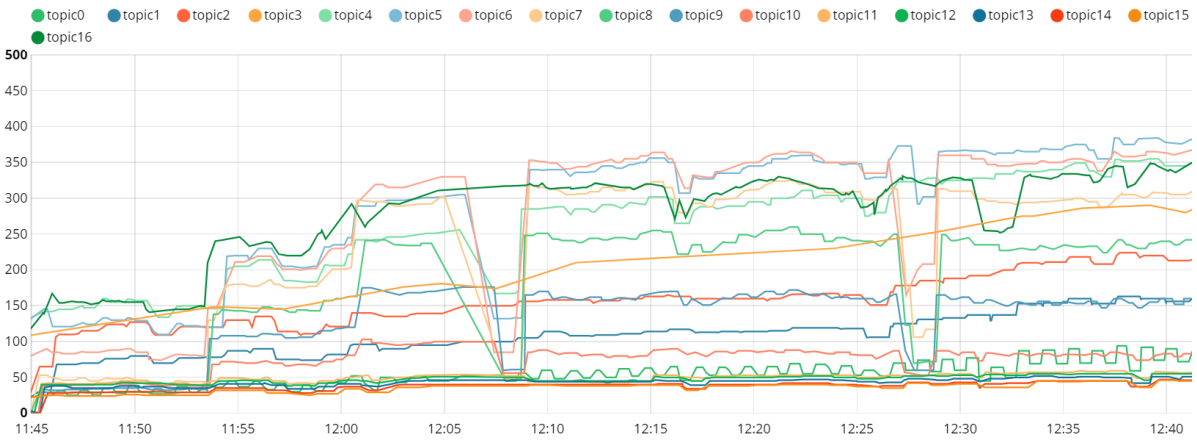


Figure 7: Snapshot of monitoring the light sensing data on the Ubidots website.

and results for testing the ability of the system to detect the main light direction, detect the blocked direction, and accurately measure light intensity.

4.1 Measuring and Transmitting Light Data

In [1], we presented the experiment that showed how the sensor measures the light level from multiple directions. For this new design, in addition to the multi-directional light sensing data, the system also collects the sensing data of a static reference sensor. We tested our device on transferring the data to the Ubidots [18] IoT platform. The data was transferred remotely through the Wifi signal. Figure 7 shows a part of the testing data monitored on the Ubidots website with multiple light sensing signals in one plot. We also have tested our system with the ThingSpeak [19] platform that has integrated MATLAB toolbox for analyzing the data.

For the data presented in Figure 7, the rotating sensor measured the light in 16 directions with $\Delta\alpha = 22.5°$ (Table 2 contains more detail of the measuring angles). It can be seen that our system is able to measure and provide light sensing data from multi-directional. In Figure 7, the rotating sensing data is labeled as a topic channel for each direction from *topic*0 to *topic*15. The static sensor channel is labeled as *topic*16. When monitoring the data, the channel topics (directions), the *Lux* values, and the measuring times can interactively be seen by selecting a data point on the graph. As shown in the figure, different sensing direction has a different value. In the beginning, when all lights were off, all sensing data was low (roughly under 150 *Lux*). Then, after we turned on all the lights (around 11:55 am timestamp, see Figure 7), the system recorded the increase of light levels in all directions. We blocked some directions of the rotating sensor at two later movements. As of the result, there were two failings of the light level at the blocked directions (in the 12:05 - 12:10 pm and the 12:25-12:30 pm periods, see Figure 7). In this experiment, the static sensor (*topic*16) was not blocked at any time, and it was used to provide the reference sensing data.

4.2 Detecting Main Light Direction

The distribution of light sources may generate an unbalance light level in the room. Our sensor can detect the major light direction that information helps to rearrange the illumination sources or indicate the location having the best light level. In this experiment, we place the rotating sensor at the center of a workroom that has several light sources. The light sources were placed around the room at different locations and distances to the sensor. We regularly turn on the lights to test the sensor performance. The sensor was set to measure in 16 directions ($d0 \ldots d15$, with $\Delta\alpha = 22.5°$). Figure 8 displays some radar charts of the measured results. For each chart, the data value as a length on angular spokes represents the magnitude of sensing value in a corresponding direction. The line connects the data values represses one full rotating revolution (360°).

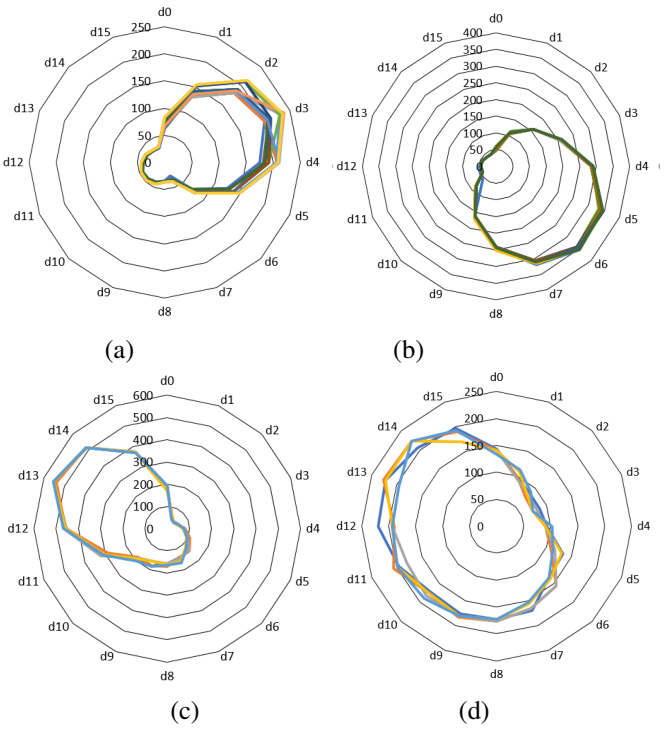


Figure 8: Radar charts of light sensing data on different main illumination directions.

The charts in Figure 8 associate with four scenarios of the lighting in the testing room. For an instant, Figure 8 (a) shows the *Lux* values measured in directions $d2$, $d3$, and $d4$ are much higher than the rest. In that case, only the light source at the direction of $d3$ was on, hence there is a peak sensing data in $d3$ (about 250 *Lux*) and two relevant highs on the adjacent angles $d2$ and $d4$. Figure 8 (b) shows the major level of light in $d5$, $d6$, and $d7$ with the peak value

of 350 *Lux*, and Figure 8 (c) shows the main level in $d12$, $d13$ and $d14$ with above 500 *Lux* peak value. These results plotted in the radar charts admit that the rotating sensor can detect the direction of the main light source. This result also can be observed from our vector of the illuminance shown in Equation 4 since the direction of the vector is influenced by the major scalar member values. Figure 8 (d) presented a test case where the light level is more consistent on the left side of the sensor ($d9$ to $d15$). In the case of a uniform light level, we may obtain consistent light sensing values in all directions.

4.3 Detecting Blocked Directions

We have performed an experiment to show how the rotating sensor can detect the temporary blocking directions with sudden changes in the sensing data. During this test, the rotating sensor was set up to run first, and then some sensing directions were temporarily blocked. Figure 9 plots the testing result. Before the time $T1$, the sensor indicates the major light level in directions $d4$, $d5$, $d6$, $d7$ and $d8$ with the high value above 300 *Lux* ($d6$). In the period of $T1$ and $T2$, when the sensor was blocked from the major light source direction, there was a sudden fall in the sensing value of the relevant angles. As plotted in Figure 9, all sensing values of $d4$ to $d8$ directions have steeply decreased to around 50 *Lux* while other values mostly unchanged. After the $T2$ moment, when the blocking object moved out, the sensing values from the main light directions got back to normal. It can be seen that we can use the directions having a temporary steeply reduction to indicate which sensing angle of the sensor was blocked. In the experiment presented in [1], we reported the result of detecting a sudden change of the natural background light, such as when opening the doors to get more sunlight.

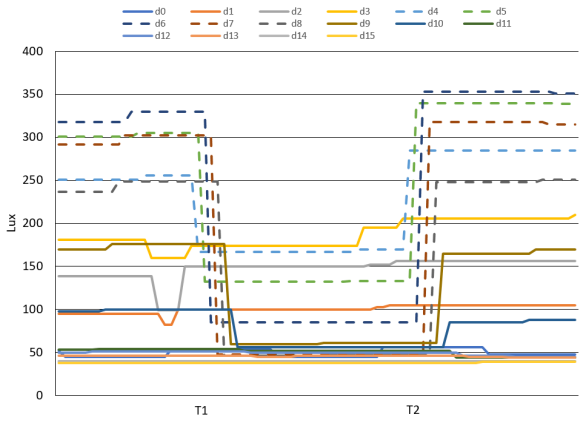


Figure 9: Sensing data of the rotating sensor when blocking some sensing angles.

4.4 Accuracy Analysis

The demand level of illumination to be controlled depends on the change in the sensing value. When using a single sensor or a network of sensors that was set at a static direction, it is possible that an unaware object temporarily blocks the sensor. Such a case makes a temporary change in the measuring data, leading to an inaccurate controlling of the illumination. The rotating sensor provides light sensing values from different directions that can help to avoid this problem.

We performed an experiment that used both static and rotating sensor. The artificial lighting sources were kept unchanged, but the static sensor would be blocked at a certain time during the experiment. Figure 10 shows the *LUX* values of the static sensor at a fixed angle and the rotating sensor rotated in eight directions (from $d0$ to $d7$). In this experiment, the rotating sensor measures light at every 45°. In the beginning, the static sensor data was around 450 *Lux* (±20), while the rotating sensor had different values at different directions with the high value in $d4$ of above 500 *Lux* and $d3$ around 430 *lux*. The directions $d0$, $d5$, and $d7$ had sensing values in between 320 and 350 *Lux*, and the direction $d1$, $d2$, and $d6$ had data in the range of 270 to 300 *Lux*. Then, we blocked the measuring direction of the static sensor. As shown in Figure 10, there is a sharp fall in the value of the static light sensing data down to 100 *Lux* and lower. If an application used just the static sensor to control the lights, it would incorrectly record a reduction in the light level and would control an unnecessarily increase of the illumination power. There was no need to change the illumination since the lights were stable throughout this experiment (except for a small variance in the natural light). For the rotating sensor, the sensing signals in the directions $d0$, $d1$, $d6$, and $d7$ were almost stable (around 300 *Lux* and 350 *Lux*). There was a small reduction in the direction $d2$ and $d5$ and a quire decreasing in the direction $d3$ and $d4$, but these sensing values were still around 250 *Lux*. When we blocked the static sensor, we approached the sensor location through the direction $d3$ and $d4$ that caused the changing of the data in these (and surrounding) directions. If we use the rotating sensing data of multiple directions for the lighting application, the control unit could use the data of the stable directions and decide to keep the illumination level the same. The prevention of an unnecessary increase in illumination power is useful to save energy costs.

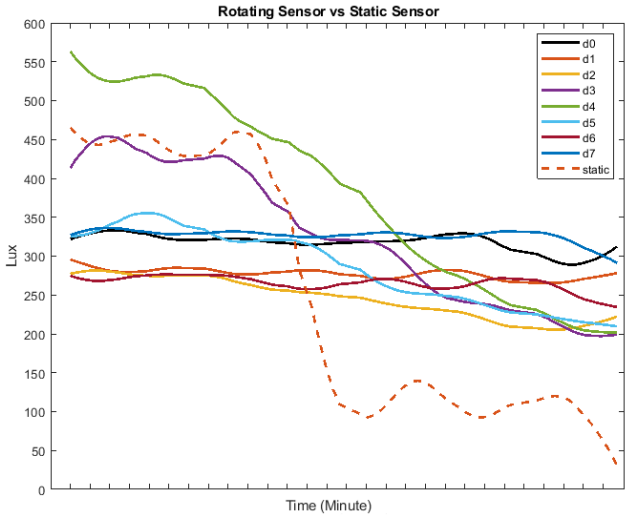


Figure 10: Light sensing values of the static sensor and the rotating sensor. The static sensor was blocked at a certain time, causing a reduction in the sensing value. The time ticks in the horizontal exist are presented at 60 seconds interval.

5 Conclusion

We have presented a rotating light sensor system that can measure light level in multi-direction rather than using server sensors positioning at different orientations. The system uses a stepper motor to rotate the sensor and track the motor steps to calculate the measuring angle. The light level values can be directly transferred to the moni-

toring system for further process. Besides, the sensor can connect with the popular IoT platforms, such as Ubidots and Thingspeak, to remotely visualize and analyze the sensing data. We also presented a new design of the sensor system with an optimal compact structure. The performance of the sensor system has been evaluated through testing experiments. The results show that the proposed sensor system is able to measure light level in different directions, detect the main light source direction, and determine the directions having temporary light changing. Moreover, even some sensing directions may be blocked, the rotating sensor still accurately measures the light level in other directions. In the future, we intend to develop a three-dimensional (3D) rotation light sensor. A motion tracking sensor shall be used to keep track of the 3D orientation. We plan to optimize the rotation system and perform more experiments of using the rotating light sensor in the light control applications.

References

[1] D. A. Hoang, T. T. Tung, C. M. Nguyen, K. P. Nguyen, "Rotating Sensor for Multi-Direction Light Intensity Measurement," in 2019 International Conference on System Science and Engineering (ICSSE), 462–467, 2019, doi: 10.1109/ICSSE.2019.8823447.

[2] A. P. Plageras, K. E. Psannis, C. Stergiou, H. Wang, B. B. Gupta, "Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings," Future Generation Computer Systems, **82**, 349–357, 2018, doi: https://doi.org/10.1016/j.future.2017.09.082.

[3] S. Chouikhi, I. El Korbi, Y. Ghamri-Doudane, L. A. Saidane, "A survey on fault tolerance in small and large scale wireless sensor networks," Computer Communications, **69**, 22–37, 2015, doi:https://doi.org/10.1016/j.comcom.2015.05.007.

[4] C. Dandelski, B. Wenning, D. V. Perez, D. Pesch, J. M. g. Linnartz, "Scalability of dense wireless lighting control networks," IEEE Communications Magazine, **53**(1), 157–165, 2015, doi:10.1109/MCOM.2015.7010529.

[5] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," IEEE Communications Surveys & Tutorials, **20**(4), 3098–3130, 2018, doi:10.1109/COMST.2018.2841349.

[6] V. Singhvi, A. Krause, C. Guestrin, J. H. Garrett Jr, H. S. Matthews, "Intelligent light control using sensor networks," in Proceedings of the 3rd international conference on Embedded networked sensor systems, 218–229, ACM, 2005, doi:https://doi.org/10.1145/1098918.1098942.

[7] J. Lu, D. Birru, K. Whitehouse, "Using Simple Light Sensors to Achieve Smart Daylight Harvesting," in Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, BuildSys '10, 73–78, ACM, New York, NY, USA, 2010, doi:10.1145/1878431.1878448.

[8] L. Yeh, C. Lu, C. Kou, Y. Tseng, C. Yi, "Autonomous Light Control by Wireless Sensor and Actuator Networks," IEEE Sensors Journal, **10**(6), 1029–1041, 2010, doi:10.1109/JSEN.2010.2042442.

[9] M. Pan, L. Yeh, Y. Chen, Y. Lin, Y. Tseng, "A WSN-Based Intelligent Light Control System Considering User Activities and Profiles," IEEE Sensors Journal, **8**(10), 1710–1721, 2008, doi:10.1109/JSEN.2008.2004294.

[10] D. Caicedo, A. Pandharipande, "Distributed Illumination Control With Local Sensing and Actuation in Networked Lighting Systems," IEEE Sensors Journal, **13**(3), 1092–1104, 2013, doi:10.1109/JSEN.2012.2228850.

[11] K. Warmerdam, A. Pandharipande, "Location data analytics in wireless lighting systems," IEEE Sensors Journal, **16**(8), 2683–2690, 2015, doi: 10.1109/JSEN.2015.2509982.

[12] X. He, A. Pandharipande, "Location-Based Illumination Control Access in Wireless Lighting Systems," IEEE Sensors Journal, **15**(10), 5954–5961, 2015, doi:10.1109/JSEN.2015.2449276.

[13] N. van de Meugheuvel, A. Pandharipande, D. Caicedo, P. van den Hof, "Distributed lighting control with daylight and occupancy adaptation," Energy and Buildings, **75**, 321 – 329, 2014, doi:https://doi.org/10.1016/j.enbuild.2014.02.016.

[14] M. Bocca, A. Luong, N. Patwari, T. Schmid, "Dial it in: Rotating RF sensors to enhance radio tomography," in 2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), 600–608, 2014, doi:10.1109/SAHCN.2014.6990400.

[15] C. Nave, "Hyperphysics," 2020.

[16] M. electronics, "BH1750 ROHM Semiconductor Datasheet," 2020.

[17] E. Systems, "ESP8266 Overview," 2020.

[18] Ubidots, "Ubidots Internet of Things and Cloud tools," 2020.

[19] thingspeak, "ThingSpeak IoT analytics platform," 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 223-233 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Dynamic Decision-Making Process in the Opportunistic Spectrum Access

Mahmoud Almasri[*,1], Ali Mansour[1], Christophe Moy[2], Ammar Assoum[3], Denis Lejeune[1], Christophe Osswald[1]

[1]*LABSTICC, UMR 6285 CNRS, ENSTA Bretagne, 2 rue F. Verny, Brest, 29806, France*

[2]*Univ Rennes, CNRS, IETR - UMR 6164, Rennes, F-35000, France*

[3] *Faculty of Science, Lebanese University, Tripoli, Lebanon*

ARTICLE INFO

Article history:
Received: 08 April, 2020
Accepted: 08 July, 2020
Online: 28 July, 2020

Keywords:
Opportunistic Spectrum Access
Cognitive Networks
Multi-Armed Bandit
Quality of Service
Priority Access

ABSTRACT

We investigate in this paper many problems related to the decision-making process in the Cognitive Radio (CR), where a Secondary User (SU) tries to maximize its opportunities by finding the most vacant channel. Recently, Multi-Armed Bandit (MAB) problems attracted the attention to help a single SU, in the context of CR, makes an optimal decision using the well-known MAB algorithms, such as: Thompson Sampling, Upper Confidence Bound, ϵ-greedy, etc. However, the big challenge for multiple SUs remains to learn collectively or separately the vacancy of channels and decrease the number of collisions among users. To solve the latter issue for multiple users, the All-Powerful Learning (APL) policy is proposed; this new policy considers the priority access and the dynamic multi-user access, where the number of SUs may change over time. Based on our APL policy, we consider as well as the Quality of Service (QoS), where SUs should estimate and then access best channels in terms of both quality and availability. The experimental results show the superiority of APL compared to existing algorithms, and it has also been shown that the SUs are able to learn channels qualities and availabilities and further enhance the QoS.

1 Introduction

Game theory represents a decision-making mathematical tool that attracts much attention, when it comes to networks for resource sharing, congestion control, transmission-rate adaptation, etc. This theory was originally and exclusively proposed for economics before being applied to many other topics, such as: financial, regulation, military, political science and also biology. The main objective for using the game theory is to study and analyze cooperative or competitive situations for rational players in order to find an equilibrium among them. When, players reach the equilibrium point, then none of them can gain more by changing its action.

Game theory is widely applied in Cognitive Radio (CR) in order to enhance the spectrum efficiency of the licensed frequency bands. Indeed, according to many recent studies, the frequency bands are not well used. On the one hand, the demands on high data rate applications and wireless devices have experienced unprecedented advancement since 1990s which makes the frequency bands more and more crowded. On the other hand, several simulations have been conducted in the United States and showed that 60 % of the frequency bands are not used [1]. Several solutions have been recommended by the Federal Communications Commission (FCC) in order to enhance the usage of the spectrum. Opportunistic Spectrum Access (OSA) in CR, represents one of the proposed solutions, where users are categorized into two groups namely: Licensed users (Primary Users: PUs) who have the right to access the frequency bands at any time, and unlicensed users (Secondary Users: SUs) that can access the frequency bands in an opportunistic manner. Usually, SUs can coexist with PUs in the same frequency bands as far as they dont cause any harmful interference to these latter. Indeed, SUs are able to access the frequency bands currently unused by PUs. SUs in OSA have many challenges in order to reduce the interference with PUs:

- **Spectrum Sensing**: A SU should sense the frequency bands and identify the available spectrum holes before

[*] Corresponding Author: Mahmoud Almasri. 2 rue F. Verny, 29806 Brest, France. email: mahmoud.almasri@ensta-bretagne.fr

making any decision. The main challenge is to gather an accurate information about the status of the spectrum (free or busy) in order to access only the unused channels without causing any harmful interference to PUs. Due to hardware constraints, delay and high energy consummation, a SU may be able to sense a portion of the frequency bands (e.g. one channel at each time slot) and decides whether the selected channel is free to transmit.

- **Spectrum Decision**: At each time slot, a SU should decide which channel to access based on past success or failure decisions. As a result, a SU can gather some information about the availability and quality of channels and build a database of the spectrum access environment. This database is used in order to make a good decision and enhance the future actions of the SU.
- **Spectrum Sharing**: In order to share the available spectrum among SUs, two main models exist: Cooperative or competitive access. In the cooperative behaviors, the users need to exchange information with each other in order to maximize their opportunities and thus decrease the interference among themselves. Despite the latter benefits of the cooperative access, each user should be informed about others decisions before making any action which may increase the complexity of the secondary network. While, in the competitive access, each SU makes an action based on its local observation. However, this lack of information exchange can increase the number of collisions among users. To solve this issue, a specific policy is required to learn the vacancy probabilities of available channels and decrease the number of collisions among users.

Table 1: List of acronyms

APL	All-Powerful Learning
CR	Cognitive Radio
DMC	Dynamic Musical Chairs
EXP3	Exponential weights for Exploration and Exploitation
FCC	Federal Communications Commission
MAB	Multi-Armed Bandit
MEGA	Multi-user ϵ-greedy collision Avoiding
OSA	Opportunistic Spectrum Access
PU	Primary User
QoS	Quality of Service
SU	Secondary User
SLK	Selective Learning of the kth largest expected rewards
TS	Thompson Sampling
UCB	Upper Confidence Bound

- Finally, in the **Spectrum Mobility**, a SU should evacuate the selected channel when a PU reappears. Moreover, a SU may badly identify its dedicated channel and then access a channel that does not correspond to its prior rank[1]. Therefore, the user should evacuate its current channel when he identifies its targeted channel.

This paper is an extension of our original work presented in [2] with a novel policy called All-Powerful Learning (APL) is proposed in order to maximize the opportunities of SUs, share the available spectrum among them, and limit the interference among PUs and SUs. Instead of only considering the availability, this paper takes into account a quality information metric, where the priority users should access only best channels with the highest availability and quality.

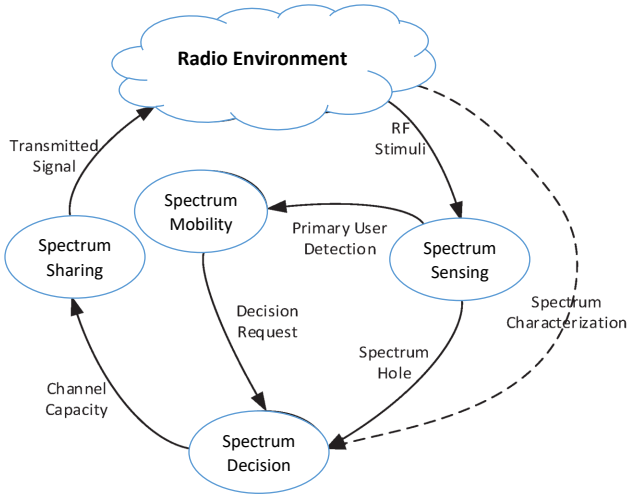


Figure 1: Cognitive cycle as introduced in [3].

2 Multi-Armed Bandit Problem

Multi-Armed Bandit (MAB) model represents one of the famous models, in game theory, that is adopted to enhance the efficiency of the licensed frequency bands. Moreover, MAB problem represents a simple case of the Reinforcement Learning (RL).

In the RL, the agent should enhance his behavior from the feedback (e.g. reward). Indeed, the RL may allow an agent to adapt to his environment by finding a suitable action to reach the best reward. The agent can maximize his reward without any prior information about his environment. However, by memorizing the states of an environment or the actions he took, the agent can make a better decision in the future. The reward feedback, also called reinforcement signal, has an important role to help an agent to learn from its environment. The RL is widely used in several domains: Robotics, Aircraft control, self- driving cars, Business strategy planning, etc. It was first developed for a single agent who should find an optimal policy that maximizes his expected reward knowing that the optimal policy depends on the environment. Unlike the case of a single agent, for multiple agents, the optimal policy

[1]Based on our APL policy, each user has a prior rank and should access the channel corresponding to its rank.

depends not only on the environment but also on the policies selected by other agents. Moreover, when multiple agents apply the same policy their approaches in such systems often fail because each agent tries individually to reach a desired result. In other words, it is impossible for all agents in a certain system to maximize simultaneously their personal reward, although find an equilibrium for the system representing a point of interest. Subsequently, it is important to find a policy for each agent in order to guarantee the convergence to an equilibrium state in which no agent can gain more when modifying its own action. In RL, Exploitation-Exploration dilemma represents an attractive problem. In order to maximize his performance (exploitation), the agent should gather some information about his environment (exploration). This is known as the Exploration-Exploitation dilemma in the reinforcement learning. If the agent spends a lot of time on the exploration phase, then he cannot maximize his reward. Similarly, when the agent focuses on the exploitation phase by exploiting his current information, then he may miss the best action that leads to the highest reward. Thus, the agent needs to balance the tradeoff between Exploration and Exploitation in order to obtain an appropriate result.

Due to its generic nature, the MAB model is widely adopted in many fields, such as: wireless channel access, jamming communication or object tracking. In such model, an agent can play a single arm at each time trying to maximize its long-term reward. To reach its goal, the agent needs to find the best arm in terms of expected reward. At each time slot, the agent can choose the current best arm (exploitation) or play other arms trying to obtain a robust estimation of their reward (exploration). Generally, an optimal policy, used by the agent, should balance between the exploitation and the exploration phases while pulling the arms.

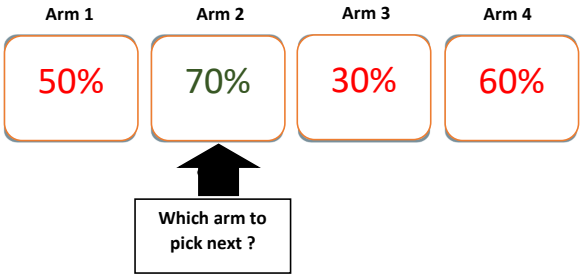


Figure 2: Several Arms with different expected reward. After a finite period of time the agent has a perception about the reward obtained from each arm.

Like most RL frameworks, the agent starts the game without any priori knowledge about the expected reward of the arms. The main goal of the agent is to find the arm with the highest expected reward. Here, we should define two classes of arms:

Optimal arm: This arm has the highest expected reward and is represented by the arm 2 in Fig. 2. The agent tries to reach this arm in order to maximize his expected reward.

Suboptimal arms: Include all other arms considered as non-optimal. Efficient MAB algorithms should be able to limit playing with suboptimal arms.

To solve the MAB problem, several algorithms have been proposed, such as: Thompson Sampling [4], Upper Confidence Bound (UCB) [5], ϵ-greedy [6], Exponential weights for Exploration and Exploitation (EXP3) [7], etc. The performance of a given MAB algorithm is usually measured by a regret that represents the gap between the reward obtained in the ideal scenario, where the user know the expected reward of each arm and often pulls the best one, and that obtained using a given MAB algorithm.

It is worth mentioning that these algorithms have been suggested for a single SU in the context of OSA where the SU is considered as an agent and the channels become equivalent to the different arms. Then, it is assumed that each channel is associated with a distinct availability probability and the SU should estimate this latter after a finite number of time slots. In this work, we first start to formulate the classical OSA as a MAB problem, in which, we consider a single Secondary User (SU) that needs to access opportunistically the frequency band. Later on, we will consider more realistic conditions that deal with the OSA (e.g. multiple users, Quality of Service, collision among users, dynamic access).

2.1 Thompson Sampling

Thompson Sampling (TS), a randomized algorithm with a bayesian spirit, represents one of the earliest algorithms proposed to tackle the MAB problem. In TS, each arm has assigned an index $B_i(t, T_i(t))$ that contains information based on the past success and failure observations. After a finite number of time slots, the index $B_i(t, T_i(t))$ will be very close to the mean reward of each arm. By selecting the arm with the highest index at each time slot, the agent often selects the best arm with the highest reward. This index achieves a trade-off between the exploration and the exploitation phases and can be defined as follows:

$$B_i(t, T_i(t)) = \frac{W_i(t, T_i(t)) + a}{W_i(t, T_i(t)) + Z_i(t, T_i(t)) + a + b} \quad (1)$$

where $W_i(t, T_i(t))$ and $Z_i(t, T_i(t))$ represent respectively the success and failure access; a and b are constant numbers.

Despite its excellent performance that can exceed the state-of-the-art MAB algorithms [8, 9, 10], TS is widely ignored in the literature. This ignorance is due to the fact that this algorithm is proposed with a lack of proof and a slight mathematical background unlike other MAB algorithms, such as: UCB or ϵ-greedy. Recently, TS has attracted more attention and is being used in several fields [11, 12, 13]. Recent studies have found a theoretical upper bound for its convergence to the best choice [14, 15, 16].

2.2 Upper Confidence Bound

Upper Confidence Bound (UCB) represents one of the famous MAB algorithms firstly proposed in [5]. Like TS, the index $B_i(t, T_i(t))$ of UCB contains two phases, the exploration

and the exploitation phases, in order to estimate the vacancy probabilities of channels and then access the best one. In the literature, several variants of UCB have been proposed to enhance the performance of the classical UCB, such as: UCB1, UCB2, UCB-tuned, Bayes-UCB, KL-UCB [8, 17, 18, 19]. UCB1 [17] represents the simplest version that balances between the complexity and the optimality.

Algorithm 1: Thompson Sampling Algorithm

Input: C, n,
C: number of channels,
n: total number of slots,
Parameters: $S_i(t)$, $T_i(t)$, $W_i(t, T_i(t))$, $Z_i(t, T_i(t))$,
$S_i(t)$: the state of the selected channel, equals one if the channel is free and 0 otherwise,
$T_i(t)$: number of times the ith channel is sensed by SU,
$W_i(t, T_i(t))$: the success access of the ith channel,
$Z_i(t, T_i(t))$: the failure access of the ith channel,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: the index assigned for the ith channel,
foreach $t = 1$ *to* n **do**
  $a_t = \arg\max_i B_i(t, T_i(t))$,
  Observe the State $S_i(t)$,
  $W_i(t, T_i(t)) = \sum_{t=0}^{n} S_i(t) 1_{a_t=i}$,
  % $1_{a_t=i}$: equal 1 if the user selects the ith channel and 0 otherwise,
  $Z_i(t, T_i(t)) = T_i(t) - W_i(t, T_i(t))$,
  $B_i(t, T_i(t)) = \frac{W_i(t,T_i(t))+a}{W_i(t,T_i(t))+Z_i(t,T_i(t))+a+b}$

Algorithm 2: UCB1 Algorithm

Input: α, C, n,
α: exploration-exploitation factor,
C: number of channels,
n: total number of slots,
Parameters: $T_i(t)$, $X_i(T_i(t))$, $A_i(t, T_i(t))$,
$T_i(t)$: number of times the ith channel is sensed up to t,
$X_i(T_i(t))$: the exploitation contribution of ith channel,
$A_i(t, T_i(t))$: the exploration contribution of ith channel,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: the index assigned for i^{th} channel,
foreach $t = 1$ *to* C **do**
  SU senses each channel once,
  SU updates its index $B_i(t, T_i(t))$,
foreach $t = C + 1$ *to* n **do**
  $a_t = \arg\max_i B_i(t-1, T(t-1))$,
  $T_i(t) + +$,
  $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau)$,
  % $S_i(\tau)$ is the observed state from channel i at τ,
  % $S_i(\tau) = 1$ if the channel i is vacant and 0 otherwise,
  $A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}}$,
  $B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t))$,

For this reason, UCB1 is the widely adopted version scheme in the context of CR to help a SU make an optimal decision [20, 21, 22, 23, 24, 25]. In UCB1, the index $B_i(t, T_i(t))$ essentially comprises two important factors: $X_i(T_i(t))$ and $A_i(t, T_i(t)))$ that represent respectively the exploitation (or the expected reward) and the exploration phases:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{2}$$

where the exploitation and the exploration factors can be expressed as:

$$X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{j=1}^{t} r_i(j) \tag{3}$$

$$A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}} \tag{4}$$

The factor $A_i(t, T_i(t))$ has an important role in learning the availability probabilities of channels by pushing the algorithm to examine the state of all available channels. Thus, after a finite time t, $X_i(T_i(t))$ of the ith channel will approximately equal to its availability probability μ_i.

In [17], the authors found an upper bound of the sum of regret (i.e. the loss of reward by selecting the worst channels) for a single agent and C arms. It has shown that the upper bound of the regret achieves a logarithmic asymptotic behavior, which means that after a finite number of time slots, the agent will be able to identify the best arm and always select it.

2.3 ϵ-greedy

One of the simplest MAB algorithms to tackle the MAB problem is referred to ϵ-greedy that was firstly proposed in [6]. A recent version of this algorithm is proposed in [17] in order to achieve a better performance compared to several previous versions (see algorithm 1). Like several MAB algorithms, ϵ-greedy contains two phases completely separated: exploration and exploitation. During the exploration phase, the user chooses a random channel in order to learn the vacancy probability of channels.
While in the exploitation phase, the user usually selects the channel with the highest expected reward $X_i(T_i(t))$. The authors of [17] have also investigated the analytical convergence of the ϵ-greedy and proved that the regret (i.e. the loss of reward by selection the worst channel) achieves a logarithmic asymptotic behavior.

3 Problem Formulation

In the previous section, we introduced the well-known MAB algorithms that help a MAB agent makes a good decision. In this section, we present the classical OSA for a single SU in order to formulate it as a MAB problem. However, MAB algorithms can represent an optimal solution for the classical OSA, as it can be seen in section 5. On the other hand, we

consider more developed scenarios compared to the classical OSA such as multiple SUs, decreasing the collisions among users and also estimating the quality of the available channels. We first present the OSA for multiple SUs in the next section and, hereinafter, we propose the new APL policy to manage a secondary network.

Algorithm 3: ϵ-greedy Algorithm

Input: C, H, n,
C: number of channels,
H: exploration constant,
n: total number of slots,
Parameters: $T_i(t)$,
$T_i(t)$: number of times the channel is sensed up to time t,
χ: a uniform random variable in [0,1],
Output: $X_i(T_i(t))$,
$X_i(T_i(t))$: the expected reward that depends on $T_i(t)$,
foreach $t = 1$ *to* n **do**
  if $\chi < min\{1, \frac{H}{t}\}$ **then**
    SU makes a random action a_t,
  else
    $a_t = \max_i X_i(T_i(t))$,
    $T_i(t) + +$,
    $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau)$,
    % $S_i(\tau)$ is the observed state from channel i at τ,
    % $S_i(\tau) = 1$ if the ith channel is vacant and 0 otherwise,

3.1 Single User Case

Let us consider a SU accesses C channels, each of which associated with a vacancy probability $\mu_i \in [0, 1]$. Let the vacancy probabilities be ordered by their availability probabilities, $\mu_1 > \mu_2 > ... > \mu_C$, which are initially unknown for the secondary user. A most important objective of the SU is to estimate the vacancy probabilities of channels after a finite time in order to access the best channel that has μ_1 as vacancy probability. At each time slot, the user can select one channel and transmit its data if available; otherwise, it should wait the next slot to sense another channel. Let the state of the ith channel at slot t be referred to $S_i(t)$: $S_i(t)$ equals 1 if the ith channel is free and 0 otherwise. Hereinafter, we consider that the obtained reward from the ith channel $r_i(t)$, at slot t is equal to its state: $r_i(t) = S_i(t)$. Let $T_i(t)$ represent the number of times to access the ith channel up to the slot t. The user should be rational by adopting a given policy in order to quickly identify the best channel. A policy selected by the SU may not be considered as optimal in term of the accuracy of the channels' vacancy estimation or the convergence speed towards the best channel. Finally, let us introduce the regret that rerepsents the gap between the reward obtained in an ideal scenario and that can be obtained using a given policy as follows:

$$R(n, \beta) = n\mu_1 - E\left[\sum_{t=1}^{n} \mu_i^{\beta}(t)\right] \tag{5}$$

where n represents the total number of time slots and $\mu_i^{\beta}(t)$ stands for the vacancy probability of the selected channel at slot t under the policy β, and $E(.)$ is the mathematical expectation.

3.2 Multi-User Case

In this section, we consider U SUs trying to learn the vacancy probabilities of the C channels and then access only the U best ones ($C > U$). When several SUs existing in the spectrum, their main challenge is to learn collectively or separately the vacant probability of channels as much as possible in order to access the best ones. Therefore, a policy selected by users should estimate the vacancy of channels as much as possible, and should also be able to decrease the collisions number among users. Therefore, let us define the regret for multiple users that takes into account both the convergence speed to the U best channels and the collision number among users as follows:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{t=1}^{n} E\left[S^{\beta}(t)\right] \tag{6}$$

where μ_k stands for the vacancy probability of the kth best channel; $S^{\beta}(t)$ represents the global reward obtained by all users at time t using the policy β and is defined as follows:

$$S^{\beta}(t) = \sum_{j=1}^{U} \sum_{i=1}^{C} S_i(t) I_{i,j}(t) \tag{7}$$

where $S_i(t)$ represents the state of the ith channel at time t: $S_i(t) = 1$ if the ith channel is available and 0 otherwise; $I_{i,j}(t)$ indicates that no collisions have appeared in the ith channel by the jth user at slot t: $I_{i,j}(t) = 1$ if the jth user is the sole occupant of the channel i and 0 otherwise. Finally, the regret that takes into consideration the channels' occupancy and the collisions number among users can be expressed by:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{j=1}^{U} \sum_{i=1}^{C} P_{i,j}(n) \mu_i \tag{8}$$

where $P_{i,j}(n) = \sum_{t=1}^{n} E\left[I_{i,j}(t)\right]$ represents the expectation of times that the jth user is the only occupant of the ith channel up to n, and the mean of reward can be given by:

$$\mu_i \approx \frac{1}{n} \sum_{t=1}^{n} S_i(t)$$

4 Multi-Priority Access

In the existing models of OSA where several SUs exist in the network, the main challenge is to learn collectively (via a cooperative learning) or separately (via a competitive learning) the

available channels while decreasing the number of collisions with each other. In our work, we focus on the competitive priority access, where the kth user should selfishly estimate the vacancy probabilities of channels in order to access the kth best one. Our proposed policy for the priority access takes into account the dynamic access where the priority users can enter/leave the network at any time. To the best of our knowledge, only the priority or the random access are considered without the dynamic access in several proposed MAB policies [24, 25, 26, 27] (a simple example for the priority dynamic access is shown in Fig. 3).

To formulate the OSA as a MAB problem, recent works extend the simple case of MAB (i.e. the case of a single agent) to consider several agents [20, 25, 26, 28, 29]. In our work, we are interested in the OSA for multiple priority access in which SUs should access the spectrum according to their ranks. Moreover, decreasing the number of collisions among SUs represents a point of interest to enhance the global performance of the secondary network. In general, when two SUs access the same channel to transmit, their data cannot be correctly received because of the interference between them. When a collision occurs among users, several proposals can be found in the literature in order to enhance their behavior in the next slots. We present below two well-known collision models in the literature that are widely used in OSA:

- ALOHA-like model: If a collision occurs between two or more users, then none of them receives a reward, despite the selected channels is free. This model may ensure the fairness among users, and no collision avoidance mechanism is used.

- Reward sharing model: If two or more users select the same channel at the same time, the colliding users share the obtained reward from the selected channel (each of them receives the same reward).

The above models can affect the methodologies used to collect the reward from the target channel while the learning phase is not affected. In our work, we consider the most widely used, ALOHA-like.

Based on the ALOHA-like, the works of [2, 20, 21, 25, 26, 27, 28, 30] proposed semi-distributed and distributed algorithms in which users cannot exchange information with each other. Liu and Zhao in [28], proposed Time-Division Fair Share (TDFS) policy and showed that the proposed algorithm may achieve an asymptotic logarithmic behavior. In such algorithm, the users access the channels with different offsets. TDFS also ensures the fairness among users; while in our work we are interested in the priority access where users access the channels based on their prior rank. In [28], TDFS policy was been used to extend UCB1 algorithm to consider multiple users. Beside TDFS, the authors of [20] proposed Random Rank policy, based on UCB1, to manage the secondary network. Random Rank represents a distributed policy (i.e. no-information exchange among users) in which the user achieves a different throughput.

The authors of [24] proposed the Selective Learning of the kth largest expected rewards (SLK) policy, based on UCB1, that represents an efficient policy for the priority access. However, SLK allows only a fixed number of users to access the available channels. So that, the dynamic access under SLK cannot be considered since this latter restricts the access. Similarly to SLK, the authors of [25] proposed the $kth - MAB$ for the priority access which is based on UCB1 and ϵ-greedy. In $kth - MAB$, the time is slotted and each slot is divided into multi sub-slots depending on the users priority ranks. For instance, the slot of SU_U is divided into U sub-slots in order to find the Uth best channel and transmit data via this channel. Therefore, the main limitation of this policy remains in the dissatisfaction of transmission time of high ranked users.

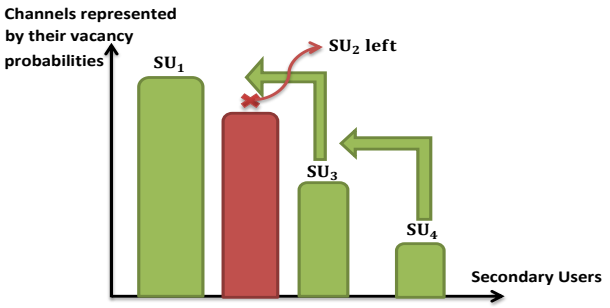


Figure 3: Priority access after a user left its dedicated channel

For the random access, several learning policies can be found in the literature, where the SU selects randomly its channel. The authors of [26] proposed the Musical Chairs policy as well the Dynamic Musical Chairs (DMC) policy for a dynamic access. In both policies, the SU selects a random channel up to time T_0 in order to estimate the vacancy of channels and the number of users, U, in the network. After T_0, the SU chooses a random channel between $\{1, ..., U\}$. The main drawback of the Musical Chairs and DMC is that the users should known the total number of transmission time slots as well as the number of available channels. Moreover, in DMC a restrict access is considered, where the users cannot leave the network during the time T_0. To find the U-best channels, the authors of [27] proposed the Multi-user ϵ-greedy collision Avoiding (MEGA) algorithm based on the ϵ-greedy algorithm proposed in [17]. However, their algorithm suffers the same drawbacks of the Musical Chairs and the Dynamic Musical Chairs and it does not consider the priority access. In order to solve all these limitations, we propose in section (4.1) a novel policy called APL for the priority dynamic access.

4.1 APL for the Priority Access

In this section, we propose a new policy for the priority access. This policy enables a secondary user to learn the vacant probabilities of channels and ensures the convergence to his dedicated channel. Moreover, it can be used with all learning MAB algorithms such as: Thompson Sampling (TS), Upper

Confidence Bound (UCB), AUCB, *e*-UCB, *e*-greedy, etc. We should highlight that our proposed policy does not require prior knowledge about the channels as in the case for other policies, such as: Musical Chair [26], SLK [24], *k*-th MAB [25], MEGA [27], etc. Indeed, existing policies to manage a secondary network suffer from one or more of the following disadvantages:

1. The number of users should be fixed and known to all users.
2. SUs should have a prior information about the number of channels.
3. Expected transmission time should be known.
4. The dynamic access is not suggested. To recall, in a dynamic access, the users can at any given time enter or leave the network.
5. Some algorithms consider a restricted dynamic access, where a SU can't leave the network during the learning or the exploration phases.
6. The vacant probabilities of channels should be static; otherwise, users cannot adapt to their environment.
7. The priority access is seldomly suggested in the literature, while the random access represents the most used model.

Unlike SLK and *k*-th MAB, our proposed policy for the priority access, called All-Powerful Learning algorithm (APL), doesn't suffer from the above mentioned drawbacks. As a matter of fact, SLK and *k*-th MAB policies suffer from the 1st, 2nd and 4th mentioned drawbacks.

In a classical priority access, each channel has assigned an index $B_i(t)$ and the highest priority user SU_1 should sense and access the channel with the highest index $B_i(t)$ at each time slot. Indeed, the best channel, after a finite number of time slots, will have the highest index $B_i(t)$.

As the second priority user SU_2 should avoid the first best channel and try to access the second best one. To reach his goal, SU_2 should sense the first and second best channels at each time slot in order to estimate their vacant probabilities and then access the second best channel if available. In this case, the complexity of the hardware is increased, and we conclude that a classical priority access represents a costly and impractical method to settle down each user to his dedicated channel. In the case of APL, at each time slot, the user senses a channel and transmits his data if the channel is available (see algorithm 4). In our policy, each SU_k has a prior rank, $k \in \{1, ..., U\}$, and his target is to access the *k*-th best channel. The major problem of the competitive priority access is that each user should selfishly estimate the vacant probabilities of the available channels. Our policy can intelligently solve this issue by making each user generate a rank around his prior rank to get information about the channels availability. For instance, if the rank generated by the *k*-th user equals 3 (considering that $k > 3$), then he should access the channel that has the third index, i.e. $B_3(t)$. In this case, SU_k can examinate the states of the *k* best channels and his target is the *k*-th best one.

Algorithm 4: APL for the priority dynamic access

```
Input: k, ξ_k(t), r_i(t),
k: indicates the k − th user or k − th best channel,
ξ_k(t): indicates a presence of collision for the k − th user
  at instant t,
r_i(t): indicates the state of the i − th channel at instant t,
  r_i(t) = 1 if the channel is free and 0 otherwise,
Initialization
k = 1,
for t = 1 to C do
    SU_k senses each channel once,
    SU_k updates his index B_i(t),
    SU_k generates a rank of the set {1, ..., k},
    k + 1,
for t = K+1 to n do
    SU_k senses a channel in his index B_i(t) according to
      his rank,
    if r_i(t)=1 then
        SU_k transmits his data,
        if ξ_k(t)=1 then
            SU_k regenerates his rank of the set {1, ..., k},
        else
            SU_k keeps his previous rank,
    else
        SU_k refrains from transmitting at instant t,
    SU_k updates his index B_i(t)
```

However, if the rank created by SU_k is different than k, then he selects a channel with one the following probabilities: $\{\mu_1, \mu_2, ..., \mu_{k-1}\}$ and he may collide with a priority user, i.e. $SU_1, SU_2, ..., SU_{k-1}$. Therefore, SU_k should avoid regenerating his rank at each time slot; otherwise, a large number of collisions may occur among users and transmitted data can be lost. So, after each collision, SU_k should regenerate his rank from the set $\{1, ..., k\}$. Thus, after a finite number of slots, each user settles down to his dedicated channel. It remains to investigate the analytical convergence of APL to verify its performance in a real radio environment.

4.2 Quality of Service

As mentioned before, UCB represents one of the popular MAB algorithms that is widely suggested in the literature, where several variants have been proposed. In [23], we proposed a new variant of UCB called the Quality of Service UCB1 (QoS-UCB1) for a single SU, where this latter is able to learn channels' vacancy and quality. To consider multiple SUs, this version of UCB is extended using the Random Rank policy proposed in [20] to manage a secondary network. It has been shown that the Random Rank policy with the QoS-UCB1 represents an optimal solution to allow users to learn separately channels' vacancy and quality. However, in this paper, we

evaluate the performance of our APL policy with QoS-UCB1 for the priority access.
Supposing that each channel has a binary quality represented by $q_i(t)$ at slot t: $q_i(t) = 1$ if the channel has a good quality and 0 otherwise. Then, the expected quality collected from the channel i up to time n is given by:

$$G_i(T_i(n)) = \frac{1}{T_i(n)} \sum_{\tau=1}^{T_i(n)} q_i(\tau) \quad (9)$$

The global mean reward, that takes into account channels' vacancy and quality, can be expressed as follows [23]:

$$\mu_i^Q = G_i(T_i(n)).\mu_i \quad (10)$$

The index assigned to the ith channel that considers both vacancy and quality $B_i^Q(t, T_i(t))$ can be defined by:

$$B_i^Q(t, T_i(t)) = X_i(T_i(t)) - Q_i(t, T_i(t)) + A_i(t, T_i(t))) \quad (11)$$

According to [23], the term $Q_i(t, T_i(t))$ of the quality factor is given by the following equation:

$$Q_i(t, T_i(t)) = \frac{\gamma M_i(t, T_i(t)) \ln(t)}{T_i(t)}$$

where the parameter γ stands for the weight of the quality factor; $M_i(t, T_i(t)) = G_{\max}(t) - G_i(T_i(t))$ being the difference between the maximum expected quality over channels at time t, i.e. $G_{\max}(t)$, and the one collected from channel i up to time slot t, i.e. $G_i(T_i(t))$. However, when the ith channel has a good quality $G_i(T_i(t))$ as well as a good availability $X_i(T_i(t))$ at time t. The quality factor $Q_i(t, T_i(t))$ decreases while $X_i(T_i(t))$ increases. Subsequently, by selecting the maximum of its index $B_i^Q(t, T_i(t))$, the user has a large chance to access the ith channel with a high quality and availability.

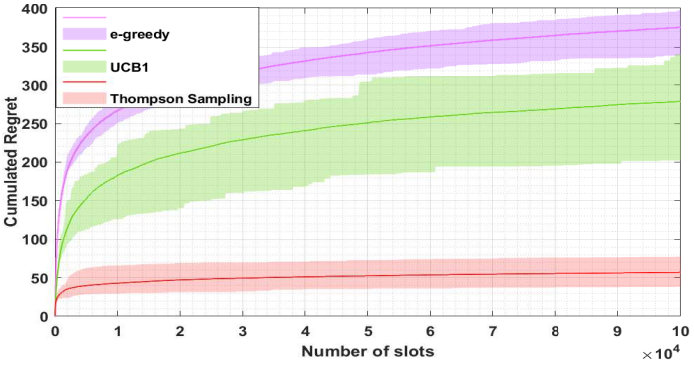


Figure 4: Evaluate the performance of TS, UCB1 and ϵ-greedy in OSA

5 Simulations and Results

In our simulations, we consider three main scenarios: In the first one, a SU tries to learn the vacancy of channels using the MAB algorithms: TS, UCB1 and ϵ-greedy in order to access the best one with the highest vacancy probability. We also compare the performance of these MAB algorithms to show which one can offer more opportunities for the SU. In a second scenario, we considered 4 SUs trying to learn the vacancy of channels with a low number of collisions. In this scenario, we show that, based on our policy APL, users reach their dedicated channel faster than several existing policies. In the last scenario, using APL with the QoS-UCB1, users should learn both vacancy and quality of channels and then converge towards channels that have a good vacancy and quality.

In our algorithm, two factors can affect the convergence: α or H while the convergence of UCB1 and ϵ-greedy are affected by α and H respectively. We consider the value of α and H for which UCB1 and ϵ-greedy achieve their best performance. According to [17], the best value of $H = \frac{c \times K}{d^2}$ (i.e. $c = 0.1$ is a constant number, $K = 9$ and $d = \min_i(\mu_1 - \mu_i) = 0.1$) and α are 90 and 2 respectively in order to ensure a balance between the exploration and exploitation phases.

Let us initially consider a SU trying to access 9 channels associated with the following vacancy probabilities:

$$\Gamma = [0.9 \;\; 0.8 \;\; 0.7 \;\; 0.6 \;\; 0.5 \;\; 0.4 \;\; 0.3 \;\; 0.2 \;\; 0.1]$$

Fig. 4 compares the regret of the SU using the three MAB algorithms: TS, UCB1 and ϵ-greedy over 1000 Monte Carlo runs. The simulation outcomes are presented with a shaded region enveloping the average regret. As we can see, the regrets of the 3 MAB algorithms have a logarithmic asymptotic behavior with respect to the number of slots, while TS produces a lower regret for all simulations. That means that the SU can quickly reaches the best channel that offers more opportunities for the user compared to other channels. In the second scenario of our simulation, we evaluate the performance of APL and its ability to make each user selects his dedicated channel after a finite number of time slots. We evaluate the performance of APL compared to the existing learning policies such as Musical Chair and SLK.

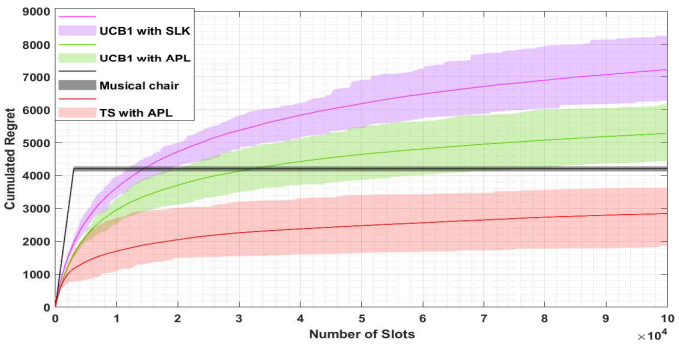


Figure 5: TS, UCB1 with APL compared to SLK and Musical Chairs

To make this comparison, we use two main performance indexes: the regret related to the access of worst channels and the percentage of times to access best channels by each user. A collision may occur when two or more users try to access the same channel. We adopt in our simulations the

ALOHA model, widely used one in OSA, in which none of the collided users receives a reward. After each collision, and based on our policy APL, the collided users should regenerate their rank. First, we consider a static setting of users, then we investigate the dynamic access in which the priority users can enter or leave the network.

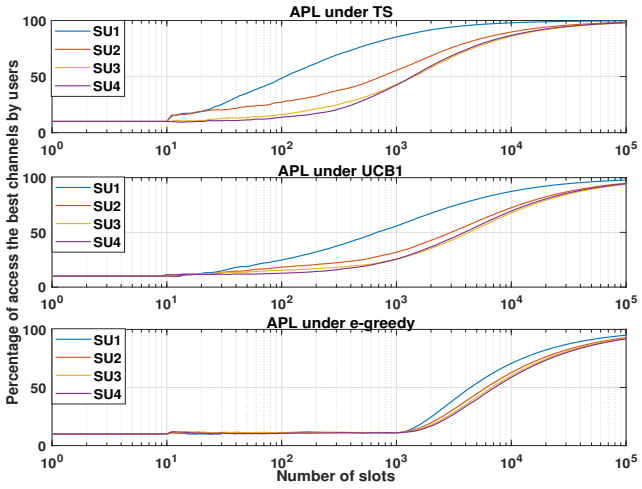


Figure 6: The percentage of times where each SU_k selects its optimal channel using the proposed approach

In Fig. 5, we compare the regret of APL to SLK and Musical Chair. APL and SLK take into consideration the priority access while Musical Chair is proposed for the random access. Despite the regret of APL and SLK has a logarithmic asymptotic behavior, the regret of Musical Chair has two parts:

- A linear part at the beginning, during the learning period, due to the large number of collisions resulting from the random selection.
- A constant part in which the users exploit the U best channels.

As we can see from Fig. 5, APL using TS outperforms Musical Chair and SLK by achieving the lower regret.

Fig. 6 shows the percentage of times that the k-th user accesses his ded- icated channel based on our policy APL up to n, $P_k(n)$. This latter is given by:

$$P_k(n) = \frac{1}{n}\sum_{t=1}^{n} 1_{(\text{if } \beta^l_{APL}(t)=k)} \tag{12}$$

where $\beta^l_{APL}(t)$ represents the channel selected at time t under APL using the learning algorithm l, such as: TS, UCB1 or ϵ-greedy. As we can see, based on our policy APL, the users are able to converge to their targeted channels: SU_1 converges to the best channel μ_1, followed by SU_2 , SU_3 and SU_4 to the channels μ_2 , μ_3 and μ_4 respectively. In addition, we can observe a fast converges of APL using TS compared to TS.

This figure clearly shows that, based on APL, the users converge to their dedicated channels: the first priority user SU_1 converges towards the best channel $\mu_1 = 0.9$, followed by SU_2, SU_3 and SU_4 towards channels $\mu_2 = 0.8$, $\mu_3 = 0.7$ and $\mu_4 = 0.6$ respectively. In addition, we can see that the users quickly reach their dedicated channels using TS and a slow one under UCB1 and ϵ-greedy.

Fig. 7 compares the regret of APL and DMC for the dynamic access where the dotted line indicates the entering and leaving of users on the network. Figures (6a) and (6b) represent respectively the cumulative and average regrets of APL, where at each entering or leaving of users, a significant increase in the regret is observed. It is worth mentioning that, in the dynamic scenario and based on APL, the user can change its current channel for two reasons:

1. When a collision occurs, SU_k should generate a random

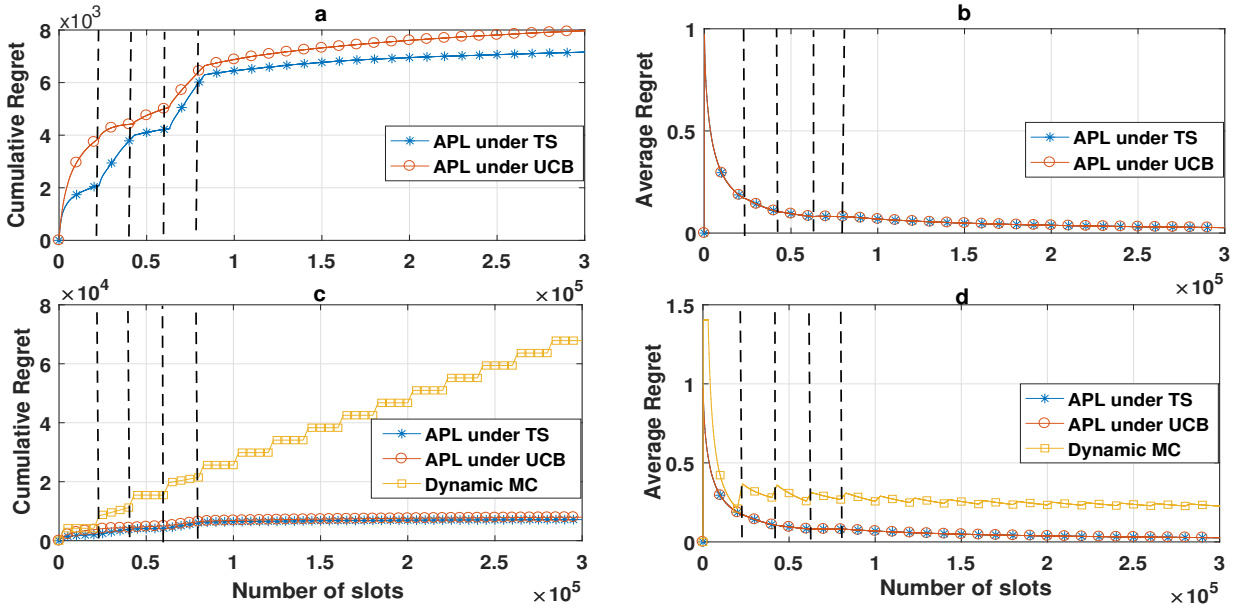


Figure 7: APL and DMC for dynamic access

rank from the set $\{1, ..., k\}$.

2. When a PU reappears in the network and accesses the current channel used by SU_k, the index of this channel decreases, and it may be overwhelmed by another channel that has a low index.

To the best of our knowledge, two policies exist in the literature that consider the dynamic access but without considering priority access: DMC [26] and MEGA [27]. The authors of [26] show that the DMC achieves better performance compared to MEGA policy. In Figures (6c) and (6d), we can see that the performance of APL outperforms the one of DMC and achieves a lower regret. However, after the dynamic access interval, our algorithm achieves a logarithmic regret although the regret of DMC keeps growing with time. Thus, the access under DMC algorithm is realized in epochs, where each one is composed of a learning phase with enough rounds of random exploration to learn the U best channels and the number of users under the dynamic access. The length of an epoch and the learning phase are T_1 and T_0 respectively. These two parameters depend on the number of channels C and the total number of slots n.

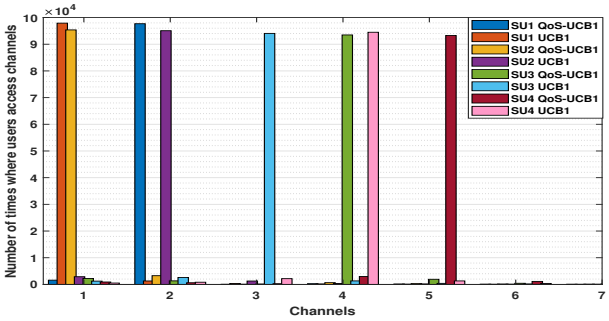


Figure 8: Access channels by priority users using APL

Let us start with the last scenario in which users are able to learn both channels' vacancy and quality using our APL policy where the empirical mean of the quality collected from channels as follows: $G = [0.7\ 0.9\ 0.2\ 0.8\ 0.8\ 0.7\ 0.7\ 0.8\ 0.8]$. Thus, the global mean reward that takes into consideration both quality and vacancy μ_Q is given by: $\mu_Q = [0.63\ 0.72\ 0.14\ 0.48\ 0.4\ 0.28\ 0.21\ 0.16\ 0.08]$.After estimating the channels' availability and quality (i.e. μ_Q) and based on our APL policy with QoS-UCB1, the first priority user SU_1 should converge towards the channel that has the highest global mean, i.e. channel 2, while the target of SU_2, SU_3 and SU_4 should be respectively channels 1, 4 and 5. On the other hand, in the case of APL with UCB1, the target of the priority users SU_1, SU_2, SU_3, and SU_4 should be respectively the channels 1, 2, 3 and 4. This result can be confirmed in Fig. 8, where the priority users access their dedicated channels using APL with QoS-UCB1 or UCB1. Fig. 9 displays the achievable regret of APL with QoS-UCB1 and UCB1 in the multi-user case. Despite the fact that the two curves have a logarithmic asymptotic behavior, we notice an improvement regret of APL with QoS-UCB1 compared to UCB1.

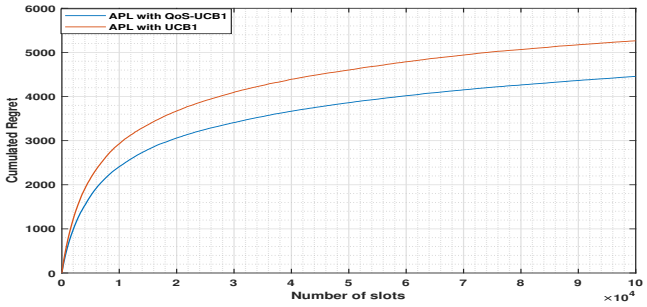


Figure 9: The regret of APL with QoS-UCB1 and UCB1

6 Conclusion

This paper deals with the Opportunistic Spectrum Access (OSA) problem in the context of Cognitive Radio (CR) for a single or multiple Secondary Users (SUs). Recently, several Multi-Armed Bandit (MAB) algorithms have been suggested to help a single SU make a good decision. To tackle the problem of OSA with several SUs, we proposed a novel policy for the priority access called All-Powerful Learning (APL) that allows several SUs to learn separately the channels' vacancy without any cooperation or a prior knowledge about the available channels. Moreover, APL considers the priority dynamic access while only the priority or the dynamic access are separately considered in several recent works, such as Selective Learning of the kth largest expected rewards (SLK), Musical Chairs, Multi-user ϵ-greedy collision Avoiding (MEGA) and $kth - MAB$. In our work, the Quality of Service (QoS) have been also investigated where SU is able to learn both quality and availability of channels and then make an optimal decision with respect to its prior rank. Like most important works in OSA, this work focuses on the Independent Identical Distributed (IID) model in which the state of each channel is supposed to be drawn from an IID process. In future work, we will consider the Markov process as a dynamic memory model to describe the state of available channels, although it is a more complex process compared to IID.

References

[1] M. Marcus, C. Burtle, B. Franca, A. Lahjouji, N. McNeil, "Federal Communications Commission (FCC): Spectrum Policy Task Force," in ET Docket no. 02-135, November 2002.

[2] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, D. Lejeune, "All-Powerful Learning Algorithm for the Priority Access in Cognitive Network," in EUSIPCO, A Corua, Spain, September 2019.

[3] S. Haykin, "Brain-Empowered Wireless Communications," IEEE Journal on Selected Areas in Commun.

[4] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," Biometrika, **25**(3), 285–294, 1933.

[5] T. Lai, H. Robbins, "Asymptotically efficient adaptive allocation rules," Advances in Applied Mathematics, **6**(1), 4–22, 1985.

[6] C. Watkins, Learning from delayed rewards, Ph.D. thesis, University of Cambridge, 1989.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, "The nonstochastic multiarmed bandit problem," SIAM journal on computing, **32**(1), 48–77, 2002.

[8] G. Burtini, J. Loeppky, R. Lawrence, "A survey of online experiment design with the stochastic multi-armed bandit," arXiv preprint arXiv:1510.00757, 2015.

[9] S. Scott, "A modern Bayesian look at the multi-armed bandit," Applied Stochastic Models in Business and Industry, **26**(6), 639–658, 2010.

[10] O. Chapelle, L. Li, "An empirical evaluation of thompson sampling," in Advances in neural information processing systems, Granada, Spain, December 2011.

[11] S. Guha, K. Munagala, "Stochastic regret minimization via Thompson sampling," in Conference on Learning Theory, 317–338, 2014.

[12] T. Kocák, M. Valko, R. Munos, S. Agrawal, "Spectral thompson sampling," in Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[13] I. Osband, D. Russo, B. V. Roy, "(More) efficient reinforcement learning via posterior sampling," in Advances in Neural Information Processing Systems, 3003–3011, 2013.

[14] S. Agrawal, N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in conf. on Learning Theory, Edinburgh, Scotland, June 2012.

[15] E. Kaufmann, N. Korda, R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in International conf. on Algorithmic Learning Theory, Lyon, France, October 2012.

[16] S. Agrawal, N. Goyal, "Further optimal regret bounds for thompson sampling," in Artificial intelligence and statistics, Scottsdale, USA, April 2013.

[17] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multi-armed Bandit Problem," Machine Learning, **47**(2), 235–256, 2002.

[18] E. Kaufmann, O. Cappé, A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in Artificial intelligence and statistics, La Palma, Canary Islands, April 2012.

[19] O. Maillard, R. Munos, G. Stoltz, "A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences," in Annual conf. On Learning Theory, Budapest, Hungary, July 2011.

[20] A. Anandkumar, N. Michael, A. Tang, A. Swami, "Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret," IEEE Journal on Sel. Areas in Com., **29**(4), 731–745, 2011.

[21] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, D.Lejeune, "Distributed Algorithm to Learn OSA Channels Availability and Enhance the Transmission Rate of Secondary Users," in ISCIT, HoChiMinh, Vietnam, September 2019.

[22] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, D. Lejeune, "Distributed algorithm under cooperative or competitive priority users in cognitive networks," EURASIP Journal on Wireless Communications and Networking, **2020**(1), 1–31, 2020.

[23] N. Modi, P. Mary, C. Moy, "QoS driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-armed Bandit Approach," IEEE Trans. on Cog. Com. & Networking, **3**(1), 1–6, 2017.

[24] Y. Gai, B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access," in GLOBECOM, Texas, USA, December 2011.

[25] N. Torabi, K. Rostamzadeh, V. C. Leung, "Rank-optimal channel selection strategy in cognitive networks," in GLOBECOM, California, USA, December 2012.

[26] J. Rosenski, O. Shamir, L. Szlak, "Multi-player bandits-a musical chairs approach," in ICML, New York, USA, June 2016.

[27] O. Avner, S. Mannor, "Concurrent bandit and cognitive radio networks," in European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Nancy, France, September 2014.

[28] K. Liu, Q. Zhao, B. Krishnamachari, "Decentralized multi-armed bandit with imperfect observations," in 2010 48th Annual Allerton Conference on Communication, Control, and Computing, Monticello, USA, October 2010.

[29] Y. Gai, B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access," in GLOBECOM, Texas, USA, December 2011.

[30] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, D. Lejeune, "Distributed Algorithm under Cooperative or Competitive Users with Priority Access in Cognitive Networks," EURASIP journal on wireless communications and networking, (Accepted).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 224-236 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Applicability of Generalized Metropolis-Hastings Algorithm to Estimating Aggregate Functions in Wireless Sensor Networks

Martin Kenyeres[*,1], Jozef Kenyeres[2]

[1]*Institute of Informatics, Slovak Academy of Sciences, Dubravska cesta 9, 845 07 Bratislava 45, Slovak Republic*

[2]*Sipwise GmbH, Europaring F15, 2345 Brunn am Gebirge, Austria*

ARTICLE INFO

Article history:
Received: 09 July, 2020
Accepted: 25 August, 2020
Online: 14 September, 2020

Keywords:
Distributed computing
Data aggregation
Consensus algorithms
Wireless sensor networks
Metropolis-Hastings algorithm

ABSTRACT

Over the last decades, numerous distributed consensus-based algorithms have found a wide application as a complementary mechanism for data aggregation in wireless sensor networks. In this paper, we provide an analysis of the generalized Metropolis-Hastings algorithm for data aggregation with a fully-distributed stopping criterion. The goal of the implemented stopping criterion is to effectively bound the algorithm execution over wireless sensor networks. In this paper, we analyze and compare the performance of the mentioned algorithm with various mixing parameters for distributed averaging, for distributed summing, and for distributed graph order estimation. The algorithm is examined under different configurations of the implemented stopping criterion over random geometric graphs by applying two metrics, namely the mean square error and the number of the iterations for the consensus. The goal of this paper is to examine the applicability of the analyzed algorithm with the stopping criterion to estimating the investigated aggregate functions in wireless sensor networks. In addition, the performance of the algorithm is compared to the average consensus algorithm bounded by the same stopping criterion.

1 Introduction

This paper is an extension of work originally presented in the Proceedings of the EUROCON 2019 - 18th International Conference on Smart Technologies [1].

1.1 Wireless Sensor Networks

Wireless sensor networks (WSNs) are a technology formed by small-size autonomous low-cost sensor nodes equipped with three basic components - a sensing subsystem, a processing subsystem, and a wireless communication subsystem [2]–[4]. These subsystems enable the sensor nodes to concurrently sense the surrounding area, to process the measured data, and to communicate with each other via wireless transmission channels [2, 3]. Moreover, the sensor nodes are supplied by an energy source, allowing WSNs to execute the programmed tasks [3]. WSNs are required to obtain a lot of information about the observed physical phenomenon (e.g., light, motion, temperature, seismic events, humidity, pressure, etc.) and to effectively deliver the measured data to the end-users [3, 5]. Furthermore, they are self-organizing systems typically consisting of many sensor nodes (i.e., hundreds or even thousands) distributed over large-scale areas (either in a random fashion or manually) whereby large geographical territories can be cooperatively monitored with very high precision [2, 6]. As WSNs are generally built-up on an ad-hoc basis, the deployment of the sensor nodes is significantly simplified [7]. In many applications, the sensor nodes are required to operate under inhospitable environmental conditions and without any human interaction; therefore, the design of the algorithms for WSNs has to be affected by these facts [2]. So, WSN-based applications have to be viable and energy-efficient so that the robustness of WSNs to potential threads is high [2]–[4]. Over the last years, WSNs have found the application in various areas such as industrial monitoring, health care, localization, security, structural monitoring, military monitoring, etc. [8].

1.2 Consensus-based Algorithms for Data Aggregation

As the sensor nodes are vulnerable to numerous threats (e.g., temperature, radiation, pressure variations, electromagnetic noise, coverage problems, a high failure rate, etc.) and can measure highly correlated or even duplicated information, complementary algorithms for data

[*]Corresponding Author: Martin Kenyeres, E-mail: martin.kenyeres@savba.sk

aggregation are applied in many WSN-based applications in order to suppress these negative factors and to optimize the overall energy consumption (see Figure 1 for an example demonstrating the difference between WSNs with and without data aggregation) [9, 10]. In addition, data aggregation algorithms can optimize also routing mechanisms, thereby reducing the total energy requirements as well [11]. In the literature, one can find several papers concerned with how to effectively bound algorithms for data aggregation whereby the communication, computation, and energy requirements of WSNs can be optimized [12]–[14]. Note that data aggregation is an important process not only in WSNs but also in other industries [15]–[18].

Over the last decades, consensus-based algorithms for data aggregation have significantly gained the attention of the world-wide scientific community [13, 17, 19]. The problem of consensus achievement poses one of the most fundamental challenges in distributed computing [20]. In WSNs, the term consensus, in general, means the task of getting a group of the sensor nodes to agree on a common value determined by the initial states of all the sensor nodes in this group [21]. The consensus is achieved either asymptotically or in a bounded time interval due to iterative mutual interactions among these sensor nodes [21]. Eventually, each sensor node knows the exact value or an estimate of the wanted aggregate function. Gutierrez-Gutierrez et al. [22] define two categories of consensus-based algorithms for data aggregation, namely the deterministic and the stochastic algorithms. In the literature, one can find many deterministic approaches, e.g., the Maximum Degree weights algorithm, the Metropolis-Hastings algorithm (MH), the Best Constant weights algorithms, the Convex Optimized weights algorithm, etc. [1, 13, 19]. Probably, the most frequently quoted stochastic consensus-based algorithms are the Push-Sum algorithm, the Push-Pull algorithm, the Broadcast gossip algorithm, the Pairwise gossip algorithm, the Geographic gossip algorithm, etc. [14, 22]–[25]. As stated in [4], the algorithms differ from each other in many aspects such as the convergence rate, the robustness to potential threats, the initial configuration, the performance in mobile systems, etc.

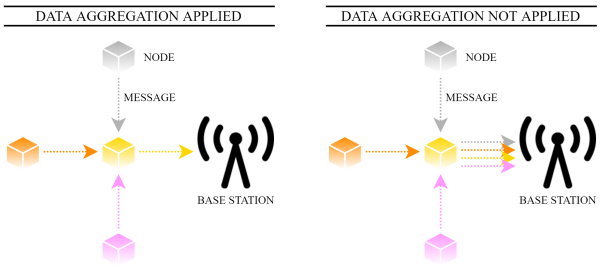


Figure 1: Comparison of WSN with/without data aggregation

1.3 Generalized Metropolis-Hastings for Data Aggregation

In this paper, we focus our attention on MH, which was proposed by Nicolas Metropolis et al. in the 1950s and extended by Wilfred Keith Hastings approximately twenty years later [1]. Since its definition, it has found the application in various areas, e.g., simulating multivariate distributions, block-at-a-time scans, acceptance-rejection sampling, etc. [1]. Lately, Schwarz et al. have defined its generalized variant (referred to as the generalized Metropolis-Hastings algorithm and labeled as GMH) and derived the convergence conditions over arbitrary graphs including critical topologies such as bipartite regular graphs [26]. This algorithm poses a multi-functional distributed linear consensus algorithm able to estimate various aggregate functions, namely the arithmetic mean, the sum, and the graph order. Its weight matrix is symmetric and doubly stochastic, and each graph edge is allocated a weight determined by the degrees of two linked vertices [1]. Thus, this algorithm represents a fully-distributed approach in contrast to many other distributed consensus-based algorithms for data aggregation that require global information for the optimal initial configuration and the proper operating (e.g., the maximum degree, the Laplacian spectrum, etc.) [19]. Therefore, this algorithm (including its original version) has significantly attracted the attention of the world-wide scientific community over the past years [1].

1.4 Summary of Contribution

In this paper, we extend the analysis from [1], where GMH for distributed averaging with the stopping criterion proposed in [27] is analyzed. The extension presented in this paper lies in an analysis of GMH also for distributed summing and distributed graph order estimation. Thus, we compare GMH for estimating different aggregate functions with a various mixing parameter and under various initial configurations of the implemented stopping criterion over 100 unique random geometric graphs (RGGs). In [1], it is identified that GMH can be applied to estimate the arithmetic mean, which motivates us to examine and to compare the applicability of GMH to estimating also other aggregate functions. In the case of distributed summing, we carry out two scenarios - either the initial inner states or the final estimates (i.e., the inner states after the consensus is achieved) are multiplied by the graph order. Thus, the experimental section consists of four scenarios. Finally, GMH for distributed summing and distributed graph order estimation is compared to our previous work where the average consensus algorithm is analyzed [8].

1.5 Paper Organization

The content of this paper is organized as follows:

- *Section 1*: is divided into five subsections:
  - *Section 1.1*: provides basic information about WSNs, i.e., a brief introduction about this technology, its purpose, constraints, and application.
  - *Section 1.2*: is concerned with consensus-based algorithms, i.e., a brief introduction into data aggregation/these algorithms, their purpose, classification, and examples.
  - *Section 1.3*: deals with GMH and introduces its history, application, purpose, and features.
  - *Section 1.4*: is focused on a summary of our contribution presented in this paper.
- *Section 2*: is concerned with a literature review focused on the application of GMH in (not only) WSNs.

- *Section 3*: provides a theoretical insight into the topic and is partitioned into two subsections:
  - *Section 3.1*: introduces the applied mathematical model of WSNs, the definition of GMH, its convergence conditions, and the analyzed functionalities of GMH.
  - *Section 3.2*: deals with the implemented stopping criterion, i.e., why to implement it, its parameters, and its principle.
- *Section 4*: is focused on the applied research methodology and the metrics for performance evaluation.
- *Section 5*: consists of the experimentally obtained results in Matlab2018b (32 figures in overall), an analysis of the depicted results, a comparison of various functionalities of GMH, and a comparison of GMH to the average consensus algorithm.
- *Section 6*: briefly summarizes the outcomes of the research presented in this paper.

2 Literature Review

In this section, we turn our attention to the applicability of MH for various purposes in (not only) WSNs.

In [29], the authors focus their attention on data acquisition in WSNs. They apply MH to determine the mixing time, which is the minimal length of a random walk to approximate a uniform stationary distribution. In the next paper [30], the authors propose an approach for estimating the model parameters for time synchronization in WSNs. Here, they apply MH in combination with Gibbs samples for estimating the parameters of the Bayesian model. The paper [31] is concerned with the tuning parameters of WSNs. In this case, MH is used to determine the acceptance probability. For higher values of the temperature, the algorithm accepts all the moves, meanwhile, stochastic hill-climbing is performed in the case of lower temperatures. The authors of [32] apply a random walk as a routing scheme and compressive sensing to recover raw data in order to optimize the communication costs and detection performance. MH is used to choose neighbors that receive a message. In [33], a novel approach based on MH that synthesizes safe reversible Markov chains is introduced. The authors of [34] propose an approach for evaluating the network loss probability of the mobile collectors for harvesting data in WSNs. As stated in the paper, the optimal movement strategy is achieved in the case of applying MH. In [35], the authors provide a Markov point process model to generate evolving geometric graphs capable of responding to external effects. In the presented model, the vertices can move according to MH-based rules, giving a random pattern of points whose distribution is similar to the Markov point process distribution. In [36], MH finds an application to obtaining a sequence of random samples from an arbitrary distribution. The biggest benefit of applying MH is that this algorithm is independent of the normalization factor. In the paper [37], a novel MH-based algorithm assumed to protect WSNs from internal attacks is proposed and analyzed. In this approach, MH is used to generate the samples from a stationary distribution. In [38], the authors propose a spatial-temporal data gathering mechanism based on MH with delayed acceptance. This approach allows harvesting compressive data by sequentially visiting small subsets of nodes along a routing path. In [13], MH is analyzed over WSNs with embedded computing and communication devices. A centralized stopping criterion based on the mean square error is applied to bound the algorithm execution in this paper. The authors of [39] propose a novel consensus-based algorithm for data aggregation over WSNs by combining MH with the convex optimized weights algorithm, thereby optimizing both the transient and the steady-state algorithm phase. In [40], MH is applied to target tracking in WSNs for surveillance tasks. The proposed approach is to convert binary detections into finer position reports by applying the spatial correlation. The mechanism for a node-specific interference problem appearing in heterogeneous WSNs is proposed in [41]. In this approach, MH is employed in its first step. In [42], it is stated that MH can be used to choose the next nodes in random walks. The paper [43] is focused on a modified MH for estimating the item, structural, and **Q** matrix parameters. In [44], the authors propose a TunaMH method based on MH (as its name evokes) for exposing a tunable tradeoff between its batch size and the convergence rate that is theoretically guaranteed. The authors of [45] propose a novel MH to provide large proposal transitions accepted with a high probability due to a significant increase in the complexity and in the dimension of the interference problems. The authors of [46] address the incipient fault detection problem, which is solved by the proposed two-step technique. In the second step, MH is used to perform the change point detection. In [47], the authors present a received signal strength-based inter-network cooperative localization approach utilizing MH. At the end of this section, as mentioned above, Schwarz et al. define the generalized variant of MH and analyze its convergence in [26].

Thus, it can be seen that our contribution is an analysis of GMH for estimating three aggregation functions with the fully-distributed stopping criterion from [27]. In none of the provided papers, such a deep analysis of GMH for data aggregation with a stopping criterion proposed primarily for WSNs is carried out.

3 Theoretical Background

3.1 Model of GMH for Estimating Average/Sum/Graph Order in WSNs

In this paper, WSNs are modeled as undirected simple finite graphs (with the graph order n) formed by two time-invariant sets, namely the vertex set **V** and the edge set **E** ($G = (\mathbf{V}, \mathbf{E})$), and so the graph order and the graph size are constant [1]. The vertex set **V** gathers all the graph vertices, which represent the sensor nodes in WSN and are distinguishable from each other by the unique index number, i.e., $\mathbf{V} = \{v_i : i = 1,2,...n\}$. The edge set **E** is formed by all the graph edges representing the direct connection between two vertices (v_i and v_j are linked by e_{ij}). Subsequently, the neighbor set of v_i, which is a set consisting of all the neighbors of the corresponding node, can be defined as follows [1]:

$$\mathcal{N}_i = \{v_j : e_{ij} \in \mathbf{E}\} \tag{1}$$

From a global view, GMH can be modeled as follows [26]:

$$\mathbf{x}(k+1) = \mathbf{W} \times \mathbf{x}(k) \tag{2}$$

From a local view, the formula (2) can be cast as follows [1]:

$$x_i(k+1) = [W]_{ii} \cdot x_i(k) + \sum_{j \in \mathcal{N}_i} [W]_{ij} \cdot x_j(k), \quad i = \{1, 2, ...n\} \tag{3}$$

Here, $\mathbf{x}(k)$[1] is a variant column vector containing the inner states of all the sensor nodes in WSN at the corresponding iteration, and $\mathbf{W}$ is the weight matrix defined for GMH as follows [26]:

$$[W]_{ij} = \begin{cases} \frac{1}{max\{d_i, d_j\}+\epsilon}, & \text{if } e_{ij} \in \mathbf{E} \\ 1 - \sum_{i \neq l} [W]_{il}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Here, d_i poses the degree of v_i, and ϵ represents the mixing parameter, which has to take a value from the following interval for the convergence achievement [26]:

$$\epsilon \in [0, 1] \tag{5}$$

Note that the convergence is obtained for $\epsilon = 0$ only unless the underlying graph is bipartite regular [26]. As stated in [26], the adjacency matrix $\mathbf{A}$ of a bipartite graph has a block structure with zero diagonal in the case of $\epsilon = 0$, which implies:

$$\mathbf{W} = \frac{1}{d} \cdot \mathbf{A} = \frac{1}{d} \cdot \begin{pmatrix} 0 & \mathbf{A}_0 \\ \mathbf{A}_0^{\mathrm{T}} & 0 \end{pmatrix} \tag{6}$$

Thus, the sensor nodes are divided into two groups, and their inner states oscillate between two values equal to the arithmetic means determined by the inner states of the sensors nodes in these groups. The mixing parameter $\epsilon = 0$ causes the equality in $\rho(\widetilde{\mathbf{L}})^2 \leq 2$ in the case of the bipartite regular graphs, resulting in the divergence of the algorithm:

$$\rho(\widetilde{\mathbf{L}}) = 2 \implies \rho(\mathbf{W} - \frac{1}{n} \cdot \mathbf{1} \times \mathbf{1}^{\mathrm{T}}) = 1 \tag{7}$$

In Figure 14, we show the evolution of the inner states over a random bipartite regular graph for four values of the mixing parameter ϵ, including $\epsilon = 0$. From the figures, it can be seen that the inner states converge to the arithmetic mean except for $\epsilon = 0$, when the algorithm diverges.

Properly configured GMH operates in such a way that the inner state of each sensor node asymptotically converge to the value of the estimated aggregated function (which determines the value of the steady-state), i.e. [1]:

$$\lim_{k \to \infty} \mathbf{x}(k) = \lim_{k \to \infty} \mathbf{W}^k \times \mathbf{x}(0) = \frac{1}{n} . \mathbf{1} \times \mathbf{1}^{\mathrm{T}} \times \mathbf{x}(0) \tag{8}$$

Here, $\mathbf{1}$ poses an all-ones vector formed by n elements, and $\mathbf{1}^{\mathrm{T}}$ is its transpose. In the following part, examples of how the inner states evolve with an increase in the iteration number for each analyzed functionality of GMH are shown.

[1]Here, k is the label of an iteration, and $\mathbf{x}(k = 0)$ represents the initial inner states

[2]$\widetilde{\mathbf{L}}$ is the weighted Laplacian matrix

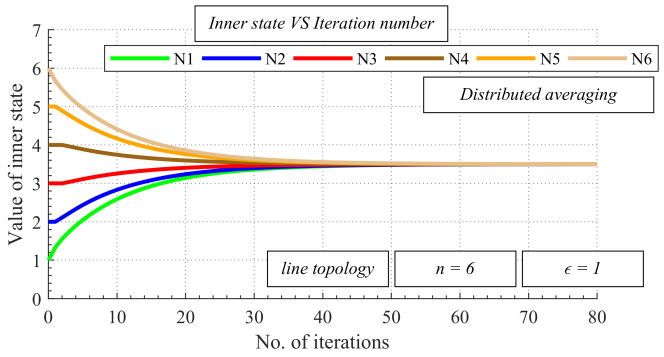


Figure 2: Inner states as function of number of iterations over line topology with graph order $n = 6$ - distributed averaging, mixing parameter $\epsilon = 1$

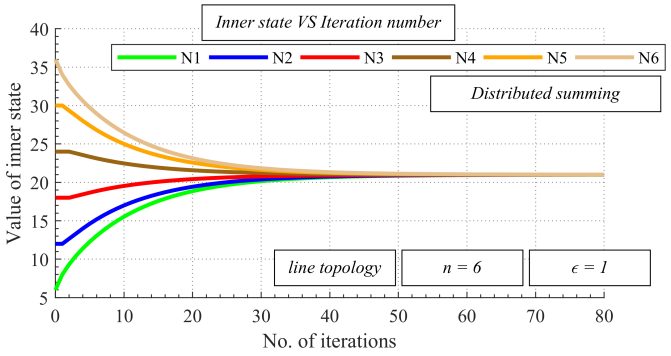


Figure 3: Inner states as function of number of iterations over line topology with graph order $n = 6$ - distributed summing, mixing parameter $\epsilon = 1$

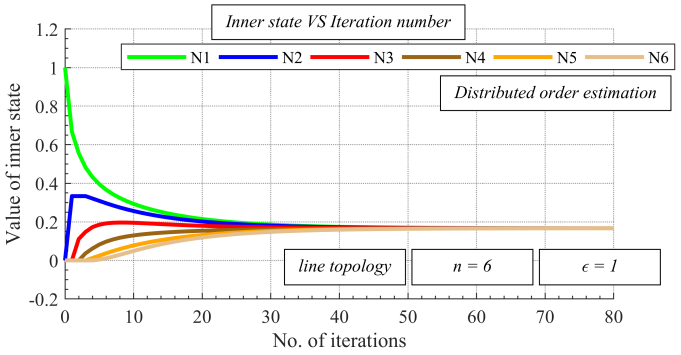


Figure 4: Inner states as function of number of iterations over line topology with graph order $n = 6$ - distributed graph order estimation, mixing parameter $\epsilon = 1$

The existence of the limit (8) is essential for GMH to operate correctly and is ensured, provided that these three convergence conditions are met [26]:

$$\mathbf{1}^{\mathrm{T}} \times \mathbf{W} = \mathbf{1}^{\mathrm{T}} \tag{9}$$

$$\mathbf{W} \times \mathbf{1} = \mathbf{1}, \tag{10}$$

$$\rho(\mathbf{W} - \frac{1}{n} \cdot \mathbf{1} \times \mathbf{1}^{\mathrm{T}}) < 1 \tag{11}$$

Here, $\rho(\cdot)$ is the spectral radius of the corresponding vector/matrix, and its value is equal to the largest eigenvalue in the absolute value from the corresponding spectrum, i.e.:

$$\rho(\cdot) = \max_i \{| \lambda_i(\cdot) |\} \tag{12}$$

Here, $\lambda_i(\cdot)$ is the ith largest eigenvalue from the corresponding spectrum.

As mentioned earlier, GMH can estimate three various aggregate functions, namely the arithmetic mean, the sum, and the graph order. The values of the initial inner states are affected by the estimated aggregate function as follows:

- *Distributed averaging*: the initial inner states are equal to (for example) independent local measurements.
- *Distributed summing*: the initial inner states are equal to (for example) independent local measurements and multiplied by the graph order n - this multiplication is carried out either before the algorithm begins or after completing the algorithm.
- *Distributed graph order estimation*: one of the sensor nodes is selected as the leader, and its initial inner state is equal to "1". The other sensor nodes initiate their inner state with "0". Once the consensus is achieved, the sensor nodes determine the inverted value from the final estimates, which is equal to an estimate of the graph order n.

3.2 Distributed Stopping Criterion for WSNs

In this section, we introduce the implemented stopping criterion for bounding the execution of GMH. As mentioned earlier, an effectively stopped execution may significantly optimize the algorithm in terms of many aspects such as the overall energy consumption, the communication amount, etc., and therefore, a proper initial configuration of the implemented stopping criterion is essential especially in energy-constrained technologies such as WSNs. In this paper, we analyze the fully-distributed stopping criterion from [27], which is determined by two constants:

- *accuracy*
- *counter threshold*

Both constants are the same at each sensor node and have to be set before the algorithm begins. Also, each sensor node has its own *counter*, which is a variable initiated by "0" at each sensor node. The principle of the stopping criterion lies in a comparison of the finite difference between two subsequent inner states at each sensor node with pre-set *accuracy*. If the finite difference is smaller than *accuracy counter threshold*-times in the array, the algorithm is locally completed at the corresponding nodes, and this node does not participate in the algorithm anymore. If not, the value of *counter* is reset regardless of its current state. See Algorithm 1 for the formalization of GMH with the stopping criterion from [27].

Algorithm 1 GMH for distributed averaging/summing/graph order estimation with bounded execution

At the beginning of the algorithm (i.e., $k = 0$), every sensor node $v_i \in \mathbf{V}$ initiates its inner state (i.e., $x_i(0)$) with a scalar value according to the estimated aggregate function:

- *Distributed averaging*: the initial inner states are equal to (for example) independent local measurements.
- *Distributed summing*: the initial inner states are equal to (for example) independent local measurements and multiplied by the graph order n - this multiplication is carried out either before the algorithm begins or after completing the algorithm.
- *Distributed graph order estimation*: one of the sensor nodes is selected as the leader, and its initial inner state is equal to "1". The other sensor nodes initiate their inner state with "0". Once the consensus is achieved, the sensor nodes determine the inverted value from the final estimates.

Also, each sensor node $v_i \in \mathbf{V}$ sets its *counter* to "0". The parameters *counter threshold* and *accuracy* are constant during the algorithm execution and set to the same value at each sensor node. Each sensor node is considered to be active until it locally completes the algorithm.

At the next iterations (i.e., $k = 1, 2, ...$), every active sensor node repeats the following steps as long as it is active:

1. sends a broadcast message containing its current inner state to (1) as well as collecting the inner states from (1).
2. updates its inner state for the next iteration by applying the update rule (2)/(3).
3. verifies whether the finite difference between two subsequent inner states is smaller than *accuracy*, i.e., $|\Delta x_i(k)| < accuracy$. If it is true, it increments its *counter* by "1", otherwise, sets it to "0".
4. verifies the condition: *counter* = *counter threshold*. If it is valid, the corresponding sensor node considers GMH to be locally completed, and this sensor node becomes inactive. If it is invalid, the sensor node repeats the previous steps.

4 Research Methodology

This section is concerned with the applied research methodology and the metrics for performance evaluation.

As already mentioned in this paper, the weight matrix of GMH is modifiable by changing the value of the mixing parameter ϵ. As stated in [26], the convergence is ensured for ϵ taking value from (5) unless the underlying graph is bipartite regular. As it is not too likely that WSNs are bipartite regular [26], we do not include them in the experimental section - GMH diverges over these topologies, which may significantly skew the presented results. Therefore, we separately analyze these critical topologies in Section 3.1. In the experimental section, the mixing parameter ϵ takes the following four values:

- $\epsilon = \{1, 2/3, 1/3, 0\}$

As mentioned above, we apply the stopping criterion from [27] to bound the algorithm. In our experiments, its configuration is selected as follows:

- *accuracy* = $\{10^{-2}, 10^{-4}, 10^{-5}, 10^{-6}\}$
- *counter threshold* = {3, 5, 7, 10, 20, 40, 60, 80, 100}

As mentioned earlier, the performance of GMH with various initial configurations is tested over RGGs with the graph order $n = 200$ - see Figure 5 for their representatives. For the sake of high research credibility, we generate 100 graphs each with a unique topology where the algorithm is examined.

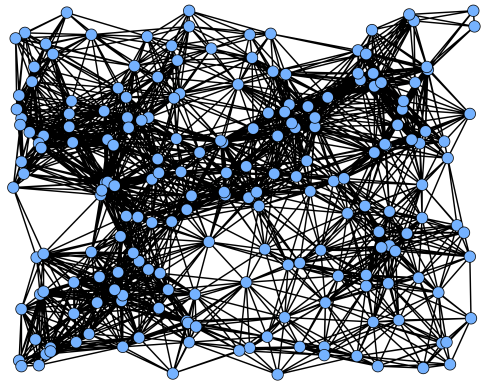


Figure 5: Representative of random geometric graphs

In the experimental part, we analyze GMH estimating three different aggregate functions, and therefore we partition the content of the experimental part into four scenarios as follows:

- ◇ *Scenario 1* - GMH estimates the arithmetic mean from all the initial inner states, which are independent and identically distributed random values of the standard Gaussian distribution, i.e.:

$$x_i(0) \sim N(0,1), \quad i = \{1, 2, ...n\} \tag{13}$$

- ◇ *Scenario 2* - the algorithm determines estimates of the sum calculated from all the initial inner states, which are IDD random values of the same distribution as in the previous scenario. The inner states are multiplied by the graph order n before GMH begins.
- ◇ *Scenario 3* - in this scenario, GMH also estimates the sum from all the initial inner states like in Scenario 2, compared to which the inner states are multiplied by the graph order n after the consensus is achieved.
- ◇ *Scenario 4* - the graph order n, i.e., the size of a network, is estimated in this scenario. One of the vertices is selected as the leader, whose initial inner state is equal to "1", meanwhile, the other ones initiate their initial inner state with "0" (in our analyses, the best-connected node is the leader), i.e.:

$$x_i(0) = \begin{cases} 1, & if\ v_i\ is\ the\ leader \\ 0, & if\ v_i\ is\ not\ the\ leader \end{cases} \tag{14}$$

The inverted value of the inner states is determined after the consensus is achieved.

To evaluate the performance of the algorithm, we apply two metrics:

- *Mean Square Error* (MSE): is used to quantify the precision of the final estimates and defined as follows [48, 49]:

$$\mathrm{MSE} = \frac{1}{n}.\sum_{i=1}^{n}\left(x_i(k_l) - \mathbf{1}^{\mathrm{T}} \times \frac{\mathbf{x}(0)}{n}\right)^2 \tag{15}$$

Here, k_l is the iteration when the consensus is achieved.

- *Convergence rate*: quantifies the algorithm speed (labeled as k_l) and is expressed as the number of the iterations for the consensus achievement among all the sensor nodes in WSN.

In all the shown figures, we depict and analyze the average of both metrics over 100 graphs for each *accuracy*, *counter threshold*, and the mixing parameter ϵ.

5 Experimental Section

In this section, we provide and discuss numerical experiments executed in Matlab2018b. As mentioned above, we extend the analysis of bounded GMH from [1], where this algorithm for distributed averaging is analyzed in terms of the estimation precision and the convergence rate. In this work, the extension lies in an analysis of GMH for distributed summing (either the initial inner states or the final estimates are multiplied by the graph order n) and GMH for graph order estimation. Also, GMH is compared to the average consensus algorithm. The experimental section consists of:

Table 1: List of figures provided in experimental section

Figure	Applied Metric	Scenario	Note
No. 6	MSE	No. 1	Cited from [1]
No. 7	Convergence rate	No. 1	Cited from [1]
No. 8	MSE	No. 2	
No. 9	Convergence rate	No. 2	
No. 10	MSE	No. 3	
No. 11	Convergence rate	No. 3	
No. 12	MSE	No. 4	
No. 13	Convergence rate	No. 4	

The first part is concerned with an MSE analysis of all the functionalities of GMH. From Figure 6, Figure 8, Figure 10, and Figure 12, it is seen that an increase in *counter threshold* and a decrease in *accuracy* result in a decrease in MSE, whereby the precision of the final estimates is increased in each scenario. Also, lower values of the mixing parameter ϵ ensure that the value of MSE is lower regardless of the estimated aggregate function. Furthermore, we can see that the algorithm achieves the highest precision in the case of estimating the graph order (i.e., in Scenario 4). The values of MSE, in this case, are from the range $[-110.7, -43.2]$[3]. The second highest precision is obtained when the arithmetic mean is estimated (Scenario 1). In this case, the values of MSE are from

[3]only MSE for $\epsilon = 0$ are provided - see Table 2 for the results obtained for the other values of the mixing parameter ϵ

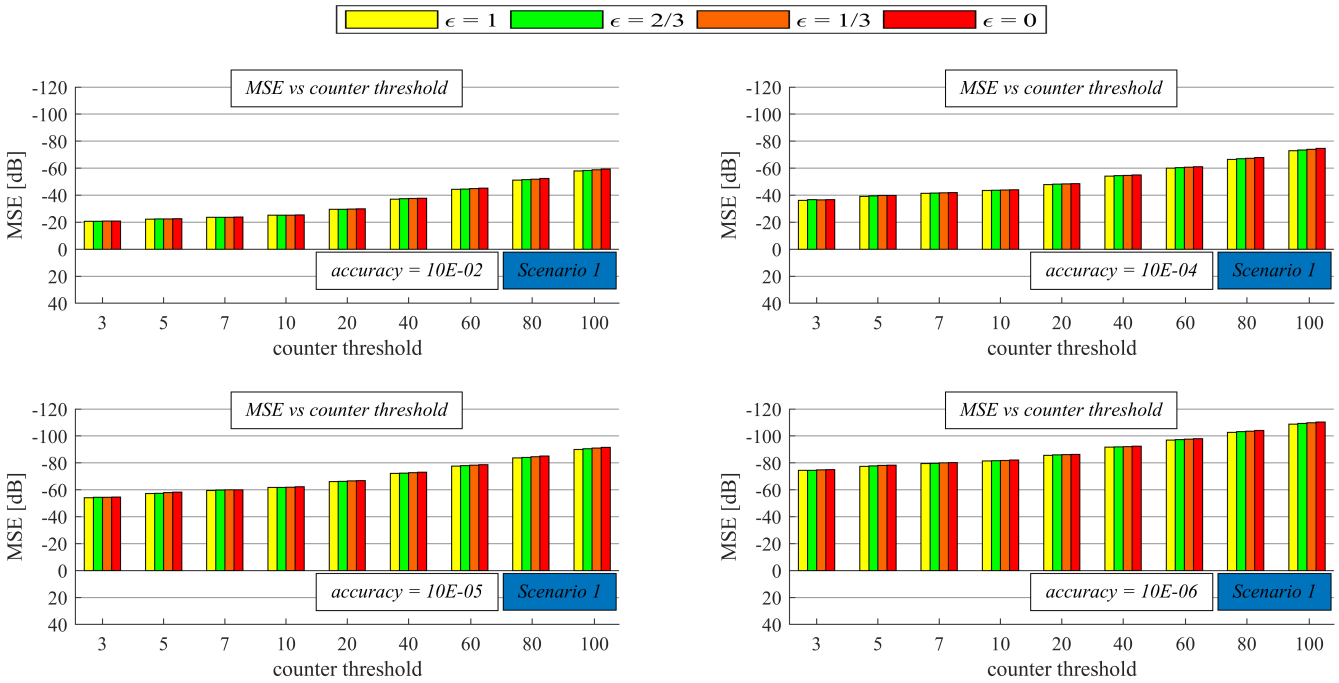


Figure 6: Estimation precision quantified by mean square error for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed averaging – Scenario 1

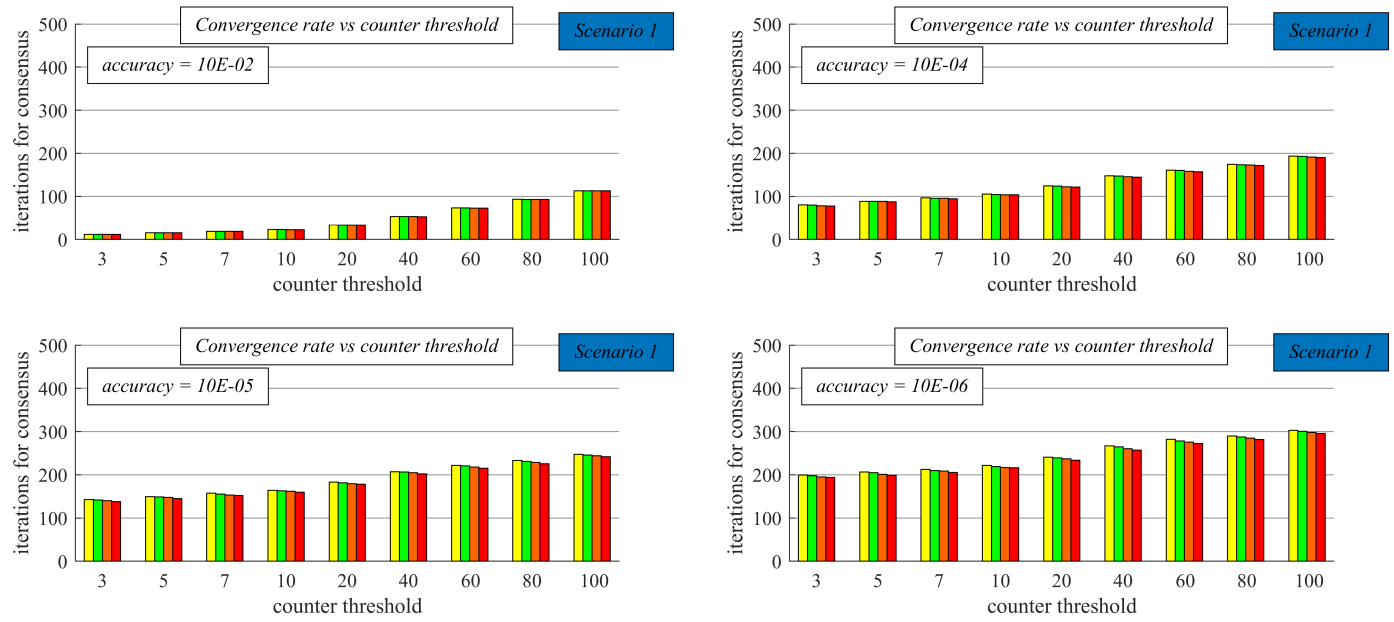


Figure 7: Convergence rate expressed as iteration number required for consensus achievement for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed averaging – Scenario 1

the range $[-110.4, -20.88]^3$. The worst estimation precision is observed in the case of sum estimation. Here, Scenario 2 (MSE is from $[-109.9, 4.288]^3$) outperforms Scenario 3 (MSE is from $[-64.36, 25.14]^3$), proving that multiplying the initial inner states with the graph order n ensures higher precision of the algorithm than multiplying the final estimates. As mentioned earlier, we analyze the average consensus algorithm for distributed summing and distributed graph order estimation in our previous work [8]. Compared to those results, it can be seen that GMH outperforms the average consensus algorithms for estimating both examined aggregate functions. For *counter threshold* = 100, *accuracy* = 10^{-6}, and ϵ ensuring the highest performance, GMH outperforms the average consensus algorithm by 12.78 dB in Scenario 2, by 77.31 dB in Scenario 3, and by 104.469 dB in Scenario 4. Moreover, an important fact

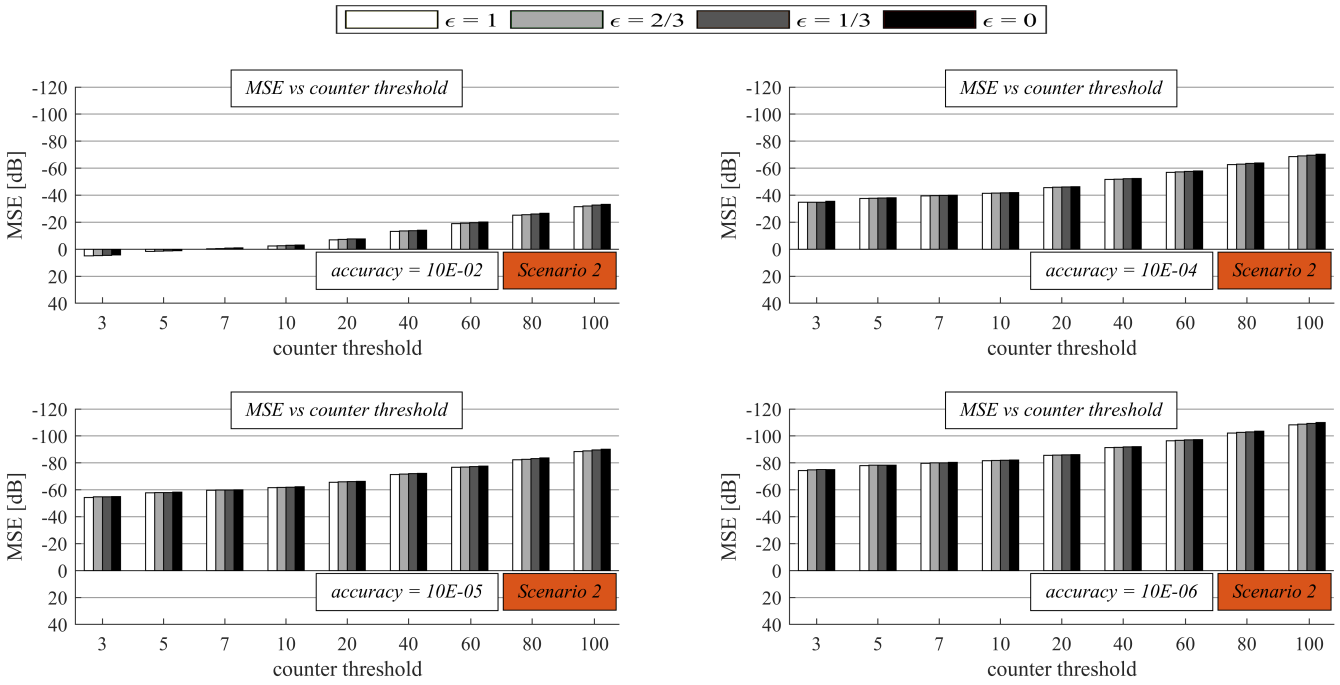


Figure 8: Estimation precision quantified by mean square error for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed summing – Scenario 2

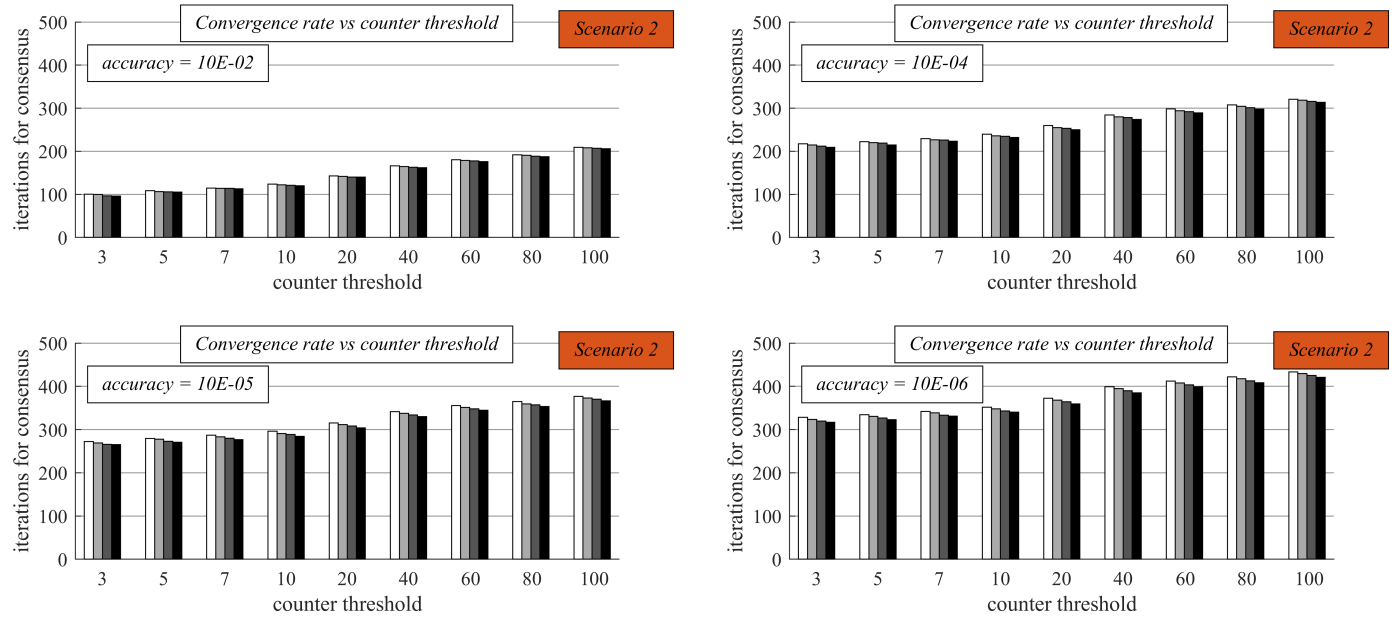


Figure 9: Convergence rate expressed as iteration number required for consensus achievement for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed summing – Scenario 2

identified in [8] is that the average consensus achieves a very low estimation precision for lower values of *counter threshold* and higher values of *accuracy* in the case of graph order estimation, making the algorithm unusable with these configurations of the stopping criterion. However, in this paper, we identify that this phenomenon is not visible in the case of GMH; therefore, this algorithm is much more appropriate for graph order estimation than the average consensus algorithm when the stopping criterion [27] is implemented. In addition, the value of the mixing parameter ϵ has a less intensive impact on MSE compared to the average consensus algorithm for distributed summing/distributed graph order estimation. In the following part, we turn our attention to the convergence rate of GMH expressed as the iteration number for the consensus achievement. From Figure 7, Figure 9, Figure 11, and Figure 13, we can

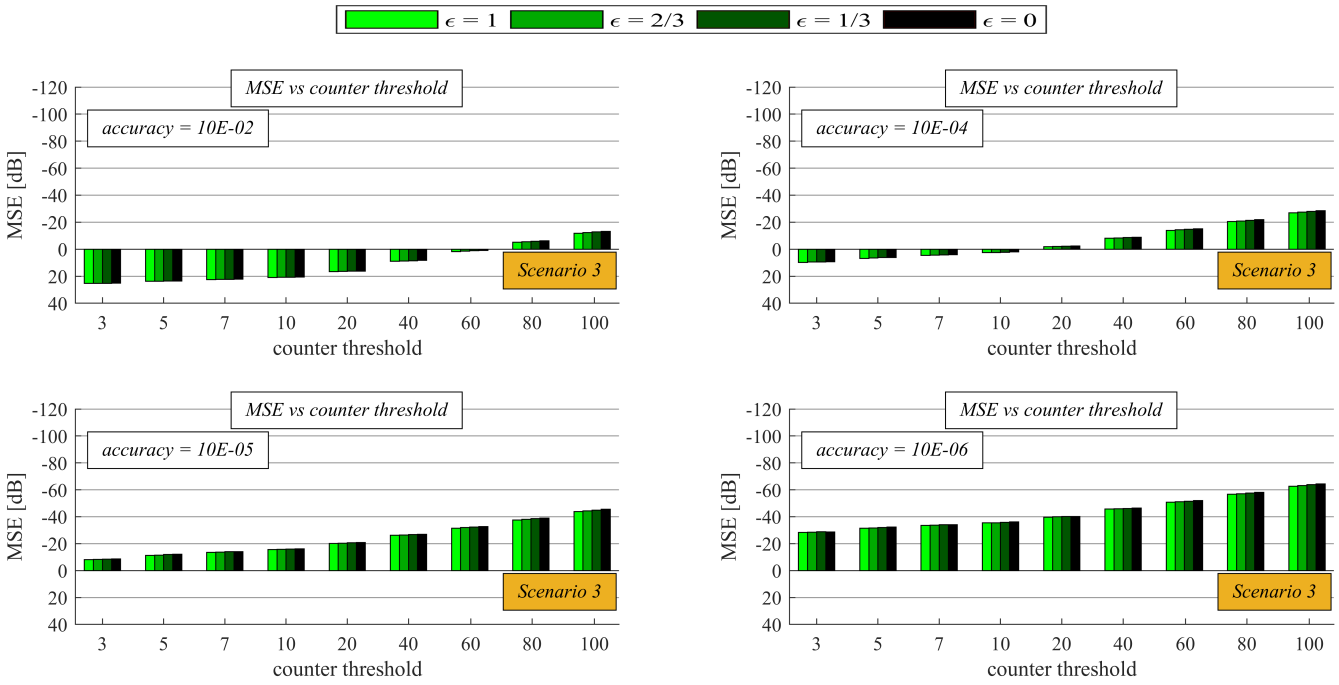


Figure 10: Estimation precision quantified by mean square error for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed summing – Scenario 3

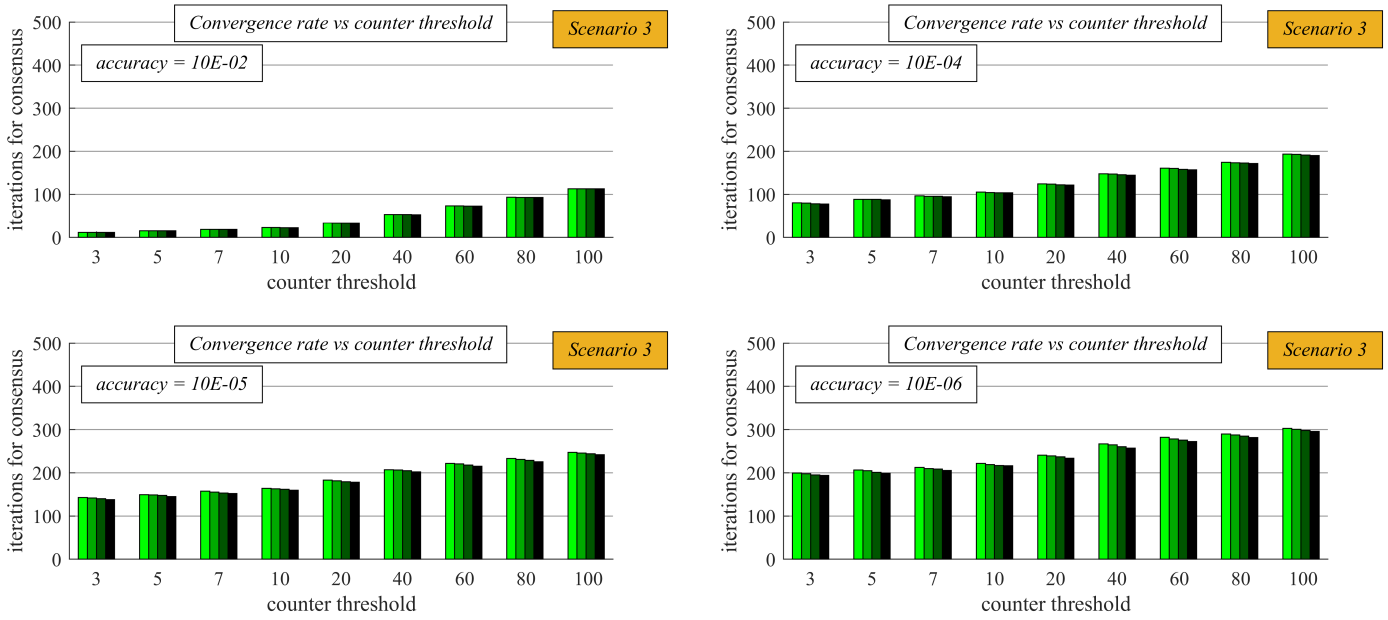


Figure 11: Convergence rate expressed as iteration number required for consensus achievement for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed summing – Scenario 3

see that an increase in *counter threshold* and a decrease in *accuracy* result in an increase in the number of the iterations required for the consensus; therefore, the convergence rate is decreased in each analyzed scenario. Again, a higher convergence rate in each scenario is obtained for lower values of the mixing parameter ϵ. Like in the previous analysis, Scenario 4 outperforms all three other scenarios also in terms of the convergence rate, and the convergence rate in this scenario is from the interval $[5, 240.5]^3$. As furthermore seen, the second-highest convergence rate is achieved by both Scenario 1 and Scenario 3 ($[11.55, 295.80]^3$); therefore, the number of the iterations for the consensus achievement is the same (with a few exceptions) in both scenarios. In contrast to the previous analysis, Scenario 2 ($[95.78, 420.9]^3$) is outperformed by Scenario 3, meaning that multiplying the final estimates ensures a higher convergence rate than multiplying the initial inner states. Compared to our pre-

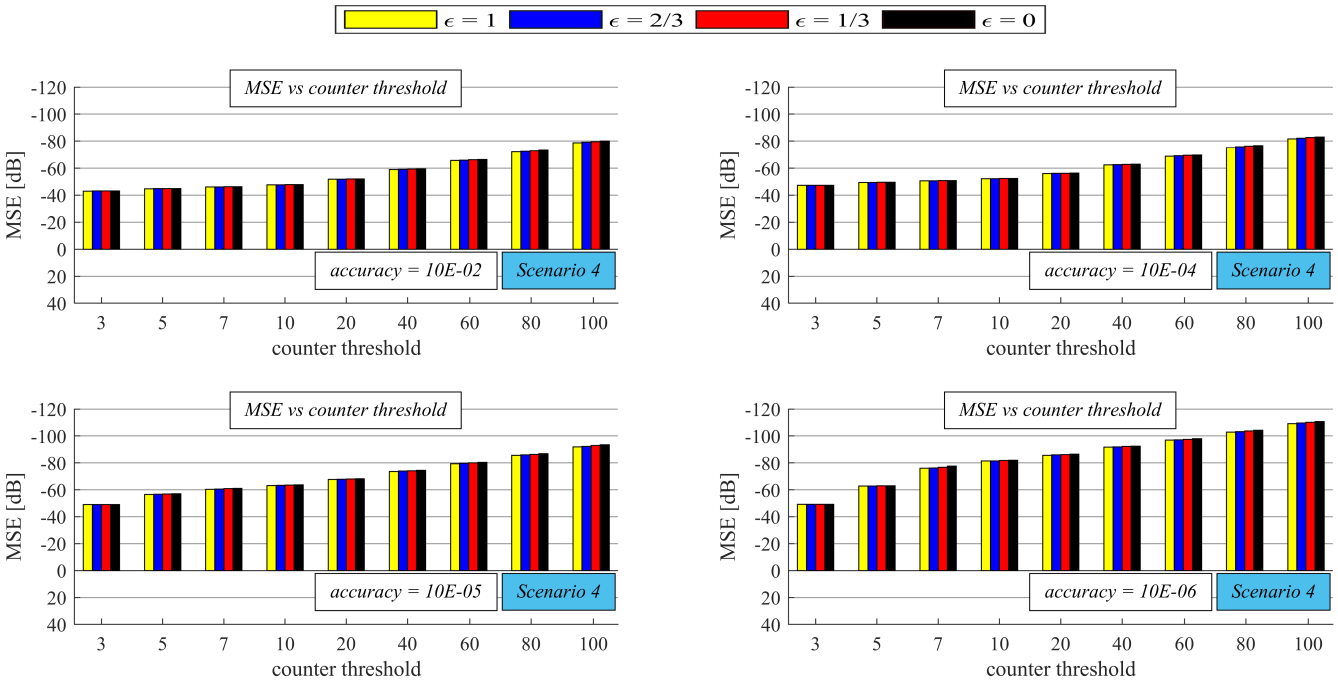


Figure 12: Estimation precision quantified by mean square error for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed graph order estimation – Scenario 4

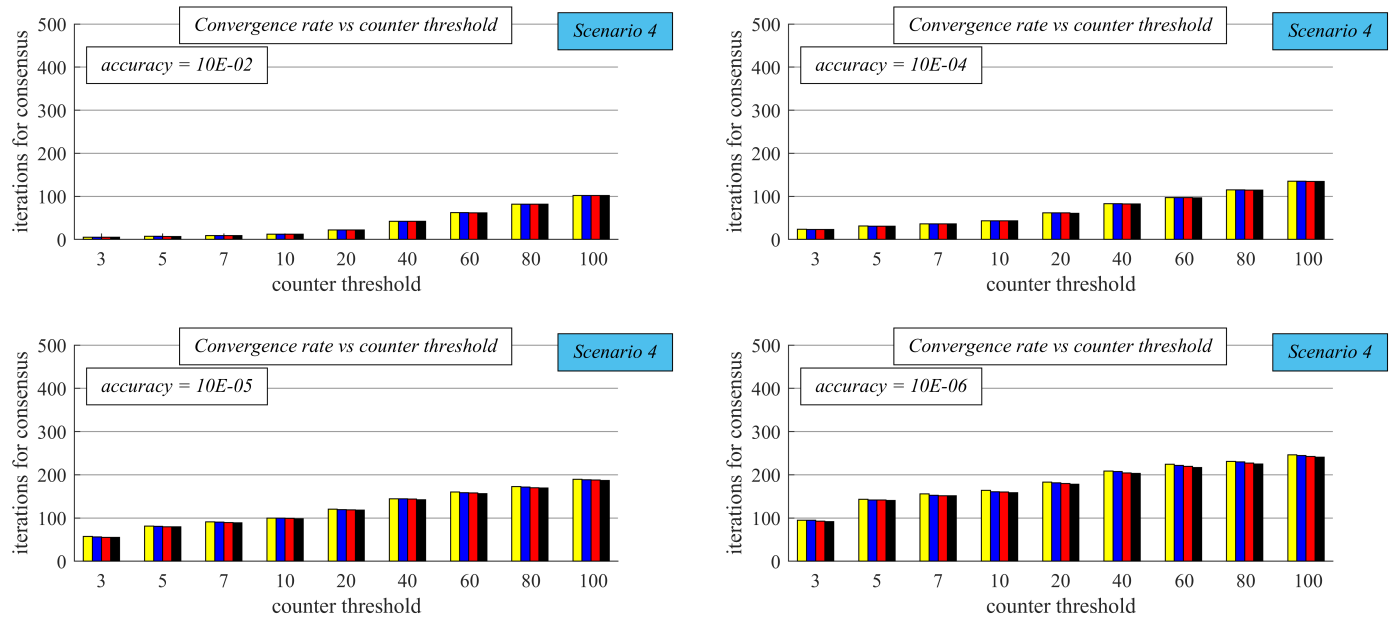


Figure 13: Convergence rate expressed as iteration number required for consensus achievement for various configurations of implemented stopping criterion and different mixing parameter – generalized Metropolis-Hastings for distributed graph order estimation – Scenario 4

vious work, GMH outperforms the average consensus algorithm also in terms of the convergence rate in both distributed summing and distributed graph order estimation. For *counter threshold* = 100, *accuracy* = 10^{-6}, and ϵ ensuring the highest performance, the convergence rate of GMH is greater than the convergence rate of the average consensus algorithm by 139.5 iterations in Scenario 2, by 87.3 iterations in Scenario 3, and by 52.5 iterations in Scenario 4. Also, like in the previous analysis, the value of the mixing parameter ϵ has only a marginal impact on the convergence rate in contrast to the average consensus algorithm examined in [8].

6 Conclusion

In this paper, we provide an extension of the analysis of GMH for distributed averaging over WSNs published in the conference proceedings of IEEE Eurocon 2019. The extension presented in this paper lies in an analysis of GMH for estimating also other aggregate functions, i.e., distributed summing and distributed graph order estimation. In our analyses, the algorithm with a different mixing parameter ϵ is bounded by a distributed stopping criterion with varied initial configurations (i.e., *counter threshold* takes the values (10^{-2}, 10^{-4}, 10^{-5}, 10^{-6}), and *accuracy* takes the values (3, 5, 7, 10, 20, 40, 60, 80, 100)). The experimental part consists of four scenarios (i.e., these functionalities are analyzed: distributed averaging, distributed summing - the initial inner states are multiplied by the graph order n, distributed summing - the final estimates are multiplied by the graph order n, a distributed graph order estimation). From the presented results, it is seen that an increase in *counter threshold* and a decrease in *accuracy* result in higher precision of the algorithm, however, the convergence rate expressed as the iteration number of the consensus achievement is decreased as a consequence. Thus, regardless of the estimated aggregate function, the highest precision of the final estimates and the lowest convergence rate is ensured for the configuration *accuracy* = 10^{-6} and *counter threshold* = 100, and vice versa the configuration *accuracy* = 10^{-2} and *counter threshold* = 3 results in the highest convergence rate and the lowest precision. Also, we identify that the optimal performance in terms of the estimation precision and the convergence rate is achieved for $\epsilon = 0$ (i.e., the lower bound of the mixing parameter ϵ), and an increase in ϵ causes lower precision and lower convergence rate. However, the value of the mixing parameter ϵ does not have an essential impact on the algorithm performance according to both applied metrics. In the experimental section, also the performances of GMH for estimating the arithmetic mean, the sum, and the graph order are compared to each other. The best performance is achieved in the case of graph order estimation in terms of both the estimation precision and the convergence rate (for $\epsilon = 0$, the values of MSE lie in $[-43.2, -110.7]$, and the convergence rate in $[5, 240.5]$). The second-best performance is achieved when the arithmetic mean is estimated (for $\epsilon = 0$, MSE is from $[-20.88, -110.4]$, and the convergence rate from $[11.55, 295.8]$). Thus, the lowest precision and convergence rate are achieved in the case of sum estimation. Here, multiplying the initial inner states instead of the final estimates ensures higher precision of the algorithm (for $\epsilon = 0$, MSE $[-109.9, 4.29]$, the convergence rate $[95.78, 420.9]$), meanwhile, multiplying the final estimates results in a higher convergence rate (for $\epsilon = 0$, MSE $[-64.36, 25.14]$, the convergence rate $[11.55, 295.8]$). The most important fact identified in this paper is that GMH achieves a high estimation precision regardless of the estimated aggregation function; therefore, this research identifies its applicability to estimating also two other aggregate functions in WSNs. In addition, we also compare GMH for distributed summing and distributed graph order estimation to the average consensus algorithm, thereby identifying that GMH is more appropriate for data aggregation over WSNs with the implemented stopping criterion than the average consensus algorithm in both applied metrics - the estimation precision and the convergence rate expressed as the iteration number necessary for the consensus achievement. GMH outperforms the average consensus (in the case that the performances of the most precise configurations are compared) by 12.78 dB and 139.5 iterations in Scenario 2, by 77.31 dB and 87.3 iterations in Scenario 3, and by 104.469 and 52.5 iterations in Scenario 4.

Thus, the final conclusion of the presented research is that the lower bound of mixing parameter ϵ ensures the highest performance in both the estimation precision and the convergence rate - so, in real-world implementations, $\epsilon = 0$ is recommended for use. Also, an increase in *counter threshold* and a decrease in *accuracy* cause the algorithm to be more precise but, on the other hand, decelerated regardless of the estimated aggregate function, meaning that the initial configuration of the stopping criterion has to be affected by the fact whether a WSN-based application requires more precise or faster data aggregation. The most important identified fact is that GMH is applicable to estimating all three examined aggregate functions in WSNs.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work was supported by the VEGA agency under the contract No. 2/0155/19 and by COST: Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO) CA 15140. Since 2019, Martin Kenyeres has been a holder of the Stefan Schwarz Supporting Fund.

References

[1] M. Kenyeres, J. Kenyeres, "Applicability of Generalized Metropolis-Hastings Algorithm in Wireless Sensor Networks," in 18th International Conference on Smart Technologies (EUROCON 2019), 1–5, 2019, doi:10.1109/EUROCON.2019.8861554.

[2] A. Al-Karaki, A. E. Kamal, "Routing Techniques in Wireless Sensor networks: A Survey," Secur. Netw, **11**(6), 6–28, 2004, doi:10.1109/MWC.2004.1368893.

[3] G. Anastasi, M. Conti, M. Di Francesco, A. Passarella, "Energy conservation in wireless sensor networks: A survey," Ad Hoc Netw., **7**(3), 537–568, 2009, doi:10.1016/j.adhoc.2008.06.003.

[4] M. Kenyeres, J. Kenyeres, "Average Consensus over Mobile Wireless Sensor Networks: Weight Matrix Guaranteeing Convergence without Reconfiguration of Edge Weights," Sensors, **20**(13), 3677, 2020, doi:10.3390/s20133677.

[5] M. A. Mahmood, W. K. G. Seah, I. Welch, "Reliability in wireless sensor networks: A survey and challenges ahead," Comput. Netw., **79**, 166–187, 2015, doi:10.1016/j.comnet.2014.12.016.

[6] W. Ye, J. Heidemann, D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," in 21st Annual Joint Conference of the IEEE-Computer-and-Communications-Societies Location (INFOCOM 2002), 1567–1576, 2002, doi:10.1109/INFCOM.2002.1019408.

[7] B. Rashid, M. H. Rehmani, "Applications of wireless sensor networks for urban areas: A survey," J. Netw. Comput. Appl., **60**, 192–219, 2016, doi:10.1016/j.jnca.2015.09.008.

[8] M. Kenyeres, J. Kenyeres, "Distributed Network Size Estimation Executed by Average Consensus Bounded by Stopping Criterion for Wireless Sensor Networks," in 24th International Conference on Applied Electronics (AE 2019), 1–6, 2019, doi: 10.23919/AE.2019.8867009.

[9] D. Izadi, J. H. Abawajy, S. Ghanavati, T. Herawan, "A data fusion method in wireless sensor networks," Sensors, **15**(2), 2964–2979, 2015, doi:10.3390/s150202964.

[10] B. Yu, J. Li, Y. Li, T. Herawan, "Distributed data aggregation scheduling in wireless sensor networks," in 28th Conference on Computer Communications (IEEE INFOCOM 2009), 2159–2167, 2009, doi:10.1109/INFCOM.2009.5062140.

[11] L. Krishnamachari, D. Estrin, S. Wicker, "The impact of data aggregation in wireless sensor networks," in 22nd International Conference on Distributed Computing Systems Workshops (ICDCSW 2002), 575–578, 2002, doi:10.1109/ICDCSW.2002.1030829.

[12] N. Al-Nakhala, R. Riley, T. Elfouly, "Distributed algorithms in wireless sensor networks: An approach for applying binary consensus in a real testbed," Comput. Netw., **79**(14), 30–38, 2015, doi:10.1016/j.comnet.2014.12.011.

[13] G. Stamatescu, I. Stamatescu, D. Popescu, "Consensus-based data aggregation for wireless sensor networks," Control. Eng. Appl. Inf., **19**(2), 43–50, 2017, doi:10.1016/j.comnet.2014.12.011.

[14] D. Kempe, A. Dobra, J. Gehrke, "Gossip-based computation of aggregate information," in 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2003), 482–491, 2003.

[15] K. Kenda, B. Kazic, E. Novak, D. Mladenic, "Streaming data fusion for the internet of things," Sensors, **19**(8), 1955, 2019, doi:10.3390/s19081955.

[16] G. B. Markovic, V. S. Sokolovic, M. L. Dukic, "Distributed hybrid two-stage multi-sensor fusion for cooperative modulation classification in large-scale wireless sensor networks," Sensors, **19**(19), 4339, 2019, doi:10.3390/s19194339.

[17] D. Merezeanu, M. Nicolae, "Consensus control of discrete-time multiagent systems," U. Politeh. Buch. Ser. A, **79**(1), 167–174, 2017.

[18] V. Oujezsky, D. Chapcak, T. Horvath, P. Munster, "Security testing of active optical network devices," in 42nd International Conference on Telecommunications and Signal Processing (TSP 2019), 9–13, 2019 doi:10.1109/TSP.2019.8768811.

[19] L. Xiao, S. Boyd, "Fast linear iterations for distributed averaging," Syst. Control. Lett., **53**(1), 65–78, 2004, doi:10.1016/j.sysconle.2004.02.022.

[20] M. J. Fischer, N. A. Lynch, M. S. Paterson, "Impossibility of Distributed Consensus with One Faulty Process," J. ACM, **32**(2), 374–382, 1985, doi:10.1145/3149.214121.

[21] Y. Zheng, Y. Zhu, L. Wang, "Consensus of heterogeneous multi-agent systems," IET Control. Theory Appl., **5**(16), 1881–1888, 2011, doi:10.1049/iet-cta.2011.0033.

[22] J. Gutierrez-Gutierrez, M. Zarraga-Rodriguez, X. Insausti, "Analysis of known linear distributed average consensus algorithms on cycles and paths," Sensors, **18**(4), 968, 2018, doi:10.3390/s18040968.

[23] S. S. Kia, B. Van Scoy, J. Cortes, R. A. Freeman, K. M. Lynch, S. Martinez, "Tutorial on Dynamic Average Consensus: The Problem, Its Applications, and the Algorithms," IEEE Control Syst., **39**(3), 40–72, 2019, doi:10.1109/MCS.2019.2900783.

[24] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, A. Scaglione, "Broadcast gossip algorithms for consensus," IEEE Control Syst., **57**(7), 2748–2761, 2009, doi:10.1109/TSP.2009.2016247.

[25] J. C. S. Cardoso, C. Baquero, P. S. Almeida, "Probabilistic estimation of network size and diameter," in 2009 4th Latin-American Symposium on Dependable Computing (LADC 2009), 33–40, 2009, doi:10.1109/LADC.2009.19.

[26] V. Schwarz, G. Hannak, G. Matz, "On the convergence of average consensus with generalized metropolis-hasting weights," in 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), 5442–5446, 2014, doi:10.1109/ICASSP.2014.6854643.

[27] J. Kenyeres, M. Kenyeres, M. Rupp, P. Farkas, "WSN implementation of the average consensus algorithm," in 17th European Wireless Conference 2011 (EW 2011), 139–146, 2011.

[28] M. Sartipi, R. Fletcher, "Energy-efficient data acquisition in wireless sensor networks using compressed sensing," in 2011 Data Compression Conference (DCC 2011), 223–232, 2011, doi:10.1109/DCC.2011.29.

[29] A.Chatterjee, P. Venkateswaran, "An efficient statistical approach for time synchronization in wireless sensor networks," Int. J. Commun. Syst., **29**(4), 722–733, 2016, doi:10.1002/dac.2944.

[30] A. Munir, A. Gordon-Ross, S. Lysecky, R. Lysecky, "Online algorithms for wireless sensor networks dynamic optimization," in 2012 IEEE Consumer Communications and Networking Conference (CCNC'2012), 180–187, 2012, doi: 10.1109/CCNC.2012.6181082.

[31] M. Nguyen, Q. Cheng, "Efficient data routing for fusion in wireless sensor networks," Signal, **1000**(8), 11–15, 2012.

[32] M. El Chamie, B. Açıkmeşe, "Safe Metropolis-Hastings algorithm and its application to swarm control," Syst. Control. Lett., **111**, 40–48, 2018, doi:10.1016/j.sysconle.2017.10.006.

[33] A. Vijayalakshmi, V. Bhuvaneswari, "Mobile agent based optimal data gathering in Wireless Sensor Networks," in 10th International Conference on Intelligent Systems and Control (ISCO 2016), 1–4, 2016, doi:10.1109/ISCO.2016.7726974.

[34] D. J. Irons, J. Jordan, "Geometric networks based on Markov point processes with applications to mobile sensor networks," in 2013 2nd International Conference on Mechanical Engineering, Industrial Electronics and Informatization (MEIEI 2013), 1–17, 2013.

[35] N. Ramakrishnan, N. Ertin, R. L. Moses, "Assumed density filtering for learning Gaussian process models," in 2011 IEEE Statistical Signal Processing Workshop (SSP 2011), 257–260, 2011, doi:10.1109/SSP.2011.5967674.

[36] M. R. Ahmed, H. Cui, X. Huang, "Smart integration of cloud computing and MCMC based secured WSN to monitor environment," in 2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace and Electronic Systems (VITAE 2014), 1–5, 2014, doi:10.1109/VITAE.2014.6934449.

[37] H. Zheng, J. Li, X. Feng, W. Guo, Z. Chen, N. Xiong, "Spatial-temporal data collection with compressive sensing in mobile sensor networks," Appl. Soft Comput., **17**(11), 2575, 2017, doi:10.3390/s17112575.

[38] V. Schwarz, G. Matz, "Nonlinear average consensus based on weight morphing," in 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), 3129–3132, 2012, doi:10.1109/ICASS.2012.6288578.

[39] S. Oh, P. Chen, M. Manzo, S. Sastry, "Instrumenting wireless sensor networks for real-time surveillance," in 2006 IEEE International Conference on Robotics and Automation (ICRA 2006), 3128–3133, 2006, doi:10.1109/ROBOT.2006.1642177.

[40] L. Martino, J. Plata-Chaves, F. Louzada, "A Monte Carlo scheme for node-specific inference over wireless sensor networks," in 2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2016), 1–5, 2016, doi:10.1109/SAM.2016.7569731.

[41] M. T. Nguyen, K. A. Teague, "Compressive sensing based random walk routing in wireless sensor networks," Ad Hoc Netw., **54**, 99–110, 2017, doi:10.1016/j.adhoc.2016.10.009.

[42] C. W. Liu, B. Andersson, A. Skrondal, "A Constrained Metropolis–Hastings Robbins–Monro Algorithm for Q Matrix Estimation in DINA Models," Psychometrika, **85**(2), 322-357, 2020, doi:10.1007/s11336-020-09707-4.

[43] R. Zhang, A. F. Cooper, C. De Sa, "Asymptotically Optimal Exact Minibatch Metropolis-Hastings," arXiv preprint arXiv:2006.11677, 2020.

[44] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, J. C. Pesquet, "Majorize-Minimize Adapted Metropolis-Hastings Algorithm," IEEE Trans. Signal Process., **68**, 2356–2369, 2020, doi:10.1109/TSP.2020.2983150.

[45] M. F. D'Angelo, R. M. Palhares, R. H. Takahashi, R. H. Loschi, L. M. Baccarini, W. M. Caminhas, "Incipient fault detection in induction machine stator-winding using a fuzzy-Bayesian change point detection approach," Appl. Soft Comput., **11**(1), 179–192, 2011, doi:10.1016/j.sysconle.2017.10.006.

[46] S. Büyükçorak, G. K. Kurt, A. Yongaçoğlu, "Inter-network cooperative localization in heterogeneous networks with unknown transmit power," in 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2017), 1–5, 2017, doi:10.1109/PIMRC.2017.8292505.

[47] V. Skorpil, J. Stastny, "Back-propagation and K-means algorithms comparison," in 8th International Conference on Signal Processing (ICSP 2006), 1871–1874, 2006, doi:10.1109/ICOSP.2006.345838.

[48] S. S. Pereira, A. Pages-Zamora, "Mean square convergence of consensus algorithms in random WSNs," IEEE Trans. Signal Processs, **58**(5), 2866–2874, 2010, doi:10.1109/TSP.2010.2043140.

Appendix

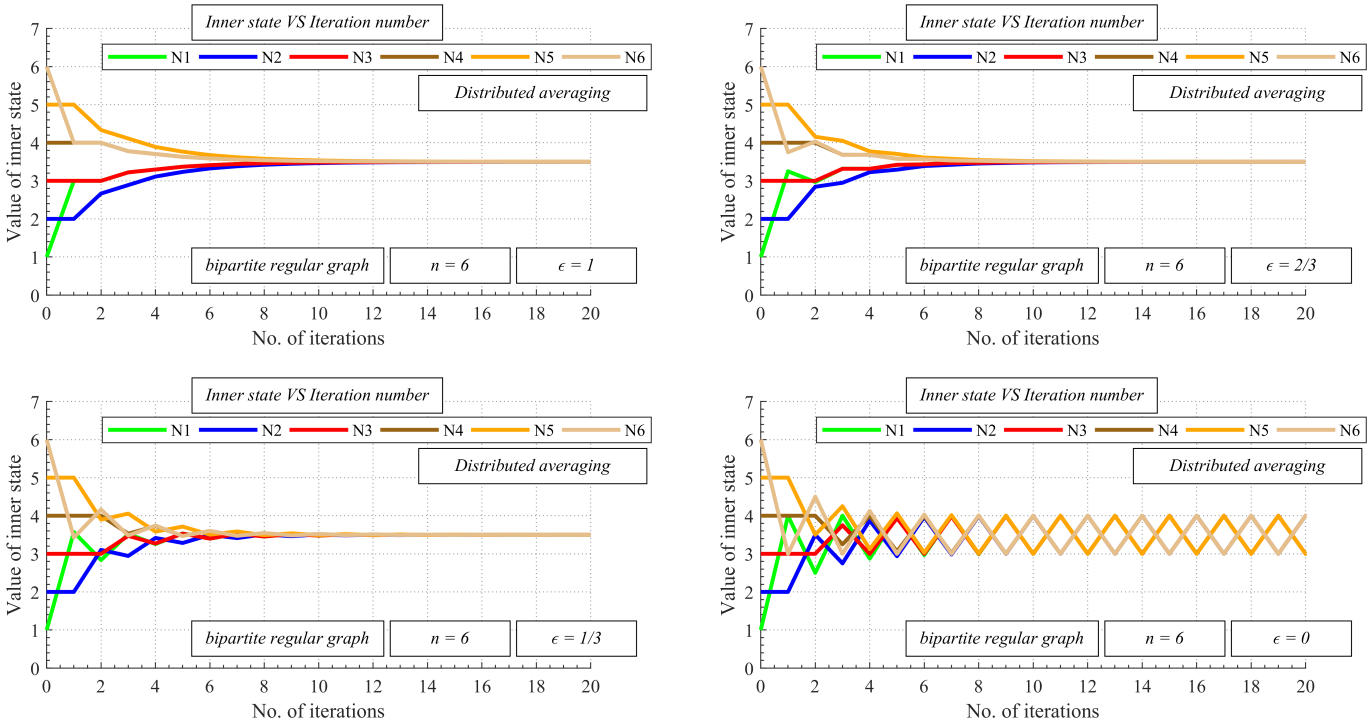


Figure 14: Evolution of inner states over bipartite regular graph for four values of mixing parameter ϵ

Table 2: Tables summarizing presented experimental results

Maximal and minimal precision of GMH quantified by MSE								
	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Max	Min	Max	Min	Max	Min	Max	Min
$\epsilon = 1$	-108.7 dB	-20.69 dB	-108.3 dB	4.89 dB	-62.68 dB	25.33 dB	-109.1 dB	-43.03 dB
$\epsilon = 2/3$	-109.2 dB	-20.73 dB	-108.8 dB	4.68 dB	-63.22 dB	25.29 dB	-109.6 dB	-43.08 dB
$\epsilon = 1/3$	-109.8 dB	-20.8 dB	-109.4 dB	4.52 dB	-63.78 dB	25.22 dB	-110.2 dB	-43.14 dB
$\epsilon = 0$	-110.4 dB	-20.88 dB	-109.9 dB	4.29 dB	-64.36 dB	25.14 dB	-110.7 dB	-43.2 dB

Maximal and minimal convergence rate of GMH quantified by iteration number								
	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Max	Min	Max	Min	Max	Min	Max	Min
$\epsilon = 1$	11.66 ite.	302.8 ite.	100.2 ite.	433.6 ite.	11.66 ite.	302.8 ite.	5.13 ite.	246 ite.
$\epsilon = 2/3$	11.61 ite.	300.5 ite.	99.87 ite.	429.6 ite.	11.61 ite.	300.5 ite.	5.09 ite.	244.5 ite.
$\epsilon = 1/3$	11.57 ite.	298.1 ite.	96.68 ite.	425.4 ite.	11.57 ite.	298.1 ite.	5.02 ite.	242.5 ite.
$\epsilon = 0$	11.55 ite.	295.8 ite.	95.78 ite.	420.9 ite.	11.55 ite.	295.8 ite.	5 ite.	240.5 ite.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 224-229 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Sensorless Control and Corrected Error Commutation of the Brushless DC Motor

Anatoly Stepanov*, Vitaly Enin

Department of Electrical and Industrial Electronics, Bauman Moscow State Technical University, Moscow, 105005, Russia

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 29 December, 2020
Online: 15 January, 2021

Keywords:
Sensorless control
Error commutation
Brushless DC motor
Simulation

ABSTRACT

This paper presents a simple method for correcting the error of commutation windings for sensorless control of a brushless DC motor with a small inductance. Switching error analysis is performed based on the phase currents of the switched-off and switched-on phases. To correct the commutatiuon signals for the inverter, a speed-independent function is used, calculated from the back-EMF, by selecting its coefficients. The back-EMF is calculated for the system obtained by transformed a three-phase system to a two-phase one, which reduces the system dimension. An algorithm for start-up the motor with sensorless control based on the method of align and damping the rotor vibrations in a stable equilibrium position is proposed. The back-EMF is approximated by a function based on a piecewise linear function and an inscribed ellipse, for a more accurate description of the shape of the back-EMF. This approximation is used when simulation the motor. The simulation confirms the effectiveness of the proposed method for correcting the commutation error in the case of sensorless control of a low-power BLDC motor.

1. Introduction

This paper is an extension of conference report "Sensorless control of low-power brushless DC motor based on the use of back-EMF," presented in "2019 International Ural Conference on Electrical Power Engineering (UralCon 2019)" [1].

In the extended version of the paper, in comparison with the conference paper, a method for correcting error of commutation winding and the start-up procedure based on the alignment of the rotor position in the stable equilibrium position of brushless DC motor is proposed, which makes it possible to improve the reliability and quality of sensorless control.

The rectifier of the electrical motors, in particular the brushless DC motors (BLDC) are widely used in automation and control systems, as well as in orientation and navigation systems as a gyroscope drive [2 – 4]. Usually, the power of the BLDC windings is produced by a semiconductor inverter, whose transistors are switched by signals from sensors installed on the motor stator.

The use of rotor position sensors has disadvantages associated with an increase in the price of the electric drive, the accuracy of their installation, which leads to errors in determining the switching moments of the windings, unreliability of the sensors in difficult environmental conditions. To eliminate these shortcomings, the methods of sensorless control of the BLDC motor have been developed.

In case of sensorless control, commutation signals are generated using back-EMF or flow, calculated from phase currents and voltages. The most widely used methods are those using back-EMF.

The most widely used methods of sensorless control based on the analysis of back-EMF are considered in [5-11]. If the neutral point of the motor is not available, the terminal voltage determines the moment when the back-EMF of the disconnected phase passes through zero (zero crossing point). Switching must be performed with a delay of 30 electrical degrees, after passing the back-EMF through zero. In [5], the least squares method is used to refine the switching points, but this does not guarantee accurate determination of the switching points, since the terminal voltage contains noise and pulses generated by the inverter. The implementation of these methods requires additional phase-shifting circuits, including a timer. The non-ideality of the back-EMF, the delay created by the low-frequency filter significantly reduces the reliability of these methods.

In [12, 13], a velocity-independent G-function calculated from the back-EMF is used to determine the commutation points. Commutation points are determined when the absolute value of this function exceeds a certain threshold value. The back-EMF contains noise and pulses generated by the inverter, which can lead

*Corresponding Author: Anatoly Stepanov, stepanov.bmstu@gmail.com

to erroneous determination of commutation points. The disadvantage of this method is also the ambiguity of choosing a threshold value.

To determine the commutation points in [14, 15], the third harmonic is used. This method requires creating a virtual third point. To form a neutral point, an additional electrical circuit is required, which is a disadvantage of this approach.

One of the problems with sensorless control is the problem of starting the motor and its operation at a low speed of rotation, in this case, the back-EMF of rotation is zero or very small. Then, determining the position of the rotor by the back-EMF of rotation is difficult or almost impossible.

In [16], when starting, the position of the rotor is initially set by the reaction of the current at magnetic saturation if the axis of the flow created by the current and the magnetic axis coincide. Only one current sensor is used. Then, after starting, the rotor is accelerated to the speed at which it is possible to switch to sensorless control of the winding switching. When the rotor position is aligned, oscillations occur in the vicinity of the equilibrium position, which reduces the reliability of this method.

Despite the large number of works on sensorless control, the problem of reliability and accuracy of determining the commutation points for the BLDC motor are relevant and require further research.

2. Mathematical model of brushless DC electric motor

Here we describe a mathematical model of a three-section electric BLDC motor. The power supply circuit of the windings of a three-section, brushless DC motor by an electronic commutator on field-effect transistors is given in Figure 1

The equations in the phase coordinates for the sections of BLDC motor with permanent magnets taking into account the mutual inductances of the sections have the form (the case of an implicit pole rotor is considered):

$$\begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix}\begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix}+\begin{bmatrix} L_s & -M_s & -M_s \\ -M_s & L_s & -M_s \\ -M_s & -M_s & L_s \end{bmatrix}\begin{bmatrix} \frac{di_a}{dt} \\ \frac{di_b}{dt} \\ \frac{di_c}{dt} \end{bmatrix}+ \quad (1)$$

$$+\begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix}=\begin{bmatrix} u_a-u_n \\ u_b-u_n \\ u_c-u_n \end{bmatrix},$$

where R_s – active resistance of section stator winding; i_a, i_b, i_c – currents in the windings of the stator sections; L_s – inductance winding sections; M_s – mutual inductance of the windings of the sections; e_a, e_b, e_c – back-EMF of the stator windings; u_a, u_b, u_c – voltage applied to the windings of sections; u_n – neutral point voltage. In wye connection $i_a + i_b + i_c = 0$.

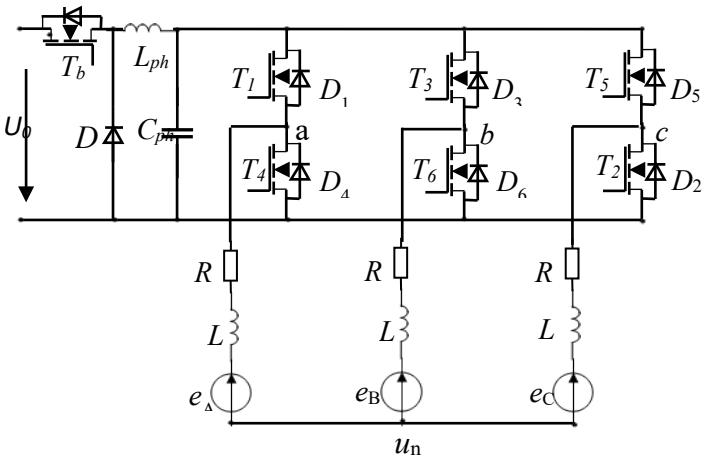


Figure 1: Three-section BDCM winding power supply circuit

Let us define the inductance of the stator along the longitudinal and transverse axes $L_d = L_q$ and the inductance L_0 for the zero sequence. The phase inductances of the windings of the sections in the rotating coordinate system d, q are defined as follows: $L_s = (L_0 + 2L_d)/3$; $M_s = (L_d - L_0)/3$.

Using the transformation (a particular form of the transformation matrix E. Clarke)

$$\begin{bmatrix} u_\alpha \\ u_\beta \end{bmatrix}=T\begin{bmatrix} u_a \\ u_b \\ u_c \end{bmatrix}, \quad (2)$$

when

$$T=\begin{bmatrix} \sqrt{2/3} & -1/\sqrt{6} & -1/\sqrt{6} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

For the transformation matrix T, the following equality holds: $T^{-1} = T^t$ (T^t is the transposed matrix), $T\,T^{-1} = E$, E – unit matrix.

Transform (2) a three-phase system (1) to a two-phase system

$$\begin{bmatrix} R_s & 0 \\ 0 & R_s \end{bmatrix}\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix}+\begin{bmatrix} L & 0 \\ 0 & L \end{bmatrix}\begin{bmatrix} \frac{di_\alpha}{dt} \\ \frac{di_\beta}{dt} \end{bmatrix}+\begin{bmatrix} e_\alpha \\ e_\beta \end{bmatrix}=\begin{bmatrix} u_{\alpha n} \\ u_{\beta n} \end{bmatrix}.$$

where $L = L_s + M_s$.

Using the transformation (2) we will express the back-EMF in the system α, β

$$e_\alpha=\frac{1}{\sqrt{6}}(e_a-e_b)-\frac{1}{\sqrt{6}}(e_c-e_a), \quad (3)$$

$$e_\beta=\frac{1}{\sqrt{2}}(e_b-e_c). \quad (4)$$

The back-EMF differences $e_{ab} = e_a - e_b$, $e_{bc} = e_b - e_c$, $e_{ca} = e_c - e_a$. They are calculated using equation (1). to calculate the back-

EMF according to (3),(4), it is not necessary to determine the voltage at the neutral point.

Back-EMF sections at a constant rotational speed vary according to the law $e_a(\theta) = k_e\omega_e\Phi'_a(\theta)$, $e_b(\theta) = k_e\omega_e\Phi'_b(\theta)$, $e_c(\theta) = k_e\omega_e\Phi'_c\theta)$, where $\omega_e = p\omega$, (ω – is the angle rotor speed, p – number of pairs of poles.

Trapezoidal function $d\Phi/d\theta$ (Φ is magnetic flux, θ electric angle) is approximated by piecewise linear function with inscribed ellipse [1]. This function $\Phi'(\theta)$ is used when simulation the BLDC motor.

Graphs of the functions $\Phi'_a(\theta)$, $\Phi'_b(\theta)$, $\Phi'_c\theta)$ for a three-phase and functions $\Phi'_\alpha\theta)$, $\Phi'_\beta\theta)$ for a two-phase system are shown in Figure 2, 3.

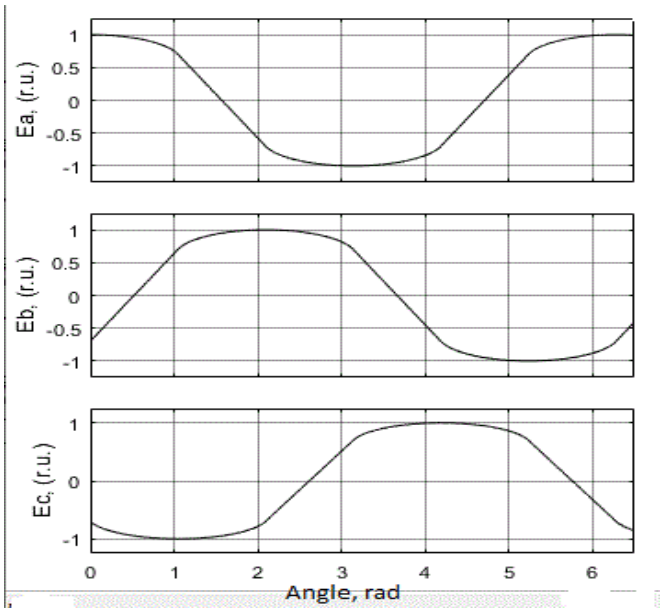


Figure 2: Trapezoidal functions $\Phi'_a(\theta)$, $\Phi'_b(\theta)$, $\Phi'_c\theta)$

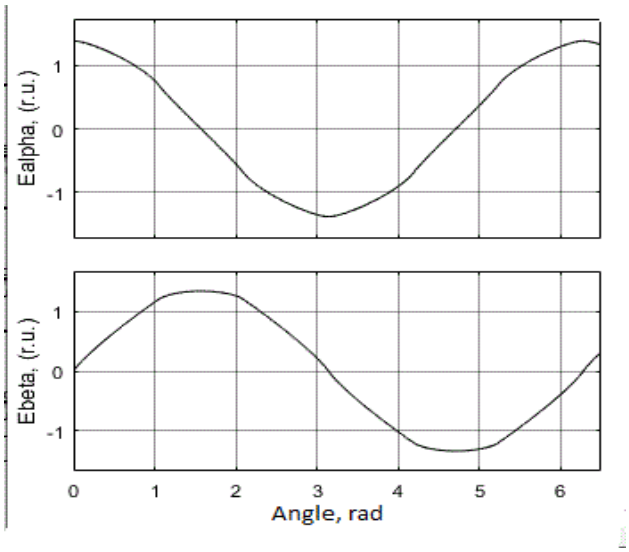


Figure 3: Functions $\Phi'_\alpha\theta)$, $\Phi'_\beta\theta)$ in system α, β

3. Algorithm sensorless control BLDC motor in system α, β

The most widespread in practice is 120 degree commutation, in which the winding is connected to the source at an interval of 120º electric degrees, then disconnected from the source at an interval of 60º electric degrees.

For 120-degree commutation, the angles of switching on and off phases , the function value is selected, which determines the value of the back-EMF of the sections at the time of switching. The dependence of the back–EMF phase in relative units expressed using the function $\Phi'(\theta)$ is shown in Figure 3, from where it follows for 120 degree commutation $\Phi'(60^\circ) = \Phi'(300^\circ)$, $\Phi'(120^\circ) = \Phi'(240^\circ)$.

To determine the commutation points, the values of the back-EMF in system α, β calculated by the formula (3), (4) are used. The ratio of the back-EMF $e_\alpha(t)$, $e_\beta(t)$ is equal to the ratio $\Phi'_a(t)$, $\Phi'_b(t)$ and does not depend on the speed of rotation of the rotor. The values of the functions $\Phi'_\alpha(\theta)$, $\Phi'_\beta(\theta)$ for commutation angles are shown in table 1.

Table 1: The value of the function $\Phi_\alpha'(\theta)$, $\Phi_\beta'(\theta)$ at the points of commutation

θ_c	$\Phi'_\alpha(\theta_c)$	$\Phi'_\beta(\theta_c)$	$\Phi'_\alpha(\theta_c)/\Phi'_\beta(\theta_c)$
$\pi/3$	$C_1(1+\Phi'_b)$	$C_2(1+\Phi'_b)$	$1/\sqrt{3}$
$2\pi/3$	$-C_1(1+\Phi'_b)$	$C_2(1+\Phi'_b)$	$-1/\sqrt{3}$
π	$-2C_1(1+\Phi'_b)$	0	$-\infty$
$4\pi/3$	$-C_1(1+\Phi'_b)$	$-C_2(1+\Phi'_b)$	$1/\sqrt{3}$
$5\pi/3$	$C_1(1+\Phi'_b)$	$-C_2(1+\Phi'_b)$	$-1/\sqrt{3}$
2π	$2C_1(1+\Phi'_b)$	0	∞

In table $C_1 = 1/\sqrt{6}$, $C_2 = 1/\sqrt{2}$.

Let's introduce the functions $H_1(t)$, $H_2(t)$

$$H_1(t) = \frac{e_\alpha^2(t) + e_\beta^2(t)}{k_1 e_\alpha^2(t) - e_\beta^2(t)},$$

$$H_2(t) = \frac{e_\alpha^2(t)}{\left(k_2 + e_\beta(t)\right)^2}$$

Coefficients k_1 and k_2 in accordance with table 1 should be equal to k_1=3 and k_2 =0. The function $H_1(t)$, $H_2(t)$ generates pulses at switching angles of 60, 120, 240, 300 electric degrees, and the function $H_2(t)$ at angles of 180 and 360 degrees. To determine points of commutation is set to the threshold value H_t. When the threshold value is exceeded ($|H_1(t)| > H_t$ or $|H_2(t)| > H_t$), the functions $H_1(t)$, $H_2(t)$ determine the commutation. The calculated back-EMF for voltage and current windings contains pulses generated by the inverter, which are smoothed by a low-frequency filter. As a result of this, a error commutation occurs. The accuracy of determining the points of commutation is also affected by the threshold value. The error of commutation can be corrected by changing the coefficients k_1, k_2. For example, for angle of commutation =240 degrees, an increase in k_1 (greater than 3) leads to a delay in the formation of the commutation signal, and when reducing k_1, the commutation signal is produced ahead of time. The estimation of the error of commutation is made by comparing the current at the time of switching off the section i_{off} (t_c) and the

current at the time of switching on the section $i_{on}(t_c + \Delta t)$ (t_c – commutation of time, Δt – delay that takes into account the time of the transient process in the winding). The coefficients k_1 and k_2 are adjusted for the difference of these currents $di = i_{off}(t_c) - i_{on}(t_c + \Delta t)$.

When correcting the error commutation, the coefficient vectors $K_1 = [k_{11}, k_{12}, k_{13}, k_{14}, k_{15}, k_{16}]$, $K_2 = [k_{21}, k_{22}, k_{23}, k_{24}, k_{25}, k_{26}]$ whose initial values are set $K_1 = [3, 3, 3, 3, 3, 3]$, $K_2 = [0, 0, 0, 0, 0, 0]$. Each element of the vector corresponds to a point of commutation 0˚, 60˚, 120˚, 180˚, 240˚, 300˚. Parameters are also set Δt, di. After determining the time of commutation, the value di is determined and the corresponding coefficient is adjusted based on the table 2. The values of the coefficients restricted to values. After several cycles of switching the windings, the error is reduced.

Table 2. The calculation of the coefficients k_1, k_2 of the error commutation corrected

θ_c	$di>0$, $\lvert di\rvert>\Delta i$	$di<0$, $\lvert di\rvert>\Delta i$
$\pi/3$	$k_1=k_1-\Delta k$, $k_2=0$	$k_1=k_1+\Delta k$, $k_2=0$
$2\pi/3$	$k_1=k_1-\Delta k$, $k_2=0$	$k_1=k_1+\Delta k$, $k_2=0$
π	$k_1=3$, $k_2=k_2-\Delta k$ 0	$k_1=3$, $k_2=k_2+\Delta k$ 0
$4\pi/3$	$k_1=k_1+\Delta k$, $k_2=0$	$k_1=k_1-\Delta k$, $k_2=0$
$5\pi/3$	$k_1=k_1+\Delta k$, $k_2=0$	$k_1=k_1-\Delta k$, $k_2=0$
2π	$k_1=3$, $k_2=k2+\Delta k$ 0	$k_1=3$, $k_2=k_2-\Delta k$ 0

For the model checking method was carried out for the model of the ideal typical signals $e_\alpha(t)$, $e_\beta(t)$, i_a, i_b, i_c . The results of the computational experiment are illustrated in Figure 4 for the angle of commutation 240 electrical degrees. The error of commutationon each cycle is reduced almost to zero.

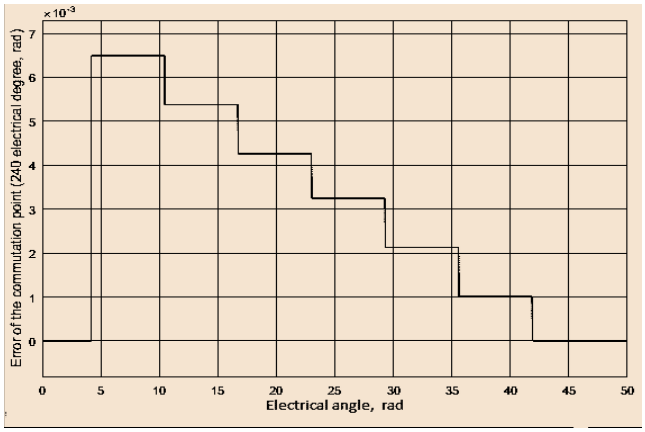


Figure 4: Error of commutation

The functional diagram of sensorless control and error of commutation correction is shown in Figure 5.

The procedure for starting the BLDC motor includes two stages: the first process of aligning the rotor in the equilibrium position, the second of accelerating up the rotor from the stable equilibrium position to the speed at which it is possible to transit to sensorless control of switching the motor windings. When the rotor is started, the rotor of BLDC is stationary or rotates at a very low speed. It is almost impossible to estimate the position of the rotor by the back-EMF of rotation, which in this case is zero or small. When switching two of the phase, the position rotor of which is unknown, acquires acceleration and moves to a stable equilibrium position. With a small damping, oscilations occur in the vicinity of a stable equilibrium position. When moving in the direction of a stable equilibrium position, the rotor accelerates, and when moving from the equilibrium position, it is slowed down. If the damping is low, the kinetic energy acquired during rotor acceleration is approximately equal to the potential energy accumulated during braking. When the rotor speed is zero and the rotor starts moving in the direction to a stable equilibrium position. The acceleration square is approximately equal to the braking square.

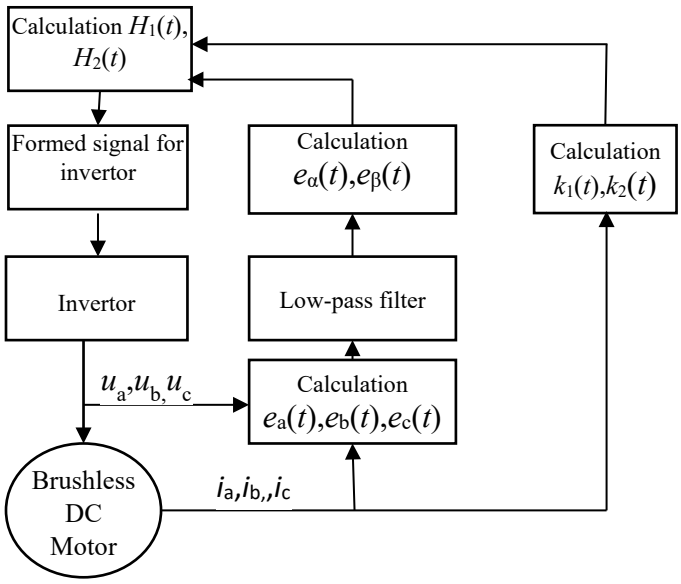


Figure 5: The functional diagram of sensor-free control BLDC motor

Stable rotor equilibrium positions is depended on various load. The area of acceleration and braking depends on the magnitude of the voltage applied to the phases. Vibration damping when the rotor position is aligned is performed by reducing the acceleration square, which is performed by reduce of the voltage on the switching phases $u = Ur$. After passing a stable equilibrium position, the voltage on the phases increases $u = Us$, which reduces the angle deviation from the equilibrium position. When the speed becomes zero, the voltage reduces $u = Ur$ (u – voltage of commutation phases). The acceleration square is reduced. The procedure is then repeated until the rotor position is aligned with the equilibrium position.

Let $S_a(\Delta\theta)$ be the acceleration area, and $S_b(\Delta\theta)$ be the deceleration area. For low damping, it must be performed $S_a(\Delta\theta) \approx S_b(\Delta\theta)$, $\Delta\theta = \theta - \theta s$.

$$\Delta\theta(k+1) = (Sa)^{-1}[S_b(\Delta\theta(k)],$$

If the maping will be compressive if $|\Delta\theta(k+1)| < |\Delta\theta(k)|$, then the iterative process converges and $\Delta\theta$ tends to zero. The rotor position is aligned relative to the stable equilibrium position.

The moments of changing the voltage on the section are determined by changing the current of the section. After switching the winding, the current at the rotor acceleration initially decreases,

and then increases when passing the equilibrium position is equal to the initial one. At this point, the voltage increases $u = U$s and a new value of the initial current is set. The current during braking first increases, and then decreases and reaches the initial value, at which point the voltage decreases $u = U$r. The process is then repeated until the rotor position is aligned (Figure 6).

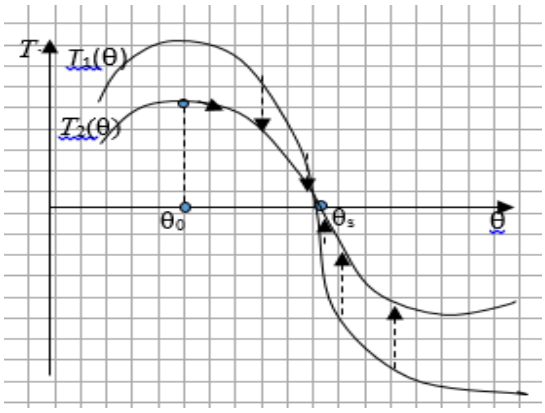


Figure 6: Aligned position of rotor

4. Simulation results

To substantiate and study the effectiveness of the proposed sensorless control, start-up and corrected error of commutation method for a low-power brushless DC electric motor with permanent magnet was simulated.

The simulation was performed using the MATLAB&Simulink system and models of a three-phase brushless DC motor and a power semiconductor bridge commutator of the SimPowerSystems expansion. A motor is an electric machine with a three-section winding on the stator and a multi-pole rotor with permanent magnets. BLDC motor parameters are given:

Rated DC Voltage – 27 V;

Rated speed, n = 2000 r/min;

Number of poles, p = 4;

Phase resistance, R = 6 Ω;

Phase Inductance, L = 0.42 mH.

When starting the motor, the position of the rotor is not determined. If one connects two sections of the motor to a DC voltage source, the rotor begins to move to a stable equilibrium position (stable point). Let $u_a = U_0$, $u_c = 0$, phase "b" is disconnected from the source.

The proposed start-up procedure involves two steps for the transition to the sensorless control of BLDC motors. In the first step, the procedure is performed to align the rotor position and to calm down oscillation in a stable equilibrium position. At this step, a constant voltage of various values Ur = 27 V, Us =50 V was applied to the switching phases (phases a, c), which provided alignment of the rotor position and damping of vibrations around a stable equilibrium position. In the second step, a constant voltage was applied to the switching phases (phases b, c). After the speed set, there was a transition to sensorless control of commutation of the BLDC motor windings. Simulation of start-up and control of commutation of the BLDC motor windings was performed using the MATLAB/Simulink system. For Figure 7 the simulation results are presented: the function $H(t)$. Function $H(t)$ is sum $H_1(t)$, $H_2(t)$. The results of modeling the motor start and transition to sensorless control are shown in Figure 8.

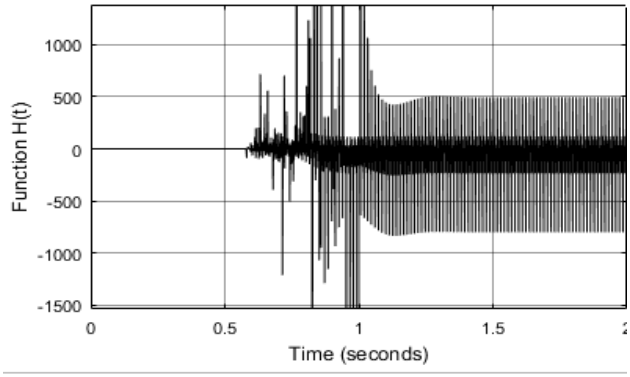


Figure 7: Function H(t)

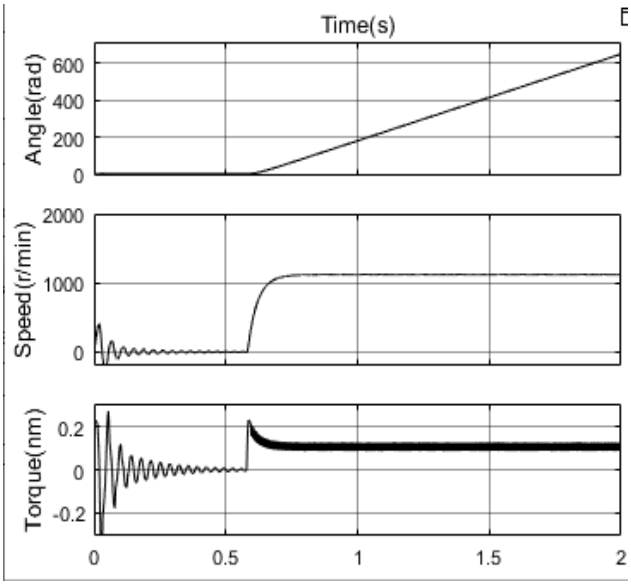


Figure 8: Start-up fnd transition to the sensorless control

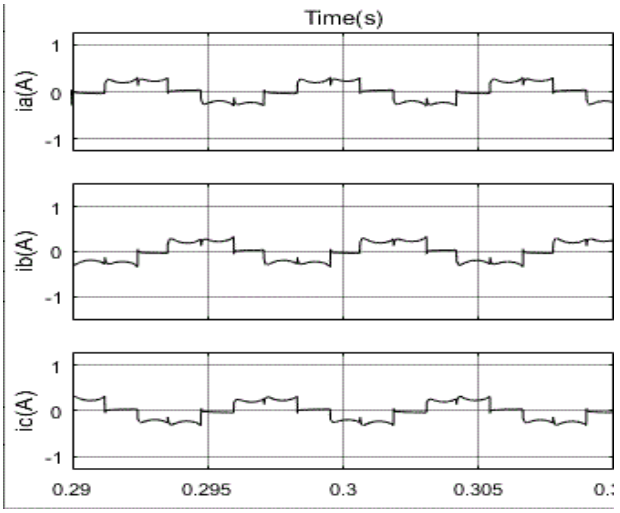


Figure 9: Corrected of phase current

To demonstrate the capabilities of the proposed method for correcting the commutation error, a simulation of BLDC motor with different speeds was performed. Error correction was performed by analyzing phase currents. Phase currents after error compensation are shown in Figure 9.

The table 3 shows the dynamics of the commutation error when it is compensated by the proposed method. At low speed=500 the initial commutation error was small. Initial commutation error was reduced to an error 0.12 electrically degree. At the speed of 1000, the initial error 3.5 electrically degree was also reduced by the correction method to 0.1 electrically degrees. At the nominal speed of rotation, the significant error of 14 elecctrically degree was reduced to 0.2 electrically degrees. These results confirm the possibility of reducing the compensation error in a simple corrected method.

Table 3: Error of commutation

t, *s*	Δθ, deg (n=500 r/min)	Δθ, deg (n=1000 r/min)	Δθ, deg (n=2000 r/min)
0.01	1.25	3.5	13.9
0.05	0.96	1.9	10.8
0.1	0.81	1.6	5.2
0.15	0.54	1.35	3.7
0.2	0.56	1.08	2.1
0.25	0.45	0.54	1.6
0.3	0.12	0.1	0.2

5. Conclusion

The paper offers a simple method for correcting the commutation error in the case of sensorless control of a BLDC motor. A start-up procedure based on align the rotor position and damping its oscilations in a stable equilibrium position is proposed. This allows to accelerate the rotor and transient to sensorless control. The proposed method can increase the efficiency of sensorless control methods for BLDC motor with small inductive.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A.V. Stepanov, V.N. Enin, “Sensorless control of low-power brushless DC motor based on the use of back-EMF,” in 2019. IEEE Internatiol Ural Conference on Electrical Power Engineering (Uralcon), 277-283, 2019, doi:10.1109/URALCON.2019.8877642

[2] A.K. Arakelyan, A.A. Afanas'ev, The rectifier of the electric machine in the systems of electric drive,” M.: Higher school, 2006.

[3] A.P. Belkovsky, “Precision electric drive with valve engines,” - Moscow: MPEI, 2010.

[4] M. Bychkov and A. Krasovsky, "Taking Into Account Nonlinear Properties of Switched Reluctance Machines in Electric Drive Control Algorithms," in 2018 X International Conference on Electrical Power Drive Systems (ICEPDS), Novocherkassk, 1-5, 2018, doi:10.1109/IWED.2018.8571868

[5] R. C. Guo, Z. Mu , J. D. Li, “Research on Position Sensorless Control System of High-Speed Brushless DC Motor,” in 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics, 62-65, 2017, doi:10.1109/IHMSC.2017.21

[6] A. Halvaei Niasar, A. Vahedi, H. Moghbelli, “A novel position sensorless control of a four-switcy, brushless DC motor drive without phase shifter,” IEEE Trans. Power Electron., **23**(6), 3079-3087, 2008, doi:10.1109/TPEL.2008.2002084

[7] H. Li, S. Zheng, H. Ren, “Self-Correction of Commutation Point for High-Speed Sensorless BLDC Motor With Low Inductace and Nonideal Back EMF,” IEEE Transactions on Power Electronics, **32**(1), 642-651, 2017, doi:10.1109/TPEL.2016.2524632

[8] W.Li, J. Fang, H. Li, J. Tang, “Position sensorless control without phase shifter for high-speed BLDC motor with low inductance and nonideal back EMF,” IEEE Trans. Power Elecnron., **31**(2), 1354-1366, 2016, doi:10.1109/TPEL.2015.2413593

[9] J.S. Park, K.L. Lee, S.G. Lee, W.H. Kim, “Unbalanced ZCP compensation method for position sensorless BLDC motor,” IEEE Trans. Power Electronics, **34**(4), 3020-3024, 2019, doi:10.1109/TPEL.2008.2002084

[10] X. Zhou, X. Chen, F. Zeng, J. Tang, “Fast commutation instant shift correction method for sensorless BLDC motor based on terminal voltage information,” IEEE Trans. Power Electronics, **32**(12), 9460-9472, 2017, doi:10.1109/TPEL.2017.2661745

[11] G. Su, J. McKeever, “Low-cost sensorless control of brushless DC motor using a frenquency-independent phase shifter,” IEEE Transaction on Power Electronics, **19**(2), 296-302, 2004, doi:10.1109/TPEL.2003.823174

[12] S. Chen, X. Zxou, G. Bai, K. Wang, L. Zhu, “Adaptive commutation error compensation strategy based on a flux linkage function for sensorless brushless DC motor drives in a wide speed range,” IEEE Trans. on Power Electron., **33**(5), 3752 -3764, 2018, doi:10.1109/TPEL.2017.2765355

[13] S. Chen, G. Liu, L. Zhu, “Sensorless Control Strategy of a 315 kW High-Speed BLDC Motor Based on a Speed-Independent Flux Linkage Function”. IEEE Trans. Ind. Electronics, **64**(11), 8607-8617, 2017, doi:10.1109/TIE.2017.2698373

[14] G. Liu, C. Cui, K. Wang, B. Han, S. Zheng, “Sensorless control for hign-speed brushless DC motor based on the line-to-line back EMF,” IEEE Trans. Power Electron., **31**(7), 4669-4683, 2016, doi:10.1109/TPEL.2014.2328655

[15] X. Song, B. Han, J. Fang, “High-precision sensorless drive for high-speed BLDC motors based on the virtual third harmonic back-EMF,” IEEE trans Power electron., **33**(2), 1528-1540, 2018, doi:10.1109/TPEL.2017.2688478

[16] W. J. Lee, S.-K .Sul, “A New Starting Method of BLDC Motors Without Position Sensor,” IEEE Trans. Ind. Appl., **42**(6), 1532-1538, 2006, doi:10.1109/TIA.2006.882668

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 245-252 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Assessing Heutagogical Elements in Learning of Engineering Education: Instrument Validation

Mimi Mohaffyza Mohamad[*,1], Alias Masek[1], Jailani Md Yunos[1], Maizam Alias[1], Nor Hidayah Hamdan[1], Andika Bagus Nur Rahma Putra[2]

[1]*Faculty of Technical and Vocational Education, Universiti Tun Hussein Onn Malaysia, Johor, 86400, Malaysia*

[2]*Department of Mechanical Engineering, Faculty of Engineering, Universitas Negeri Malang, Jl. Semarang No.5 Malang, 65145, Indonesia*

ARTICLE INFO

Article history:
Received: 12 May, 2020
Accepted: 11 August, 2020
Online: 14 September, 2020

Keywords:
Validation
Instrument
Heutagogy approach
Self-determined learning
Rasch Measurement Model

ABSTRACT

Practically level of design element (i.e., explore, sharing, connect) is an essential of heutagogical approach. The self-determined learning process can be at ease with the implementation of these elements, and the critical step is reliability to measure teaching and learning feedback. Although various instruments were proposed in the literature to assess heutagogy elements, the specific potential Rasch Measurement Model to determine the practicality levels of heutagogy element is less emphasized. This paper aimed to validate the research instrument (six constructs with 65 items). The instrument was administered to N=175 students for a pilot study. The Rasch model was conducted to examine reliability (0.93, 0.94) with α = 0.97, separation index (3.75, 4.01) for item and person, respectively. Besides, item fit (three-item dropped), polarity and standardized correlation residual (no overlapping items). The findings have shown that the instrument has high validity and reliability for use in measuring the practical level of heutagogy elements.

1. Introduction

Pedagogical or andragogical approach is generally related to the role of teachers and learners in the traditional education system. As the future of education has offered many new opportunities in the learning process, the continuous transfer of learning from pedagogy to andragogy and finally arriving in heutagogy shows that change must take place in the classroom compulsory [1]. With the rapid development of technology, the education system is increasingly shifting to a self-directed approach by heutagogical method [1,2,3]. From this trend, students can obtain valuable information from different sources and accessible from any platform (i.e., online or mobile system). Therefore, the heutagogy approach focuses on life-long learning and self-determined learning methods. There are multiple contexts of studies that have been discussed in the literature, such as related to necessity [4], practice [5,6,7] and teacher's training [8] of heutagogy learning approach. Most studies have found that the heutagogical approach can be an essential learning style for producing students who are well-prepared for the complexities of today's workplace.

In the context of measurement, evaluation of the heutagogical learning approach's implementation strategies in the teaching and learning process is a vital issue to be taken into account. Implementation and learning strategies need to be aligned with the design principle of the heutagogical learning environment (i.e., exploration, creation, reflection, connection, assessment and sharing) [9]. To date, a number of validation instruments, tests, experiments and studies have been developed to evaluate the effectiveness of the heutagogical learning from a different point of view [4, 6, 9]. The validation process consists of a systematic procedure, including a pilot study. The pilot study's main goal is to improve the quality of the item and increase confidence in the interpretation of the data [10]. While every student has a different style of preferred learning ways, a pilot study is required to determine the capability of the developed instrument to analyze the practicality of heutagogy data. During data analysis, the need to identify the element of heutagogical approach that best fits the different categories of student's profiles, especially in engineering education, arises as to the main issue

Based on the above premises, an instrument was developed in this study, involving the student's reflection to determine the practical level of heutagogical elements applied in learning, based

[*]Corresponding Author: Mimi Mohaffyza Mohamad, Tel: +6074534193, Email: mimi@uthm.edu.my

on the design elements proposed [9]. This research instrument was developed in the questionnaire form. It is a practical way to gather data, collect results easily, and gather information from a large number of respondents [11]. When involving a self–assessment test, the psychometric tests element must be considered during the instrument development to ensure the high reliability and validity of the instruments.

In this phase, a set of questionnaire with a Likert scale was used to obtain feedback from students. However, using only a descriptive analysis of the Likert scale data does not provide sufficient information and concrete findings [12], since the Likert scale is commonly ordinal data that categorizes the respondent's point in general (i.e., less accurate interoperation). The Rasch Measurement Model analysis was therefore conducted to improve response accuracy. The Rasch model analysis can convert the latent traits, either nominal or ordinal data into interval and ratio data by creating a logit scale. This can be useful and provide scientific measurement for this study. The objectives of the study are as follows:

- to identify the validity and reliability of the heutagogy elements in learning of engineering education using Rasch Measurement Model
- to identify the point measure correlation in the construct of the heutagogy elements in the learning of the engineering education which are in the acceptable range
- to identify the fit items in the construction of the items heutagogy elements within the acceptable range.

2. Literature Review

2.1. Heutagogy Learning Approach

Heutagogy, or the study of self-determined learning, has gained interest in the field of engineering education aimed at 21st-century skill requirements and industrial needs [9]. The increasing number of education institutions, for example, technical universities, have started to explore and implement a self-determined learning environment for their students. The main goal of heutagogy is to teach lifelong learning. Heutagogical approach differs from other traditional methods (i.e., pedagogy and andragogy). In comparison, these three learning approaches can be differentiated based on the teaching and learning environment and its implementation. Generally, figure 1 shows the progression of three teaching approaches.

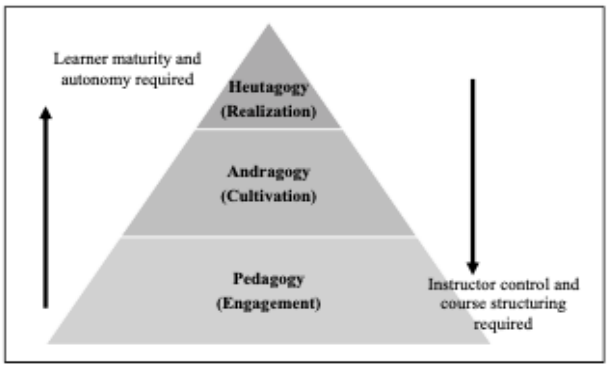


Figure 1: Progression of Learning Approach

At the first level, pedagogy is mainly focusing on teacher-led learning, where teachers have to decide on the overall learning content and process (i.e., what students will learn and how they will learn). In the first level, students depend on instruction from the teachers. In the second level, andragogy is taking place where teachers play an important role as a mentor in andragogical environments and cultivate the students' ability to self-directed learning. At the third level, entering at heutagogy learning level 3 encourages students to fully control the learning environment, where students solve their learning problems with their own decisions. Therefore, it is important for students as well to equip themselves with problem-solving skills [13]. In this case, the teacher is only preparing the learning context in general [1].

2.2. Heutagogical Study and Instrument

Heutagogy is an extension of the pedagogy and andragogy approach which places the learner, rather than the teacher or institution or even the curriculum, at the center of the learning process [9]. The goal of heutagogy is to prepare responsible, self-capable, proactive, competent students who are ready to face the actual workplace environment. Most heutagogical studies in engineering education, vocational education, and pre-service teacher professions have found that heutagogy is a credible response to the critical issues that learners face within the workplace and have designed their learning environments based on the approach. For example, within the vocational education, [14] the innovation of training module based on heutagogy has been developed as acceleration for increasing the pedagogical supremacy of professional education lecturers in the industrial revolution. The findings reported that heutagogy provides a learning framework that addresses the needs of vocational students, who must learn in an ever-changing environment that is both complex and helps them to become lifelong learners.

Table 1 summarizes the previous relevant heutagogical studies that can be categorized into two research directions; namely the category/principle and the type of environment. The category or principle perspective, in which heutagogy is defined as the center of the teaching and learning process and becomes an active agent of all learning experiences, from planning and execution to learning assessment [2, 9, 15]. The main goals of these studies are to understand the perceptions, needs and experiences of learners or their characteristics or to develop a learner-centered curriculum. There are previous study that have illustrated the finding of their small-scale study on heutagogy approach, which can be one of the key elements for enhancing the curriculum, encouraging students to determine what, how, when and where they learn and enabling educators to be more flexible in delivering their curriculum [16].

Next, the research also indicated several different types of the learning environment (i.e., traditional face-to-face), online, or a combination could support self-determined learning [9]. With the availability of an online platform, implementing self-determined learning activities can be more convenient as an online platform can connect, share, create, and collect more resources, that has been highlighted in several previous studies [17]. This has already been indicated in some of the research studies [18-20]. Critical thinking can be developed among students [21] to solve learning problems [22] and researcher found that reflective independent and critical thinking skills can be developed [3].

Table 1 : Summary of Related Works

Research Direction	Themes	Element focused	Instrument/ Methodology
Category/ Principle	Learner-centered and learner-determined learning	Perceptions [21]	Semi-structured group interviews
		Heutagogical needs and experiences of the learner [4]	Interviews, literature review, documentation, and observation
		Curriculum development [15,22]	Group discussion
			Semantics search and discussion
			Questionnaire survey
		Strategy and implementation [23]	Standardized and open-ended interviews
	Self-reflection	Motivation [2]	Phenomenological interviews
Type of environment	Online platform, mobile application	ICT literacy [19]	Questionnaires survey
		Practices [17]	Questionnaire, semi-structured interviews,
		Critical thinking MOOC [20]	Trello online discussion tool
	Blended Learning	Practices and technology adaptation [8,24]	Workshop
			Quasi-experimental design, semi-structured interview

in heutagogy learning program by using a blended approach. It is also noted that the affordance of mobile social media tools can facilitate student-determined learning experience (heutagogy) in authentic contexts [3].

On the other hand, interview questions were mostly used in the standard instruments, and the results encourage two-way communication. Besides, it allows participants (i.e., lecturer or students) to freely express their needs, perceptions and experiences in the implementation and practice of self-determined learning [21], the interview questions are useful in profoundly exploring the participants' attitude, experiences. For example, [19] selected 138 high school teachers to determine online learning competence through a heutagogy approach. The result indicated that all the heutagogy elements (i.e., exploration, creation, collaboration, connection, sharing and reflection of experience) were identified as important. However, one of the neglected remaining issues is on how the heutagogy approach instrument can be best validated? Therefore, this study was conducted to validate the proposed instrument used for heutagogy learning approach among engineering students.

3. Methodology

Figure 2 shows an overview of the validation process of heutagogical element in learning of engineering education. There are five main steps conducted, starting with the development of the questionnaire, data collection, data processing, data analysis, and reporting findings. First of all, the research instrument in the form of a questionnaire survey was developed. The questionnaire consists of two sections; Section A includes the students' Demographic profiles, and in Section B, a total of 65 items to gauge students' learning style. In section B, six main constructs/elements of heutagogical learning style have been evaluated namely, (a) explore, (b) create, (c) collaboration, (d) connect, (e) share and (f) reflection.

Furthermore, the questionnaire was distributed through the online platform and the student portal. Students have to rate items based on a scale ranging from 1 (never) to 5 (always). Students' response represents the percentage of the practice such as 1 (0 percent practice), 2 (1-20 percent practice), 3 (21 percent-50 percent practice), 4 (51 percent-80 percent practice) and 5 (>80 percent practice), respectively. Respondents were given approximately fifteen minutes to complete the questionnaire. A total of 175 students from University Tun Hussein Onn Malaysia, (UTHM) were randomly selected for research sampling. From the responses obtained, the collected data underwent a pre-processing process, such as determining the missing data and converting it to the appropriate format (i.e., convert excel to. prn text file). Next, the data analysis was performed using the Rasch measurement model, which focuses on four main analyses, including: (i) reliability and separation index; (ii) item fit and polarity. The flow of the data analysis was illustrated in figure 2.

3.1. Rasch Model in Measuring Heutagogical

Equation 1 shows the adaptation of the mathematics model underlying the Rasch measurement model [23,24,25]. This principle indicates that students have a higher probability of responding to the less difficult item and a lower probability of accurately responding to the more difficult item.

$$P_r\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \quad (1)$$

where β_n is the ability of student n and δ_i is the difficulty of heutagogy item *i*. Besides, $\beta_n - \delta_i$ is the probability of student's response for each heutagogy item. In this case, the $P_r\{X_{ni} = 1\}$ is the success probability or also known as logit value upon the interaction between the relevant students and the assessment item. Thus, this study has simply the Equation (1) into a logarithmic function, as shown in Equation (2).

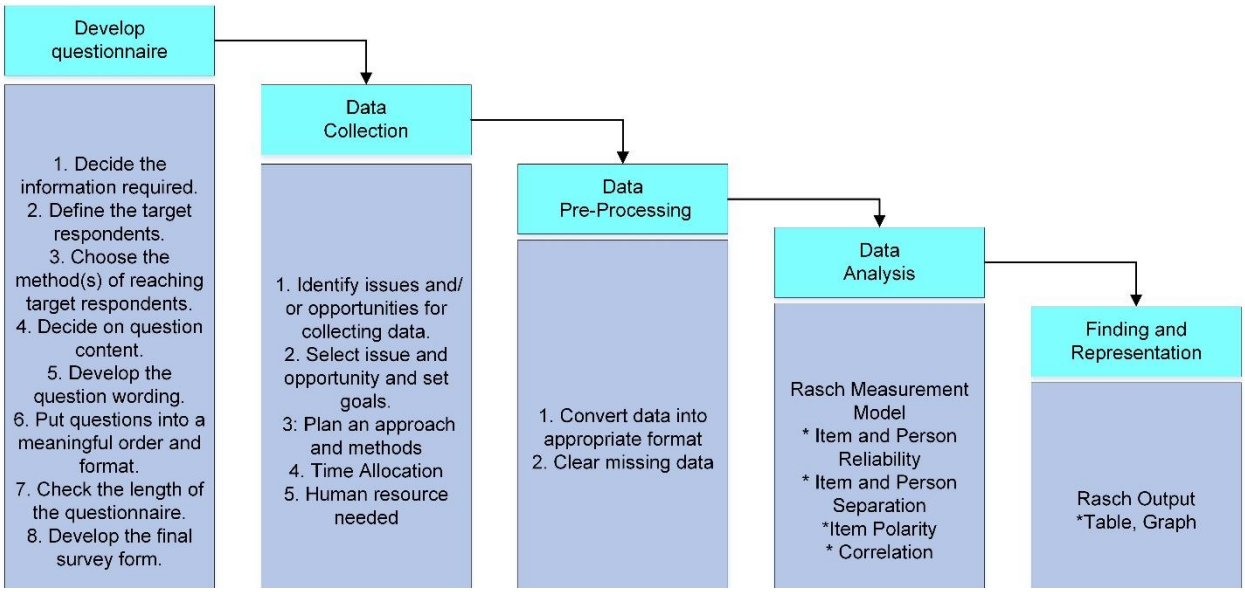


Figure 2: Data analysis flow

$$logit\left(\frac{P}{1-P}\right) = \beta_n - \delta_i \quad (2)$$

The probability of students' success for each heutagogical item demonstrates the relationship between students' ability and the difficulty of given items in the heutagogy element.

3.2. Analysis Using Rasch Measurement Model

Based on the aforementioned Rasch model analysis, Table 1 shows the summary of the statistical criteria and the acceptable value for further study. The following analysis in Table 2 was examined in order to assess the validity and suitability of the item in the Rasch Model.

a. Reliability and Separation Index. Reliability refers to the consistency of the measure, which can be obtained through the Cronbach Alpha values and the internal reliability index [26]. The value of Cronbach Alpha (α) is between 0 (no internal reliability) and 1 (perfect internal reliability) with a minimum score of 0.70 [27]. While the separation index indicates the statistically distinct measurement level of an item's difficulty or a person's ability [27].
b. Item polarity. Item polarity was evaluated by using a point–measure correlation coefficient (PTMEA CORR). The values of the high and positive items range from 0.3 to 0.8, representing items that are working in the right direction to measure each of the developed constructs [28]. However, the negative values or zero indicated a conflicted relationship between items within the construct. An item that exceeds the 0.3-0.8 interval is recommended for removal [29].
c. Item fit. In the Rasch model, item infit and outfit mean-square fit statistics (MNSQ) were commonly used to determine the item fit assessment. The infit statistics are inlier-sensitive fit statistics, which indicate the responses of items that are close to the person's ability level [30]. Outfit statistics are referred to as outlier-sensitive fit statistics, which represent unexpected responses to items outlying from the person's ability level. The item is considered to be consistent with the Rasch measurement when MNSQ reaches the expected value of 1 and always be positive.
d. The MNSQ optimal value of each item must be located within 0.5–1.5 [30]. The values of less than 0.5 or 1.5–2.0 do not bring efficiency to building measurements, but do not decline [31]. Therefore, any individual item with MNSQ more than 2.0 will be suggested for removal from the present study.

Table 2 : Summary of Statistical

Criteria	Statistical Info	Result	
		Value	Indicator
Item Misfit	Person & Item Reliability [26]	0.9-1.0	Very good and effective level of consistency
		0.7-0.8	Good and Acceptable
		0.6-0.7	Acceptable
		< 0.6	The item needs to be repaired
		< 0.5	The item needs to be dropped
	Separation [32]	Value ≥ 2	Acceptable
Item Validity	Item Polarity [26]	PTMEA CORR > 0	Acceptable
Item	Item Fit [26]	Total Mean Square infit and outfit in range 0.6-1.4	Acceptable
		Total Mean Square infit and outfit in range 0.6-1.5	Acceptable

e. Standardized measurement of the residual correlation value. This analysis was conducted to determine the overlapping of the items. High residual correlation for pair items indicated that item is not independent (i.e., items are the same characters or combination of the shared item). As for the pairing item's correlation, value that is greater than 0.7 indicates a high correlation value, and only one item must be maintained. In contrast, the other items should be dropped [32,33].

4. Results and Discussion

4.1. Demography Profile

Table 3 presented the demographic profile of the respondents. There are three main demographic profiles collected, such as gender, year of study, and faculty members. All the information was presented in the cross-tabulation output, which is faculty versus year of study and gender versus faculty. Gender information is included in this study as it is a way of looking at how heutagogy element impacts learning experiences between different groups of male and female students.

Table 3: Demography Profile

Criteria	Year of Study				Gender	
Faculty	1st	2nd	3rd	4th	Male	Female
FKAAS	0	1	0	0	0	1
FPTP	23	43	68	1	56	79
FPTV	9	2	17	11	13	106
Total	32	46	85	12	69	106

4.2. Reliability and Validity

Table 4 shows the statistical summary of the reliability and separation index of the survey instrument. The value of 0.97 obtained in this study indicates that the reliability of the instrument used was high and adequate. In item measurement, the reliability value of 0.93 indicates that items of the instrument are very sufficient to measure the learning heutagogical elements. The person's reliability indicates the probability of a person's response results of 0.94 when the same test is performed. Hence, this heutagogical learning instrument is at a high level of reliability. Meanwhile, for the separation index (SI), both item and person measurements were obtained acceptable value $\geqslant 2$, SIitem = 3.75, and SIperson = 4.01.

Table 4 : Reliability and Separation Index

TE	Item Measurement		Person Measure		α
	Reliability	Separation	Reliability	Separation	
65	0.93	3.75	0.94	4.01	0.97

4.3. Mean and Mean Logit

Table 5 shows the mean and mean logits for each construct in this developed instrument. The higher mean score is obtained for 'explore' construct, $\bar{x}$ = 4.018, which indicated that the students agreed that the explore item suggested is working well in their learning process. This is matter because the student needs to be provided with free and large opportunity to explore the variety of learning pathways, information, and resources during the learning process [9, 15]. However, the value of the connect construct means obtained, $\bar{x}$ = 3.804 still indicates the high level of student's agreement with the connecting element. This matters because the creation connections are comfortable with the affordance of an online platform (i.e., social media, video conferencing), which allows students to have the opportunity to share new knowledge with others who have the same interest [34].

On the other hand, the measure means were obtained to determine the item agreement among the students. The negative value item indicated that students were more comfortable to agree on the item. In this case, explore (- 0.24 logit) and share (- 0.03 logit) constructs were more comfortable obtaining a high degree of agreement among the students.

Table 5 : Mean and Mean Logit

No	Construct	TE	$\bar{x}$	SD	Mean Logit		
					Min	Max	Total
1	Explore	10	4.01	18.9	-0.54	0.31	-0.24
2	Create	11	3.87	52.5	-0.75	1.19	0.03
3	Collaboration	12	3.90	31.9	-0.61	0.72	0.00
4	Connect	10	3.80	45.9	-0.82	1.01	0.19
5	Share	12	3.92	28.2	-0.64	0.48	-0.03
6	Reflection	10	3.88	33.0	-0.45	1.00	0.05
	Total	65					

* TE- Total Item, SD - Standard Deviation, $\bar{x}$ = Mean

4.4. Correlation

Standardized measurement of the residual correlation value is to determine whether there are overlapping items. High residual correlation for the two items showed that the item is not independent, either because the item has the same characteristics among each other or because it combines several different dimensions that are shared. If the correlation value of the two items above 0.7 shows a high correlation value and only one item must be maintained [31]. In contrast, the other items should be dropped. Based on Table 6, no overlapping pairs have been found, and no items have been dropped.

Table 6 : Large standardized residual correlation used to identify the dependent item

Correlation	Entry Number	Item	Entry Number	Item
0.56	51	SH51	52	SH52
0.50	38	CN38	39	CN39
0.46	43	CN43	45	SH45
0.45	26	CL26	35	CN35
0.43	53	SH53	58	RF58
0.42	64	RF64	65	RF65
0.42	25	CL25	26	CL26
0.41	23	CL23	24	CL24
0.41	60	RF60	61	RF61
0.41	8	EX8	9	EX9

4.5. Item Fit and Item Polarity

Due to space limitation, Table 6 presented only the estimation of difficulty parameter and model fit estimations of misfit items. Mean square information (MNSQ) and the point measure correlation (PTMEA-CORR) aimed to determine the item's suitability for the developed instrument. For the MNSQ assessment, the list presented that MNSQ infit range is at 0.67 logit – 1.90 logit. As the MNSQ infit acceptable range is 0.50 logit – 1.5 logit, three items RF59 = 1.90 logit SH53, item SH44 = 1.58 logit, were found exceed the acceptable range.

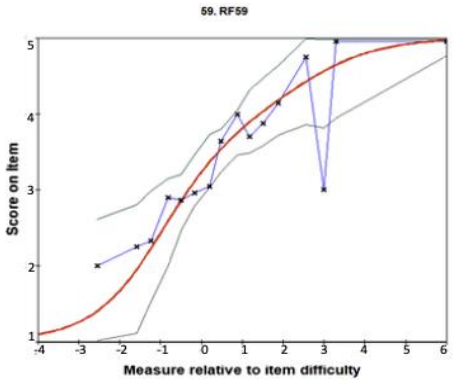


Figure 3: The ICC Curve for Misfitting Item, RF 59

Based on the output presented in Table 5, an example of the ICC curve for misfitting items was generated in Figure 3. The red curve is the expected ICC if the data fitted the Rasch model. The blue curve is the observed or empirical ICC. The grey line in the outskirts of the red curve is the confidence interval constructed from an estimate and its standard error. The empirical ICC for misfitting item RF59 has a slightly large deviation from the expected ICC. This fact is reflected in RF59 with a large outfit-MNSQs value of 2.02. However, Figure 4 presented a good-fit item, SH47, in which the outfit-MNSQs value, 0.90 was obtained and is within an acceptable range. In addition, PTMEA-CORR can also be used to accept or abolish the items tested. The acceptable range depends on the purpose of the instrument. However, the negative value of the PTMEA-CORR items does not measure what should be measured and should be dropped. In this study, all tested PTMEA-CORR items were positive (0.43 logit -0.63 logit) and were therefore retained.

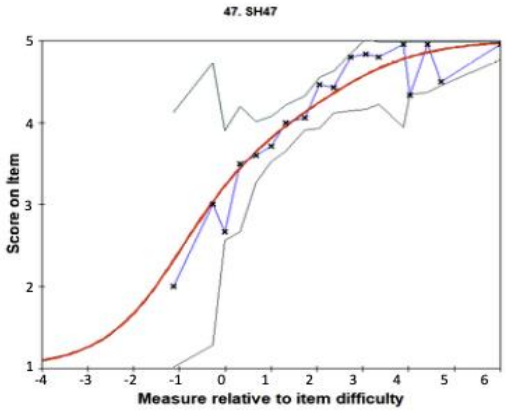


Figure 4: The ICC Curve for Good-fit Item,47

The detail of the item fit explained in Table 7, which shows fit index item that measured in infit/outfit, and items RF59, SH83, and SH44 are more than 1.4 logit values, and this item needs to revise.

4.6. Variable Map

Based on the above findings, this list of logit measurement information for each item is useful for understanding each item's position in the variable map. Figure 5 shows the distribution of the item's difficulty. In this case, item CR12 received a high measure of logit value, +1.99 indicating the most difficult item to respond to. In contrast, item CN35 indicated the easier item to be answered as the measured value obtained, - 0.82 logit.

Table 7 : Partially Outcome of Item Fit

Measure	Model S.E	Infit		Outfit		PT-Measure		Item
		MNSQ	ZSTD	MNSQ	ZSTD	CORR	EXP	
1.00	0.10	1.90	6.70	2.02	7.3	0.43	0.57	RF59
0.48	0.10	1.83	5.9	1.84	5.0	0.43	0.54	SH53
0.32	0.11	1.58	4.3	1.52	4.0	0.44	0.53	SH44
1.06	0.10	1.38	3.2	1.49	4.0	0.50	0.57	CR20
1.01	0.10	1.40	3.3	1.47	3.8	0.46	0.57	CN42
0.31	0.11	1.41	3.2	1.45	3.5	0.30	0.53	EX7
0.41	0.10	1.32	2.6	1.34	2.7	0.46	0.58	CN43
1.19	0.09	1.26	2.3	1.32	2.8	0.51	0.46	CR12
-0.75	0.12	1.30	2.4	1.23	1.9	0.44	0.53	CR13
0.38	0.11	1.30	1.7	1.17	1.4	0.53	0.52	SH45
0.11	0.11	1.20	0.9	1.19	1.6	0.43	0.49	CL31
-0.32	0.12	1.11	0.9	1.17	1.4	0.41	0.49	EX1
-0.30	0.12	1.06	0.7	1.15	1.3	0.46	0.51	EX3

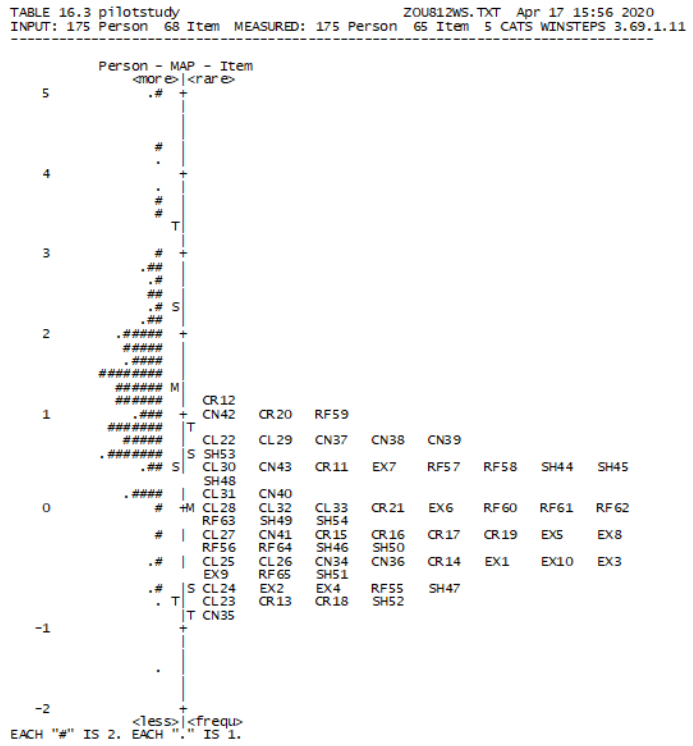


Figure 5 : Variable Map (Item – Person)

5. Conclusion

An instrument for gauging a heutagogy element in learning for the engineering education area was successfully developed and validated using the Rasch measurement model. Using this model, the validity and reliability of items have been proven using the data gathered from the pilot study, to fit the engineering education students. In this study, we have provided two implications of the findings based on The Rasch model framework provides systematics procedures for validating the developed instrument (e.g., Reliability, construct validity, item fit, separation index). Besides, the Rasch measurement model provides an impact on discovering the meaning of students' measure, which heutagogy element applied in students' learning using from the context of the instrument's items. The first contribution in this context, a huge implication, is probably on the potential for individual reflection and feedback practices. Practitioners in schools and teacher educators can be referred to the Wright Map/Variable Map that explicitly uses the heutagogy element in the learning process. Teachers can prepare the best strategies and implementation process for important aspects of the heutagogy element to students. Another implication from this study identified potential strategies for curriculum development and the teaching process. A total of 175 students can generally use the heutagogical elements at least by sharing information with other engineering or non-engineering students through an online platform. In this case, students can discover more learning experiences, learn new things, and identify mutual students with similar interests. This repeating process can lead to life-long learning and future collaboration. We are currently working on a test within the different categories of students, from pre and post-secondary subjects, using the developed instrument to explain the changes in the critical heutagogical elements. The findings will be reported in the subsequent publication.

Acknowledgment

The authors would like to thank all the associate editors and referees for their thorough reading and valuable suggestions, which led to the improvement of this paper. The Research Management Center (RMC), Universiti Tun Hussein Onn Malaysia and Ministry of Higher Education (MOHE) are acknowledged for the financial support received for this research (Research Fund: Development of Heutagogical Learning Module for Engineering Education Student , Grant Code K134, MyRIVET). Universitas Negeri Malang is recognized as a research collaborator and researchers from both universities Tee Tze Kiong, Yee Mei Heong, Lee Ming Foong, Razali Hassan, Haris Anwar Syafrudie, Ahmad Mursyidun Nidhom, Azhar Ahmad Smaragdina and research assistant Anies Faziehan Zakaria.

References

[1] J. Md. Yunos, H. A. Syafrudie, M. Alias, R. Hassan, M. H. Yee, "Heutagogical Teaching And Learning Activities Preferences For Malaysia Technical University Lecturer: Instrument Validation," International Journal of Advanced Science and Technology, **29(**4), 3875-3886, 2020.

[2] Blaschke L.M. Self-determined Learning (Heutagogy) and Digital Media Creating integrated Educational Environments for Developing Lifelong Learning Skills. In: Kergel D., Heidkamp B., Telléus P., Rachwal T., Nowakowski S. (eds) The Digital Turn in Higher Education. Springer VS, Wiesbaden. 129-140,2018. https://doi.org/10.1007/978-3-658-19925-810

[3] N. Canning, "Playing with heutagogy: Exploring strategies to empower mature learners in higher education," Journal of Further and Higher Education, **34**, 59-71, 2010. https://doi.org/10.1080/03098770903477102

[4] V. Narayan, J. Herrington, and T. Cochrane, "Design principles for heutagogical learning: Implementing student-determined learning with mobile and social media tools," Australasian Journal of Educational Technology, **35**, 2019. 2019. https://doi.org/10.14742/ajet.3941

[5] A. B. N. R. Putra, H. A. Syafrudie, J. M. Yunos, A. M. Nidhom, A. A. Smaragdina, and A. I. Sembiring, "Analysis of the Necessity for Heutagogical Approach Through 4Cs Skills as Innovation for Vocational Lectures in the Education 4.0," in 1st Vocational Education International Conference (VEIC 2019), 2019, 362-371.

[6] H. Praherdhiono, E. P. Adi, and R. N. Devita, "Understanding of Digital Learning Sources with the Heutagogy Approach using the K-Means and Naive Bayes Methods," in 2018 4th International Conference on Education and Technology (ICET), 2018, 23-27.

[7] K.-T. Wong, N. binti Abdullah, and A. Hamdan, "Mobile-Heutagogical Practices among Student Teachers: Its Pedagogical Affordances and Challenges," International Journal of Interactive Mobile Technologies (iJIM), 14, 130-143, 2020. https://doi/10.3991/ijim.v14i02.11819

[8] W. Kung-Teck and M. A. Yeop, "Modelling the Factor Influencing the Implementation of Mobile-Heutagogical Practices among Teachers: An Application of Invariance Multi-Group Structural Model," International Journal of Learning, Teaching and Educational Research, **18,** 1-16, 2019. https://doi/10.26803/ijlter.18.12.1

[9] A. S. Preece and P. K. Hamed, "Andra-Heutagogy: A New Approach For Teacher Training," International Journal of Education and Pedagogy, **2**, 98-105, 2020.

[10] L. M. Blaschke and S. Hase, "Heutagogy: A holistic framework for creating twenty-first-century self-determined learners," in The future of ubiquitous learning, ed: Springer, 25-40, 2015. https://doi/10.1007/978-3-662-47724-3

[11] Z. A. Hassan, P. Schattner, and D. Mazza, "Doing a pilot study: why is it essential?," Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia, **1**(70) 2006.

[12] E. M. Mitropoulou, I. Tsaousis, D. Xanthopoulou, and K. V. Petrides, "Development and Psychometric Evaluation of the Questionnaire of Ethical Leadership (QueL)," European Journal of Psychological Assessment, 2019.

[13] J. Arbuckle, "Amos (Version 7.0) SPSS, Chicago," IL,2006.

[14] T. K. Tee, S. Saien, F. Rizal, Sukardi, Risfendra, M. H. Yee, M. M. Mohamad, W. Othman, M. N. A. Azman and N. Azid, "Design And Technology Teacher In TVET: A View On Thinking Style And Inventive Problem-Solving Skill," Journal of Technical Education and Training, **12**(1),197-203, 2020. https://doi.org/10.30880/jtet.2020.12.01.021

[15] A. Putra, H. Syafrudie, A. Nidhom, A. Smaragdina, J. M. Yunos, and A. Sembiring, "The innovation of module training based heutagogy as an acceleration for increasing pedagogical supremacy of vocational education lecturers in the industrial revolution 4.0," in Journal of Physics: Conference Series, 012043, 2020. https://doi.org/10.1088/1742-6596/1456/1/012043
[16] S. Hase, "Learner defined curriculum: Heutagogy and action learning in vocational training," Southern Institute of Technology Journal of Applied Research, 1-10, 2011.
[17] M. Snowden and J. Halsall, "Exploring the application of a self-determined approach to learning," International Journal of Innovation and Learning, **22**, 293-303, 2017. https://doi.org/10.1505/IJIL2017.086730
[18] A. Chimpololo, "An analysis of heutagogical practices through mobile device usage in a teacher training programme in Malawi," 2019.
[19] J. Purnomo, "ICT literacy of high school mathematics teacher: online learning competence with heutagogical approach," in Journal of Physics: Conference Series, 2019
[20] M. Rathakrishnan and A. Raman, "Heutagogy Approach Using Trello Online Learning on the Critical Thinking Skills Amongst Students With Different Learning Styles," in Redesigning Higher Education Initiatives for Industry 4.0, ed: IGI Global, 41-55, 2019. http://doi.org/10.4018/978-1-5225-7832-1.ch003
[21] N. Azid, H. Ridzuan, T. K. Tee and M. H. Yee, "Malaysia and China Students' Feedback on the Implementation of Critical Thinking Pedagogy: A Case Study," International Journal of Advanced Science and Technology, **29**(3), 227-237, 2020.
[22] M. H. Yee, N. Hamdan, B. C. Kok, T. K. Tee, N. Azid, "Development of Integrated Creative and Critical Thinking Module in Problem-Based Learning to Solve Problems," International Journal of Scientific & Technology Research, **9**(3), 6567-6571, 2020.
[23] M. Gribbins and V. Cook, "Workshop: Emerging technologies and practices that promote student-centered learning," in Proceedings of the Twelfth Midwest Association for Information Systems Conference (MWAIS), Springfield, Illinois, 2017.
[24] T. G. a. F. Bond, C.M., Applying the Rasch model: Fundamental Measurement in the Human Sciences Second ed. New Jersey: Lawrence Erlbaum Associates., 2007.
[25] M. Gribbins and V. Cook, "Workshop: Emerging technologies and practices that promote student-centered learning," in Proceedings of the Twelfth Midwest Association for Information Systems Conference (MWAIS), Springfield, Illinois, 2017.
[26] G. Rasch, "Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago Press," 1960.
[27] H. Othman, N. A. Ismail, I. Asshaari, F. M. Hamzah, and Z. M. Nopiah, "Application of Rasch measurement model for reliability measurement instrument in vector calculus course," Journal of Engineering Science and Technology, vol. 10, pp. 77-83, 2015.
[28] A. M. Talib, F. O. Alomary, and H. F. Alwadi, "Assessment of Student Performance for Course Examination Using Rasch Measurement Model: A Case Study of Information Technology Fundamentals Course," Education Research International, vol. 2018. https://doi.org/10.1155/2018/8719012
[29] W. Fisher, "Rating Scale Instrument Quality Criteria. Rasch Measurement Transactions", **21**(1), 1095, ed, 2007.
[30] J. M. Linacre, "A user's guide to FACETS rasch-model computer programs,", **18**, 2018, 2014.
[31] M. B. Paulsen and L. W. Perna, "Higher Education: Handbook of theory and research" vol. 31: Springer, 2016.
[32] M. Illhan and G Nese, A Comparison of Difficulty Indices Culculated for Open-Ended Items According to Classical Test Theory and Many Facet Rasch Model. Eurasioan Journal of Educational Research **75**, 99-114, 2018. https://doi:10.14689/ejer.2018.75.6
[33] T. Bond and C. M. Fox, Applying the Rasch model: Fundamental measurement in the human sciences: Routledge, 2015.
[34] J. P. Meyer, Applied measurement with Metrik: Routledge, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 247-255 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

IT GRC Smart Adviser: Process Driven Architecture Applying an Integrated Framework

Meriyem Chergui[1,*], Aziza Chakir[2]

[1]*Computer sciences Department, Higher National School of Electricity and Mechanics, Hassan II University, Casablanca, 20000, Morocco*

[2]*Economic sciences and management Department, Faculty of Economic and Social Legal Sciences, Hassan II University, Casablanca, 20000, Morocco*

ARTICLE INFO

Article history:
Received: 04 August, 2020
Accepted: 21 October, 2020
Online: 10 November, 2020

Keywords:
IT Governance
IT GRC frameworks
Artificial Intelligence
Knowledge management
Information System
Business Processes
IT processes

ABSTRACT

This article is a continuation of the work presented in the Third World Conference on Smart Trends in Systems Security and Sustainability (WorldS4). In this version we focus on matching IT and business processes from different management levels by using IT-Governance (ITG) Risk and Compliance frameworks in a smart way. In fact, every information system (IS) has two main interfaces with two key systems: Operating System (OS) and Decision System (DS). Every system has his own frameworks and best practices for efficient management. IT Governance is the ability to control IT strategy implementation by ensuring business and IT alignment for profitable digital services. Also, the variety of IT Governance frameworks for each hierarchical level in the enterprise, should be efficiently deployed together in order to have coherent recommendations for IS, OS and DS users. It is why we use COBIT 5 as a strategic IT Governance framework able to match and integrate other frameworks and best practices dealing with IT services, IT risks and compliance as well. We based the proposed IT GRC smart adviser on artificial intelligence and knowledge management as technical axes. This article presents, IT Governance frameworks, their matching problems and related works. Then it proposes a smart architecture with new functionalities to match COBIT processes with other frameworks in order to cover different management levels. The technical contribution of this article is to propose and implement an IT GRC smart adviser based on many frameworks and best practices to optimize and improve IT Strategies. The simulation was presented, the obtained results were compared to human experts' decisions and big similarities were observed. The aim of this article is to integrate knowledge from many IT GRC frameworks for better IT governance in a smart and easy way.

1. Introduction

Information system Governance is nowadays different from classical management of software and hardware. Many users with different profiles are interacting directly or indirectly with it. They use IS to enrich their data sources through operation systems deployed in different gadgets for simple use. They use it to answer daily questions as well such as production rate, new clients number and providers' expectations. The created data is the most valuable resources ever. It enables CEO and other Business offices to take the right decision. But good IT governance allows to take this decision in the right time through the right IT solution.

The real challenge of top managers is to use digital solution to create value. They should also manage risk, and comply with regulations. The decision is no longer IT responsibility. It achieves business and technical offices as well. The use of one IT governance framework in its document version is no longer enough: It is true that they contain very advanced know-how and feedback, but it requires experts, time and a large budget to take advantage of them. These three elements: time, budget and knowledge are very expensive in the current market due to its severe competition and the law of the most powerful.

As a result, digital tools are mandatory to facilitates the governance of the information system. It is also a need for

*Corresponding Author: Meriyem Chergui, Email: chergui.meriyem@gmail.com

performance management system to ensure the right and efficient use of these frameworks.

Consequently, in this new multi-user, multi-partner environment, the CIO's missions remain information system strategy definition, IT resources management and technical problems resolution as well. Meanwhile, with Big data and cloud computing matters, we notice the emergence of new technical and business risks that should be managed effectively by IT and business together. To perform these tasks, many IT governance frameworks should be deployed in easy to use digital solutions as we did in [1]. But the main problem is how to integrate them in the same personalized repository for every company. How can we design an IT Governance, risk and compliance framework that matches the company's needs with optimal time, minimum budget and good software quality insurance? How can we integrate Data Governance into IT Governance in a fluid and value-creating way by taking in consideration Governance critical success factors [2]?

To answer these question, we propose an intelligent IT Governance model based on COBIT 5 as a business driven repository for corporate and IT governance. In fact, COBIT 5 offers with a common, non-technical and technology-independent language a management framework able to align governance and management standards. By providing a simple architecture to structure guidance documents, Cobit 5 contains knowledge previously distributed in different standards like COBIT 4, Risk IT, Val IT, ITAF and BMIS. It also allows practical matchmaking with different management levels' frameworks like ITIL, CMMI, ISO 27000 family and COSO.

In this article, we propose an update of the smart IT Governance adviser architecture [1] to allow IT services, IT risk management and compliance efficient management, in addition to matching stakeholders' motivations with real business goals [2].

We will add new layers communicating with the semantic model, to link stakeholders and digital solutions via IT Governance ontology designed and created in [3]. We also propose a smart matching between every COBIT process and its correspondence with other frameworks. In IT Governance jargon we call that "Appling a unique and integrated standard principal" for better alignment as presented in [4].

After the introduction in the first section, the second section presents the most pertinent IT Governance Risk and Compliance frameworks and then exposes an overview of their matchings with COBIT 5. The third section gives a brief state of art of Knowledge management as a technical axe. The fourth section analyses briefly related works. The fifth section presents the smart IT Governance adviser updates and the knowledge management system architecture. The fifth section, discusses the new model, its uses and its extensions, the sixth section details the model implementation and results analysis. In the last section, the article is concluded and work perspectives are listed.

2. IT GRC frameworks

In the ITG literature, numerous frameworks are presented, with advantages and limitations of each of them. Thus, each framework offers specific recommendation according to its domain. In practice, ITG should makes these repositories coexist for better results, since they present a necessary complementarily for pertinent decisions. We present some of the main standards in terms of enterprise management, services, projects management and IT security to show how IT governance frameworks drive accountability and process maturity [5]. In what follows, we present the most used frameworks namely ITIL, CMMI, ISO27001 and IT BSC.

2.1. ITIL

IT Infrastructure Library (ITIL), is an approach documentation for IT service management, to support business organizations users. It is created by the British government in 1990 [6]. In 2004, ITIL v2 was created with 9 books, focusing on the link between technology and business, and based on processes to provide the right services to customers. Many companies are using ITIL to improve IT services control and good results are as for processes improvements. ITIL service is a way to deliver value to a customer in an easy way and with low cost and risk. In 2007, ITIL V3 was implemented on five best practice books, offering supplements by sector or by market as well as generic models (process maps). This version is not subject to certification, and it offers continuous improvement of the services offered to IT Department clients. The latest version is ITIL V4. its main advantage is the integration of new agile practices and Devops method. It has developed 34 practices, in 3 themes instead of 26 processes in 5 categories.

- 14 in "General Management practices",
- 17 in "Service management practices"
- 3 in "Technical management practices".

ITIL V4 edition book introduces new like Service Value System (SVS) and the four-dimensional model.

2.2. ISO 27001

ISO 27001 is a framework to implement information security management system (ISMS). Released in November 2005, it also provides a model for operation, monitoring, and review the ISMS. It also enables the creation and implementation of good IT service management, a s presented in [7]. The risk management approach ISO 27001 uses is a top-down and neutral method. It is a planning process with six parts:

- Security policy definition
- WSIS scope definition.
- Risk assessment performance.
- Risk identification.
- Goals and controls implementation
- Applicability declaration.

ISO 27001 deals with responsibility management, internal audit and continuous improvement without specific information security controls. It should be noted that in the accompanying code for each practice, a checklist must be taken into consideration.

The last version is ISO27001 with 2014 and 2015 corrections. It highlights information security management systems requirements with special focus on information technology and security techniques.

2.3. IT Balanced Scorecard

Balanced Scorecard (BSC) [8] is the framework explaining company vision and strategy. It suggests action plans to give internal processes feedback with continues improvement. The BSC, according to its authors Robert Kaplan and David Norton, allows for traditional financial results, but these results highlight the past, which was normal in the industrial era, with long-term investments. There are insufficient customer relationships. These financial elements are insufficient, however, to control companies in the digital age, which should be based on their future value by investing in customers, suppliers, employees, processes, technology and innovation. ". The BSC offers a management tool, organized on four aspects:

- Financial aspect: define the financial benefit provided to shareholders
- Customer aspect: market and customer expectations;
- Internal process aspect: potential improvements to be made to the company's internal processes, which bring value;
- Future construction aspect: mobilization of human resources and infrastructure for improvement and learning.

It is assumed that this approach leads to a good vision of corporate Governance.

2.4. CMMI

The Capability Maturity Model (CMM) [9] is a project management framework developed by the Software Engineering Institute (SEI) of Carnegie Mellon University. It describes software development maturity practices and principals. With an ad hoc path, CMMI proposes to software development organizations a process improvement capability. IT contains a list of models like MM Software, CMM systems engineering and integrated development of CMM products. These models were merged and extended into CMM integration known as "CMMI" with either a stage view or continuous view. As for the staged view, it offers five organizational maturity level (initial, managed, defined, quantitatively managed and optimization). As for the continuous view, it includes six levels of processing capacity (incomplete, executed, managed, defined, quantitatively managed and optimization).

2.5. IT GRC Frameworks matching with COBIT

For better IT Governance, IT department must deal with each task by deploying it in its real context and link it with a project or a process. IT processes are the starting points while governing an information system. For example: deploying an ERP; computer backup organization; servers' security; or new IT supplier integration;

An IT process is defined as a set of activities around a specific event to measure results for customers [10].

While analyzing the IT GRC frameworks before, we remark that process-driven aspect is almost in all them. IT changes approaching components into active components inside the organization. The IT process is measurable in effectiveness and efficiency by defining activities to implement, responsibilities and controls. At this level, each repository focuses on one or more concepts to find more priority and to zoom on either governance risk or compliance.

Meanwhile, process management is highly recommended in the agile environment we are living in. This type of management allows [11]:

- Requirements understanding
- Processes added value;
- Process performance and efficiency measurement;
- Positive communication between stakeholders.
- Top Management easy and measurable decisions.

Technically speaking, IT process is easily designed as a simple or complex function or procedure with inputs and outputs. In this article encapsulate best practices and frameworks processes into computerized executable functions.

a) ITIL –COBIT

COBIT identifies what the CIO should do, ITIL prescribes how it should be done to optimize the use of IT resources. ITIL and COBIT have several common points (for example the BAI10 process in COBIT is equivalent to configuration management process in ITIL).

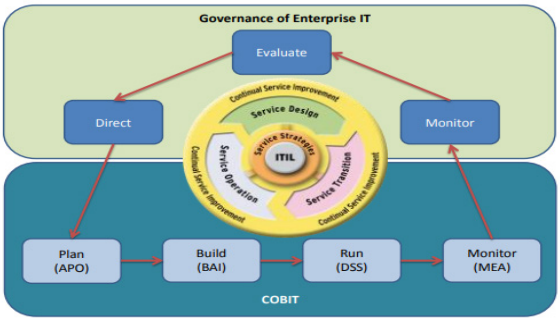


Figure 1: Matching COBIT with ITIL

COBIT 5 proposes the convenient IT investment to link priorities with objectives in both IT and business context (cascade principle). We have already seen it in [1]. The cascade of COBIT 5 objectives makes it possible to align IT objectives and company objectives, which makes it possible to refine and prioritize IT processes. the cascade of COBIT 5 objectives allows stakeholders to answer the following two questions:

- How to prioritize areas of interest when applying ITIL v3 processes?

- How to define the right maturity level and the efficient processes for better business decisions?

One last thing, COBIT 2019 is also correctly matched with ITIL v4 which allow us to upgrade our system regarding future IT GRC Frameworks versions.

b) CMMI –COBIT

Traditional mapping is designed to be used in one direction. Indeed, a user searching for an element in COBIT model finds the element or elements linked to the appropriate CMMI model. Others by the reverse probably exist in several places on the map (several elements of CMMI for a COBIT element) which is not easy to isolate to determine what is linked exactly. These traditional cards are unidirectional and they are valid for the correspondence of all IT GRC repositories. In this perspective, COBIT 5 CMMI Practices Pathway Tool, has emerged as a common Excel tool made by ISACA and CMMI, intended for users who want to navigate freely from COBIT 5 to CMMI and vice versa[12]. The dynamic and permanent correspondence between CMMI and COBIT makes it possible to limit maturity in the different levels of governance, which allows self-assessment and continuous development.

c) ISO27001 -COBIT

There is a mapping between COBIT 5 and ISO 27001 domains: “COBIT 5 for information security” this repository presents a mapping guide aiming at implementing the domain area of COBIT process with ISO27001 controls[13].

3. Knowledge Management

3.1. Definition

Knowledge Management (KM,) is a multidisciplinary research area offering methods, techniques and assistant tools: creation, acquisition, sharing and use of knowledge in an organization, with a view to 'improve individual, collective or organizational learning [14]. Knowledge is defined in a complementary way by raising one of his systemic aspects namely:

- Ontological (what is knowledge),
- Functional (what it does),
- Genetics (genetics here has the meaning of genesis not inheritance)
- Transformational (where it comes from and what it becomes)
- Teleological (why).

In artificial intelligence context, different types of knowledge are identified:

- Procedural knowledge to explain how can we solve a given problem is solved. It indicates how to complete a task;
- Declarative knowledge: it is statements describing objects and concepts and focusing on problem resolution indications.
- Meta Knowledge is specific domain knowledge explanation. For example, what is suitable at a given situation e to solve a given task;
- Heuristic knowledge: also called surface knowledge. It is empirical knowledge acquired by an expert through his past experience, it is used to guide reasoning;
- Structured knowledge: describes a mental model of the expert in the form of a structure (concepts, sub-concepts, objects, etc.).

3.2. Design methods

There are 3 approaches of Knowledge management modeling:

a) Bottom-up modeling approach:

The bottom-up or data-driven approach is based on a step of elicitation of expert knowledge (interviews, document analysis, task analysis, etc.) followed by a conceptualization step. Example: KOD and MIKE [14]. One of the advantages of the bottom-up approach is that it leaves experts free to express their perceptions and their tasks without constraining them. In addition, developed models correspond to existing points of view. However, the high cost in terms of time and expertise in the elicitation process, the difficulty of re-use as well as the complexity of the validation process are major drawbacks of the approach.

b) Top-down modeling approach:

This approach is called Model-Driven Approach. The development of knowledge models in this approach consists of finding pre-existing generic models and adapting them to the field and the application concerned. Two sub-approaches can be distinguished in this family: the select-and-modify approach and the modeling-by-composition approach from library elements. Example: CommonKADS [15], MASK [16]. The main advantages of this approach in addition to providing a support for knowledge acquisition are models generality and reusability. But adapting an existing model to a specific application is a very difficult task. In the literature, we also find mixed approaches which tried to take advantage of the two approaches like [17] and [18].

3.3. COBIT knowledge Base

Applying a single integrated framework is the third principle of COBIT 5 that proposes a knowledge management for COBIT alignment with other standards and frameworks for enterprise complete coverage. As shown in figure 2, Cobit 5 has a knowledge base of current guidance and contents collected from old versions and other standards and frameworks. This knowledge base enriched by enablers such as data, IT processes, resources and ethics allow different stakeholders to find answers to their questions about best practices to deal with different IT Governance situations in different levels (strategy, services providing, risk management…etc.). In fact, simplicity, guidance materials structuring and consistent production are the main advantages of this architecture that integrates different frameworks. knowledge in one shot [19]. IT Governance repositories are a set of knowledge sources capitalized by experts for better Information system management in different types of organizations. In this perspective, we are working on a smart IT Governance Knowledge

Management system that will be interfaced with the adviser in future version.

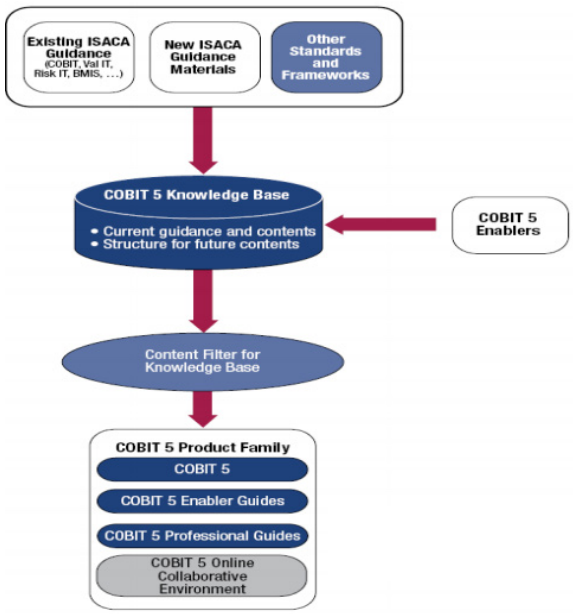


Figure 2: COBIT 5 Knowledge base

4. Related works

To assess the state of art of IT GRC platforms, a rigorous evaluation of the fourteen main suppliers [20] is presented in table 1. A comparison of the strengths of each of the solutions was established from the Forrester description, functionalities presented on the official sites and the demos available on the net, as long as the solutions are all proprietary.

Table 1: GRC existing platforms

Vendor	Product	Version	Strong points
ACL	ACL	ACL Analytics 12.5 ACL Analytics 12.5 Exchange 6.5	- Financial control analysis and testing -Audit solution - Community of practitioners -Simple user interface, mobile support and analytical integration.
Enablon	Enablon 8	V8.4	-Operational risks - Legal and regulatory aspect - Environmental, health and safety management - Corporate social responsibility initiatives, litigation, a responsible supply chain, air quality management or public service data.
IBM	IBM Ope,Pages	V7.3	-Financial control management -Operational risk management
LogicManager	LogicManager	V17	-Humain advisors availability
			-Fastest average deployment times
MetricStream	MetricStream M7 GRC PLatform	V7.0	-Energy, health care, insurance, food and beverage and automotive -Good user interface
Nasdaq	BWise	V5	-Middle market and information security organizations
NAVEX Global	PolicyTech NAVE Engage RiskRate EthicsPoint	Sep 1, 2017 July 31, 2017 August 30, 2017 August 15, 2017	-Compliance -Data Management
Riskonnect	Riskonnect Integrated Risk Management	GRC 2017.2	-Support 72 languages -GRC workflow and dashboards -Adapted to Salesforce.com clients
RSA Archer	RSA Archer	V6.2	- IT GRC - Financial risks and regulatory change management
Rsam	Rsam GRC	V9.2	-Data collection from different platforms -Risk analysis and management
SAI Global	Compliance 360	Sept 7, 2017	-Health care and IT GRC
SAP	SAP solutions for GRC	V10.1	- Big data and predictive analytics -Audit tool
ServiceNow	ServiceNow GRC	Jakarta	-IT Services management -IT GRC -Risks to information security and IT operations
Thomson Reuters	Thomson Reuters Connected Risk	Sept 7, 2017	-Risk and Compliance -Financial control management -Managing regulatory changes

The comparative analysis and the measurement of the effectiveness of these solutions in relation to the ITG dependency factors allow us to deduce that:

- In most cases, it is GRC solutions without IT GRC functionalities,
- The majority of these solutions even if they have IT GRC functionalities, it is ERP solution module to be deployed with the rest of the modules.
- The cost of setting up is high (by consulting the list of their references on the market and their targets).

- Need for external advice for better operation (possibly advice from the seller of the solution).
- There is a discrepancy with regard to the ITG implementation in the IS of various companies (size, turnover, type, maturity, location, etc.).
- They do not concern all of Information System stakeholders and all areas of the company at the same time.

5. Proposed model

This project started by EAS team [21] to propose an IT GRC digital solution with five intelligent modules for business and IT alignment by mapping processes. It also manages IT risks and measures business impacts. This was done by choosing for every process the right framework(s). There were many limitations for this architecture namely as far as execution time is concerned and also many communication problems among distributed systems which cause the decrease of IT Governance service's quality. IT is the reason why in[1], we implement the strategic level via a smart advisor to optimize the IT Governance semantic model. This solution main advantage was the "All in one" service and the easy use for final users. The advisor is a simple chatbot to which IT manager, Business Manager, CEO or even a project manager can ask IT Governance questions and he is answered simply and quickly.

In this version, we kept the same semantic engine model that interprets user real time questions using test mining. The same IT Governance ontology [3] to match concepts and relations but we replace our knowledge base by COBIT knowledge base. Cobit knowledge base, as shown before, is fed by Cobit but also the other repositories (ITIL, CMMI and ISO27001 at first) with processes matching.

This change allows the semantic engine not only to answer via Cobit guidelines but also ISO27001 for risk management, ITIL for services Governance and CMMI for IT project management in implicit way. The unique Knowledge base improves the performance of the system and ensures the efficiency of responses by arriving at a more delicate granularity for each process.

Indeed, the smart adviser is mainly made of an expert system applying cascading principles (1st principal of Cobit 5) to unify objective and value creation definition among different stakeholders. By updating the semantic engine, we also unified guidelines in hierarchical way.

Every stakeholder will find deeper guidelines depending on what he is asking for: a top manager will be satisfied by Cobit suggestions but CIO or a business manager will need more details that he will find properly in ISO27001, ITIL or even CMMI. So, by deploying the cascading principle and the single unified principle, the IT GRC smart adviser deep dive enterprise needs to suggest more operational recommendations.

The business driven architecture has a chatbot to ask stakeholder questions about eventual risks, resources management, value creation…etc) according to users' answers the smart advisor cascade his business request into enterprise goals. Then, depending on the questioned profile, the smart advisor asks for more details to match enterprise goals with IT goals.

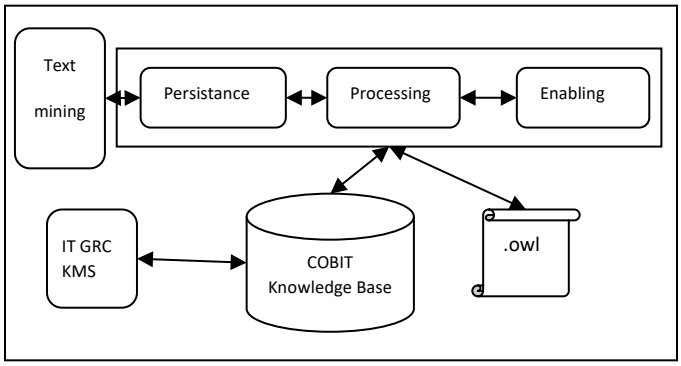


Figure 3: Semantic engine update

Every company description and characteristics should be previously configured in the knowledge base facts to define IT Governance road map. The proposed solution also allows cascading IT decision according to COBIT 5, through the semantic engine as shown in figure 2. After that, IT related goals are matched to enablers in a new abstraction level compared to [13]. The proposed semantic engine is now a five layers' system:

- Text mining layer annotate stakeholder answers.
- Persistence component storages stakeholder needs in the knowledge base
- Processing component compares stakeholder's request to Cobit 5 suggestions from knowledge base.
- Enabling component chooses the best facilitator (s) to implement.
- IT GRC KMS is a knowledge management system to feed Cobit Knowledge base with matched processes and guidelines.

As for Cobit Knowledge base, it is the set of rules and facts related to each COBIT process and domain one by one with: CMMI, ISO27001 and ITIL. In this level we are not updating the ontology since the system will interpret user questions through COBIT cascading.

6. Discussion

In previous work [1], The smart adviser contains a matchmaker agent between the semantic engine described above and IT GRC business driven configuration.

The matchmaker verifies first the company static parameters to initiate the IT GRC road map. It also enables the stakeholder to ask questions and have answers from the semantic engine in Chatbot way. Meanwhile, matchmaking research and priority algorithm detailed in [22] match every request with best guidelines according to a corresponding rate. The indexing algorithm does the following:

- Text indexing according to the words that compose it.
- Similarity test with inverted index.
- Returning similar texts with rank for every user query.
- Knowledge extraction from Cobit Knowledge Base.

Table 3: Proposed solution and state of the art comparison

Solution	Simple user interface	Risks management	- Legal and regulatory aspect	Human adviser	Deployment times	information security	IT GRC	Workflow and dashboards	Data management
ACL	X								
Enablon		X	X						
IBM		X							
LogicManager				X	X				
MetricStream	X								
Nasdaq						X			
Riskonnect	X							X	
RSA Archer		X	X				X		
Rsam		X							X
SAI Global							X		
SAP								X	X
ServiceNow		X	X				X		
Thomson Reuters		X	X						
NAVEX Global			X						X
Proposed solution	X	X	X	X	X	X	X	X	X

The smart advisor update is done in this step: Once similarity ranks are fixed, an inference engine questioned the rules base according to a maximization algorithm. The rules base communicates simultaneously with the fact base and the inference engine to give the convenient answer according to COBIT first and later to ITIL, CMMI and ISO27001 (see figure 3). The answer is returned to the matchmaker agent who sends it to the active adviser. One last thing is that two or three advisors could be handled for the same request successively for IT Governance, risk management or project management as well.

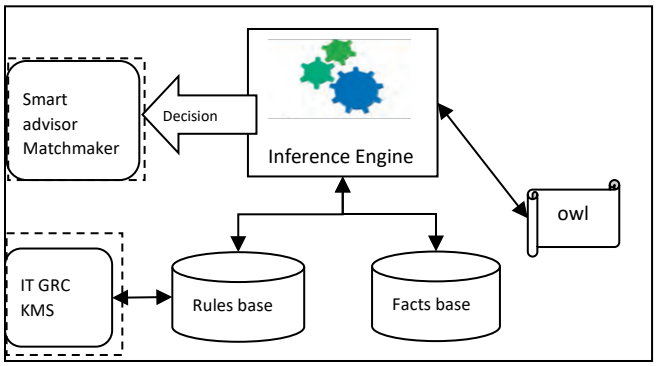


Figure 4: Cobit Knowledge Base architecture

In table 2 we compare the proposed IT GRC smart adviser to solutions presented in table 1:

To summarize, the update of the smart IT Governance adviser understand stakeholders' needs by annotating their questions and finding answers in Cobit 5 cascading model, via a text mining layer. This component is mainly managed by the processing entity connected with the Cobit knowledge base to confront them at this last base of rules to elaborate a roadmap with many granularity levels (IT Governance, risk management, services managements and IT project management. Cobit knowledge base is mainly made of facts bases and rules base linked to IT GRC Knowledge management system (that we will detail in a future work). Rules base is regularly actualized through IT GRC new repositories and expectations. As for Facts base, it is actualized by the organization information system parameters. The inference engine feeds the smart advisor with the appropriate decisions after semantic analysis and primitive matching. In addition, for every level facilitator implementation is also given, with a balanced squared card dimension, the result road map enables a prioritization of processes implementation.

As for the processing and analysis steps compared to [1], we:

- update the semantic engine,
- implement the COBIT knowledge base,
- link the KB with the matchmaker agent,
- implement the chatbot
- annotate answers
- test the adviser
- compare its results with a human expert's decisions
- evaluate the adviser performance

7. Implementation

To validate experimentally the proposed architecture, we implement and improve different versions of the platform. Let's give an overview of the IT Governance platform evolutions:

The first and second versions were java simulators, the third, fourth and fifth version were web application where, in addition to the proposed functionalities, communicate with the global EAS-IT GRC platform[23]. In the first version, we simply validated the communication of different intelligent agents with predefined business goals. In the second version[24], we added to the first version kernel, the knowledge base where we stored the basic information and the different mediation rules and we implemented an agent launch interface to simulate exchanges. We have carried out tests to validate results. In the third version, we have implemented almost all of the features offered with interfaces of potential users;

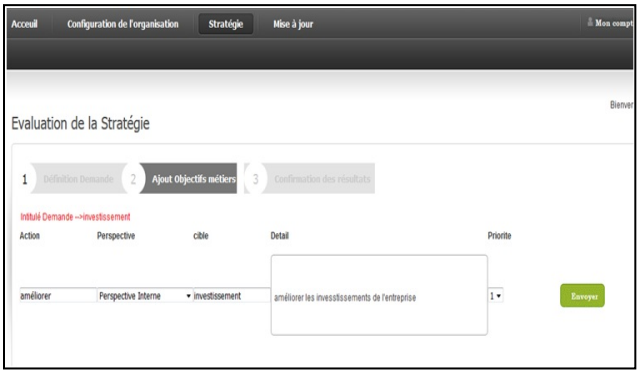
- IT Governance Ontology of information systems governance that we deployed in the mediation layer for the interpretation of offers and requests;
- Global comparison algorithm to improve the expert mediation system;

- Update repository with version 5 of Cobit;
- Request logging and learning to develop the proposed offers.

As for the fourth version presented in [1], we simplify user web interface and encapsulate IS Layer in a configuration agent for better device coordination in multiple access case. We also optimize the semantic engine model, to extract IT Governance knowledge.

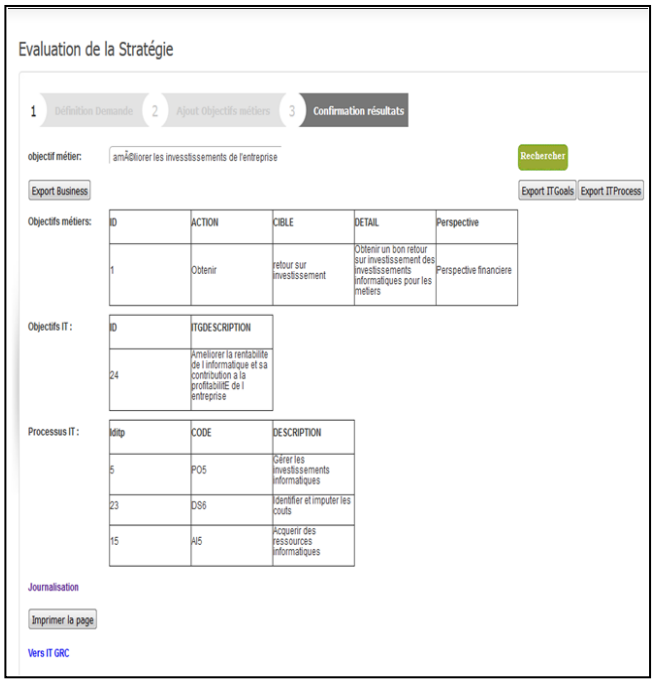
In the fifth version, presented in this paper, we add two main contributions namely the semantic engine knowledge management updater and the knowledge base detailed architecture.

Here is the fifth version IT Governance user query interface:



After a secure authentication and configuration questions interface, figure 5 shows user requests about the IT strategy or business strategy related to digital issues. An order of priority for every request is defined by the user (value from 1 to 5).

Figure 6 shows results as recommended business objectives, IT objectives and IT processes to focus on. A report is generated with other details such as facilitators, controls and metrics, maturity models, key activities and the responsibility matrix.

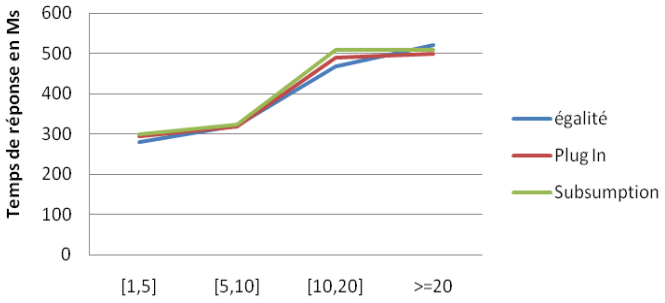


The application presents the detail of the request and the corresponding offers before allowing the user to edit the report, this primary summary allows him to decide whether it is necessary to start detailed ITG processing or obtained results are sufficient.

Many functional tests have been done. The corresponding framework of every objective is calculated via the priority function of the semantic engine. It also depends on the attribute and search criteria in the knowledge base. Let's compare as example the results of the platform with experts' decision about the request: "obtain relevant information and make strategic decisions in relation to risks" (experts proposed framework is ISO27001).

The results of the platform are presented in table 3. The request corresponds to several business objectives (OB). Each OB is detailed on one or more IT objectives and each IT objective is explained with several IT processes. According to results in table 3, we can deduce that the adviser chooses ISO27001 like the expert. IT objectives are recommended with priority gap due to semantic engine priority calculation; an IT Objective could be proposed by the expert but appears in the solution with a lower priority value.

As for evaluating the adviser response time to enhance its performance to give business managers rapid recommendations. It was made on a PC with Windows 10 operating system Intel Core i7 processor, 2.66 GHz and 16 GB of memory. Three comparison matchmaking modes were tested, namely equality, plug-in and Subsumption. We varied the number of requests according to four intervals [1, 5] [5, 10] [10.20] and> 20. The number of results per request were also taken into consideration.



Obtained results are presented in figure 7. The adviser is efficient since obtained results are in milliseconds. Governance advisor [1] is then presented trough a smart semantic model updater with IT Governance knowledge system architecture to deal with frameworks matching problem. These two contributions were added to implement other IT Governance frameworks for a 360° IT Strategy view: strategic alignment, risk and resources management, performance management and compliance. We chose COBIT 5 as IT Governance generic framework and semantic Web, and knowledge management as technical background to implement this solution.

As for limitations, the actual version of IT GRC adviser does not deep dive compliance recommendations. Compliance processes design should be done according to regional regulations. The same remark is available for specific risks namely financial and health risks analysis. As for the state of progress and perspectives of this work, we are working on:

- IT Governance Knowledge management system
- Functional tests and proof of concepts.

These perspectives enable the IT GRC advisor to be more performing so as to achieve industrialization level. The difference is also minimal between the 3 matchmaking mods which means that it is indifferent to requests complexity. The correlation of response time with requests number is linear per part, which means the algorithms complexity is reasonable with the increase of iterations.

8. Conclusion and perspectives

In this article, we raise IT Governance frameworks matching problem. After an introduction, we present IT Governance frameworks and their correspondences with COBIT5 as strategic level repository. We also listed existing IT Governance platforms with their strong and week points analysis. An update of the IT Governance advisor [1] is then presented trough a smart semantic model updater with IT Governance knowledge system architecture to deal with frameworks matching problem. These two contributions were added to implement other IT Governance frameworks for a 360° IT Strategy view: strategic alignment, risk and resources management, performance management and compliance. We chose COBIT 5 as IT Governance generic framework and semantic Web, and knowledge management as technical background to implement this solution.

As for limitations, the actual version of IT GRC adviser does not deep dive compliance recommendations. Compliance processes design should be done according to regional regulations. The same remark is available for specific risks namely financial and health risks analysis. As for the state of progress and perspectives of this work, we are working on:

- IT Governance Knowledge management system
- Functional tests and proof of concepts.

These perspectives enable the IT GRC advisor to be more performing so as to achieve industrialization level.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Chergui, A. Chakir, H. Medromi, "Smart IT governance, risk and compliance semantic model: Business Driven architecture," in Proceedings of the 3rd World Conference on Smart Trends in Systems, Security and Sustainability, WorldS4 2019, 2019, doi:10.1109/WorldS4.2019.8903997.

[2] I. Alhassan, D. Sammon, M. Daly, "Critical success factors for data governance: a theory building approach," Information Systems Management, **36**(2), 98–110, 2019.

[3] C. Meriyem, S. Adil, M. Hicham, "IT Governance ontology building process: example of developing audit ontology," International Journal of Computer Techniques, **2**(1), 134–141, 2015.

[4] T. Huygh, S. De Haes, A. Joshi, W. Van Grembergen, "Answering key global IT management concerns through IT governance and management processes: A COBIT 5 View," in Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

[5] A. Joshi, L. Bollen, H. Hassink, S. De Haes, W. Van Grembergen, "Explaining IT governance disclosure through the constructs of IT governance maturity and IT strategic role," Information & Management, **55**(3), 368–380, 2018.

[6] H. Gunawan, "Strategic Management for IT Services Using the Information Technology Infrastructure Library (ITIL) Framework," in 2019 International Conference on Information Management and Technology (ICIMTech), IEEE: 362–366, 2019.

[7] A.S.O. Parra, L.E.S. Crespo, E. Alvarez, M. Huerta, E.F.M. Paton, "Methodology for Dynamic Analysis and Risk Management on ISO27001," IEEE Latin America Transactions, **14**(6), 2897–2911, 2016, doi:10.1109/TLA.2016.7555273.

[8] J.-H. Shin, J.-S. Choi, I.-H. Kim, "Development of IT BSC-based Assessment System to Measure BIM Performance for Architectural Design Firms," Journal of the Architectural Institute of Korea Planning & Design, **32**(1), 3–12, 2016.

[9] U. Durak, A. D'Ambrogio, T. Gerlach, "Applying IEEE Recommended Practice for Distributed Simulation Engineering and Execution Process for Modeling and Simulation Based Airborne Systems Engineering," in AIAA Scitech 2020 Forum, 896, 2020.

[10] T. Clauss, "Measuring business model innovation: conceptualization, scale development, and proof of performance," R&D Management, **47**(3), 385–403, 2017.

[11] N. Slack, A. Brandon-Jones, Operations and process management: principles and practice for strategic impact, Pearson UK, 2018.

[12] M. Ramakrishnan, A. Shrestha, A. Cater-Steel, J. Soar, "IT service management knowledge ecosystem–literature review and a conceptual model," 2018.

[13] R. Almeida, R. Lourinho, M.M. da Silva, R. Pereira, "A model for assessing COBIT 5 and ISO 27001 simultaneously," in 2018 IEEE 20th Conference on Business Informatics (CBI), IEEE: 60–69, 2018.

[14] H. de J.G. Antunes, P.G. Pinheiro, "Linking knowledge management, organizational learning and memory," Journal of Innovation & Knowledge, **5**(2), 140–149, 2020.

[15] S.-A.K. Kaya, N. Messaoudi, M. Bouhadida, A. Bennour, "Le Knowledge Management: Socle de Construction de Memoire de Projet," Réformes Economiques et Intégration En Economie Mondiale, **424**(3911), 1–21, 2016.

[16] M.S. Saleh, O. Ismail, A. Kamel, H. Hassan, "From CommonKADS to SOA Environment: An Adaptation Model," Arabian Journal for Science and Engineering, **43**(12), 7605–7619, 2018.

[17] P.G.C. Hector, J.-L. Ermine, V. Ribiere, A. Bennet, "A knowledge-based development model for primate cities of the developing world," International Journal of Knowledge-Based Development, **9**(4), 386–419, 2018.

[18] V. Bombardier, P. Charpentier, B. Almecija, "Design of a log sorting X-ray system based on domain ontologies," IFAC-PapersOnLine, **49**(12), 479–484, 2016.

[19] J.J. Kirton, B. Warren, "G20 Governance of Digitalization," International Organizations Research Journal, **13**(2), 16–41, 2018.

[20] R. Murphy, "The Forrester Wave™: Governance, Risk, And Compliance Platforms, Q1 2016," Cambridge, USA, 2016.

[21] M. Chergui, H. Medromi, A. Sayouti, "Intelligent audit of information systems by inter- organizational workflow," 2014 International Conference on Next Generation Networks and Services (NGNS), 1–6, 2014, doi:10.1109/NGNS.2014.7090888.

[22] M. Chergui, A. Chakir, H. Medromi, M. Radoui, A new approach for modeling strategic IT governance workflow, 2017, doi:10.1007/978-981-10-1627-1_22.

[23] A. Chakir, M. Chergui, H. Medromi, A. Sayouti, "An approach to select effectively the best framework IT according to the axes of the governance IT, to handle and to set up an objective IT," 2015 Third World Conference on Complex Systems (WCCS), 1–8, 2015, doi:10.1109/ICoCS.2015.7483311.

[24] A. Chakir, M. Chergui, J.F. Andry, "A smart updater it governance platform based on artificial intelligence," Risk, **8**, 9, 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 3, 249-258 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Improved Nonlinear Fuzzy Robust PCA for Anomaly-based Intrusion Detection

Amal Hadri[*,1], Khalid Chougdali[2], Raja Touahni[1]

[1]*LASTID Laboratory, Faculty of Science, Ibn tofail University, Kenitra, Morocco*

[2]*GREST Research Group, National School of Applied Sciences (ENSA), Kenitra, Morocco*

ARTICLE INFO

Article history:
Received: 05 February, 2020
Accepted: 10 May, 2020
Online: 28 May, 2020

Keywords:
KDDCup99
Principal Component Analysis
PCA
NFRPCA
IDS
Feature extraction methods
NSL-KDD

ABSTRACT

Among the most popular tools in security field is the anomaly based Intrusion Detection System (IDS), it detects intrusions by learning to classify the normal activities of the network. Thus if any abnormal activity or behaviour is recognized it raises an alarm to inform the users of a given network. Nevertheless, IDS is generally susceptible to high false positive rate and low detection rate as a result of the huge useless information contained in the network traffic employed to build the IDS. To deal with this issue, many researchers tried to use a feature extraction methods as a pre-processing phase. Principal Component Analysis (PCA) is the excessively popular method used in detection intrusions area. Nonetheless, classical PCA is prone to outliers, very sensitive to noise and also restricted to linear principal components. In the current paper, to overcome that we propose a new variants of the Nonlinear Fuzzy Robust PCA (NFRPCA) utilizing the popular data sets KDDcup99 and NSL-KDD. The results of the conducted experiments demonstrated that the proposed approaches is more effective and gives a promising efficiency in comparison to NFRPCA and PCA.

1 Background

Thanks to the major shift in technology tools in the twenty first century, the complexity of network security has greatly increased, which gave birth to highly developed attacks. There are several traditional security methods like firewalls, data encryption and user authentication. Those techniques are insufficient to protect the network systems against all the existing threats. As a consequence, they may be less effective in detecting several dangerous attacks. Therefore, we need to strengthen our systems by adding more powerful systems such as intrusion detection system (IDS). The IDS protects the network systems by preventing the eventual damages that could be caused by an intrusion. Commonly, there is 2 principal categories of IDS, misuse-based and anomaly-based. The misuse-based method aim to classify an attack via comparing its signature with the attacks currently existing in a database of signatures of attacks and produce an alarm if any malicious activity is detected. The most two well-known misuse detection methods are STAT [1] and Snort [2]. This technique has proved its effectiveness in detecting the attacks stored in the datasets but it can not detects new intrusions or attacks and maintaining the databases is very expensive. Hence anomaly-based detection was initiated by Anderson [3] & Denning [4], the fundamental idea behind this concept is to specify the normal behaviour or model and generate an alarm if the difference between an observation and the defined model surpasses a threshold already defined. The uniqueness of this concept is its capability to categorize new and unusual intrusions.

Nonetheless, the current network traffic data, which are usually tremendous, are in fact a big challenge to anomaly based IDS. This type of traffic may decrease the whole detection mechanism and lead frequently to a falsified classification accuracies. This kind of huge dataset often have redundant and noisy data, that can be very difficult to model.

To address that, many feature extraction techniques have been used to increase network IDS efficiency. For instance, the paper [5] used a Discrete Differential Evolution to recognize the most important features. The detection accuracies were enhanced significantly. Likewise, the work [6] presented an IDS that can detect several attacks by exploring just a small number of features, the algorithm utilized is called Ant Colony Optimization algorithm. In [7] the authors used a feature selection method called cuttlefish

[*]Corresponding Author Amal Hadri, LASTID Laboratory, Faculty of Science, Ibn tofail University, Kenitra, Morocco, amal.hadri@uit.ac.ma

which suppress the noisy and redundant data and simultaneously guarantee the quality of data. The authors of [8] proposed to utilize the Principal Component Analysis (PCA) and Fuzzy Principal Component Analysis (FPCA) as a pre-processing step, before applying the k nearest neighbor (KNN) classifier, the same authors suggest in another publication [9] an improved variant method called Robust Fuzzy PCA. The acquired results show the promising performance of the technique proposed with regard to network attacks detection, as well as false alarms reduction.

Neverthless, PCA [10]-[12] and its linear variants are known to be sensitive to noise and outliers, which can impact on the deriving principal component (PCs) [13], therefore they effect as well the results of classification. In addition to that, PCA allows uniquely a linear dimensionality reduction [14]. Hence, in the case of complicated structures like nonlinear structures, the data will not be correctly expressed in a linear space, linear variants of PCA will not be the optimal solution. To deal with this issue, NFRPCA (Nonlinear Fuzzy Robust PCA) [15] was suggested to calculate PCs by utilizing a non-linear technique.

Nevertheless, this method is based on the L_2-norm that is highly sensitive to outliers and it also squares the error, and so the model can have a much bigger error. So as to tackle this issue, we introduce a new variant of NFRPCA called L_p-norm NFRPCA. Note that this paper is an extension of work originally published in 2018 IEEE 5th International Congress on Information Science and Technology (CiSt) [16], in that work we suggested a variant of NFRPCA employing L_1-norm rather than the classical Euclidian norm. In this paper, we propose another variant of NFRPCA using L_p-norm, we conducted new experiments besides the ones previously proposed in [16].

The remainder of this paper is structured as follows: Section 2 deliver a brief presentation of PCA, Section 3 will present an overview of NFRPCA. We present the suggested techniques in Section 4, Section 5 is dedicated to give shortly an overview of the two popular datasets namely KDDcup99 and NSL-KDD. Section 6 presents the conducted experiments and discuss the results and conclusions are summarized in section 7.

2 Principal Component Analysis Method

Principal Component Analysis (PCA) [17] is as an exploratory data analysis tool that involves a dataset with observations on variables, it was employed extensively in several research areas. The principal concept of PCA is to change data into a downsized form and preserve most of the initial variance from the original data at the same time. In other terms, PCA major role is to change variables n that were correlated into uncorrelated state d, the uncorrelated variables d are usually called the principal components (PCs) [14] [18].

Consider we already have a data set of M vectors $v_1,\ v_2,\ v_3, ...,\ v_M$ where each vector is represented by N features. To obtain the PCs we comply to the strategy explained through the steps underneath:

- Compute the mean μ of the data set

$$\mu = \frac{1}{M}\sum_{i=1}^{M} v_i \tag{1}$$

- Determine the deviation from the mean

$$\theta_i = v_i - \mu \tag{2}$$

- Compute the covariance matrix of the corresponding data set:

$$C_{n*n} = \frac{1}{M}\sum_{i=1}^{M} \theta_i {\theta_i}^T = \frac{1}{M} A A^T \tag{3}$$

where $A = [\theta_1, \theta_2, \theta_3, ..., \theta_n]$

- Assume that U_k is the kth eigenvector of C, λ_k the related eigenvalue and let's say that the $U_{n*d} = [U_1, U_2, ..., U_d]$ is the matrix of these eigenvectors. Hence:

$$CU_k = \lambda_k U_k \tag{4}$$

- Sort the Eigenvalues in decreasing order, hence pick the eigenvectors (known as principal components PC_i)that have the largest Eigenvalues. We can compute the number of PCs that we could keep as follow:

$$\tau = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{5}$$

- Consider t as a new sample column vector, t is projected on the reduced subspace covered by the PC_i :

$$y_i = U_i^T t \tag{6}$$

3 Nonlinear Fuzzy Robust pca (nfrpca) Method

The Nonlinear fuzzy robust principal component analysis employed in the current paper was suggested initially by Luukka in [15]. It was inspired from the robust principal component techniques that Yang & Wang proposed in [19] that fundamentally introduced by Xu & Yuilles techniques [20] where PCA learning rules are associated to energy functions and they proposed a cost function in regard to outliers. In Yang & Wang's proposed methods the cost function was modified to be fuzzy and it involved Xu & Yuilles techniques as a particular case. We introduce briefly these methods in this section. Xu and Yuille [20] suggested an objective function, based on $u_i \in \{0, 1\}$:

$$E(U, w) = \sum_{i=1}^{n} u_i e(x_i) + \eta \sum_{i=1}^{n} 1 - u_i \tag{7}$$

The variables are defined as follows: η is the threshold, $U = \{u_i | i = 1, , n\}$ is the membership set, & $X = \{x_1, x_2, ..., x_n\}$ is the dataset. The principal objective is to minimize $E(U, w)$ with regard of u_i & w. It should be noted that w is a continuous variable and u_i is a binary

variable that engender optimization challenging to resolve with a gradient descent technique. To solve the problem employing the gradient descent technique the minimization problem is simplified into maximization of Gibbs distribution as underneath:

$$P(U,w) = \frac{exp(-vE(U,w))}{Z} \tag{8}$$

Where Z is the partition function confirming $\sum_U \int_w P(U,w) = 1$. The measure $e(x_i)$ might be, e.g. one of the below functions

$$e_1(x_i) = \left\| x_i - w^T x_i w \right\|^2 \tag{9}$$

$$e_2(x_i) = \|x_i\|^2 - \frac{\left\|w^T x_i\right\|^2}{\|w\|^2} = x_i^T x_i - \frac{w^T x_i x_i^T w}{w^T w} \tag{10}$$

To minimize $E_1 = \sum_{i=1}^n e_1(x_i)$ and $E_2 = \sum_{i=1}^n e_2(x_i)$, the gradient descent rules are

$$w^{new} = w^{old} + \alpha_t[y(x_i - u) + (y - v)x_i] \tag{11}$$

$$w^{new} = w^{old} + \alpha_t(x_i y - \frac{w}{w^T w} y^2) \tag{12}$$

Where $y = w^T x_i$, $u = yw$,$v = w^T u$ and α_t is the learning rate. The nonlinear case of PCA was presented as underneath :

$$e_3(x_i) = \left\| x_i - w^T g(y) \right\| \tag{13}$$

And $y = x_i\, w$ and g could be selected as nonlinear functions. The weight updating in this situation is

$$w^{new} = w^{old} + \alpha_t(x_i e^T w^{old} F + e_3(x_i) g(y)) \tag{14}$$

And: $F = \frac{d}{dy}(g(y))$.

The cost function proposed by Yang and Wang:

$$E = \sum_{i=1}^{n} u_i^{m_1} e(x_i) + \eta \sum_{i=1}^{n} (1 - u_i)^{m_1} \tag{15}$$

subject to $u_i \in [0,1]$ and $m_1 \in [1, \infty)$. Now u_i is the membership value of x_i associated to the data cluster and $(1 - u_i)$ is the membership value of x_i associated to the noise cluster and m_1 is the fuzziness variable. And $e(x_i)$ is the error between the class center and x_i.

As u_i is a continuous variable now, the complexity of an amalgamation of continuous and discrete optimization could be obviated and the gradient descent technique can be employed.

The gradient of E(15) is calculated with regard to u_i.

By choosing $\frac{\delta E}{\delta u_i} = 0$, we obtain

$$u_i = \frac{1}{1 + \left(\frac{e(x_i)}{\eta}\right)^{\frac{1}{(m_1 - 1)}}} \tag{16}$$

Replacing this membership back, we get

$$E = \sum_{i=1}^{n} \left(\frac{1}{1 + \left(\frac{e(x_i)}{\eta}\right)^{\frac{1}{m_1 - 1}}} \right)^{m_1 - 1} e(x_i) \tag{17}$$

The gradient with regard to w would be

$$\frac{\partial E}{\partial W} = \left(\frac{1}{1 + \left(\frac{e(x_i)}{\eta}\right)^{\frac{1}{(m_1 - 1)}}} \right)^{m_1} \left(\frac{\partial e(x_i)}{\partial w} \right) \tag{18}$$

Consider

$$\beta(x_i) = \left(\frac{1}{1 + \left(\frac{e(x_i)}{\eta}\right)^{\frac{1}{(m_1 - 1)}}} \right)^{m_1} \tag{19}$$

where m_1 is the fuzziness variable. If $m_1 = 1$, the fuzzy membership downsizes into hard membership and could be picked following the concept:

$$u_i = \begin{cases} 1 \; if \;\; e(x_i) < \eta \\ 0 \;\; otherwise \end{cases}$$

right now η is a though threshold in this situation. The setting of m_1 has no rule. We sum up NFRPCA steps in Algorithm1.

Algorithm 1 NFRPCA algorithm

Step 1: At first set the count of iteration $t = 1$, bound of iteration T, learning coefficient $\alpha_0 \in (0, 1]$ soft threshold η to a very small positive value then initialize in random way the weight w.

Step 2: As long as t is smaller than T, do steps 3-9.

Step 3: Calculate $\alpha_t = \alpha_0(1 - t/T)$, set $i = 1$ and $\sigma = 0$.

Step 4: As long as i is smaller than n, do steps 5-8

Step 5: Calculate $y = w^T x_i$, $u = yw$ and $v = w^T u$.

Step 6: Calculate $g(y)$, $F = \frac{d}{dy}(g(y))$, $e_3(x_i) = x_i - w^{old} g(y)$, then the weight is updated:

$w^{new} = w^{old} + \alpha_t(x_i e^T w^{old} F + e_3(x_i) g(y))$

In [15] $g(y)$ was selected to be a sigmoid function such as $g(y) = tanh(10y)$, F is the first derivative of $g(y)$.

4 The Proposed Methods

4.1 Nonlinear Fuzzy Robust L_1-norm PCA method

We can clearly remark that, NFRPCA technique utilize an Euclidian norm for computing the reconstruction error in (13), and it is widely recognized that the Euclidian norm usually squares the reconstruction error, thus the approach have a much bigger error. Consequently, this could falsify the results and deteriorate the quality of solutions. For the sake of addressing this issue, the paper [16] suggest utilizing L_1-norm to compute the reconstruction error.

The reconstruction error equation could be re-written as below:

$$e_3'(x_i) = \left\| x_i - w'^T g(y') \right\|_1 \tag{20}$$

$y' = x_i w'$ and g could be picked as nonlinear functions. Here the weight updating is

$$w^{new'} = w^{old'} + \alpha_t(x_i e^{T'} w^{old'} F' + e_3'(x_i) g(y')) \tag{21}$$

Where :$F' = \frac{d}{dy'}(g(y'))$.
Equally as in algorithm 1, we utilize updating weight to compute the PCs.

Algorithm 2 L_1-NFRPCA algorithm

Step 1: At first set the count of iteration $t = 1$, bound of iteration T, learning coefficient $\alpha_0 \in (0, 1]$ soft threshold η to a very small positive value then initialize in random way the weigh w.
Step 2: As long as t is smaller than T, do steps 3-9.
Step 3: Calculate $\alpha_t = \alpha_0(1 - t/T)$, set $i = 1$ and $\sigma = 0$.

Step 4: As long as i is smaller than n, do steps 5-8

Step 5: Calculate $y' = w'^T x_i$, $u' = y'w'$ and $v' = w'^T u'$.

Step 6: Calculate $g(y')$, $F' = \frac{d}{dy'}(g(y'))$, $e'_3(x_i) = x_i - w'^{old} g(y')$, then the weight is updated: $w'^{new} = w'^{old} + \alpha_t(x_i e'^T w'^{old} F' + e'_3(x_i) g(y'))$

In the proposed algorithm $g(y')$ was picked to be sigmoid like the function $g(y') = tanh(10y')$, & F' is the first derivative of $g(y')$. We use the term L_1-norm NFRPCA to refer to this technique in the rest of this paper.

4.2 Nonlinear Fuzzy Robust L_p-norm PCA method

The technique that we suggest is L_p-norm NFRPCA where we propose to substitute the L_2-norm with generalized L_p-norm in the computation of the reconstruction error, in order to minimize the large error that L_2-norm can cause. The reconstruction error equation would be re-written as follow:

$$e^{"}_3(x_i) = \left\| x_i - w^{"T} g(y") \right\|_p \tag{22}$$

Where $0 > p >= 2$.
And $y" = x_i w"$ and g could be selected as nonlinear function. Here the weight updating is

$$w^{new"} = w^{old"} + \alpha_t(x_i e^{T"} w^{old"} F" + e^{"}_3(x_i) g(y")) \tag{23}$$

Where: $F" = \frac{d}{dy"}(g(y"))$.

In the same way as in algorithm 1, we use the updating weight to compute the PCs. Finally the main steps of the proposed method is summarized in Algorithm 3.

Algorithm 3 L_p-norm NFRPCA algorithm

Step 1: At first set the count of iteration $t = 1$, bound of iteration T, learning coefficient $\alpha_0 \in (0, 1]$ soft threshold η to a very small positive value then initialize in random way the weight $w"$.
Step 2: As long as t is smaller than T, do steps 3-9.
Step 3: Calculate $\alpha_t = \alpha_0(1 - t/T)$, set $i = 1$ and $\sigma = 0$.

Step 4: As long as i is smaller than n, do steps 5-8

Step 5: Calculate $y" = w^{"T} x_i$, $u" = y"w"$ and $v" = w^{"T} u"$.

Step 6: Calculate $g(y")$, $F" = \frac{d}{dy"}(g(y"))$, $e"_3(x_i) = x_i - w"^{old} g(y")$, then the weight is updated: $w"^{new} = w"^{old} + \alpha_t(x_i e"^T w"^{old} F" + e^{"}_3(x_i) g(y"))$

The function $g(y")$ was picked to be sigmoid function, such as $g(y") = tanh(10y")$, and $F"$ is the first derivative of $g(y")$. We use the term L_p-norm NFRPCA to refer to this technique in the rest of this paper.

5 The Simulated Datasets and its Preprocessings

5.1 Datasets

In the experiments we conducted, the popular public intrusion data sets were used, they are called KDDcup99 and NSL-KDD. We present them shortly in the next subsections.

5.1.1 KDDCup99 Dataset

KDDCup99 [21, 22] is a dataset that contains many TCPdump raws, that was captured during 9 weeks. This dataset was prepared and still conducted by Intrusion Detection Evaluation Program called DARPA. The main aim of KDDCup99 is to establish a generalized data set to examine researchs in intrusion detection field.

The training data set represents 4 gigabytes of data compressed (most of it is binary TCP dump), it contains 4,898,431 records, while the test dataset contains around 311,029 connection records and each record in dataset contains exactly 41 features. The attacks existing in KDDCup99 are categorized as underneath:

- Denial of Service (DoS): cyber-attack where the perpetrator attempt to consume network or machine resources to make it unavailable or limited to its legitimate users.
- Remote to Local (R2L): Where the attacker control the remote machine pretending that he is a user of the system, by exploiting the system vulnerabilities.
- User to Root (U2R): by exploiting the vulnerabilities and the flaws existing in a machine or on a network, the hacker tries to start accessing from a normal user account as to get the root access privilege to the system.
- Probing: The intruder tries to collect useful information of all services and machines existing in the same network to exploit it later.

5.1.2 NSL-KDD Dataset

The NSL-KDD [23] data set has been created to alleviate a couple of the major shortcoming of the KDDCup99 data set. This new version has some improvement compared to KDDCup99 data set and has solved a few of its fundamental issues. The main advantages of NSL-KDD are as follow:

- It suppressed redundant instances in the trainining set.
- The test dat set does not contain replicate records.

- The current version allow us to use the whole data set. Correspondingly, the random selection method will be needless because of the reduction of the number of instances in both train set and test set. Consequently, the accuracy and consistency of the evaluation and review of research works will increase.

- Every complexity level group involve a couple of instances that is oppositely corresponding to the percentage of instances in the KDDCup99 dataset. Therefore, we obtain more realistic examination of various machine learning techniques.

5.2 Normalization Phase

The normalization phase is more than important, because it allow us to apply the techniques cited above on the data set in a correct manner. To perform effectively this phase, we replaced the discrete values with continuous values for all the discrete attributes existing in the data sets through the same process used in [10], the process is briefly clarified as follow: for every attribute y that acceptes x variant values. The attribute y is illustrated by x coordinates contains zeros and ones. E.g., the attribute of the protocol type, that acceptes three, values i.e. tcp, udp or icmp. Utilizing this logic, these values are modified to the following coordinates (1,0,0), (0,1,0) or (0,0,1).

6 Experiments and Discussion

In the current part of this paper, several experiments were performed to examine the efficiency of the suggested method utilizing KDDcup99 and NSL-KDD databases, for the sake of evaluating the effectiveness of the suggested technique. We compute the measures: detection rate (DR), F-measure and false positive rate (FPR) described underneath:

$$DR = \frac{TP}{TP + FN} * 100 \tag{24}$$

$$FPR = \frac{FP}{FP + TN} * 100 \tag{25}$$

$$FMeasure = \frac{2 * TP}{2 * TP + FP + FN} * 100 \tag{26}$$

When :

- An intrusion is successfully predicted we call it a True positives (TP).

- An intrusion is wrongly predicted we call it a False negatives (FN).

- A normal connection is wrongly predicted we call it a False positive (FP).

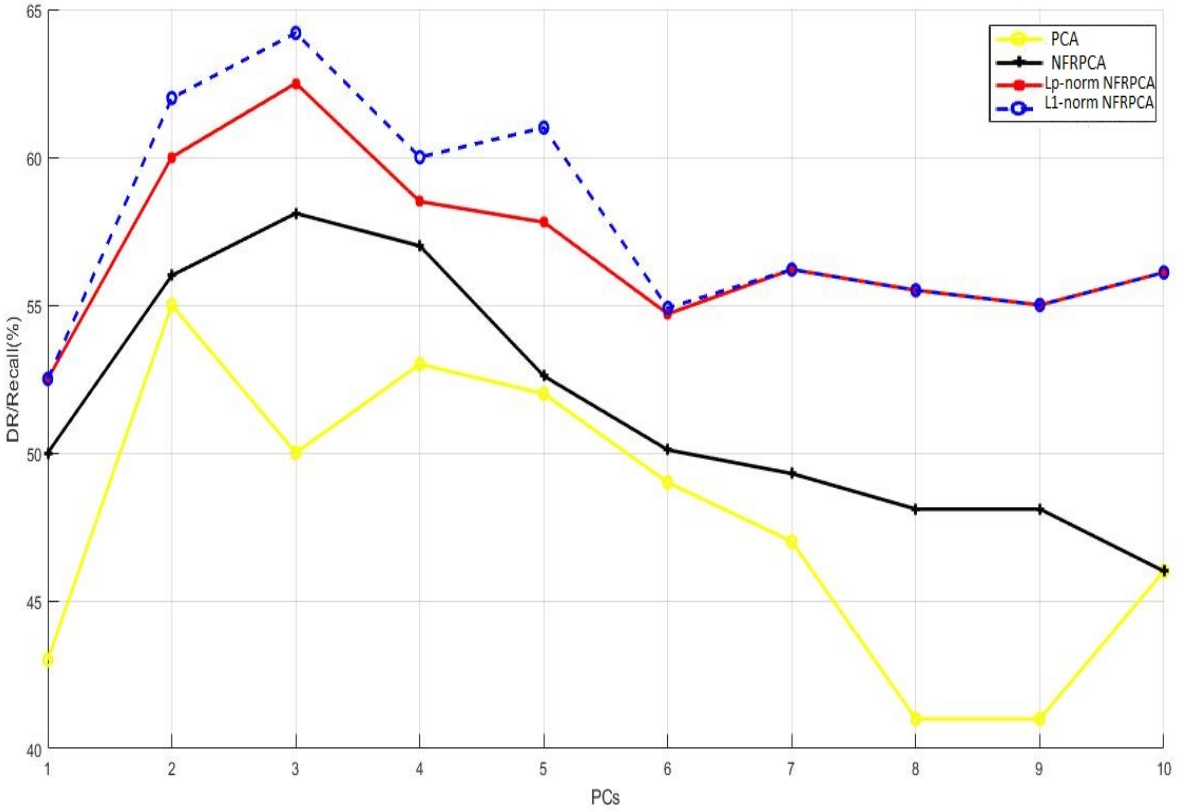


Figure 1: detection rate vs. PCs using KDDcup99 dataset

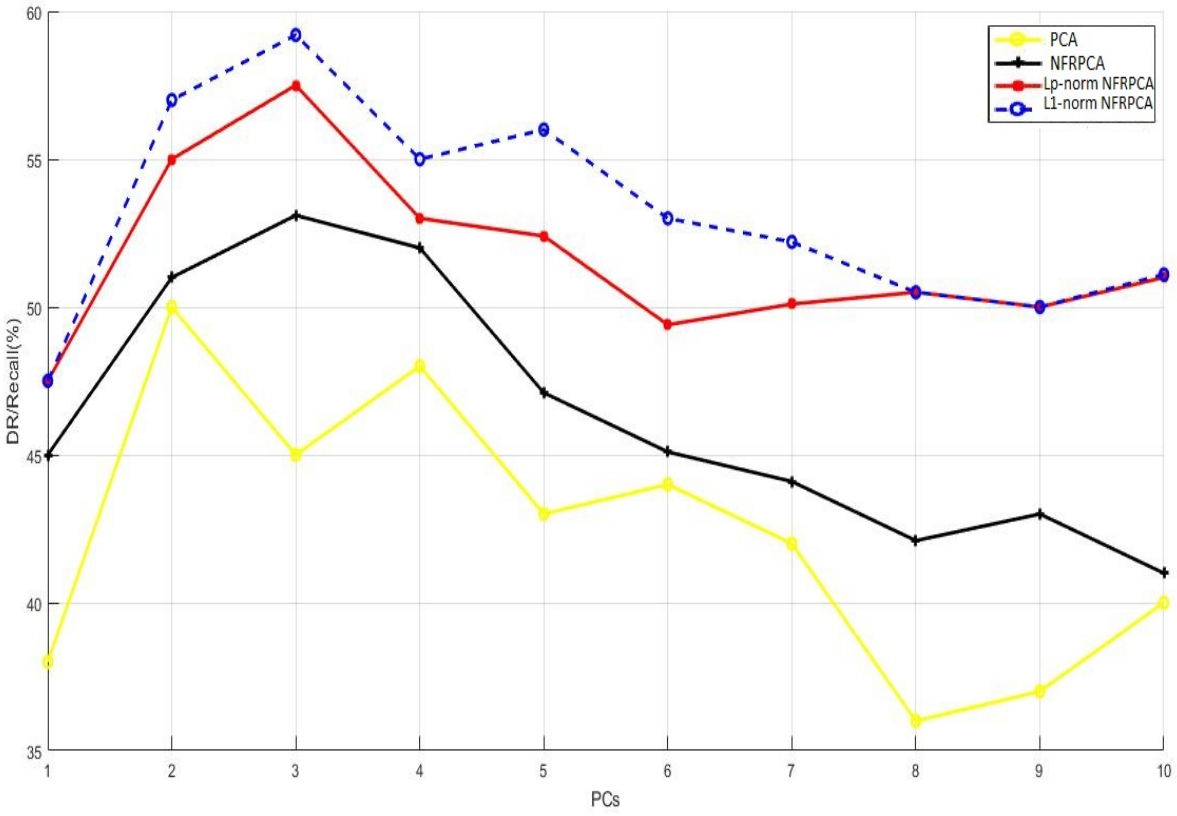


Figure 2: detection rate vs. PCs using NSL-KDD dataset

- A normal connection is correctly predicted we call it a True negatives (TN).

The classifier that was utilized in all our conducted experiments is the nearest neighbor, and for the sake of obtaining more trustworthy results we computed the average of twenty runs. The highest robust feature extraction technique must have the highest DR and F-measure rates and as much as possible the lowest FPR rate.
The simulation settings used in our first experiments are as follow: for the training sample we randomly selected 1000 normal, 100 DOS, 50 U2R, 100 R2L and 100 PROBE as for the test sample we have this structure: 100 normal data, 100 DOS data, 50 U2R data, 100 R2L data and 100 PROBE also choosed randomly using the test database. So as to both KDDcup99 and NSL-KDD data sets the settings of the simulation are similar. Also the value of p is set to 0.5 in all our experiments.

During our first experiments and to choose the ideal number of prinicipal components(PCs) for all the feature extraction techniques which helps drastically to raise F-measure and detection rate (DR) and decrease FPR, examine and make a comparaison between PCA, NFRPCA, L_1-NFRPCA and L_p-norm NFRPCA. To achieve that, we have performed PCA, NFRPCA ,L_1-NFRPCA and L_p-norm NFRPCA to train our model first. Consequently, we obtained the PCs. The number of principal components PCs represent the dimension of the new downsized samples. Furthermore, the test sample is proojected on the new downsized subspace established via the PCs.

The aim of our first experiment is to compute the measures detection rate, F-measure and false positive rate for each single principal component through choosing 10 of 41 principal component and changing their number iteratively throughout the test. We can observe clearly in the Figure.1 and Figure.2 in both datasets KDDCup99 and NSLKDD the L_1-NFRPCA and L_p-norm NFRPCA takes the lead over the original NFPRCA and the linear PCA which

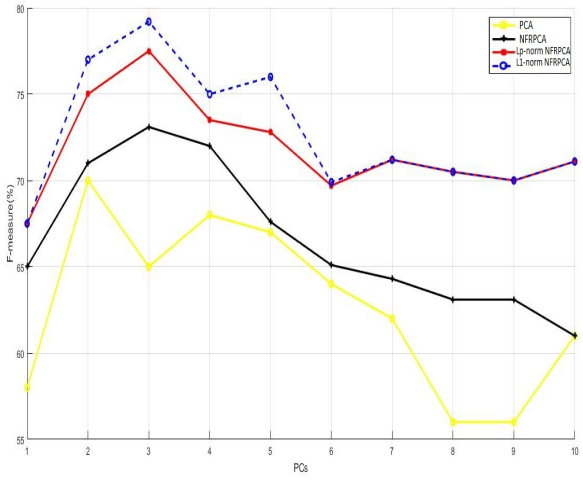


Figure 3: F-measure vs. PCs using KDDcup99 dataset

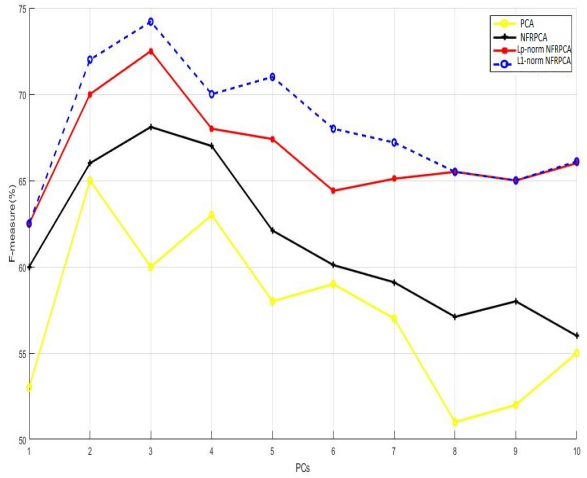


Figure 4: F-measure vs. PCs using NSL-KDD dataset

Following the previous concept, we computed the F-measure by increasing the number of PCs. As we can see from the Figure.3 and Figure.4, for both datasets the L_1-NFRPCA and L_p-norm NFRPCA got the highest values for the F-measure comparing to the original NFRPCA and the classical PCA which support our previous results.

And so as to calculate the false positive rate (FPR), all the algorithms cited above were implemented, but only the proposed algorithms (L_1-NFRPCA and L_p-norm NFRPCA) who has the lowest FPR for both datasets, as we see clearly in the figures (Figure 5 and Figure 6).

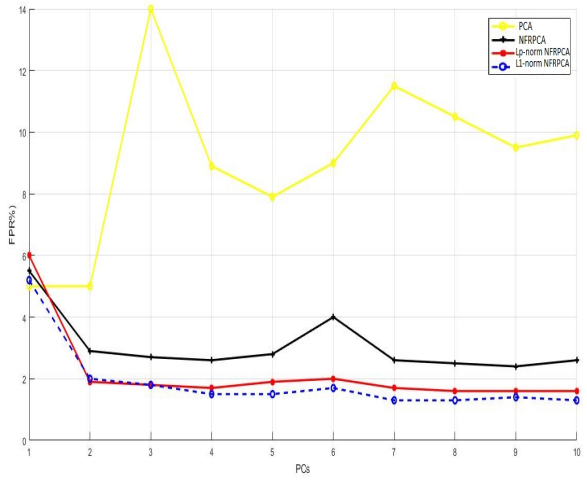


Figure 5: FPR vs. PCs using KDDcup99 dataset

Table 1: DR of Attacks for PCA, NRFPCA, L_1-NRFPCA and L_p-norm NFRPCA using KDDCup99 dataset

	Method	DOS	U2R	R2L	Probing
DR (%)	PCA	68,6656	8,6923	4,7734	92,1121
	NRFPCA	72,1123	14,4615	4,1165	90,2345
	L_1-NRFPCA	73,1993	15,9815	4,1775	91,8325
	L_p-norm NRFPCA	74,2314	16,1111	4,5556	92,1211

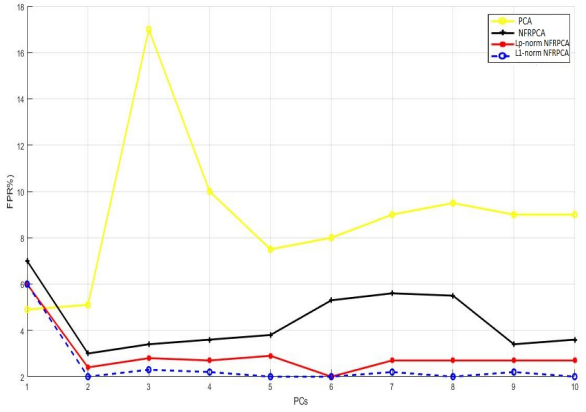


Figure 6: FPR vs. PCs using NSL-KDD dataset

In the second stage of our experiments, we intend to examine all the techniques cited above under a wide range of different training dimensionnality, and examine their impact on the DR, FPR, and F-measure. To achieve that, the structure of the test data set was kept intact by fixing it at 100 normal connections, 100 DOS, 50 U2R, 100 R2L, and 100 PROBE.

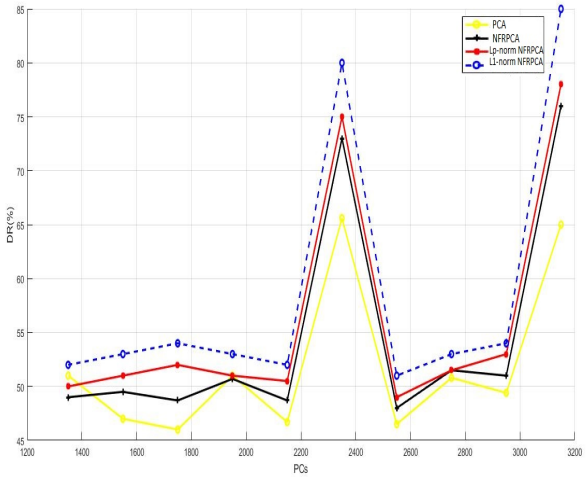


Figure 7: detection rate vs. Training data using KDDcup99 dataset

Concerning DR, Figure.7 and Figure.8 assert that the proposed methods produce a detection rate higher than the original ones. It proves that the methods are very powerful in differentiating between normal connections and attacks.

In Figure.9, Figure.10, we can clearly see that the L_1-NFRPCA achieve at least 5% improvement over L_p-norm NFRPCA, 10% over original NFRPCA and the classical PCA, the new approaches surpasses permanently the original techniques. In terms of FPR,the Figure.11 and Figure.12 show that the L_1-NFRPCA still gives the lowest FPR even under different dimensionnality. These results support the great capability of the new approaches to classify the connections autonomously of the training samples size.

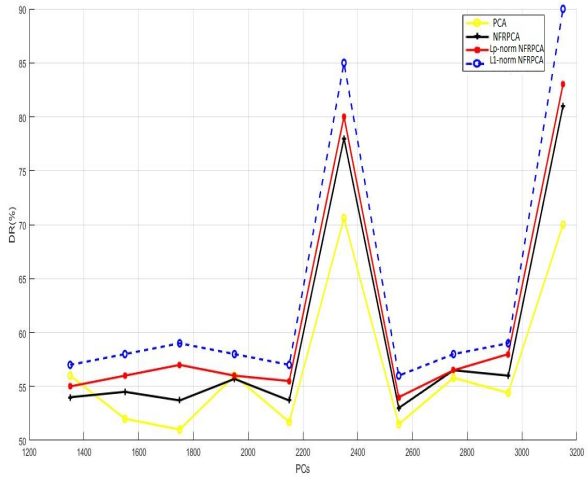


Figure 8: detection rate vs. Training data using NSL-KDD dataset

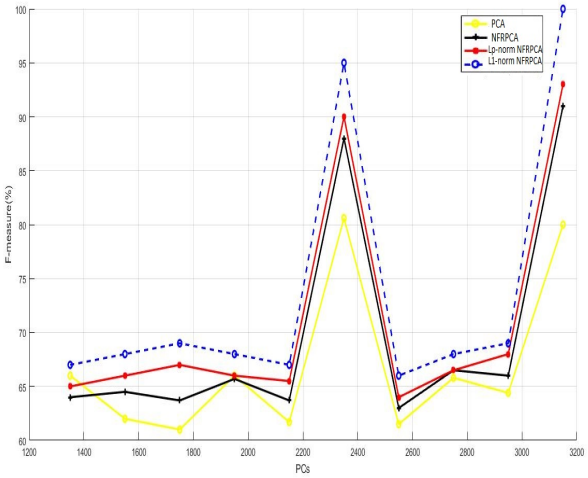


Figure 9: F-measure vs. Training data using KDDcup99 dataset

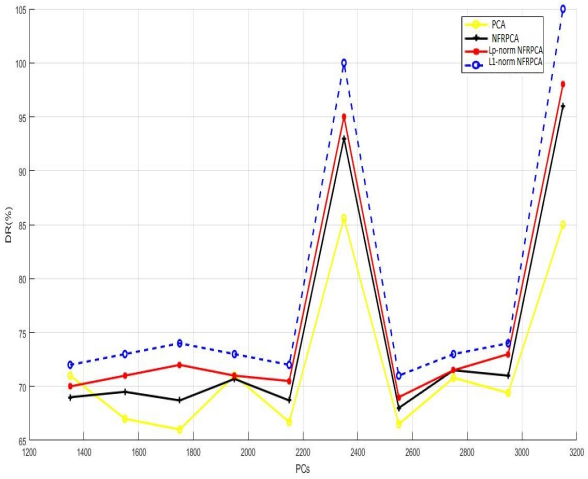


Figure 10: F-measure vs. Training data using NSL-KDD dataset

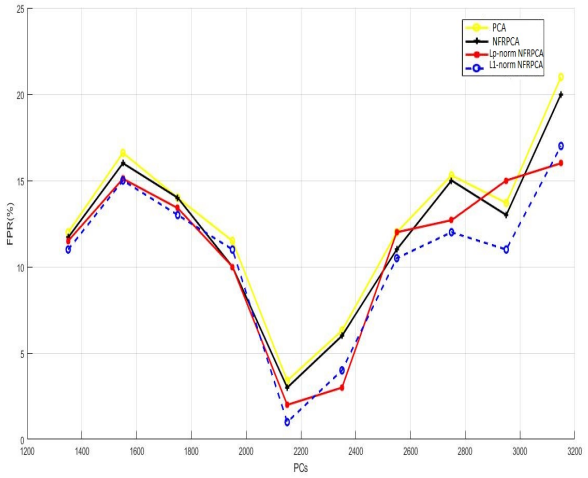


Figure 11: FPR vs. Training data using KDDcup99 dataset

Table 2: DR of Attacks for PCA, NRFPCA, L_1-NRFPCA and L_p-norm NFRPCA using NSL-KDD dataset

	Method	DOS	U2R	R2L	Probing
DR (%)	PCA	67,5546	7,9723	4,8872	93,1121
	NRFPCA	71,1211	13,6415	4,1435	90,3478
	L_1-NRFPCA	72,1341	13,9995	4,1435	91,3698
	L_p-norm NRFPCA	73,2215	15,1123	4,7656	93,1128

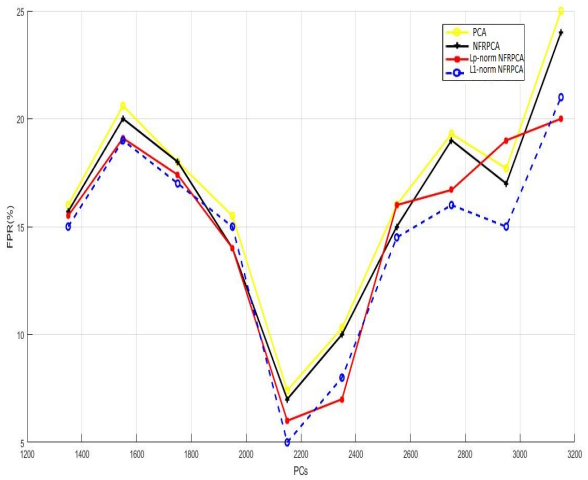


Figure 12: FPR vs. Training data using NSL-KDD dataset

The last figures (Figure.13 and Figure.14) exhibit the coorelation between CPU time and the number of principal components. As it is indicated clearly we observe that increasing the number of principal compenents PCs engender a huge consuming time. In addition to that, we can observe clearly in the figures that the suggested techniques are computationally speedy than the original algorithms.

To obtain higher precise results, we did an experiment in which we compared side by side the DR of every single category of attacks for PCA, NRFPCA, L_1-NFRPCA and L_p-norm NFRPCA. According to Table I and Table II. It is obviously clear that the DR of the

L_1-NRFPCA and L_p-norm NFRPCA for U2R and DOS attacks are often the highest compared to U2R and DOS attacks of NRFPCA and PCA.

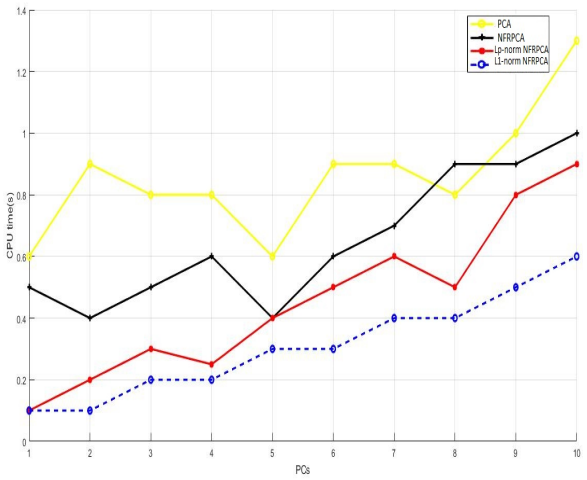


Figure 13: CPU time(s) vs. PCs using KDDCup99 dataset

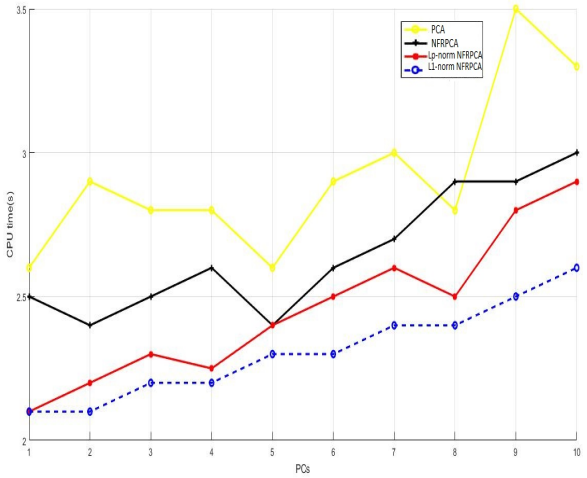


Figure 14: CPU time(s) vs. PCs using NSL-KDD dataset

7 Conclusion

As several linear statistical techniques Principal Component Analysis (PCA) has many shortcoming, it is limited just to Gaussian distribution and it has basically a high sensitivity to noise. In addition to that, principal components are frequently damaged by outliers, therefore feature extraction utilizing PCA are not credible if outliers exits in data. To tackle this issue, we proposed an effective new variants of nonlinear feature extraction techniques called Nonlinear Fuzzy Robust PCA for anomaly-based intrusion detection. The experiments performed on the popular databases (KDDcup99 and NSL-KDD), approved the effectiveness of the suggested approaches, the New variants outperfom NFRPCA and PCA in the detection of the most categories of attacks and in reducing the false positive alarms.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work is supported by CNRST-MOROCCO under the excellence program, grant no. 15UIT2016.

References

[1] E. Spafford and S. Kumar, *A software architecture to support misuse intrusion detection,* in Proceedings of the 18th National Information Security Conference, 1995, pp. 194-204.

[2] B. Caswell and J. Beale, *Snort 2.1 intrusion detection*. Syngress, 2004.

[3] J. P. A. Co, *Computer Security Threat Monitoring and Surveillance,* 1980.

[4] D. E. Denning, *Intrusion-Detection Model, IEEE Trans. Softw. Eng.*, vol. SE-13, no. 2, pp. 222-232, 1987.

[5] E. Popoola and A. O. Adewumi, *Efficient feature selection technique for network intrusion detection system using discrete differential evolution and decision.*," International Journal Network Security, vol. 19, no. 5, pp. 660-669, 2017.

[6] M. H. Aghdam and P. Kabiri, *Feature selection for intrusion detection system using ant colony optimization,*" International Journal Network Security, vol. 18, no. 3, pp. 420-432, 2016.

[7] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, *A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,*" Expert Systems with Applications, vol. 42,no. 5, pp. 2670-2679, 2015.

[8] A.Hadri, K.Chougdali and R.Touahni,*Intrusion detection system using PCA and Fuzzy PCA techniques*, in the proceeding of the Interntional Conference on Advanced Communication Systems and Information Security (ACOSIS), 17-19 October 2016, Marrakesh, Morocco.

[9] A.Hadri, K.Chougdali and R.Touahni, *Identifying intrusions in computer networks using Robust Fuzzy PCA*, in the proceeding of the IEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), 30 October-3 November 2017,Hammamet , Tunisia.

[10] Y. Bouzida and N. Cuppens-boulahia, *Efficient Intrusion Detection Using Principal Component Analysis. pp. 381-395. 2004*

[11] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang, *A novel anomaly detection scheme based on principal component classifier,* Miami Univ. Dept Electr. Comput. Eng. Tech. Rep, 2003.

[12] W. Wang and R. Batitti, *Identifying intrusions in computer networks with principal component analysis,* 2006, p. 8.

[13] L. Xu and A. L. Yuille, *Robust principal component analysis by self-organizing rules based on statistical physics approach,* IEEE Trans. on Neural Net., vol. 6, no. 1, pp. 131-143, 1995.

[14] R. L. Kashyap, M. J. Paulik, N. Loh, A. Automation, A. K. Jain, G. M. Jenkins, S. Francisco, and L. Sirovich, *Application of the Karhunen-Lokve Procedure for the Characterization of Human Faces,* vol. 12, no. 4, 1990.

[15] P. Luukka, *A New Nonlinear Fuzzy Robust PCA Algorithm and Similarity Classifier in Classification of Medical Data Sets,* Int. J. Fuzzy Syst., vol. 13, no. 3, pp. 153-162, 2011.

[16] Amal HADRI, Khalid Chougdali, and Raja Touahni."A Network Intrusion Detection based on Improved Nonlinear Fuzzy Robust PCA." 2018 IEEE 5th International Congress on Information Science and Technology (Cist).IEEE, 2018.

[17] M. Ringnr,*What is principal component analysis?,"* vol. 26, no. 3, pp. 303-304, 2008.

[18] J. Shlens, M. View, and I. Introduction, *A Tutorial on Principal Component Analysis,* 2014.

[19] T. N. Yang and S. D. Wang, *Robust algorithms for principal component analysis,* Pattern Recognit. Lett., vol. 20, pp. 927-933, 1999.

[20] L. Xu and A. L. Yuille, *Robust principal component analysis by self-organizing rules based on statistical physics approach,* IEEE Trans. Neural Net., vol. 6, no. 1, pp. 131-143, 1995.

[21] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, *A Detailed Analysis of the KDD CUP 99 Data Set,* no. Cisda, pp. 1-6, 2009.

[22] [Online] Available: http://kdd.ics.uci.edu/databases/kddcup99/task.html.

[23] [Online] Available: http://www.unb.ca/cic/research/datasets/nsl.html.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 253-261 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Differential Evolution based Hyperparameters Tuned Deep Learning Models for Disease Diagnosis and Classification

Sathyabama Kaliyapillai*, Saruladha Krishnamurthy

Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, 605014, India

ARTICLE INFO

Article history:
Received: 21 June, 2020
Accepted: 05 August, 2020
Online: 14 September, 2020

Keywords:
Deep Learning
Disease Diagnosis
Feature Selection
Healthcare
Differential evolution
Parameter Tuning
LSTM.

ABSTRACT

With recent advancements in medical filed, the quantity of healthcare care data is increasing at a faster rate. Medical data classification is considered as a major research topic and numerous research works have been already existed in the literature. Presently, deep learning (DL) models offers an efficient method for developing a dedicated model to determine the class labels of the respective medical data. But the performance of the DL is mainly based on the hyperparameters such as, learning rate, batch size, momentum, and weight decay, which need expertise and wide-ranging trial and error. Therefore, the process of identifying the optimal configuration of the hyper parameters of a DL is still remains a major issue. To resolve this issue, this paper presents a new hyperparameters tuned DL models for intelligent medical diagnosis and classification. The proposed model is mainly based on four major processes namely pre-processing, feature extraction, classification and parameter tuning. The proposed method makes use of simulated annealing (SA) based feature selection. Then, a set of DL models namely recurrent neural network (RNN), gated recurrent units (GRU) and long short term memory (LSTM) are used for classification. To further increase the classification performance, differential evolution (DE) algorithm is applied to tune the hyperparameters of the DL models. A detailed simulation analysis takes place using three benchmark medical dataset namely Diabetes, EEG Eye State and Sleep stage dataset. The simulation outcome indicated that the DE-LSTM model have shown better performance with the maximum accuracy of 97.59%, 88.52% and 93.18% on the applied diabetes, EEG Eye State and Sleep Stage dataset.

1. Introduction

At present times, healthcare sector becomes more common where massive amount of medical data plays a major role. In this view, for instance, precision healthcare aims to assure the proper medication is offered to the appropriate patients promptly by considered different dimensions of patient's information, comprising variability in molecular trait, atmosphere, electronic health record (EHR) and standard of living [1]. The higher accessibility of medical details brought numerous chances and issues related to healthcare researches. Particularly, discover the interconnections between diverse set of data exist in dataset still remains a basic issue to design effective medicinal models using data driven techniques and machine learning (ML). Earlier studies have focused on linking many data sources for building joint knowledge bases which can be utilized for predictive analysis and discovery. Though former techniques exhibit noteworthy performance, prediction models using ML models are not mainly employed in healthcare sector [2]. Actually, it is not possible to completely utilize the available healthcare data due to its sparsity, high dimensional heterogeneity, temporal dependence, and irregularities. These problems become even more difficult by distinct medicinal ontologies employed for data generalization which frequently include disagreements and inconsistencies. In some cases, the identical clinical phenotype can be defined in several ways along the data [3].

For instance, in EHR, a patient affected by 'type 2 diabetes mellitus' could be detected by the use of laboratory test reports of haemoglobin A1C >7.0, occurrence of 250.00 ICD-9 code, 'type 2 diabetes mellitus' declared in the free text medical notes, etc. As a result, it is not trivial to harmonize every medical concept to develop a higher-level semantic structure and comprehend their relationships [4]. A widespread method in healthcare researches is to have medical experts for specifying the phenotypes to utilize

*Corresponding Author: Sathyabama Kaliyapillai, Email: sathii_manju@pec.edu

in an adhoc way. At the same time, the supervised description of the feature space performs poor scaling and ignores the chances of discovering effective patterns.

On the other hand, representation learning approaches permit automatic discovery of representations required to predict it from the actual data. Deep learning (DL) techniques are representation-learning approaches with many stages of representation, attained by integrating simpler but nonlinear models [5]. DL models exhibited better results and attained significant attention in natural language processing, computer vision, and speech recognition. DL models have been introduced in healthcare sector due to its better performance in various fields and fast developments of technical enhancements [6]. Several works have also been carried out on DL model for biomedical sector. For instance, Google DeepMind has initiated schemes to utilize its knowledge in medicinal field. In contrast, DL models have not been broadly validated for a wide range of medicinal issue which can get advantages from its abilities [7] . DL have several ways which can be supportive in healthcare, namely multi-modality data, capability of handling complex, superior performance, and end-to-end learning scheme includes characteristic learning etc. Stepping up these works, the entire DL research study should address many issues describing the features of healthcare info requires to upgrade the patterns and tools which allows DL to combine through healthcare workflows and medical decision support [8] . At the same time, the performance of the DL is mainly based on the hyper parameters such as, learning rate, batch size, momentum, and weight decay, which need expertise and wide-ranging trial and error. As a result, the procedure to determine the optimum configuration of the hyper parameters of a DL is still remains a major issue**.**

In this view, this paper presents a new hyperparameters tuned DL models for intelligent medical diagnosis and classification. The proposed model is mainly based on four major processes namely pre-processing, simulated annealing (SA) based feature extraction, recurrent neural network (RNN), gated recurrent units (GRU) and long short term memory (LSTM) based classification and differential evolution (DE) based parameter tuning. A detailed simulation analysis takes place using three benchmark medical dataset namely Diabetes, EEG Eye State and Sleep stage dataset.

2. Related Works

Numerous research methods apply DL to forecast diseases from the status of the patient [9]. Use a 4-layer CNN to identify block of heart failure and chronic obstructive pulmonary disease and demonstrate topmost beneficiary measures. RNN with word embedding, pooling and Long Short-Term Memory (LSTM) concealed unit were applied in Deep Care, an end-to-end deep dynamic network that conclude existing infection declare and calculates upcoming healthcare results. The authors planned to sense LSTM part with a degrade events in handling uneven timed actions (i.e. usually longitudinal EHR). Besides, they included medicinal involvement in the techniques to finding the shapes in a dynamic manner. DeepCare estimates the threat on diabetes in future and psychological healthcare of patient cohort, intervention recommendation, and modelling the evolution of disease [10].

RNN with Gated Recurrent Unit (GRU) are utilized by to design Doctor AI, an end-to-end model which make use of patient's record to calculate diagnose and medication for successive encounter. The valuation indicated notably higher recall compared to low baseline and good believes by adjusting the resulting method from one body to another without losing large accuracy rate [11]. In other aspects, [12] designed a model to study deep patient representation from the EHR via a 3-layer Stacked Denoising Autoencoder (SDA). They apply this novel version on disease to predict the risk by means of random forest as classification model. The validation have been implied on 76,214 patients comprise seventy eight diseases from various medical domains and temporal window (within 1 year). As a finally result, drastically the deep representation direct to better prediction than utilizing raw EHR or conventional representation learning algorithms (e.g. Principal Component Analysis (PCA), k-means). Likewise, they illustrated that result considerably gets better after accumulating a logistic weakening layer on top of the last AE to modify the whole supervised network.

Likewise, [13] use RBM to discover representation of EHR which exposed a novel concept and established an improved approach for predicting accuracy ratio on diseases count . DL was tested to model nonstop time signals, like laboratory result, towards the usual recognition of definite phenotype. For instance,[14] utilized RNN by LSTM to identify the pattern in multivariate time series of experimental dimensions. Especially, a technique has been trained to classify 128 diagnoses from 13 but unevenly sample experimental dimensions from patients in pediatric Intensive Care Unit (ICU). The result shows a major improvement comparing with some sturdy baselines includes multilayer perceptron skill on the hand-engineered type. Use SDA regularize with earlier facts based on ICD-9 for discovering a featured pattern of physiology in medical time sequence [15].

Use a 2-layer stack AE (without regularization) to design a longitudinal series of serum uric acid measurement to differentiate the uric-acid signature of gout and sharp leukemia [16]. Evaluate CNN and RNN with LSTM unit to foresee the beginning stage of disease only from lab-test measures, presenting enhanced output than logistic regression with handcrafted, medically related characteristics [17]. Neural language deep model are applied to EHR, specifically to study embedded representations of the medicinal proposal, equally as diseases, medication and laboratory test that can be used for testing and estimate. As a pattern, use RBM to find out abstraction in ICD-10 cryptogram on a cohort of 7578 logical health patients for forecasting risks of suicide [18]. A wide plan related to RNN obtains capable effects in eliminating protected health details from medical observations to influence the usual de-identification of free-text patient summary. The calculation for not planned patient readmissions after the discharge in recent times receives concentration as well. In this field, presented Deepr, an end-to-end architecture which supports on CNN, which detect and merge medical pattern in the longitudinal patient EHR for stratifying medicinal threats [19]. Deepr carried out fine in forecasting the readmission in six months and have ability to notice significant and interpretable clinical patterns.

3. Objectives

Differential Evolution (DE) is an optimization algorithm. This study aims to apply DE to tune hyper parameter settings of the

deep learning models. The convergence of DE algorithm is evaluated to select optimal hyperparameters on the basis of certain searching process such as crossover, mutation and selection. The enhancement of the accuracy and performance of the proposed model is also determined in comparison with Random search algorithm.

4. The Proposed Model

The working process of the presented model is shown in Figure 1. As shown in Figure, the input data undergoes pre-processing, feature extraction, classification and parameter optimization. Initially, the input data is pre-processed to remove the unwanted data and transform it to a compatible format. Then, SA-FS process takes place to select the useful subset of features. Then, DL models are applied to carry out the classification process. Finally, DE is applied for the parameter optimization of the DL models.

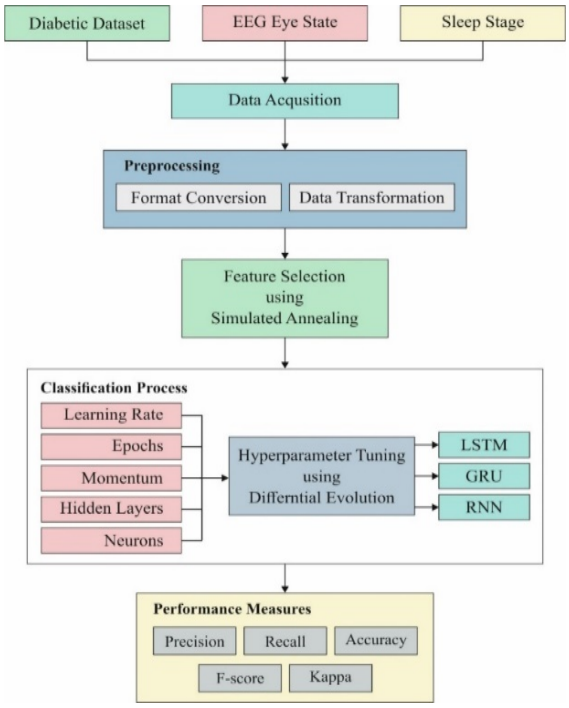


Figure 1: Block diagram of proposed method

4.1. Pre-processing

The major task of data pre-processing is converting the original input data into most intelligible format. As the practical input data is incomplete, there is maximum probability of having error filled data. Thus, data pre-processing is mainly used for transforming the actual data into understandable format which can be applied for next computation. In this approach, pre-processing is carried out in 2 phases such as, Format conversion as well as Data transformation. Initially, format conversion task is conducted when any kind of data type is transformed into .arff format. Then, data transformation is processed with diverse sub-processes as given in the following.

Normalization: This process is applied for scaling the data measures within the specific grade as (-1.0 to 1.0 or 0.0 to 1.0)

Attribute Selection: Novel attributes are obtained from the given set of variables which is applied for mining task in future.

Discretization: Actual measures of mathematical attributes would be interchanged by conceptual levels.

4.2. SA based Feature Selection Process

SA is defined as a processing model of annealing operation. For physical content, the premise should be filled with massive amount of energy and periodically during the cooling operation. When a solution is applied with minimum criterion value, then it is managed using a degree of minimization as well as a current temperature $T(q)$ for all $Q(1 \leq q \leq Q)$ iterations [20]. Followed by, the temperature is slowly reduced while limiting the likelihood of accepting inferior solutions. Some of main components of SA for FS are listed as follows.

- The method to generate initial subset;
- The selection of a temperature rank which is defined by $\text{maximum}\,T(1)$, $\text{minimum}\,T(Q)$, temperature, and cooling scheme.

5. Methodologies

In this section, three DL models namely RNN, GRU and LSTM which are applied for classification process has been discussed in the following subsections.

5.1. RNN based Classification Model

RNN belongs to the NN where the output from existing step is induced as input to the next step. In classical NN, every input and output are autonomous; however, the prediction of a data is processed using previous data and no requirement of memorizing the previous data. Hence, RNN is used to resolve these issues with the help of a hidden layer. The most remarkable feature of RNN is hidden state that saves few data regarding the sequence. RNN is composed of a “memory” that records every detail that has to be determined. It applies similar parameters for every input as it processes the similar operation on hidden layers and produces the better outcome. Finally, the difficulty of parameters is reduced on contrary to alternate NN. Here, NN performs the following operations such as RNN transforms the autonomous activations to dependent activations by generating the similar weights and biases for all layers, and minimize the complications of parameters and remember every existing output by providing the result as input to the subsequent hidden layer. Thus, 3 layers are combined where weights and bias of each hidden layer is identical, that forms an individual recurrent layer.

Training through RNN

The training process involved in RNN is listed as follows.

- A single time step of input is given to the network.
- Determine the present state with the help of recent input and existing states.
- The present h_t forms h_{t-1} for future time step.

- The same is repeated on the basis of problem and combine the data from previous states.
- After completing the last current state then it computes the attained result.
- Then result is related to original output where the desired result and an error are produced.
- The error undergoes back-propagation to the network and improves the weights so that RNN it trained.

5.2. LSTM based Classification Process

LSTM performs the learning operation over a range of prolonged time intervals. It solves the diminishing gradients problem by replacing periodical neuron using complex structural method named as LSTM unit. The LSTM consists of four neural network layer and three gates (input gate, forget gate and output gate) that are used to control the flow of information. These gates are employed using logistic function in order to calculate the values between 0 and 1. The key elements of LSTM are provided in the following and illustrated in Figure 2.

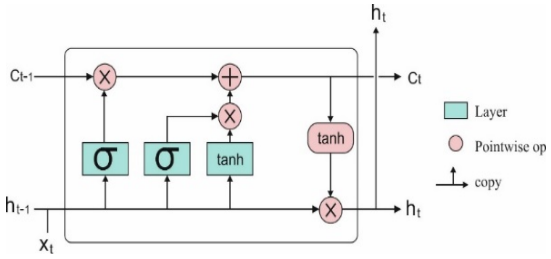


Figure 2: LSTM Model

Input Gate (i_t)

An input gate controls a new value flow into the memory.

Forget gate (f_t)

A forget gate removes the information from the cell state which is no longer required to process. It performs scaling on the internal state of the cell in prior to adding it as input to the cell through the self-recurrent connection of the cell, so adaptively forgetting or resetting the cell's memory

Output Gate (o_t)

It is also a multiplicative unit determines the next hidden state from the current cell state.

Cell State (c_t)

Cell state contains memory unit which maintains all relevant information required for processing. After the gates are closed, data is trapped inside a memory cell. It activates the error signals flowing over several time steps with no consideration of vanishing gradients. The equation of the LSTM units are given below.

$$z_t = tanh(W^z x_t + R^z h_{t-1} + b^z) \quad (1)$$
$$i_t = \sigma(W^i x_t + R^i h_{t-1} + b^i) \quad (2)$$
$$f_t = \sigma(W^f x_t + R^f h_{t-1} + b^f) \quad (3)$$
$$o_t = \sigma(W^o x_t + R^o h_{t-1} + b^o) \quad (4)$$
$$c_t = z_t \odot i_t + s_{t-1} \odot f_t \quad (5)$$
$$h_t = tanh(c_t) \odot o_t \quad (6)$$

5.3. GRU based classification model

The Vanishing-Exploding gradients issue can be solved using RNN. The major scheme is LSTM. A model with minimum popularity but highly productive variations is named as GRU. In contrast to LSTM, it has 3 gates which do not retain the Internal Cell State. The data saved in Internal Cell State in an LSTM recurrent unit is embedded into hidden state of GRU. The collected data is provided to the subsequent GRU. Some of the various gates of a GRU is defined in the following:

Update Gate (z)

It computes the previous knowledge that has to be conveyed to future processing. It is analogous to Output Gate in LSTM recurrent unit.

Reset Gate (r)

It evaluates the older knowledge which has to be discarded. It is analogous to the integration of the Input Gate and Forget Gate in an LSTM recurrent unit.

Current Memory Gate ($\bar{h}_t$)

It is highly applied for GRU process. It is combined into Reset Gate with Input Modulation Gate is a sub portion of Input Gate and applied to establish non-linearity into input and make the input Zero-mean. An alternate reason to make a sub-part of Reset gate is to limit the effect of existing data on recent information which is applied for next computation.

The fundamental task of GRU is related with RNN as defined with few variations between 2 models. Figure 3 shows the structure of GRU model. The inner working process of GRU has gates that change the recent input and existing hidden state. The working of GRU has been discussed as follows. Initially, the input as current input and older hidden state are assumed as vectors. Then, the measures of 3 various gates are computed as follows.

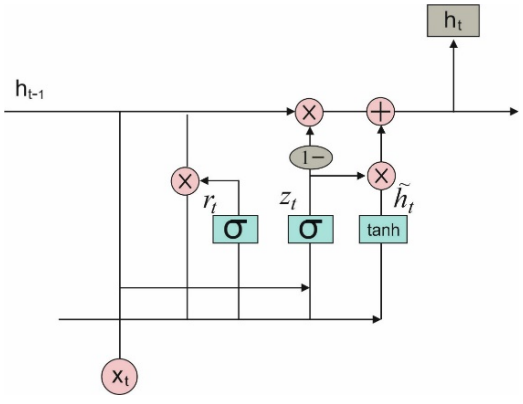


Figure 3: Structure of GRU model

Initially for all gates, determine the parameterized recent input and previous hidden state vectors by processing element-wise multiplication among the assumed vector and the respective

weights. Then, Use the concerned activation function for a gate element-wise on parameterized vectors. The list of gates is provided with the activation function. The task of calculating the recent Memory Gate is different from alternate process. Initially, the Hadamard product of Reset Gate as well as previous hidden state vector has been examined. Then vector undergoes parameterization and included to parameterize recent input vector.

$$\overline{h}_t = \tanh\left(W \odot x_t + W \odot (r_t \odot h_{t-1})\right) \quad (7)$$

The present hidden state is calculated by same vector and dimensions of input are defined. These vectors are indicated by 1. Then compute the Hadamard product of update gate and existing hidden state vector. Also, produces a novel vector by reducing the update gate from ones and determine the Hadamard product of vector present in recent memory gate. Consequently, include 2 vectors for reaching recent hidden state vector.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \bar{h}_t \quad (8)$$

6. DE based Parameter Optimization Model

The main aim of DL classification model is to optimize the hyper-parameters namely epochs, learning rate, momentum, hidden layers, and neurons with the application of DE method and results in optimal medical data classification. Here, the parameters exist in these methods are Batch size and count of hidden neurons. The DE model is initialized from first solutions that is produced randomly and tries to enhance the accuracy of emotion classification. The Fitness Function (FF) of DL approach is applicable to perform the estimation and provide the accuracy of medical data classification [21]. DE approach was initially coined by Storn. Generally, DE is mainly applied for parameter optimization as well as real value functions. It is a population oriented searching that has been employed extensively for frequent searching process. Currently, the efficiency of DE is represented in various applications like strategy adaptation for global numerical optimization and FS for healthcare diagnosis. Similar to Genetic Algorithm (GA), DE models employs the crossover and mutation; however, the equation is improved in explicit manner. The optimization in DE is composed of 4 phases: initialization, mutation, crossover and selection.

Firstly, the primary population $S_i^t = \{s_{1,i}^t, s_{2,i}^t, , s_{D,i}^t\}, i = 1, .., Np$ has been produced randomly, in which Np implies the population size and t means the recent iteration, where initialization is 0. It is defined for all types of problem space with specific integer ranges constrained by upper as well as lower bounds: $s_j^{low} \le s_{j,i} \le s_j^{up}$, for $j = 1$, 2, D, where D denotes the dimension of problem space. The main objective of this model is to optimize the hyperparameters of RNN, GRU and LSTM networks, hence, the dimension of the problem is $2(D = 2)$. Every vector is combined and develops into candidate solutions$_i^t$, named as target vector, to the multi-dimensional optimization issue. Secondly, in mutation, there are 3 unique vectors $s_{j,p}^t, s_{j,r}^t, s_{j,q}^t$ from population S_i^t are decided arbitrarily to produce novel donor vector v with the help of mutation function

$$v_{j,i}^t = s_{j,p}^t + F_i * \left(s_{j,r}^t - s_{j,q}^t\right) \quad (9)$$

Where F_i depicts the constant lies between $[0, 2]$ defines the mutation factor. It is apparent that, the value of $v_{j,i}^t$ for all dimensions are measured adjacent to integer value. Then, in crossover, the trial vector has been determined for all D dimension of target vector s_i^t and every D dimension of donor vector v_i^t using binomial crossover. Here, the dimension of trial vector is managed by c_r , cross over rate, that is a user based constant from $[0, 1]$.

$$u_{j,i=}^t \begin{cases} v_{i,j}^t \; if \; r \le c_r \; or \; j = J_{rand} \\ s_{i,j}^t \qquad otherwise \end{cases} \quad (10)$$

where J_{rand} refers the even distributed random integer within $[1, D]$, and r_i represents the shared random value from $[0, 1]$. Finally, in selection, the target vector s_i is related with trial vector u_i , and maximum accuracy is elected and applied for future generation. It is clear that, the selection operation is attained using different deep learning technique, and a vector with effective fitness value is admitted for next iteration. Hence, the last 3 phases are repeated until meeting the termination condition. The entire process of DE based parameter tuning of DL models is provided in Algorithm 1.

Algorithm 1: Differential Evolution with Deep Learning Models (LSTM, GRU, RNN)
Initialization; Population Size (PS), Crossover Rate (CR), Scale Factor (SF), Termination Criteria (TC)

Begin Procedure
while (TC Not Met) **do**
For Each individual, target vector, in the PS;
Mutation: Choose three individuals from the PS arbitrarily and generate a donor vector using Eq. 9
Crossover: Calculate the trial vector for the ith target vector using Eq. 10
Selection: Apply LSTM, GRU, RNN Algorithm as fitness function f and evaluate

$$\text{If } f(s_i^t) \le f(u_i^t) \text{ then } s_i^{t+1} = u_i^t$$
$$\text{Else } s_i^{t+1} = s_i^t$$

End For
End While
End Procedure

7. Performance Validation

7.1. Dataset used

The performance of the proposed model undergo validation against three benchmark medical dataset namely diabetes, EEG EyeState (UCI Machine Learning Repository) and Sleep Stage dataset (physionet.org). The details of the dataset are shown in Table 1. The first diabetes dataset includes a total of 101766 instances with the existence of 49 features. Besides, the number of classes in diabetes dataset is two, where 78363 instances comes under positive class and remaining number of 23403 instances falls into negative class.

Table 1: Dataset Description

Dataset Type	Description	Values
Diabetes	Number of Instances	101766
	Number of Attributes	49
	Number of Class	2
	Number of Positive Samples	78363
	Number of Negative Samples	23403
	Data source	[UCI Repository]
EEG Eye State	Number of Instances	14980
	Number of Attributes	15
	Number of Class	2
	Number of Class 1	82527
	Number of Class 2	6723
	Data source	[UCI Repository]
Sleep Stage	Number of Subjects	25
	Sex Distribution Male	21
	Sex Distribution Female	4
	Average Age	50
	Average Weight	95
	Average Height	173
	Number of Class	5
	Name of the Class	Wake/S1/S2/SWS/REM
	Data source	[Physionet.org]

Next, the second EEG Eye Sate dataset comprises a total of 14980 instances with 15 features. Among the total number of instances, a set of 82527 instances comes under class 1 and the remaining 6723 instances falls into class 2. The duration time of every recording was 117 seconds. Next, diverse eye states monitored for each recording were added manually. Consequently, the corpus dataset was developed with 14980 instances. Every instance has 14 EEG features and an eye-state class (either 0 for open, or 1 for closed). Additionally, the count of instances with open-eye class in the corpus dataset is 8257 (55.12%), at the time of number of closed-eye type instances is 6723 (44.88%). Moreover, the dataset is applied in massive studies. Hence, 3 of the instances' (2 open states, and one closed state) values were outliers and it is applicable to remove from classification process. Finally, the Sleep stage dataset from UCI repository is composed of 25 instances, including 21 male and 4 female instances. It consists of the entire night PSG recordings of 25 persons with sleep-disordered breathing. Every individual recording has 2 EEG channels, two EOG channels, and one EMG channel as well as an annotation file with complete onset time and period of each hypopnea event.

7.2. Results analysis

Table 2 provides the feature selection results of the proposed SA model on the applied dataset. The table values indicated that the SA based model has chosen a collection of 16 features out of 49 features on Diabetes dataset and 9 features out of 15 features on the applied EEG Eye State dataset.

Table 2: Selected features of Simulated Annealing for Applied Dataset

Dataset	Selected Features
Diabetes	4,5,6,14,17,18,19,20,21,22,23,24,27,29,30,31
EEG Eye State	1,3,4,5,7,9,10,11,12

Table 3: Average epoch rate, learning rate, momentum, hidden layers, and neurons by DE and RS optimization models

Methods	Epochs	Learning Rate	Momentum	Hidden Layers	Neurons
DE	100 200 300	0.0001 0.001 0.01	0.0 0.5 0.9	2 4 6	32 64 128
Random Search	100 200 300	0.0001 0.001 0.01	0.0 0.5 0.9	2 4 6	32 64 128

Table 3 tabulates the different hyperparameters tuned by DE and Random Search (RS) optimization algorithms. The optimal hyperparameters determined by the proposed method are number of epochs: 100, learning rate: 0.001, momentum: 0.5, hidden layers: 4 and number of neurons: 64.

Table 4: Performance Analysis of Proposed Method for Applied Datasets

Methods	Measures	Precision	Recall	F-Measure	Accuracy	Kappa
DE-LSTM	**Diabetes**	97.46	96.20	97.23	97.59	92.40
	EEG Eye State	90.13	89.40	90.30	88.52	86.83
	Sleep Stage	92.94	91.65	92.31	93.18	89.58
RS-LSTM	**Diabetes**	96.15	95.62	96.29	95.89	90.39
	EEG Eye State	88.72	88.56	89.71	87.11	85.09
	Sleep Stage	90.63	89.73	90.66	91.49	87.48

Table 4 and Figure 4 (a) (b) (c) shows the classifier outcome of the DE-LSTM model on the applied three datasets to investigate the impact of parameter tuning process. The table values showed that the DE-LSTM model has attained better results over the RS-LSTM on the applied dataset. It can be ensured from the values that the DE-LSTM model has attained a higher accuracy of 97.59% on the applied diabetes dataset whereas the RS-LSTM model has reached to a lesser accuracy of 95.86%. Similarly, on the applied

EEG Eye State dataset, the proposed DE-LSTM model has exhibited better results with the accuracy of 88.52% whereas a slightly lesser accuracy of 87.11% has been achieved by the RS-LSTM model. Likewise, on the applied Sleep Stage dataset, the proposed DE-LSTM has shown its superior performance with the maximum accuracy of 93.18%. Figure 5 shows the loss graph of the RS and DE models on the applied diabetes dataset. The figure indicated the loss rate of the proposed model gets reduced with an increase in number of epochs.

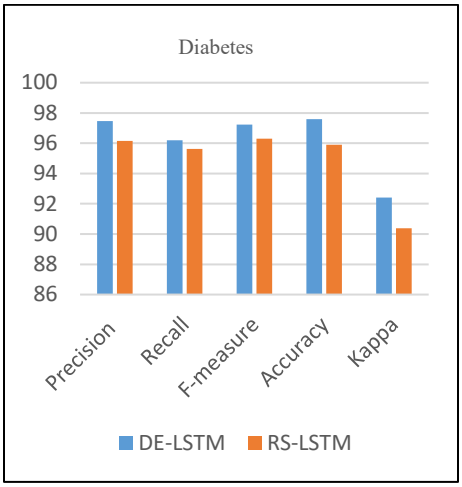


Figure 4 (a) Diabetes

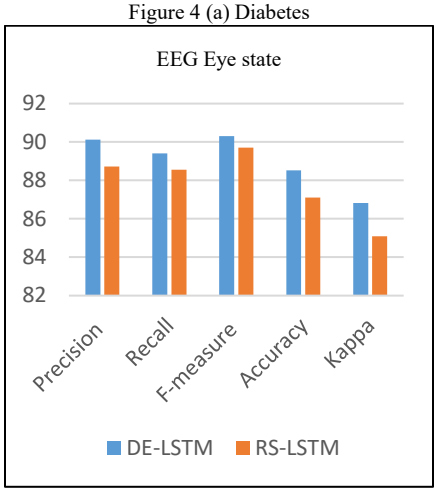


Figure 4 (b) Eye state

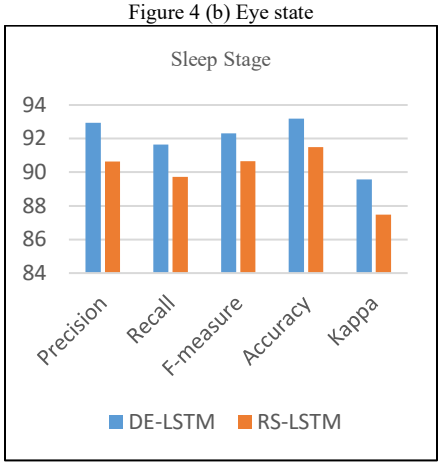


Figure 4 (c) Sleep stage dataset.

Figure 4. Classification results analysis of DE-LSTM and RS-LSTM models (a) Diabetes, (b) EEG Eye State and (c) Sleep stage dataset.

Table 5: Comparison of Proposed DE-LSTM with State of Arts for Diabetes Datasets in terms of Accuracy

Classifiers	Accuracy (%)
DE-LSTM	97.59
DE-GRU	96.34
DE-RNN	92.43
RS-LSTM	95.89
SA-LSTM	95.48
GRU	91.54
RNN	86.87
LSTM	91.75
MLP	81.90

Table 5 and Figure 6 offer a comparative analysis of the proposed models on the applied diabetes dataset in terms of accuracy. The

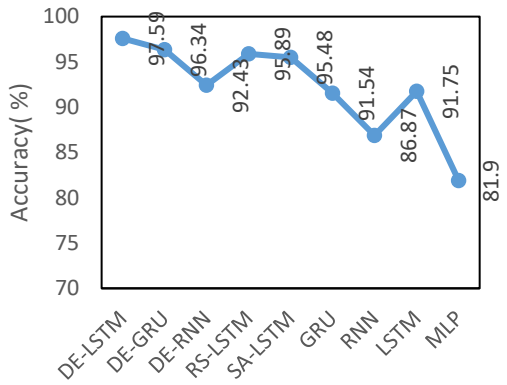


Figure 6. Accuracy analysis of various models on Diabetes Datasets

Table 6: Comparison of Proposed DE-LSTM with State of Arts for EEG Eye State in terms of Accuracy

Classifiers	Accuracy (%)
DE-LSTM	88.52
DE-GRU	87.14
DE-RNN	85.23
RS-LSTM	87.11
SA-LSTM	84.49
GRU	80.65
RNN	78.94
LSTM	81.78
Linear SVM	55.12
HMM	54.12

experimental values indicated that the RNN and MLP models have stated ineffective performance by attaining minimal accuracy values of 86.87% and 81.90% respectively. Next to that, the GRU and MLP models have reached to slightly higher and closer accuracy values of 91.54% and 91.75% respectively. Then, the DE-RNN model has accomplished better results with the accuracy of 92.43%. Simultaneously, the RS-LSTM and SA-

LSTM models have accomplished manageable results with the closer accuracy values of 95.89% and 95.48% respectively. On continuing with, the near optimal accuracy of 96.34% has been offered by the DE-GRU model. However, the DE-LSTM model has outperformed all the compared methods with the maximum accuracy of 97.59%.

Table 6 and Figure 7 provide a comparative analysis of the presented models on the applied EEG Eye State with respect to accuracy. The experimental values implied that the Linear SVM as well as HMM methods have shown inferior function by reaching lower and same accuracy values of 55.12%. Then, the RNN and GRU models have attained better and nearby accuracy values of 78.94% and 80.65% correspondingly. Followed by, the LSTM and SA-LSTM approaches have achieved manageable accuracy values of 81.78% and 84.49% respectively. At the same time, the DE-RNN and RS-LSTM models have obtained considerable results with the nearby accuracy values of 85.23% and 87.11% correspondingly. Along with that, the near optimal accuracy of 87.14% is provided by the DE-GRU model. Thus, the DE-LSTM model has performed well than other approaches with the higher accuracy of 88.52%.

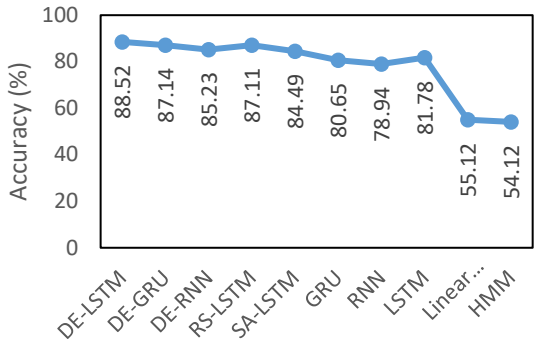


Figure 7: Accuracy analysis of various models on EEG Eye State

Table 7: Comparison of Proposed DE-LSTM with State of Arts for Sleep Stage Dataset in terms of Accuracy

Classifiers	Accuracy (%)
DE-LSTM	93.18
DE-GRU	91.57
DE-RNN	87.93
RS-LSTM	91.49
GRU	85.42
RNN	82.19
LSTM	86.45
CNN	78.23
5C-CNN	83.20
7C-CNN	87.50
9C-CNN	89.00
11C-CNN	90.12
DBN	72.20
SAE	77.70
RBF	81.70

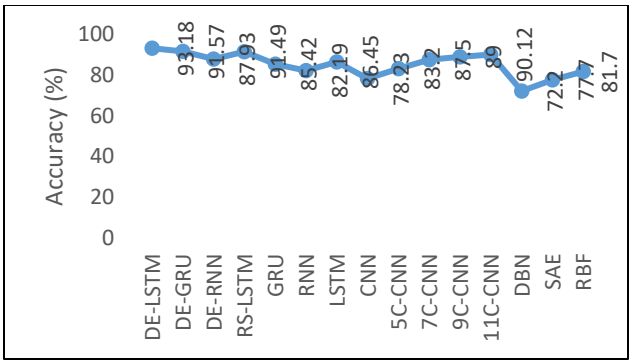


Figure 8: Accuracy analysis of various models on Sleep Stage Dataset

Table 7 and Figure 8 show a comparative analysis of the proposed method on the given Sleep Stage Dataset in terms of accuracy. The experimental values showcased that the DBN model has revealed poor performance by accomplishing lower accuracy values of 72.20%. Then, the SAE, CNN and RBF models have attained reasonable performance and closer accuracy values of 77.70%, 78.23% and 81.70% respectively. Next to that, the RNN, 5C-CNN and GRU methods have achieved appreciable results with the accuracy values of 82.19%, 83.20% and 85.42% respectively. Meantime, 7C-CNN, LSTM and DE-RNN frameworks have attained better results with the near accuracy values of 86.45%, 87.50% and 87.93% correspondingly. Concurrently, the 9C-CNN, 11C-CNN and RS-LSTM methodologies have performed better results with accuracy values of 89%, 90.12% and 91.49% respectively. In line with this, the closer and best accuracy of 91.57% has been provided by the DE-GRU model. Therefore, the DE-LSTM model functions an outstanding performance of related models with optimal accuracy of 93.18%. From the above mentioned tables and figures, it is evident that the proposed DE-LSTM model can be employed as an appropriate medical data classification model. Besides, it is ensured that the inclusion of hyper parameter tuning technique helps to improvise the classification performance.

8. Conclusion

This paper has presented a new hyperparameters tuned DL models for intelligent medical diagnosis and classification. The proposed model involves different processes namely pre-processing, feature extraction, classification, and parameter optimization. Initially, the input data is pre-processed to remove the unwanted data and transform it to a compatible format. Then, SA-FS process takes place to select the useful subset of features. Then, DL models are applied to carry out the classification process. Finally, DE is applied for the parameter optimization of the DL models. A detailed simulation analysis takes place using three benchmark medical dataset namely Diabetes, EEG Eye State and Sleep stage dataset. The simulation outcome indicated that the DE-LSTM model have shown better performance with the maximum accuracy of 97.59%, 88.52% and 93.18% on the applied diabetes, EEG Eye State and Sleep stage dataset. In future, the performance of the proposed models can be improved by the use of clustering techniques and learning rate scheduler.

Conflict of Interest

The authors declare no conflict of interest.

References

[1]. A. Gottlieb, G.Y. Stein, E. Ruppin, R.B. Altman, R. Sharan, "A method for inferring medical diagnoses from patient similarities," BMC Medicine, **11**, 194, 2013, doi:10.1186/1741-7015-11-194.

[2]. R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," Briefings in Bioinformatics, **19**(6), 1236-1246, 2017, doi:10.1093/bib/bbx044.

[3]. I. Tobore, J. Li, L. Yuhang, Y. Al-Handarish, A. Kandwal, Z. Nie, L. Wang, "Deep learning intervention for health care challenges: Some biomedical domain considerations," Journal of Medical Internet Research, **7**(8), e11966, 2019, doi:10.2196/11966.

[4]. R. Bellazzi, B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International Journal of Medical Informatics,**77**(2), 81-97, 2008, doi:10.1016/j.ijmedinf.2006.11.006.

[5]. G. LeCun, Y., Bengio, Y., Hinton, "Deep learning. Nature," **521** (7553), 436–444, 2015, doi:10.1038/nature14539.

[6]. T. Young, D. Hazarika, S. Poria, E. Cambria, " Recent trends in deep learning based natural language processing [Review Article], " IEEE Computational Intelligence Magazine, **13**(3), 55-75, 2018, doi:10.1109/MCI.2018.2840738.

[7]. J. Powles, H. Hodson, "Google DeepMind and healthcare in an age of algorithms," Health and Technology, **7**(4), 351-367, 2017, doi:10.1007/s12553-017-0179-1.

[8]. A. Qayyum, J. Qadir, M. Bilal, A. Al Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," IEEE Reviews in Biomedical Engineering, 2020, doi:10.1109/RBME.2020.3013489.

[9]. Y. Cheng, F. Wang, P. Zhang, J. Hu, "Risk prediction with electronic health records: A deep learning approach," in 16th SIAM International Conference on Data Mining, 432-440, 2016, doi:10.1137/1.9781 611974348.49.

[10]. T. Pham, T. Tran, D. Phung, S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," Journal of Biomedical Informatics, 16, 218-229, 2017, doi:10.1016/j.jbi. 2017.04.001.

[11]. E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks.," JMLR Workshop and Conference Proceedings, 301-318, 2016.

[12]. R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Scientific Reports, 6(1), 1-10, 2016, doi:10.1038/srep26094.

[13]. Z. Liang, G. Zhang, J.X. Huang, Q.V. Hu, "Deep learning for healthcare decision making with EMRs," in Proceedings IEEE International Conference on Bioinformatics and Biomedicine(BIBM),556-559, 2014, doi:10.1109/BIBM.2014.6999219.

[14]. Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, 2016.

[15]. Z. Che, D. Kale, W. Li, M.T. Bahadori, Y. Liu, "Deep computational phenotyping," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 507-516, 2015, doi:10.1145/2783258.2783365.

[16]. T.A. Lasko, J.C. Denny, M.A. Levy, "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data," PLoS ONE, **8**(6), e66341, 2013, doi:10.1371/journal.pone.0066341.

[17]. D. Kollias, A. Tagaris, A. Stafylopatis, S. Kollias, G. Tagaris, "Deep neural architectures for prediction in healthcare," Complex & Intelligent Systems, 4(2),119-131, 2018, doi:10.1007/s40747-017-0064-6.

[18]. T. Tran, T.D. Nguyen, D. Phung, S. Venkatesh, "Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)," Journal of Biomedical Informatics, **54**, 96-105, 2015, doi:10.1016/j.jbi.2015.01.012.

[19]. P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, "Deepr: A Convolutional Net for Medical Records," IEEE Journal of Biomedical and Health Informatics, **21**, 22-30, 2017, doi:10.1109/JBHI.2016. 2633963.

[20]. R. Panthong, A. Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm," in Procedia Computer Science, **72**, 162-169, 2015, doi:10.1016/j.procs. 2015.12.117.

[21]. B. Nakisa, M.N. Rastgoo, A. Rakotonirainy, F. Maire, V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," IEEE Access, **6**, 49325-49338, doi:10.1109/ACCESS.2018.2868361.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 256-260 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Multi Operated Virtual Power Plant in Smart Grid

Yevhen Fediv, Olha Sivakova, Mykhailo Korchak*

Lviv Polytechnic National University, Department of Electric Power Engineering and Control Systems, Institute of Power Engineering and Control Systems, Lviv, 79012, Ukraine

ARTICLE INFO

Article history:
Received: 23 August, 2020
Accepted: 14 October, 2020
Online: 10 November, 2020

Keywords:
Virtual power plant
Smart grid
AC voltage controller
GTO thyristor
Reactive power
IoT

ABSTRACT

In order to balance the power in intelligent distribution networks (Smart Grid or Microgrid), it is proposed to organize a «multi operated virtual power plant». The resources of active and reactive power for which can be obtained using an AC voltage controller with a phase-angle control for regulation of operating modes of the ohmic load of consumers, for example, distributed systems of electric space heating or electric water heating, etc. A method is proposed and the results of the analysis of the phase-angle control modes by gate turn off (GTO) thyristors of the AC voltage regulator, which provides the generation of virtual reactive power by consumers of active power, are presented. The process equipment of such a virtual power plant is fully suitable both for the dynamic production of a virtual resource of active power to balance, for example, the power of dynamic distributed renewable energy sources (virtual power plant (VPP) mode), and to regulate reactive power to ensure adequate voltage levels and increase stock stability of operation of electric load units (virtual reactive power plant (VRPP) mode). References 24, figures 8.

1. Introduction

This article is an extension and continuation of the work presented in the materials of the IEEE conference [1], which shows the possibility of power electronics to obtain a resource of reactive power from consumers-regulators with an active nature of consumption. Improving the efficiency of power supply, development in the direction of intellectualization according to the concept of Smart Grid is based on modern advances in power electronics [2-4].

In addition to balancing active power with frequency stabilization, an important operational task in Smart Grid systems is to ensure proper voltage levels in power grid units, increase the stability margin of power generation and consumption systems, and optimize operating costs in the transportation and distribution of electricity.

The basis for solving these problems is to ensure the balance of reactive power in the system, the availability of appropriate resources and the ability to quickly involve them at different levels of transmission, conversion and distribution of electricity [5, 6]. In the Smart Grid concept, the problem of reactive power compensation is highlighted as key. Different methods of reactive power compensation are considered in [7, 8].

*Corresponding author: Mykhailo Korchak., misha.korkirov@gmail.com +380981712257

The universal and most efficient centralized source of reactive power at present is the Static Compensator (Statcom) device [9], which is primarily intended for flexible power transmission systems or powerful industrial power supply systems.

Problems of using Statcom in low-voltage networks and their interaction with distributed energy sources are considered in articles [10,11], where special attention is paid to the use of photovoltaic inverters as reactive power sources.

In [12-14], the use of static compensator D-STATCOM for unbalanced radial power distribution systems is investigated in order to ensure voltage stability within acceptable limits. This review of known works addresses issues related to the optimal choice of power, control methods, installation location, etc. of reactive power sources as physical objects installed by the power supplier. However, currently in the Smart Grid an important role is given to virtual energy sources

In the classical sense, a VPP is a distributed power plant that combines the resources of active power generation from distributed energy resources (DER) as a reserve at the disposal of transmission and distribution operators [15]. VPPs in Microgrid are able to increase resource efficiency and maintain the necessary balance of generation and consumption in the power system. Distributed renewables and drives in VPPs are not virtual but real

resources. In fact, virtual (feigned) generation of electricity will be equivalent to refusing or limiting its consumption. For example, by integrating water heating or electric heating systems into a grid and remotely controlling them with electricity consumption, you can align the active power consumption schedule at a particular distribution network node without significant consequences for household or industrial consumers themselves.

Another possibility is the participation of customers in virtual power plants to provide reactive energy. [16-20] discusses the provision of reactive power services from DER using a market mechanism and a plan to convert Microgrid into VPP, which is beneficial for consumers, prosumers and utility operators.

Involving distributed sources to solve the problem of reactive power balancing, by analogy with a virtual power plant, can be combined into virtual reactive power sources. In [21, 22] the problem of reactive power is not allocated separately but included in the list of tasks for the organization of VPP. [23] presents the concept of a dynamic VRPP as an ancillary service to reasonably combine decentralized power distribution sources into a coordinated pool to provide reactive power for the needs of power transmission system operators. Common in the known concepts is the involvement in the VRPP real distributed sources of reactive power.

2. Purpose and task of the research

The aim of the work is to study the problem of balancing active and reactive power in distribution networks with the involvement of consumers' resources in the form of integrated virtual power plant (VPP-VRPP), where reactive power resource is obtained by regulating the receivers used for VPP**.** This is especially true for facility power systems that are not encouraged to install their own sources to compensate for reactive power.

The ranges of complex regulation by means of power electronics of consumers with active character of loading for reception of a virtual resource of active and reactive power by regulation according to needs of operators of power supply systems are substantiated in the work.

3. Finding a solution to the problem

Figure 1 shows the equivalent circuit diagram of one phase VPP-VRPP, which shows the AC voltage regulator, which includes a source of sinusoidal voltage with internal inductance, a block of power dual-operation (GTO) thyristors VS1 and VS2, a limiting diode suppression diode) VD, as well as resistive resistance R of water heater or electric heating.

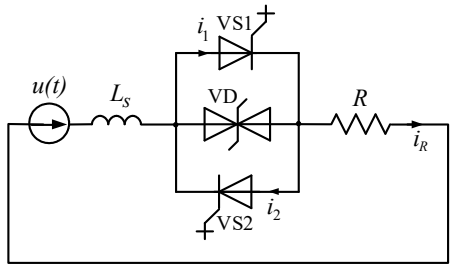


Figure 1: Equivalent power supply circuit of the heater

To protect against the effects of rigid switching of thyristors inductive overvoltage emissions in the interval between the closing of the thyristors, a two-way TVS diode VD is designed.

To obtain the energy characteristics in the steady state VPP-VRPP, we assume that the thyristors VS1, VS2 are ideal switches, not taking into account the diode VD, inductance $L_S = 0$, and voltage of the power supply system is sinusoidal as shown in (1)

$$u(t) = U_m \sin(\omega t + \varphi_u) \quad (1)$$

where U_m is power supply voltage amplitude; φ_u is the initial phase of the supply voltage; ω is angular velocity.

To obtain the power in the circuit of the resistor, controlled by key semiconductor elements, we have to analyze two characteristic modes.

3.1. Mode of consumption of virtual reactive power from the power supply network.

This mode is characteristic for traditional single-operation (semi-controlled) thyristors or SCR (Semiconductor Controlled Rectifier). For its realization thyristors VS1 and VS2 is being opened by the controlling impulses with a delay of an angle α relative to the natural angle of their entry into operation $\alpha_0 = 0 \pm \pi k$, where $k = 0, 1, 2\ldots$, and closing them at the natural time of closing $\beta_0 = \pi + \pi k$ (Fig. 2).

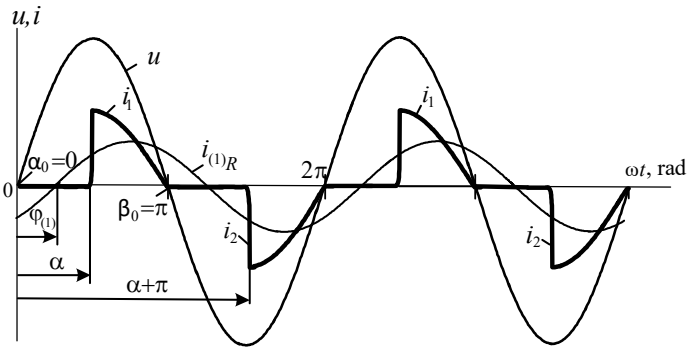


Figure 2: VRPP current and voltage time diagrams in the mode of reactive power consumption

As a result, we obtained the shift of the main harmonic of the source current $i_{(1)R}$ towards the delay relative to the voltage of the grid. It is equivalent to the consumption of virtual reactive power by electrical circuit, in which reactive elements are not presented.

It should be noted that the value of the reference times of the thyristor control angles (natural value of the angles of opening and closing of the thyristors) are characteristic of the purely active nature of the circuits with key elements.

To evaluate the energy effect of the above method of regulating the operation of the electric receiver, we decompose the current i_R in the Fourier series. *n*-harmonic component of the relative current is descripted in expression (2)

$$i_n(\omega t) = I_{nm*} \cdot \sin(n\omega t + \varphi_n) \quad (2)$$

can be obtained as (3)

$$I_{nm*} = \frac{I_{nm}}{I_b} = \sqrt{A_{n*}^2 + B_{n*}^2} \tag{3}$$

where I_b is the current amplitude in the mode of full conductivity of thyristors according to (4)

$$I_b = I_{m,R} = U_m / R \tag{4}$$

A_{n*} and B_{n*} are the coefficients for the decomposing to the Fourier series is written using (5) and (6):

$$\begin{aligned} A_{n*} &= \frac{1}{\pi}\int_{\alpha}^{\pi} i_R(\omega t)\cos n\omega t\, d(\omega t) + \frac{1}{\pi}\int_{\pi+\alpha}^{2\pi} i_R(\omega t)\cos n\omega t\, d(\omega t) \\ &= \frac{1}{2\pi}\left\{\frac{\cos(n\pi)+\cos\left[(1-n)\alpha\right]-\cos(2n\pi)+\cos\left[(1-n)(\pi+\alpha)\right]}{1-n}\right. \\ &\left.+\frac{\cos(n\pi)+\cos\left[(1+n)\alpha\right]-\cos(2n\pi)+\cos\left[(1+n)(\pi+\alpha)\right]}{1+n}\right\} \end{aligned} \tag{5}$$

$$\begin{aligned} B_{n*} &= \frac{1}{\pi}\int_{\alpha}^{\pi} i_R(\omega t)\sin n\omega t\, d(\omega t) + \\ &\frac{1}{\pi}\int_{\pi+\alpha}^{2\pi} i_R(\omega t)\sin n\omega t\, d(\omega t) \\ &= \frac{1}{2\pi}\left\{\frac{-\sin\left[(1-n)\alpha\right]-\sin\left[(1-n)(\pi+\alpha)\right]}{1-n}\right. \\ &\left.+\frac{\sin\left[(1+n)\alpha\right]+\sin\left[(1+n)(\pi+\alpha)\right]}{1+n}\right\} \end{aligned} \tag{6}$$

For the 1st harmonic ($n=1$) obtained equations (7) and (8):

$$A_{1*} = \frac{1}{2\pi}\left[\cos(2\alpha)-1\right] \tag{7}$$

$$B_{1*} = \frac{1}{2\pi}\left[\sin(2\alpha)+2(\pi-\alpha)\right] \tag{8}$$

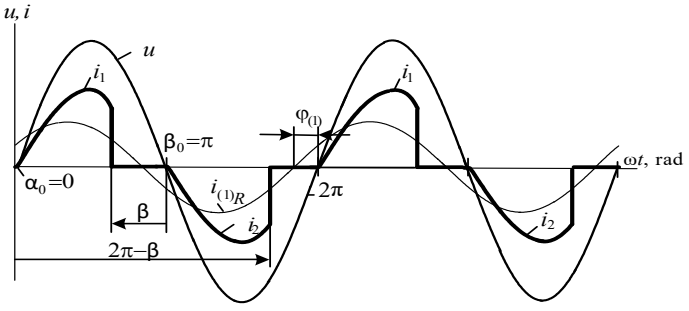


Figure 3: VRPP current and voltage time diagrams in the mode of reactive power generating

3.2. Mode of virtual reactive power generation in the power supply network.

This mode is dual to the previous one. For its realization thyristors VS1 and VS2 is being opened in the natural moments of their entry into operation $\alpha_0 = 0 + \pi k$, where $k = 0,1,2\ldots$, and closing with an advance on the angle β relative to the natural angle β_0 of closure (Fig. 3). As a result, we obtained the shift of the main harmonic of the source current $i_{(1)R}$ in the direction of advance relative to the grid voltage, which is equivalent to generation of reactive power by resistance into the power supply system.

The formulas obtained by the authors for the decomposition coefficients in the Fourier series of the resistor current take the form (9) and (10).

$$\begin{aligned} A_{n*} &= \frac{1}{\pi}\int_{0}^{\pi-\beta} i_R(\omega t)\cos n\omega t\, d(\omega t) \\ &+\frac{1}{\pi}\int_{\pi}^{2\pi-\beta} i_R(\omega t)\cos n\omega t\, d(\omega t) \\ &= \frac{1}{2\pi}\left\{\frac{1-\cos\left[(1-n)(\pi-\beta)\right]-\cos\left[(1-n)(2-\beta)\right]-\cos(n\pi)}{1-n}\right. \\ &\left.+\frac{1-\cos\left[(1+n)(\pi-\beta)\right]-\cos\left[(1+n)(2\pi-\beta)\right]-\cos(n\pi)}{1+n}\right\} \end{aligned} \tag{9}$$

$$\begin{aligned} B_{n*} &= \frac{1}{\pi}\int_{0}^{\pi-\beta} i_R(\omega t)\sin n\omega t\, d(\omega t) \\ &+\frac{1}{\pi}\int_{\pi}^{2\pi-\beta} i_R(\omega t)\sin n\omega t\, d(\omega t) \\ &= \frac{1}{2\pi}\left\{\frac{\sin\left[(1-n)(\pi-\beta)\right]+\sin\left[(1-n)(2\pi-\beta)\right]}{1-n}\right. \\ &\left.-\frac{\sin\left[(1+n)(\pi-\beta)\right]+\sin\left[(1+n)(2\pi+\beta)\right]}{1+n}\right\} \end{aligned} \tag{10}$$

For the 1st harmonic ($n=1$) we obtained equations (11) and (12):

$$A_{1*} = \frac{1}{2\pi}\left[1-\cos(2\beta)\right] \tag{11}$$

$$B_{1*} = \frac{1}{2\pi}\left[2(\pi-\beta)+\sin(2\beta)\right] \tag{12}$$

4. Generalization of the results of the analysis of control regimes

In both of the above modes of thyristor control with a change in the value of the control angles and there is a simultaneous change in both active and reactive power, the value of which can be calculated by expressions (13) and (14).

$$P_{1*} = \sqrt{A_{1*}^2 + B_{1*}^2}\cos(\varphi_{1u} - \varphi_{1i}) = I_{1m*}\cos\varphi_1 \tag{13}$$

$$Q_{1*} = I_{m1}\sin\varphi_1 \tag{14}$$

where, P_{1*}, Q_{1*}

$\sqrt{A_{1*}^2 + B_{1*}^2} = I_{1m*}$ – the amplitude of the fundamental harmonic of the current, $\varphi_1 = \arctan(B_1/A_1)$ – the initial phase of the fundamental harmonic of the current, and $\varphi_{1u} = 0$.

Figure 4 shows the results of the calculation of VPP-VRPP control characteristics by reactive power. The maximum generation of reactive power is $|Q_{1*}| \approx 0,32$ p.u. different signs are reached at the control angles of the boards $\alpha = \beta = \pi/2$. In case of regulating the VPP-VRPP by changing the opening angles $\alpha = \text{var}$ of thyristors and constant closing angles $\beta = \text{const} = 0$ from the mains consumes excess reactive bias power ($Q_1 > 0$), and by constant angles of thyristor opening ($\alpha = \text{const} = 0$) and adjusting the angle $\beta = \text{var}$ change we get the mode of generating reactive bias power to the mains ($Q_1 < 0$).

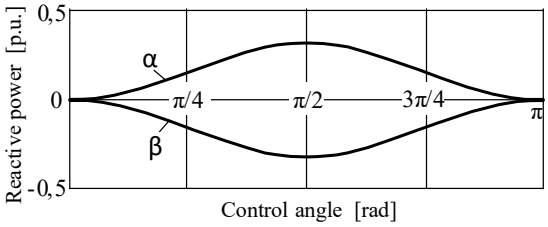


Figure 4: VPP-VRPP control characteristics by reactive power

The obtained characteristics are symmetric with respect to the angle $\alpha = \beta = \pi/2$, but in the mode of reactive power control the preference should be given to the range of angles $\alpha = \beta = 0 \ldots \pi/2$ of regulation of the boards, because in the range of angles $\alpha = \beta = \pi/2 \ldots \pi$ we get a significant harmonic current distortion (Fig. 5)

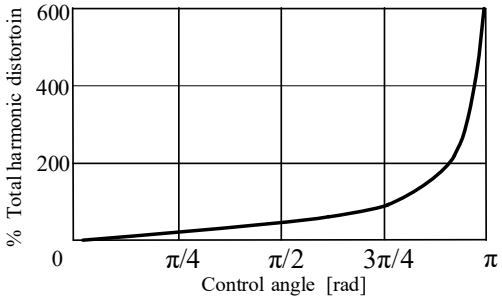


Figure 5: Coefficient of harmonic current distortion

Calculated by expression (15)

$$THD_I(\%) = \frac{I_{dis}}{I_1}100 = \frac{\sqrt{\sum_{n\neq 1} I_n^2}}{I_1}100, \qquad (15)$$

The value of the harmonic distortion coefficients of the current consumed from the mains for the above-recommended range of adjustment of the control angles is $THD_I = (0...36,6)\,\%$. For the mode of full conductivity of thyristors, there is no distortion, and for the modes $\alpha = \beta = \pi/2$ the harmonic distortion of VPP-VRPP current is $THD_I \approx 36,6\,\%$.

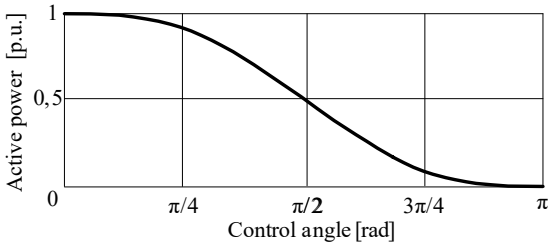


Figure 6: VPP-VRPP control characteristic by active power

The change in the reactive power of the VPP-VRPP in the process of regulation is accompanied by a simultaneous change in the active power consumed by the electric receiver (Fig. 6).

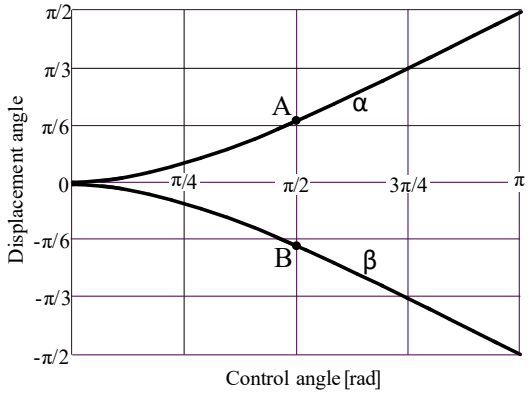


Figure 7: Displacement angle for two ways of regulation (modes) VRPP

As you can see from Fig. 4 and fig. 6, in the process of regulation, with increasing value of reactive power, the value of active power decreases, which generally contributes to the tasks to be solved by virtual power plants, the control effect on the active $dP/d\alpha < 0$, $dP/d\beta < 0$ and reactive $dQ/d\alpha > 0$, $dQ/d\beta > 0$ power of the resistive electrical receiver is different in signs.

We obtained modes in which the generation of reactive power to balance it in the load center is accompanied by a simultaneous decrease in active power consumption, which contributes to both the normalization of the voltage level (VRPP task) and active power balancing (VPP task).

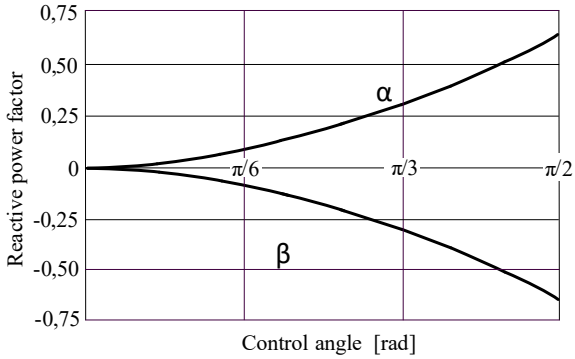


Figure 8: Reactive displacement power factor

In the mode of maximum reactive power $\alpha = \beta = \pi/2$, the value of the displacement angle of the main harmonic of the current is $\varphi_1 \approx \pm 32,5^\circ$ (points A and B on the corresponding characteristics of Fig. 7), and the coefficient of reactive displacement power (Fig. 8) $tg\varphi_1 = \pm 0,64$.

5. Conclusions

The article considers a comprehensive approach to solving the problems of active and reactive power balances in Microgrid networks, using hidden (virtual) resources, which are organized in the form of a virtual active-reactive power plant (VPP-VRPP). Using distributed resources of consumers and their active role in the operation of Microgrid.

Resources of active and reactive power are received from one source - ohmic receivers of consumption with involvement of means of power electronics as regulating elements.

By feeding the ohmic load from voltage regulators with GTO thyristors, the distribution network operator, if necessary, reduces the consumption of active power and adjusts the value and sign (consumption from grid or grid generation) of reactive power.

The range of change of control angles of thyristors of a voltage regulator is chosen for reasons of reduction of negative influence of nonlinear electric loading on quality of the electric power in Microgrid.

The use of integrated multi-controlled VPP-VRPP can be recommended for damping transients and ensuring the stability of Microgrid operation in modes with frequency and voltage value close to critical, to prevent avalanche-like changes.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank both the authorities of the Lviv Polytechnic National University especially leadership of the Institute of Power Engineering and Control Systems for their support during the preparation of this paper. The authors also thank the anonymous reviewers, whose comments significantly improved the content of the paper.

References

[1] Y. Fediv, O. Sivakova, M. Korchak, "Model of Virtual Source of Reactive Power for Smart Electrical Supply Systems," in 2019 IEEE 20th International Conference on Computational Problems of Electrical Engineering (CPEE), 1-4, 2019, doi:10.1109/CPEE47179.2019.8949159.

[2] G. Benysek, M.P. Kazmierkowski, J.Popczyk, R. Strzelecki, "Power electronic systems as a crucial part of Smart Grid infrastructure – a survey," Bulletin of the polish academy of sciences technical sciences, **59(**4), 2011, doi:10.2478/v10175-011-0058-2.

[3] B.K. Bose, "Power Electronics in Renewable Energy Systems and Smart Grid: Technology and Applications," Wiley-IEEE Press, 2019.

[4] B. K. Bose, "Power Electronics in Smart Grid and Renewable Energy Systems," in Proceedings of the IEEE, **105**(11), 2007-2010, 2017, doi:10.1109/JPROC.2017.2752538.

[5] J. Ali, S. Massucco, G. Petretto, "Reactive power provision to TSO/DSO by aggregators and conventional generators," 2017 IEEE International Conference on Smart Grid Communications (SmartGridComm), 486-491, 2017, doi:10.1109/SmartGridComm.2017.8340740.

[6] K. Wang, M.L. Crow, "Electricity Transmission, Distribution and Storage Systems," Woodhead Publishing Series in Energy, 2013.

[7] J. Ali, S. Massucco, F. Silvestro, "STATCOM application for voltage profiling of a distribution grid with high penetration of distributed energy resources," Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER 2018), 1-7, 2018, doi:10.1049/cp.2018.1883.

[8] J. Ali, S. Massucco, F. Silvestro, "Aggregation Strategy for Reactive Power Compensation Techniques-Validation," Energies **2019**, 1-13, 2019, doi:10.3390/en12112047.

[9] Q. Wang, B. Wang, W. Xu, J. Xu, "Research on STATCOM for reactive power flow control and voltage stability in microgrid," 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, 2474-2479, 2018, doi:10.1109/ICIEA.2018.8398126.

[10] S. Hasan, R. Luthander, J. de Santiago, "Reactive Power Control for LV Distribution Networks Voltage Management," in 2018 IEEE PES Innovative Smart Grid Technologies Conference Europe *(*ISGT-Europe*)*, 1-6, 2018, doi:10.1109/ISGTEurope.2018.8571817.

[11] D. Sailaja, V.B. Kumar, "A Voltage Regulator for Power Quality Improvement in Low-Voltage Distribution Grids," Journal of Applied Science and Computations, **5**(7), 461-468, 2018, DOI:10.1109/TPEL.2017.2693239.

[12] A.R. Guptaa, A. Kumar, "Impact of D-STATCOM placement on improving the reactive loading capability of unbalanced radial distribution system," Procedia Technology 25, 676-683, 2016, doi: 10.1016/j.protcy.2016.08.170.

[13] W. Guoyou, Z. Xilin, "Study on hybrid reactive power compensation in low-voltage distribution system," International Conference on Energy and Environmental Protection (ICEEP 2016), 618-622, 2016, doi:10.2991/iceep-16.2016.106.

[14] L. E. Christian, L. M. Putranto, S.P. Hadi, "Design of Microgrid with Distribution Static Synchronous Compensator (D-STATCOM) for Regulating the Voltage Fluctuation," 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), 48-52, 2019, doi:0.1109/SEGE.2019.8859860.

[15] M.M. Othman, Y.G. Hegazy, A.Y. Abdelaziz, "A Review of Virtual power plant Definitions, Components, Framework and Optimization," International Electrical Engineering Journal (IEEJ), 2010-2024, **6**(9), 2015.

[16] M. Korchak, O. Sivakova, Y. Fediv, «Virtual source of reactive power in electricity supply systems of household consumers,» Energy Engineering and Control Systems, 9–14, **5**(1), 2019 doi:10.23939/jeecs2019.01.009.

[17] D. Pudjianto, P. Djapic, G. Strbac, B. Stojkovska, A. R. Ahmadi, I. Martinez, "Integration of distributed reactive power sources through Virtual Power Plant to provide voltage control to transmission network," 25thInternational Conference on Electricity Distribution (CIRED-2019), 1-5, (2055), 2019. http://dx.doi.org/10.34890/933.

[18] J. Ali, S. Massucco, F. Silvestro, A. Vinci, «Participation of Customers to Virtual Power Plants for Reactive Power Provision,» Conference: 2018 53rd International Universities Power Engineering Conference (UPEC), 1-6, 2018, doi:10.1109/UPEC.2018.8541978.

[19] K. L. Anaya, M. G. Pollitt, "Reactive Power Procurement: Lessons from Three Leading Countries," Cambridge Working Papers in Economics, 2018.

[20] J. Naughton, H. Wang, S. Riaz, M. Cantoni, P. Mancarella, "Optimization of Multi-Energy Virtual Power Plants for Providing Multiple Market and Local Network Services," Electric Power Systems Research, 189, 2020, doi:10.1016/j.epsr.2020.106775.

[21] Levent Yavuz, Ahmet Önen, S.M. Muyeen and Innocent Kamwa, "Transformation of Microgrid to Virtual Power Plant - a Comprehensive Review," IET Generation Transmission & Distribution, **13**(11), 1994-2005, 2018, doi:10.1049/iet-gtd.2018.5649.

[22] R. Bertram, A. Schnettler, "A control model of virtual power plant with reactive power supply for small signal system stability studies," 2017 IEEE ManchesterPowerTech, Manchester, 1-6, 2017, doi:10.1109/PTC.2017.7980838.

[23] F. Muuß, G.A.N. Hemdan, M. Kurrat, D. Unger, B. Engel, "Dynamic virtual reactive power plant in active distribution networks," IEEE Eindhoven PowerTech (PowerTech, 697-702, 2015, doi**:**10.1109/PTC.2015.7232356.

[24] IEEE Standard Definitions for the Measurement of Electric Power Quantities Under Sinusoidal, Nonsinusoidal, Balanced, or Unbalanced Conditions, IEEE Std., 1459-2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 3, 259-263 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Enhanced Collaborative Constellation for Visible Light Communication System

Manh Le Tran, Sunghwan Kim*

School of Electrical Electronics, University of Ulsan, 93, Daehak-ro, Nam-gu, Ulsan, 44610, Korea

ARTICLE INFO

Article history:
Received: 03 April, 2020
Accepted: 05 May, 2020
Online: 28 May, 2020

Keywords:
Maximum-likelihood detection
Constellation design
Wireless optical communication

ABSTRACT

Visible light communication (VLC) that simultaneously gives illumination and information transmission abilities, is recognized as a hopeful competitor for prospective wireless networks. Furthermore, a channel adaptive collaborative constellation (CASCC) with the capacity of adapting according to the channel condition to enhance the bit error rate (BER) while concurrently improving the versatility of the receiver mobility was regarded to be further effective than the contemporary CC. Nonetheless, the early CASCC barely presents modest performance enhancement in a strong correlation channel. Hence, by this study, we provide a design to form a channel-adaptation CC, namely the enhanced CC (ECC) for VLC systems. More specifically, from the basic constellation points, we form the optimization problem of efficient size that can be solved by any convex optimization solving technique. Also, the computational simulation outcomes confirm that the ECC is more beneficial than preceding constellations in term of BER for different channels. Moreover, we also provide the result comparison of the proposed scheme with other schemes in the imperfect channel condition. Overall, by effectively reducing the distance among the constellation points, a significant signal-to-noise ratio gain can be achieved.

1 Introduction

As a substitute competitor for radio frequency (RF) communications, visible light communication (VLC) has lately brought significant attention of researchers, accompanied by the growing popularity of the emerging solid-state lighting like the laser diode (PD) and then light-emitting diode (LED) to realize the conventional illumination sources thanks to their extraordinary efficiency and massive bandwidth, exceptional security, inexpensive cost, independence from the spectral licensing problem, simple implementation into existing infrastructure [1, 2]. In conventional VLC systems that convey information based on LEDs, because of the low hardware-cost and simplicity of implementation, the intensity-modulation (IM) with direct-detection (DD) designs are employed. Furthermore, the communications systems with optical multiple-input multiple-output (MIMO) techniques [3–6] can be engaged to obtain various data transmission rates [7]. Though, MIMO-VLC systems, which heavily depend on channel correlations, severely influence the achievable performance [8]. On the other hand, various modulation schemes have been proposed for MIMO-VLC, such as repetition code (RC), spatial modulation (SM), and spatial multiplexing (SMP) [9]. In RC the same signal is emitted simultaneously from all transmitters. In SM, the conventional signal constellation diagram is extended to the spatial dimension, in which only one transmitter is activated at any symbol duration. In SMP, independent data streams are simultaneously emitted from all transmitters.

Recently, an appealing design named collaborative constellation (CC) concerning MIMO-VLC systems with various numbers of PDs and LEDs was introduced [10]. To outlining the transmitted constellation of varied LED spaces, the Euclidean distances (EDs) were simultaneously maximized following the constrains in average transmit power. Later, to considerably enhanced the design of CC, the channel adaptive space CC (CASCC) was proposed in [11] by taking the channel influences on the system model into consideration. More particularly, in expressions including four primary points, CASCC design intended to a 2×2 MIMO systems and later prolonged to higher-order constellations. Noticeably, the idea about CASCC mentioned in [11] can accommodate to different channels among variations in positions of both LEDs and PDs.

In this study, we offer a practical algorithm to outline a multilayer CC (ECC) for MIMO-VLC system utilizing simultaneously the space-collaboration and channel-adaptation ideas [3]. Furthermore, the introduced approach attends to produce the constellations of each size based on particular primary points. Besides, with the

*Corresponding Author Sunghwan Kim, School of Electrical Electronics, University of Ulsan, 93, Daehak-ro, Nam-gu, Ulsan, 44610, Korea, sungkim@ulsan.ac.kr

help of the multilayer constellation scheme, the offered design can minimize the amount of constraint presented in optimization problems. More precisely, our offered design analyzes the distance of every pair of points; also, the form of the constellation in the PDs does not have to be distinct as in CASCC. Our proposed scheme considers the distance of any two constellation points and the shape of constellation in receiver space does not have to be a diamond shape. Hence, effectively design the adaptive collaborative constellation and achieve SNR gain in comparison with CASCC, both under perfect and imperfect CSI. Consequently, ECC efficiently produces the adaptive constellation and gains a notable SNR enhancement in comparing to CASCC including different designs.

2 Transmission model

In this study, the introduced principle of the ECC for MIMO-VLC systems, comparable to [10, 11], have the numbers of LEDs and PDs are $n = 2$, respectively. Furthermore, we denote C that contains $m = 2^r$ symbols as the constellation at the transmitting side, that is $C = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ with r bits as the transmission rate. Moreover, with $x_\alpha^{(1)}$ denotes the signal value transmitted from 1th-LED while $x_\alpha^{(2)}$ denotes the signal value transmitted from 2th-LED, the overall transmitted signal $\mathbf{x}_\alpha$ is represented as $\mathbf{x}_\alpha = (x_\alpha^{(1)}, x_\alpha^{(2)})^T \in C$. Since we have the number of constellation points in C should be a power of two, hence we at the beginning present detail procedure to design a constellation with several layers, so-called $\mathcal{X}$. After the constellation $\mathcal{X}$ is initially formed, by picking a subset of constellation point in $\mathcal{X}$, the constellation set C is optimally obtained. This is clear since the proposed constellation $\mathcal{X}$ with the number of layers is l, will have $l(l+1)/2$ as the cardinality.

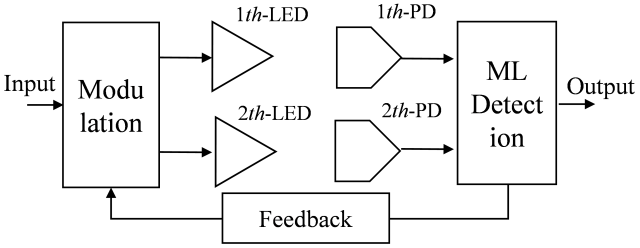


Figure 1: System model.

Following the constellation introduced in [10], a set of constellation points that contained in a number of g layers is first described . It can bee seen that, in the g-th layer of constellation (g=1, ...,l), as shown in Fig.1 there are$g$ points. Moreover, begin by the primary points $\mathbf{x}_{11} = (0,0)^T$, $\mathbf{x}_{21} = (x_{21}^{(1)}, x_{21}^{(2)})^T$, and $\mathbf{x}_{22} = (x_{22}^{(1)}, x_{22}^{(2)})^T$, the g layers constellation set can be obtained with $\mathbf{x}_{\alpha p} = (x_{\alpha p}^{(1)}, x_{\alpha p}^{(2)})^T$ denotes the p-th constellation point in the $\alpha-$th layer of the proposed constellation set, and $x_{\alpha p}^{(1)}, x_{\alpha p}^{(2)}$ denote the signal values that can be transmitted simultaneously by both 1th-LED and 2th-LED, respectively. Moreover, the at the transmitter side, the constellation set $\mathcal{X}$ will be expressed by

$$\mathcal{X} = \left\{ \mathbf{x}_{\alpha p} | \mathbf{x}_{\alpha p} = \left(x_{\alpha p}^{(1)}, x_{\alpha p}^{(2)}\right)^T = \left[(\alpha - p) x_{21}^{(1)} + (p-1)\, x_{22}^{(1)}, (\alpha - p) x_{21}^{(2)} + (p-1)\, x_{22}^{(2)} \right]^T \right\} \quad (1)$$

with $1 \le p \le \alpha, 1 \le \alpha \le l$.

The line-of-sight (LOS) VLC systems, which comprise two LEDs and two PDs is analyzed in this study. Moreover, the received signal $\mathbf{y}$ at the receiver, with $\mathbf{x} = \left(x^{(1)}, x^{(2)}\right)^T$ are the conveyed signals and the i-th LED signal is denoted as $x^{(i)}$, similar to, [11], can be rewritten as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \quad (2)$$

Besides, we assume that $\mathbf{n}$ is the $n_t \times 1$ additive white Gaussian noise with zero mean and σ^2 is the variance. Moreover, the assumption that the LEDs' sufficient knowledge about the channel matrix $\mathbf{H}$ of size 2×2 through feedback channel is acceptable. Also, the channel element h_{ji} between the j-th PD and the i-th LED is described by [1]

$$h_{ji} = \begin{cases} \frac{(k+1)A_r}{2\pi\rho_{ji}^2} \cos^k(\phi) \cos(\psi) & , 0 \le \psi \le \psi_{1/2} \\ 0 & , \psi > \psi_{1/2} \end{cases} \quad (3)$$

where the angle-of-emergence with the LED axes and the angle-of-incidence with its normal axes can be denoted as ϕ and ψ, respectively. Also, the receiving area of the PDs can be generally expressed by A_y. $\psi_{1/2}$ indicates the field-of-view semi angle of the PDs. The distance from the j-th PD to the p-th LED is denoted as ρ_{ji}. The Lambertian order k will be described by $k = \frac{-\ln 2}{\ln(\cos \Phi_{1/2})}$ with the half-power semi angle of the LEDs is $\Phi_{1/2}$.

On the other hand, with $y_{\alpha p}^{(1)}, y_{\alpha p}^{(2)}$ denoting signals received by PD_1 and PD_2, the received symbol at both PDs $\mathbf{y}_{\alpha p}$ can be expressed as follows

$$\mathbf{y}_{\alpha p} = \mathbf{H}\mathbf{x}_{\alpha p} \Leftrightarrow \begin{bmatrix} y_{\alpha p}^{(1)} \\ y_{\alpha p}^{(2)} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_{\alpha p}^{(1)} \\ x_{\alpha p}^{(2)} \end{bmatrix} \quad (4)$$

where the corresponding transmitted constellation $\mathbf{x}_{\alpha p} \in \mathcal{X}$, $\mathcal{X} = [\mathbf{x}_{11}, \mathbf{x}_{21}, \mathbf{x}_{22}, \mathbf{x}_{31}, \mathbf{x}_{32}, \mathbf{x}_{33}, ..., \mathbf{x}_{l1}, \mathbf{x}_{l2}, ...\mathbf{x}_{ll}]$ and the channel matrix $\mathbf{H}$ of size 2×2. Therefore, the constellation in the PDs will be expressed as $\mathcal{Y} = [\mathbf{y}_{11}, \mathbf{y}_{21}, \mathbf{y}_{22}, \mathbf{y}_{31}, \mathbf{y}_{32}, \mathbf{y}_{33}, ..., \mathbf{y}_{l1}, \mathbf{y}_{l2}, ...\mathbf{y}_{ll}]$. Additionally, we have $|\mathcal{Y}| = |\mathcal{X}|$ with $\mathcal{Y} = \mathbf{H}\mathcal{X}$. Any symbol belong to $\mathcal{X}$, as previously remarked in [11], will be exclusively assigned to $\mathcal{Y}$. As a result, $\mathcal{Y}$ is

$$\mathcal{Y} = \left\{ \mathbf{y}_{\alpha p} | \mathbf{y}_{\alpha p} = (y_{\alpha p}^{(1)}, y_{\alpha p}^{(2)})^T = \left[(\alpha - p) y_{21}^{(1)} + (p-1)\, y_{22}^{(1)}, (\alpha - p) y_{21}^{(2)} + (p-1)\, y_{22}^{(2)} \right]^T \right\}. \quad (5)$$

With a l layers-constellation set, the cardinality of it is calculated as $l(l+1)/2$. Furthermore, there is the need to bound the number of layers l by using the constrain that $l(l+1)/2 \ge M$ with a $M = 2^y$ points constellation C. Therefore, we arrive at the condition that $l = \left\lceil \frac{\sqrt{8M+1}-1}{2} \right\rceil$. Besides, to maintain a r bits constellation C from the primary constellation-points $\mathbf{x}_{11}, \mathbf{x}_{21}, \mathbf{x}_{22}$, following steps are performed to obtain the full constellation:

Input: matrix $\{\mathbf{x}_{11}, \mathbf{x}_{21}, \mathbf{x}_{22}\}$ where $\mathbf{x}_{11} = (0,0)^T$, size of constellation.

Output: the optimal constellation set C

- Compute g
- Assign $\mathcal{X} = \{\mathbf{x}_{11}, \mathbf{x}_{21}, \mathbf{x}_{22}\}$
- *form the full $l(l+1)/2$ symbols*
  - For $\alpha = 3$ to l
    - For $p = 1$ to α

- Set $\mathcal{X} \leftarrow \mathcal{X} + \left\{\mathbf{x}_{\alpha p} | \mathbf{x}_{\alpha p} = \left(x_{\alpha p}^{(1)}, x_{\alpha p}^{(2)}\right)^T = \left[(\alpha - p)x_{21}^{(1)} + (p-1)\, x_{22}^{(1)}, (\alpha - p)x_{21}^{(2)} + (p-1)\, x_{22}^{(2)}\right]^T\right\}$

- End

- End

Choose the set C from the set $\mathcal{X}$ to be the optimal constellation set.

- For $\mathbf{x}_{\alpha p} \in \mathcal{X}$

- Choose $M = 2^y$ constellation symbols that are nearest from (0,0) to make the set C.

- End

- With the power level requirement, normalize the set C.

- Output the set C as the final optimal constellation set.

3 Proposed design

In this section, the set of primary constellation points is adequately decided to maximize the minimum Euclidean distance between any two constellations on the receiver side. The optimization problem which can be manipulated to create a l layers constellation $\mathcal{X}$ that enhance the distance between each pair of points from the obtained constellation under the total average transmitting power $P_{\max}$. More specifically, the optimization problem purpose is to determine optimal primary points $\mathbf{x}_{11}, \mathbf{x}_{21}, \mathbf{x}_{22}$. Also, at the receiver, the constellation is $\mathcal{Y} = \mathbf{H}\mathcal{X}$. Then, the resultant primary points will then be used to acquire l layers constellation. Moreover, we set $\mu = \alpha - \beta$; $\gamma = p - q$ with the assumption that $\alpha \geq \beta \geq 1$ without any loss of generality. We also define in this study the minimum Euclidean distance $d_{\mathcal{Y}}$ among any two constellation points $\mathbf{y}_{\alpha p}, \mathbf{y}_{\beta q}$. In here, $\mathbf{y}_{\alpha p}, \mathbf{y}_{\beta q}$ are the elements of a particular $\mathcal{Y}$.Consequently, the primary points $\mathbf{x}_{11}, \mathbf{x}_{21}, \mathbf{x}_{22}$, based on the aforementioned criterion, would be optimally designed. From equation (5), The Euclidean distance among any two points, $\mathbf{y}_{\alpha p}, \mathbf{y}_{\beta q} \in \mathcal{Y}$ will be expressed as

$$
\begin{aligned}
d^2\left(\mathbf{y}_{\alpha p}, \mathbf{y}_{\beta q}\right) &= \left\|\mathbf{y}_{\alpha p} - \mathbf{y}_{\beta q}\right\|^2 = \left\{\left[(\alpha-\beta)-(p-q)\right]y_{21}^{(1)} \right. \\
&\left. + (p-q)y_{22}^{(1)}\right\}^2 + \left\{\left[(\alpha-\beta)-(p-q)\right]y_{21}^{(2)} + (p-q)y_{22}^{(2)}\right\}^2 \\
&= \left[\,(\mu-\gamma)\, y_{21}^{(1)} + \gamma y_{22}^{(1)}\right]^2 + \left[(\mu-\gamma)\, y_{21}^{(2)} + \gamma y_{22}^{(2)}\right]^2 \\
&= \left\| \begin{bmatrix} \mu-\gamma & \gamma & 0 & 0 \\ 0 & 0 & \mu-\gamma & \gamma \end{bmatrix} \begin{bmatrix} y_{21}^{(1)} \\ y_{22}^{(1)} \\ y_{21}^{(2)} \\ y_{22}^{(2)} \end{bmatrix} \right\|_2^2 \\
&= \left\|\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{H}\mathbf{t}\right\|_2^2 = \mathbf{u}^T\mathbf{H}^T\mathbf{v}_{(\mu-\gamma)\gamma}^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{H}\mathbf{u} \\
&= \mathbf{u}^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{u} \qquad (6)
\end{aligned}
$$

where $\mathbf{u} = \begin{bmatrix} x_{21}^{(1)} & x_{22}^{(1)} & x_{21}^{(2)} & x_{22}^{(2)} \end{bmatrix}^T$, $\mathbf{v}_{(\mu-\gamma)\gamma} = \begin{bmatrix} \mu-\gamma & \gamma & 0 & 0 \\ 0 & 0 & \mu-\gamma & \gamma \end{bmatrix}$,
and $\mathbf{v}_{(\mu-\gamma)\gamma} = \mathbf{H}^T\mathbf{v}_{(\mu-\gamma)\gamma}^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{H}$.

In this paper, we consider the set $\mathbf{\Phi} = \{(\mu-\gamma, \gamma)\,|\gamma = p - q; \mu = \alpha - \beta; 1 \leq p \leq \alpha; 1 \leq \beta \leq \alpha \leq l; 1 \leq q \leq \beta; \{\alpha, p\} \neq \{\beta, q\}\}$. Consequently, the optimization problem can be formed as

$$
\begin{aligned}
&\mathcal{P}{-}1: \quad \max \min \mathbf{u}^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{u}; (\mu-\gamma, \gamma) \in \mathbf{\Phi} \\
&\text{s.t.:} \quad P_{\mathcal{X}} \leq P_{\max}
\end{aligned}
$$

and can be reformed as follows

$$
\begin{aligned}
&\mathcal{P}{-}2: \quad \max p \\
&\text{s.t.:} \quad \mathbf{u}^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{u} \geq p; (\mu-\gamma, \gamma) \in \mathbf{\Phi} \\
&\qquad\quad P_{\mathcal{X}} \leq P_{\max}
\end{aligned}
$$

Still, to determine the global solutions of $\mathcal{P}{-}2$ is challenging because $\mathcal{P}{-}2$ is a non-convex optimization problem. Accordingly, the convex relaxation strategy can be engaged in order to approximate $\mathcal{P}{-}2$. More specifically, by linearizing $\mathbf{u}^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{u}$ in $\mathbf{u}_k$, the relaxed optimization problem can be presented as

$$
\begin{aligned}
&\mathcal{P}{-}3: \quad \max p \\
&\text{s.t.:} \quad 2\mathbf{u}_k^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{u} - \mathbf{u}_k^T\mathbf{v}_{(\mu-\gamma)\gamma}\mathbf{u}_k \geq p; (\mu-\gamma, \gamma) \in \mathbf{\Phi} \\
&\qquad\quad P_{\mathcal{X}} \leq P_{\max}
\end{aligned}
$$

Furthermore, because $\mathcal{P}{-}3$ surely is a convex optimization problem, the optimization package helpful in solving convex optimization problems can be engaged to solve (P3). Besides, the initial primary constellation points are assumed as $\mathbf{x}_{11} = (0,0)^T, \mathbf{x}_{21} = (0.1, 0.1)^T, \mathbf{x}_{22} = (0.2, 0.2)^T$. Moreover, to efficiently overcome and lessen the complexity cost of solving the design optimization problem, we try to reduce the total number of constraints by defining the set $\mathbf{\Phi}$ as $\mathbf{\Phi} = \{(0,1); (1,0); (a,-b) | 1 \leq a, b \leq (l-1)\,; \gcd(a,b) = 1\}$. Moreover, to calculate the greatest common divisor between a and b, we denote a function as gcd(a,b). With the M"obius function $\mu(k)$, the cardinality of the constrain set is $|\mathbf{\Phi}| = 2 + \sum_{k=1}^{l-1} \mu(k) \left\lfloor \frac{l-1}{k} \right\rfloor^2$.

4 Generalized Constellation Design for MIMO-VLC systems

Because of the restrictions of this study, the constellations designing problem is chiefly centered on 2×2 MIMO VLC systems. This limitation was mentioned in [10, 11] because of the mathematical complexity and large volume of numerical contents. Therefore, we shortly illustrate the idea in designing the ECC for larger numbers of LEDs and PDs. Perceive that optimization can be acquired, and the equivalent design rule would be achieved without any problem. Nevertheless, the main restriction in getting the optimal points for a larger numbers of LEDs and PDs rests in huge complexities of the optimization-problem solving when high quantities of variables and constraints are examined.

Consequently, in here a MIMO-VLC system with transmitter constellation C of $M = 2^r$ symbols is considered and can be denoted as $C = [\mathbf{a}_1, ..., \mathbf{a}_M]$ where $\mathbf{a}_i = \left[a_i^{(1)}, a_i^{(2)}, ..., a_i^{(N_T)}\right]^T \in C$ is a signal vector that can be conveyed through multiple LEDs. With n_t transmitters, we define a multidimensional layered constellation that each N_i-th-dimension (N_i-D) layer comprises all ($N_i - 1$)-D layers; for example, with $n_t = n_r = 2$, an 1D-layer can belong to a line, as in Fig.1; with $n_t = n_r = 3$, a 2D-layer consists of all 1D-layers that lie in it, and every 1D-layers is a line that can be similarly constructed as in $n_t = n_r = 2$; and so on the proposed constellation for any

$n_t = n_r$ can be built by generating all constellation-points belong to (n_t-1)D-layers. Similar to the 2×2 case, a constellation-point can be represented as $\mathbf{x}_{\alpha_1\alpha_2...\alpha_{n_t}} = \left(x^{(1)}_{\alpha_1\alpha_2...\alpha_{n_t}}, x^{(2)}_{\alpha_1\alpha_2...\alpha_{n_t}}, ...x^{(n_t)}_{\alpha_1\alpha_2...\alpha_{n_t}}\right)^T$ where the signal transmitted by the p-th LED will be rewritten as

$$x^{(i)}_{\alpha_1\alpha_2...\alpha_{n_t}} = (\alpha_1 - \alpha_2)x^{(i)}_{211...1} + ... + (\alpha_{n_t-1} - \alpha_{n_t}) x^{(i)}_{222...2} + (\alpha_2 - 1)\, x^{(i)}_{211...1} + ... + (\alpha_{n_t} - 1)x^{(i)}_{222...2} \quad (7)$$

where the primary points of the constellation here are $\mathbf{x}_{11...1} = (0, 0, ..., 0)^T$, $\mathbf{x}_{211...1} = \left(x^{(1)}_{211...1}, x^{(2)}_{211...1}, ..., x^{(n_t)}_{211...1}\right)^T$, ..., $\mathbf{x}_{222...2} = \left(x^{(1)}_{222...2}, x^{(2)}_{222...2}, ..., x^{(n_t)}_{222...2}\right)^T$ and $1 \le \alpha_{n_t} \le \alpha_{n_t-1} \le ... \le \alpha_2 \le \alpha_1 \le L$. Moreover, there will be $l!/n_t!$ points in the constellation. Then, with any transmitted constellation $\mathbf{x}_{\alpha_1\alpha_2...\alpha_{n_t}} \in \mathcal{X}$, $\mathcal{X} = [\mathbf{x}_{111...1}, \mathbf{x}_{211...1}, ..., \mathbf{x}_{n_tn_tn_t...n_t}]$, we define the corresponding received constellation $\mathcal{Y}$ at the n_r PDs by vector $\mathbf{y}_{\alpha_1\alpha_2...\alpha_{n_t}}$ using equation (4) and $|\mathcal{Y}| = |\mathcal{X}|$. It is observed that the constellation when $n_t = n_r = 2$ can be consider as a particular case. Even though not detailedly considered in this article, the constellation C o higher dimensions will be produced from a collection of primary points $\{\mathbf{x}_{211...1}, \mathbf{x}_{221...1}, ..., \mathbf{x}_{222...2}\}$ by a generalization of the algorithm in previous section. Following that, the optimization problem of the constellation forming can be readily determined in an equivalent way similar to $\mathcal{P}-0$**.** The method of simplifying and solving the optimization problem is conceivable by orderly extending the method outlined in this article without any mathematical challenge. Nonetheless, this article focuses on the 2×2 VLC-MIMO scheme as earlier studies [10, 11] and the expansion to the system with a higher numbers of PDs and LEDs will be presented in later study.

5 Simulation results

In this sector, the performances of the proposed ECC are presented in comparison with various conventional constellation design techniques in MIMO-VLC. Moreover, the CC and also the CASCC designs are utilized in systems with two LEDs and two PDs. Furthermore, $A_r = 1\text{cm}^2$, $\Phi_{1/2} = 60^o$, and $\psi_{1/2} = 60^o$. Besides, the full transmitting power value is the same between the constellations to ensure between the compared schemes the fairness. As mentioned in [10, 11], SNR value can be determined with the normalized channel matrix $\mathbf{H}$ and $\mathbf{x}$. We generate a channel realization that

$$\mathbf{H} = \begin{bmatrix} 1.0000 & 0.8000 \\ 0.2000 & 0.3000 \end{bmatrix}$$

is the channel matrix.

5.1 BER Performance With Perfect CSI

Initially, the BER performances of CC, ECC, and CASCC are shown in Fig. 2 for a low bit rate and in Fig. 3 for a high bit rate. More particularly, in the simulations, we fixed the values of the transmission rate r to 4, 8, 10, and 12 bits. It is perceived that when related to different designs such as CC and CASCC, ECC presents more beneficial BER performances. An SNR gain of around 2dB in relation to CASCC can be achieved in ECC. Nonetheless, ECC and CASCC produce remarkable performance gains through employing the spatial source. Besides, by considering the influence of channel coefficients in the optimization problem, the offered ECC and CASCC exceptionally relieve the association among LEDs. Moreover, ECC seeks to lessen the influence of channel correlation by further efficiently enhancing the Euclidean distances among symbols in the constellation under the equivalent power constraint in corresponding with CASCC. Hence, the offered ECC in overall can obtain more excellent performance for MIMO-VLC in comparison among different constellations.

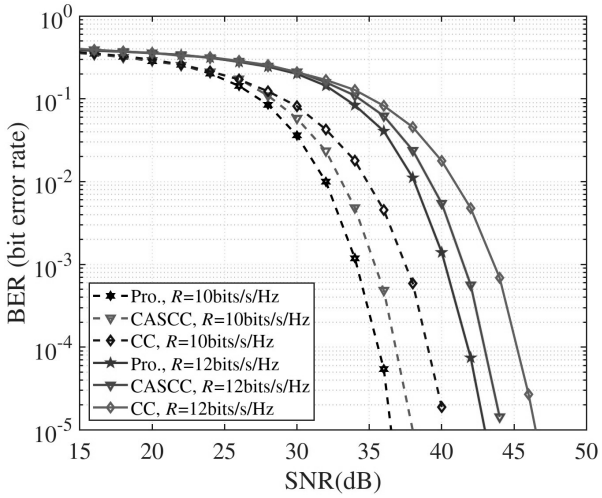


Figure 2: BER performances comparison where r = 4, 8 bits.

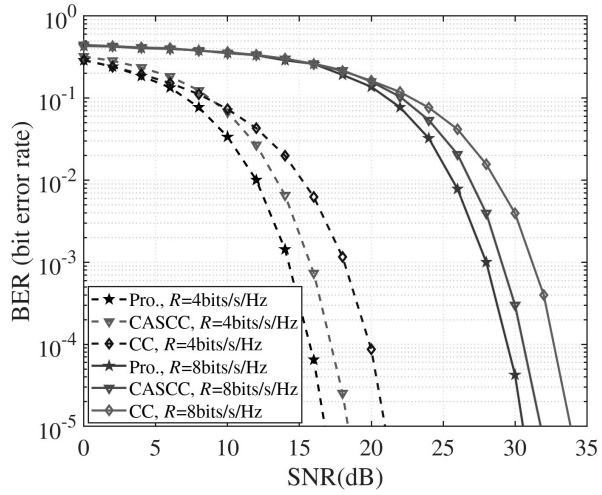


Figure 3: BER performances comparison where r = 10, 12 bits.

5.2 BER Performance With Imperfect CSI

Meanwhile, the channels in VLC for any particular transmitter and receiver fixed positions or specifications are deemed deterministic. In fact, the hypothesis of ideal CSI at the transmitter side is generally not vigorously practical for the indoor VLC environment [12]. To favorably obtain the conveyed symbol with a necessary degree of certainty, the information of each channel coefficient at received PDs would be of significant concern. Consequently, to assess and

produce insights on the impairment of incomplete CSI on MIMO-VLC system performance by the introduced ECC, in this section, suitable methods to imitate the channel estimating imperfection required to be carried out. Without the loss of generalization, the channel matrix evaluated at the receiver can be expressed as

$$\mathbf{H}_{\epsilon} = \mathbf{H} + \boldsymbol{\epsilon} \tag{8}$$

where $\mathbf{H}_{\epsilon}$ is the estimated version of $\mathbf{H}$ with the estimation error values expressed in a matrix $\boldsymbol{\epsilon}$ of dimensions $n_{\mathrm{r}} \times n_{\mathrm{t}}$. The channel estimation error $\boldsymbol{\epsilon}$ which is independent of $\mathbf{H}$ and elements follow $\mathcal{N}\left(0, \sigma_{\epsilon}^2\right)$.

In later simulations, we illustrate the situation at r=4 bits in channel $\mathbf{H}_2$. To set the overall impact of the estimating error on the deploying of systems, the power of the estimating error σ_{ϵ}^2 was set to 0.05 value for every SNR values, i.e., approximating 6.25% of the channel gain power. Particularly with ECC and CASCC that take the incomplete CSI to the design approach. Furthermore, regard that the estimation error will likewise worsen the performance of the MIMO-VLC system that engages all modulation as discussed designs such as CC, ECC, and CASCC; because ML detector is employed. Fig.4 presents the performance comparison of ECC with different designs under profoundly CSI error where $\sigma_{\epsilon}^2 = 0.05$.

It is clear that because of the ineffective utilization of energy that causes such small Euclidean distances in constellation-points, the performances of conventional schemes become worse. Consequently, while channel estimation errors are significant, both of them operate inadequately. In contrast, ECC notwithstanding outperforms various schemes.

Moreover, since the inadequate CSI can directly influence the design approach of both ECC and CASCC, we can easily obtain a thorough glimpse of those two designs in some incomplete CSI conditions in Fig.4. According to Fig.4, ECC can obtain at least 2dB of SNR gain in comparison with CASCC in moderate to great CSI imperfection. It is also deserving of noticing that the performance of both designs quickly deteriorates when CSI error rising.

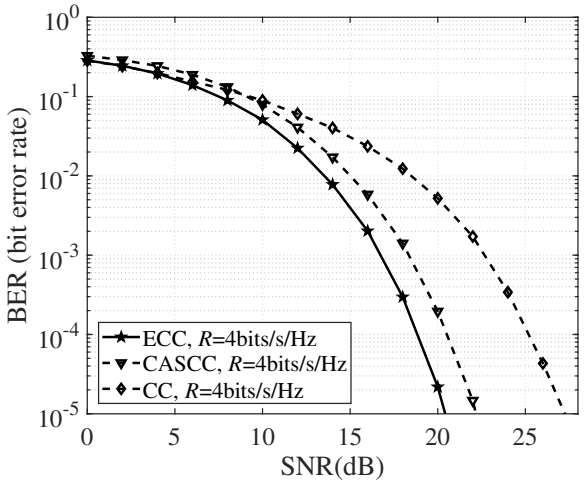


Figure 4: BER comparisons at r = 4 bits for incomplete CSI where $\sigma_{\epsilon}^2 = 0.05$.

6 Conclusion

In this study, a designing procedure concerning a constellation, particularly ECC, was offered for MIMO-VLC schemes. More specifically, the proposed scheme effectively exploits the layered structure of collaborative constellation and overcomes the aforementioned drawbacks of CASCC. By using simulations validation, the introduced method can give constellation, which achieves better BER results in comparison with several well-known constellation designs for VLC systems.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the Research Program through the National Research Foundation of Korea (NRF-2019R1A2C1005920).

References

[1] A. Jovicic, J. Li, T. Richardson, "Visible light communication: Opportunities, challenges and the path to market", IEEE Commun. Mag., 51(12), 26-32, Dec. 2013. https://doi.org/10/gfxf7x

[2] S. Dimitrov, H. Haas, Principles of LED Light Communications, en. Cambridge University Press, Mar. 2015. ISBN:978–1–107–04942–0

[3] M. L. Tran, S. Kim, "Novel bit mapping for generalized spatial modulation in VLC systems", IEEE Photonics Technol. Lett., 31(15) 1257-1260, Aug. 2019. https://doi.org/10/gf4gp4

[4] M. Le Tran, S. Kim, T. Ketseoglou, E. Ayanoglu, "LED Selection and MAP Detection for Generalized LED Index Modulation", IEEE Photonics Technol. Lett., 12(8), 1254-1260, Aug. 2018. https://doi.org/10/gfr979

[5] M. Simon, V. Vilnrotter, "Alamouti-type space-time coding for freespace optical communication with direct detection", IEEE Trans. Wirel. Commun., 4(1), 35-39, Jan. 2005. https://doi.org/10/d6hvvk

[6] M. L. Tran, S. Kim, "Joint power allocation and orientation for uniform illuminance in indoor visible light communication", Opt. Express, OE, 27(20), 28575-28587, Sep. 2019. https://doi.org/10.1364/OE.27.028575

[7] L. An, H. Shen, J. Wang, Y. Zeng, R. Ran, "Energy Eciency Optimization for MIMO Visible Light Communication Systems", IEEE Wirel. Commun. Lett.,12(10) 1322-1325, 2019. https://doi.org/10.1109/LWC.2019.2958802

[8] Q. Wang, Z. Wang, L. Dai, "Multiuser MIMO-OFDM for Visible Light Communications", IEEE Photonics J., 7(6), 1-11, Dec. 2015. https://doi.org/10/gfxdm6

[9] M. L. Tran, S. Kim, "Layered adaptive collaborative constellation for MIMO visible light communication", IEEE Access, 6, 74895-74907, 2018. https://doi.org/10.1109/access.2018.2883346

[10] Yi-Jun Zhu, Wang-Feng Liang, Jian-Kang Zhang, Yan-Yu Zhang, "Space-collaborative constellation designs for MIMO indoor visible light communications", IEEE Photonics Technol Lett, 27(15), 1667-1670, Aug. 2015. https://doi.org/10/gfxdpb

[11] K. Xu, H.-Y. Yu, Y.-J. Zhu, H.-B. Cai, "Channel-adaptive space-collaborative constellation design for MIMO VLC with fast maximum likelihood detection", IEEE Access, 5, 842-852, Jan. 2017. https://doi.org/10/gfxdj9

[12] C. Zhang, H. Du, Z. Wu, "Robust Signal Recovery for MIMO VLC System with Incomplete Channel", IEEE Commun. Lett., 16(8), 3327–3331, Jan. 2019. https://doi.org/10.1109/LCOMM.2019.2957115

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 262-268 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Novel Demand Side Management by Minimizing Cost Deviation

Vikas Anand Vatul[1], Arputha Aravinth [1], Narayanan K[*,1], Gulshan Sharma[2], Tomonobu Senjyu[3]

[1]*EEE Department, SASTRA Deemed University, Thanjavur, Tamil Nadu, India*

[2]*Department of Electrical Power Engineering, Durban University of Technology, South Africa*

[3]*Power and Energy System Control Laboratory, University of the Ryukyus, Japan*

ARTICLE INFO

Article history:
Received: 20 July, 2020
Accepted: 01 September, 2020
Online: 17 September, 2020

Keywords:
Power Systems
Energy System
Demand Side Management
Aggregated Energy
Pricing Model
Smart Grid
Renewable Energy

ABSTRACT

In the recent times power shortage has been a major setback to deal for the effective operation of power systems. Bridging the gap between generation and demand is known as Demand Side Management (DSM). For an effective DSM strategy to be implemented, it is crucial that both utility and customers be involved. By DSM, the energy generated is used more effectively. This reduces the burden of the utility to invest on additional generation. In this work, a DSM strategy has been performed on two systems: (i) on RTS 24 bus system with wind energy sources distributed at some nodes of the system (ii) on an institutional load with installed solar power plant. A generic DSM strategy to effectively utilize the generated energy and to minimize the utility bills for the customer has been proposed. An instantaneous billing scheme has been proposed. By implementing the instantaneous billing scheme, customers can be persuaded to change their consumption behavior, matching the demand with available generation. The results obtained are promising, with a resulting flat load profile and reduced utility bills for the customer.

1. Introduction

Demand Side Management (DSM) has been in practice from the 20th century. Initially, DSM was implemented by replacing older equipment with a newer one for improved efficiency. DSM effectively reduces the gap between supply and demand with the help of various services. DSM is performed by suggesting adjustment in the electricity consumption of the customer to produce desired changes in the power distribution system. Some DSM strategies used to alter customers' load profiles involve load shedding, valley filling, peak clipping, flexible load shape, strategic load growth, strategic conservation and load shifting [1]. The final load profile for each system depends on the operational requirement of the system. Performing DSM has the same objective to minimise customers' utility bills and to reduce Peak to Average Ratio (PAR) by taking up different pricing models in many researches. By performing DSM, electrical infrastructure can be utilised effectively, and new investments on the same can be deferred. Price billing schemes are an important aspect of performing DSM. In recent times, instantaneous load billing scheme has been used to perform DSM. By implementing this billing scheme, the price at each hour can be changed depending on the aggregated load at that hour.

A quadratic cost function has been used with the motive of reducing the total cost and minimize PAR with a simple billing mechanism for the subsequent period is proposed in [2]. This does not consider peak classification during the day, making it unfair to the customer by billing them similar at all times of the day based only on energy consumption. Some pricing models have been introduced and mapping between retail users and wholesale electric prices is portrayed in [3]. The motive was to improve utilization of cumulative generated energy and reduce energy cost by wholesale mapping of energy requirement [3]. A polynomial cost function with time-dependent coefficients has been used in [4]. The focus was to find the aggregated energy consumption and the selfish customers would modify their energy consumption in order to reduce their energy prices. The time slots have been classified in [4] as on peak, mid-peak and off-peak. Coefficients are time-dependent and are assigned based on the peak

[*]Corresponding Author Narayanan. K, EEE Department, SASTRA Deemed University, Thanjavur, Tamil Nadu, India. Mob: +91-9790258910 Email: narayanan.mnit@gmail.com

classification. The coefficients are based on the currency used in the billing.

A system with industrial, commercial and residential loads have been considered to perform DSM in [5] and hourly costs were paralleled with the costs in [6]. In [5], a logarithmic pricing model was used to compute the energy consumption price and a load shifting based on average cost was devised to perform DSM. The hourly prices were lower when compared to the strategy by [6] which uses a heuristic algorithm to minimise utility costs. [2]-[6] are pivotal to this work as a mathematical pricing model for energy consumption is a significant element for performing DSM.

DSM using strategic conservation is portrayed in [7]. In the study, some customers with the presence of Distributed Generators (DGs) were considered. The key highlights of the study were bi-directional energy flow from customer to grid, incentives and balancing, mass-produced renewable energy. In [8], discharge of Plug-in Hybrid Electric Vehicle (PHEV) during peak demand was used for DSM. Since different utilities use different pricing methodologies, the concept of cost efficiency as a metric to determine the consumption efficiency is explained in [9]. Home Load Management Systems (HLM) used to achieve residential load management, without focusing on electricity pricing has been discussed in [10]. The concept of coordinated demand response of residential customers in a smart grid has been highlighted. A price model as a ratio of total connected load at that instant to the total aggregated load through all time slots is proposed in [11]. It emphasizes on analysing a realistic situation by considering realistic consumer behaviour to determine the participation of the consumer in DSM. Though the work mentions to account for the realistic situation, deviation of load demand from the forecasted data has been over looked while performing DSM.

A standalone, Hybrid energy source with a combination of wind and solar energy has been designed to electrify a village in India in the study performed in [12]. The design aspects included solar irradiation and wind speed and the objective was to optimize the size of the energy source without affecting the performance of the same. In [13], planning of distributed sources such as distributed generators and shunt capacitors to achieve objectives on technical, economical and societal levels are discussed. Expansion of a micro-grid due to increasing load demand are discussed in [14] with wind turbines, Solar energy source, and batteries. The expansion is carried out in an isolated micro-grid and load forecasting with uncertainties are considered in the research.

In this work, two systems are considered: (i) IEEE 24 bus Reliability Test System (RTS 24) with integrated Windmills [15] and (ii) an institutional energy consumption model with integrated solar panels. Any deviation in the load demand from the forecasted base load is taken into account in the proposed pricing model. This makes the proposed pricing model realistic and practical since load demands are dynamic and seldom match with the forecasts. A polynomial cost function in which coefficients are based on off-peak, mid-peak and on-peak hours from a previously referenced research is also considered and analysed. Load shifting algorithm has been developed to reduce the deviation of hourly load demand from the fixed base load. This has proved to reduce consumers aggregated energy cost for the day and to acquire a flat load profile. This will also reduce PAR and enhance the efficiency of the power system.

2. System Model

Two systems were considered to apply the proposed method. It has been assumed that all the consumers were equipped with HLMs (Home load management systems) or smart meters. IEEE 24-bus reliability test system (RTS 24) with wind energy source of 200MW at six different points in the grid is considered as system 1. Hourly load demand for System 1 is acquired from [15] and the wind scenarios are acquired from [16]. An institution in which a connected load of 4000kW and decentralised solar power plant of an installed capacity of 1.25MW is considered as system 2. Since, an institutional load is considered, all loads are assumed to be shift able during any period of the day.

The RTS – 24 bus system has been considered since it has installed wind energy, which is operational for the entire 24 hours of the day. Since load considered in the RTS 24 bus system is more flexible in contrast to the institutional load considered, it offered a better scope for analysis.

Assuming load demand for future 'm' operational periods has been forecasted to perform Demand Side Management. The consumption of energy for 'm' time slots can be formulated for any random customer 'j' as

$$p_j = (p_j^1, \ldots\ldots, p_j^m) \tag{1}$$

where p_j^m is the consumption of energy at the mth hour by customer 'j'. The energy profile of the consumer is subject to a maximum and minimum constraint.

$$p_{min}^m < p^m < p_{max}^m \tag{2}$$

where p_{max}^m is the maximum energy consumption limit at hour 'm' and p_{min}^m is the minimum energy consumption limit at hour 'm'.

Energy generation from the renewable energy source is also assumed to be forecasted beforehand to perform DSM. The renewable energy output can be formulated as

$$g_j = (g_j^1, \ldots\ldots, g_j^m) \tag{3}$$

where g_j^m is the energy output from the installed renewable source at the mth hour by customer 'j'. It is assumed that the solar plant installed in System 2 can generate energy only during the day time.

3. Problem Formulation

The aim of the study is to minimise energy consumption bills of the consumer without affecting their routine in terms of energy usage.

The suggested DSM considers systems with installed renewable sources. By considering generation from renewable sources, the effective demand from utility reduces. Thus, the cost will also reduce in this case. Energy consumption from the utility can be mathematically computed by

$$E_j^m = p_j^m - g_j^m \quad (4)$$

Where E_j^m is the energy requirement from the utility for customer 'j' at the mth hour.

A pricing model has been proposed and is compared with a realistic polynomial pricing model [4]. The pricing models have been incorporated in the proposed Demand Side Management. The polynomial cost function [4] used was

$$Ch(Lh) = a(Lh^b) + c \quad (5)$$

where Lh is the hourly load in kWh or MWh, price parameters are given as

$$a > 0\ ;\ b \geq 1\ ;\ c \geq 0 \quad (6)$$

4. Proposed Price Model

The suggested pricing model is based on the fixed base load of the system for the particular day and deviation of hourly energy consumption from the base load at each time period. The suggested pricing model can be formulated as

$$\gamma(h) = x \times (\beta) + y \times (\delta) \quad (7)$$

$$\delta = p^m - \beta \quad (8)$$

$$x, y > 0 \quad (9)$$

where p^m is the energy consumption of the consumer at the mth hour and 'x' & 'y' are the cost coefficients. 'y' is a time dependent factor and is assigned based on peak timings. 'δ' is the deviation of hourly energy consumption from the base load 'β'. This cost function is more appropriate for DSM, as utility bills of the consumers can be minimised by offsetting the deviation of the energy consumption at that time slot by different methods and services. This price model has been developed from the polynomial cost function and indicates energy consumption prices in cents per kWh or MWh used.

The proposed pricing model possesses important features that are essential for an energy pricing model to be used in DSM. It is a convex cost function; it rapidly increases with an increase in hourly energy consumption.

A generic DSM algorithm has been proposed for any system with forecasted energy generation from the renewable energy source, forecasted load demand at each time slot and a base load of the system.

5. Proposed DSM and Load Shifting Model

Proposed DSM includes energy generation from renewable source installed in the system. By doing so, energy consumption of the consumer from the utility will reduce. This will reflect on the utility bills of the customer for the same energy requirement. The above-discussed cost functions are incorporated to evaluate the utility bills for the customers against their energy usage. A load shifting algorithm has been developed to minimise the hourly deviations at certain time slots. Flowchart for the proposed DSM is shown in Figure 1.

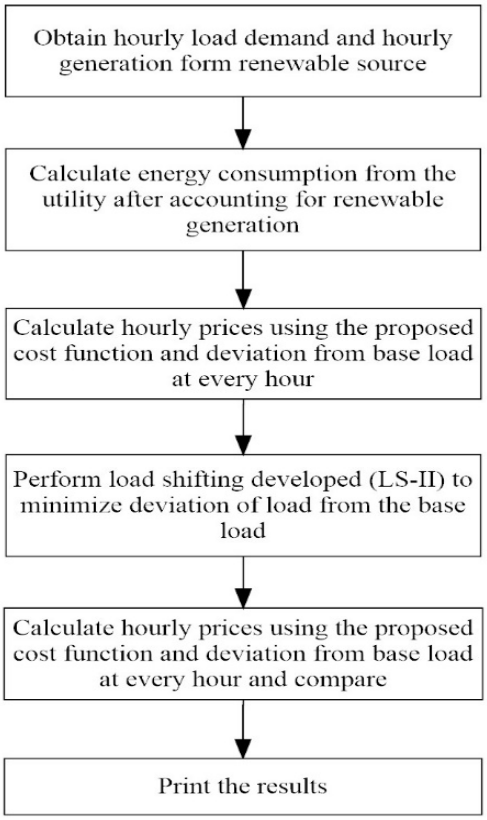


Figure 1: Proposed DSM

A load shifting method is developed to offset the deviation of the load demand from the fixed base load in the time slot under consideration. Load shifting is one among many DSM strategies, where loads from certain time slots of the day are shifted to other time slots to reduce peak demand without creating a new peak at a different time slot. Controllable and non-controllable (critical loads) loads are the two types of loads present in a system. Controllable loads are loads that can be shifted without any constraint and non-controllable loads are those that cannot be shifted [1].

Peaks timings are considered from [4] and each time in a day has been classified as On-peak, Mid-peak, and Off-peak based on the load forecast at that hour. The developed load shifting algorithm aggregates the loads exceeding the base load at each time slot within the peak timings and equally spreads the loads within the peak classifications.

Time slots at on-peak, mid-peak and off-peak is taken as m1, m2, m3 respectively. Fixed base loads at on peak, mid-peak and off-peak are assumed as B1, B2, B3 respectively. The aggregated load at on peak, mid-peak and off-peak periods are given as

$$S1 = \sum[p(m) - B1] \quad (10)$$
$$S2 = \sum[p(m) - B2] \quad (11)$$
$$S3 = \sum[p(m) - B3] \quad (12)$$

where S1, S2, S3 are the aggregated load that exceeds the base load in each peak period and $p(m)$ is the load demand of the consumer in m^{th} hour.

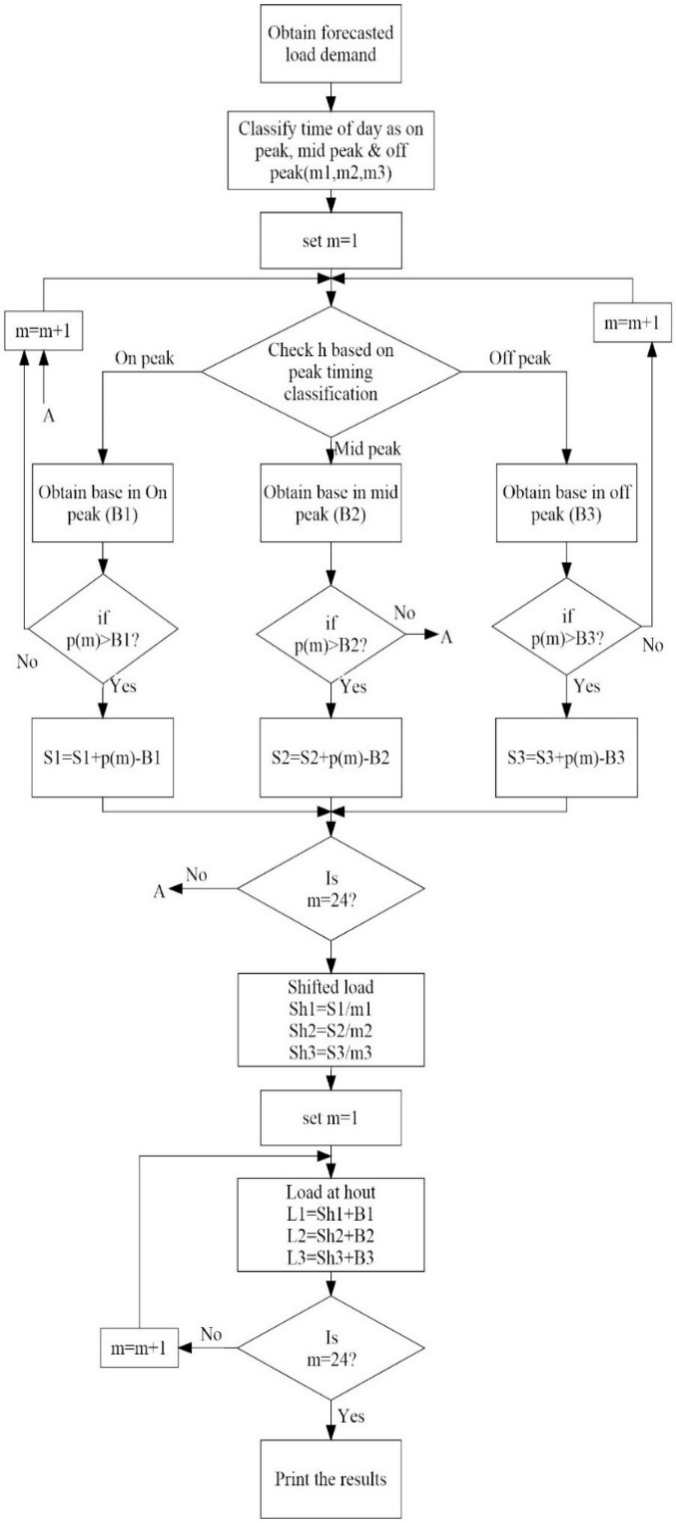


Figure 2: Load Shifting Algorithm

Load to be shifted from on-peak, mid-peak and off-peak can be formulated as

$$Sh1 = S1/m1 \quad (13)$$
$$Sh2 = S2/m2 \quad (14)$$
$$Sh3 = S3/m3 \quad (15)$$

where $Sh1, Sh2\ and\ Sh3$ depict the loads that needs to be shifted within on-peak, mid-peak and off-peak hours respectively. The load demand of the consumer at each peak period after performing the shifting can be formulated as

$$L1 = Sh1 + B1 \quad (16)$$
$$L2 = Sh2 + B2 \quad (17)$$
$$L3 = Sh3 + B3 \quad (18)$$

here, $L1, L2\ and\ L3$ represent the load demand at on-peak, mid-peak and off-peak hours after performing load shifting.

The developed load shifting algorithm is represented as a flowchart below in Figure 2.

6. Numerical Results

Numerical results of the proposed DSM are presented and are compared with polynomial pricing model [4]. The entire time frame under consideration has been classified into on peak, mid-peak and off-peak hours. 'a' in the polynomial cost function and 'y' in the proposed cost function is set as 0.003 from 12 Midnight to 7AM (Off-peak), as 0.004 from 7AM to 4PM & 10PM to 12 Midnight (Mid-peak) and 0.005 from 4PM to 10PM (On-peak), 'b' and 'c' are set as 1 and 0 respectively for all M∈m.

The proposed DSM has been carried out on System 1 (RTS 24). System data [15], Load management including the renewable generation and load shifting is shown in Figure 3.

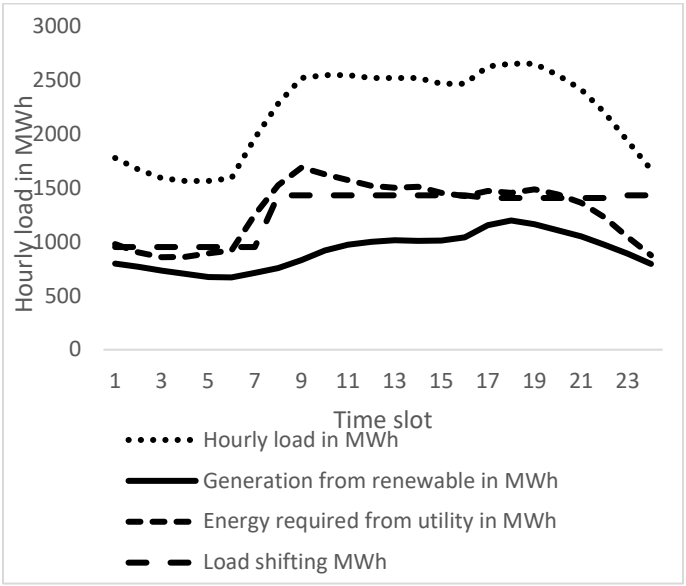


Figure 3: Proposed DSM on System 1

For System 2, peak timing has been classified in a slightly different manner taking into account the consumption pattern of the institutional load. Off-peak from 12 Midnight to 6 AM & 10PM to 12 Midnight; mid-peak from 9AM to 5PM and on-peak from 6AM to 9AM & 5PM to 10PM.

The proposed DSM was carried out on System 2. System data and Load management after including energy generation from the renewable source and load shifting are shown in Figure 4.

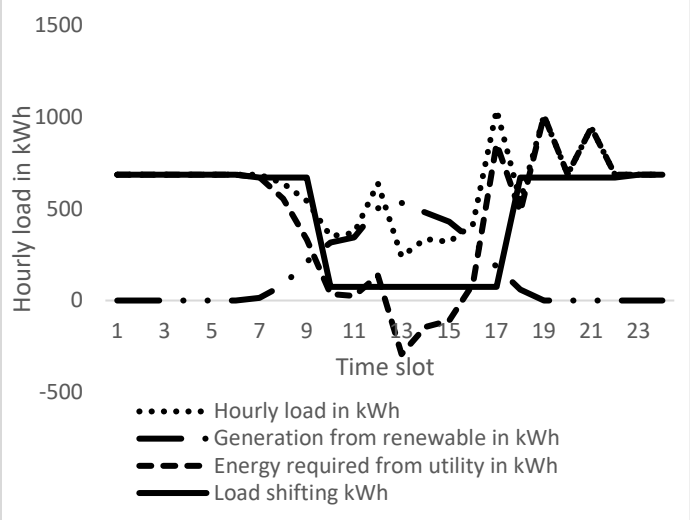


Figure 4: Proposed DSM on System 2

Hourly energy prices for System 1 and System 2 are calculated using the proposed pricing model and compared with the polynomial cost function [4]. Energy prices at each time slot for System 1 before and after load shifting, by polynomial and proposed cost functions are shown in Table 1.

Table 1: Hourly energy prices by polynomial and proposed cost function on System 1

Time slots	Hourly load in MWh	Polynomial cost function [4]		Proposed cost function	
		Before load shifting	After load shifting	Before load shifting	After load shifting
1	978.6	2.9358	2.8543	6.0634	4.7572
2	901.2	2.7036	2.8543	5.8312	4.7572
3	857.2	2.5717	2.8543	5.6993	4.7572
4	860.2	2.5807	2.8543	5.7083	4.7572
5	891.6	2.6748	2.8543	5.8024	4.7572
6	920.4	2.7612	2.8543	5.8888	4.7572
7	1250.8	3.7524	2.8543	6.88	4.7572
8	1524.4	6.0976	5.7223	7.6614	6.6738
9	1688.1	6.7525	5.7223	8.3163	6.6738
10	1625	6.4999	5.7223	8.0637	6.6738
11	1572.8	6.2911	5.7223	7.8549	6.6738
12	1519.6	6.0783	5.7223	7.6421	6.6738
13	1502.3	6.0092	5.7223	7.573	6.6738
14	1508.3	6.0332	5.7223	7.597	6.6738
15	1452.2	5.8087	5.7223	7.3725	6.6738
16	1424.1	5.6963	5.7223	7.2601	6.6738
17	1470.3	7.3516	7.0313	7.3516	7.0313
18	1453	7.2648	7.0313	7.2648	7.0313
19	1486.7	7.4337	7.0313	7.4337	7.0313
20	1437.5	7.1877	7.0313	7.1877	7.0313
21	1362.1	6.8107	7.0313	6.8107	7.0313
22	1227.9	6.1396	7.0313	6.1396	7.0313
23	1045	4.18	5.7223	5.7438	6.6738
24	874.8	3.499	5.7223	5.0628	6.6738
Total:	30834.1	125.1141	125.1132	164.2091	148.9

It can be observed from Table 1 that with the given cost function using the proposed DSM, hourly costs during on-peak periods are scaled to a price highest of the three followed by mid-peak and off-peak periods. This is detected because of the development of coefficients in the proposed cost function pertaining to peak timings. These coefficients influence the cost of electric power during a given time of the day.

Energy prices at each time slot for System 2 before and after applying the load shifting algorithm are given in Table 2.

Table 2: Hourly energy prices by polynomial and proposed cost function on System 2

Time slots	Hourly load in kWh	Polynomial cost function [4]		Proposed cost function	
		Before load shifting	After load shifting	Before load shifting	After load shifting
1	686.4	2.0592	2.0592	2.5392	2.2082
2	686.4	2.0592	2.0592	2.5392	2.2082
3	686.4	2.0592	2.0592	2.5392	2.2082
4	686.4	2.0592	2.0592	2.5392	2.2082
5	686.4	2.0592	2.0592	2.5392	2.2082
6	686.4	2.0592	2.0592	2.5392	2.2082
7	672.2	4.0335	4.0272	3.7935	3.9527
8	555.1	3.3304	4.0272	3.0904	3.9527
9	331.6	1.9898	4.0272	1.7498	3.9527
10	35.3	0.1412	0.2979	0.3812	0.3724
11	24.4	0.0976	0.2979	0.3376	0.3724
12	136.6	0.5465	0.2979	0.7865	0.3724
13	-292.9	-1.1714	0.2979	-0.9314	0.3724
14	-144.8	-0.5793	0.2979	-0.3393	0.3724
15	-109.6	-0.4382	0.2979	-0.1982	0.3724
16	90	0.3601	0.2979	0.6001	0.3724
17	856.8	3.427	0.2979	3.667	0.3724
18	484.2	2.9054	4.0272	2.6654	3.9527
19	1008	6.048	4.0272	5.808	3.9527
20	688	4.128	4.0272	3.888	3.9527
21	944	5.664	4.0272	5.424	3.9527
22	686.4	4.1184	4.0272	3.8784	3.9527
23	686.4	2.0592	2.0592	2.5392	2.2082
24	686.4	2.0592	2.0592	2.5392	2.2082
Total:	11456.5	51.0746	51.0744	54.9146	52.2664

It can be observed from Table 2 that renewable energy generation from the solar installations is higher than the load demand of 13th, 14th and 15th hour. By performing the proposed load shifting ensures effective usage of the generated renewable energy during all time slots in the day. The shiftable loads during other time slots avail the excess energy beyond the demand in these time slots.

It can be observed from Table 1 and Table 2, by the proposed pricing model, aggregated hourly cost after shifting is lesser than the aggregated hourly cost before shifting. The overall cost has been reduced by reducing the hourly deviation by load shifting at certain time slots.

Cost-benefit of the proposed cost function is depicted in Table 3. It is observed that there is a significant cost benefit with the proposed cost function after shifting during the mid-peak and off-peak hours in both the systems where as there is a meagre change in the aggregated cost on the on-peak hours.

It can be observed from Figure 3 and 4, by the proposed DSM, the energy requirement is not compromised, while it is being met efficiently. By performing the proposed DSM, a flat load profile can be obtained for certain timeslot thus improving Peak to Average Ratio.

It is a superior substitute to previously existing instantaneous energy billing models. The customers will not have to downsize their consumption but have to shift non-essential consumption to reap benefit in the proposed DSM. Participation of customers in DSM would increase, to minimize their deviation from the already predicted fixed base load with the knowledge of the deviation from base load from the HLM.

Table 3: Cost-benefit with the proposed Cost function on system 1 and 2

Cost-Benefit of proposed cost function on System 1		
Peak Classification	Before Load Shifting	After Load Shifting
On Peak	42.1881	42.1878
Mid Peak	80.1476	73.4118
Off Peak	41.8734	33.3004
Cost-Benefit of proposed cost function on System 2		
On Peak	30.2975	31.6216
Mid Peak	4.3035	2.9792
Off Peak	20.3136	17.6656

7. Conclusion

For a successful DSM, pricing model and billing mechanism are crucial. The proposed pricing model based on the deviation of hourly load demand from the fixed base load has been observed to be a better alternative over the other realistic polynomial cost function.

Renewable generation is unforeseeable and, in most cases, influenced by external factors. The generation from the renewable sources can be higher than the system requirements during some periods in a day and by performing load shifting, generation from renewables can be effectively utilised. By adopting the proposed algorithm, the energy requirement of the system from the utility decreases.

The total cost is reduced after load shifting is performed by the proposed cost function. This is due to the fact that by load shifting, the deviation of the hourly load demand from the fixed base load is reduced at certain time slots and increases in some to have a balanced shifting, but generated energy is effectively utilised. Therefore, the aim to reduce consumers' utility bills are met and peak to average ratio improves by incorporating the proposed DSM.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank SASTRA Deemed University for sharing their consumption data and energy generation from their installed solar power plant.

References

[1] B.R. Gupta, 'Generation Of Electric Energy' (S. Chand Limited, 2009)
[2] A.H. Mohsenian-Rad et al., "Autonomous Demand Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid," IEEE Trans. Smart Grid, **1**(3), 320–331, 2010, doi: 10.1109/TSG.2010.2089069
[3] P. Samadi et al., " Advanced Demand Side Management for the Future Smart Grid Using Mechanism Design," IEEE Trans. Smart Grid, **3**(3), 1170–1180, 2012, doi: 10.1109/TSG.2012.2203341
[4] H.H. Chen, "Autonomous Demand Side Management with Instantaneous Load Billing : An Aggregative Game Approach," IEEE Trans. Smart Grid, **5**(4), 1744–1754, 2014, doi: 10.1109/TSG.2014.2311122
[5] B. Saravanan, "DSM in an area consisting of residential, commercial and industrial load in smart grid," Front. Energy, **9**(2), 211–216, 2015, doi: https://doi.org/10.1007/s11708-015-0351-0
[6] T. Logenthiran et al., "Demand side management in smart grid using heuristic optimization," IEEE Trans. Smart Grid, **3**(3), 1244–1252, 2012, doi: 10.1109/TSG.2012.2195686
[7] I. Atzeni et al., "Demand-Side Management via Distributed Energy Generation and Storage Optimization," IEEE Trans. Smart Grid, **4**(2), 1–11, 2012, doi: 10.1109/TSG.2012.2206060
[8] B. Gao et al., "Game-theoretic energy management for residential users with dischargeable plug-in electric vehicles," Energies, **7**(11), 7499–7518, 2014, doi: https://doi.org/10.3390/en7117499
[9] J. Ma et al., "Residential Load Scheduling in Smart Grid: A Cost Efficiency Perspective," IEEE Trans. Smart Grid, **7**(2), 771–784, 2015, doi: 10.1109/TSG.2015.2419818
[10] A. Safdarian, "Optimal Residential Load Management in Smart Grids: A Decentralized Framework," IEEE Trans. Smart Grid, **7**, (4), 1836–1845, 2016, doi: 10.1109/TSG.2015.2459753
[11] Y. Wang et al., "Load Shifting in the Smart Grid: To Participate or Not?," IEEE Trans. Smart Grid, **7**(6), 2604–2614, 2016, doi: 10.1109/TSG.2015.2483522
[12] S. Rangnekar et al., "Sizing and performance analysis of standalone wind-photovoltaic based hybrid energy system using ant colony optimisation," IET Renew. Power Gener., **10**(7), 964–972, 2016, doi: 10.1049/iet-rpg.2015.0394
[13] N. Kanwar et al., "Optimal distributed resource planning for microgrids under uncertain environment," IET Renew. Power Gener., **12**(2), 244–251, 2018, doi: 10.1049/iet-rpg.2017.0085
[14] Z. Wang et al., "Optimal expansion planning of isolated microgrid with renewable energy resources and controllable loads," IET Renew. Power Gener., **11**(7), 931–940, 2017, doi: 10.1049/iet-rpg.2016.0661
[15] C. Ordoudis et al., "An Updated Version of the IEEE RTS 24-Bus System for Electricity Market and Power System Operation Studies," Technical University of Denmark. 2016.
[16] Data for Stochastic Multiperiod Optimal Power Flow Problem', https://sites.google.com/site/datasmopf/, accessed April 2018.

Appendix

System 2 under consideration was of an institution. The system demand was obtained from SASTRA deemed University on a working day. The renewable utilization from the installed solar plant was also obtained for the same day. It has a connected load of 4000kW and decentralised solar power plant with an installed capacity of 1.25MW.

Table 4 System Data for System 2

Hours	Hourly load in kW	Utilization from Solar plant in kW
1	686.4	0
2	686.4	0
3	686.4	0
4	686.4	0
5	686.4	0
6	686.4	0
7	686.4	14.155
8	640	84.94
9	544	212.36
10	352	316.706
11	368	343.592
12	640	503.37
13	240	532.86
14	336	480.83
15	320	429.56
16	416	325.974
17	1040	183.25
18	544	59.7733
19	1008	0
20	688	0
21	944	0
22	686.4	0
23	686.4	0
24	686.4	0

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 3, 364-369 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Trajectory Tracking Control of a DC Motor Exposed to a Replay-Attack

Reda El Abbadi*, Hicham Jamouli

The Laboratory of System Engineering and Decision Law, National School of Applied Sciences Ibn Zohr University Agadir, Morocco.

ARTICLE INFO

Article history:
Received: 24 April, 2020
Accepted: 18 May, 2020
Online: 28 May, 2020

Keywords:
Networked control system
Replay-attack
Data packet dropout
Delay
Linear matrix inequalities

ABSTRACT

This paper investigates the trajectory tracking control (TTC) problem of a networked control system (NCS) against a replay-attack. The impact of data packet dropout and communication delay on the wireless network are taken into account. A new mathematical representation of the NCS under network issues (packet dropout, delay, and replay-attack) is proposed, the resulting closed-loop system is written in the form of an asynchronous dynamical system. Linear matrix inequalities (LMIs) formulation and a cone complementary linearization (CCL) approach are used to calculate the controller gain F_1 and the trajectory tracking gain F_2. Finally, a DC motor simulation with MATLAB is carried out to demonstrate the effectiveness of our approach.

1 Introduction

This article is an extended version of a conference paper presented in 2019 at the International Conference on Control and Fault-Tolerant Systems [1].

Networked Control System (NCS) is a new generation of systems where the control loop is closed through a communication network. The defining feature of an NCS is that system states and control law exchange between the components of the system (sensors, actuators, and controllers) in the form of information packages via a wireless network. The use of wireless network diminishes system wiring, simplifies maintenance and diagnosis, and improves the system agility [2, 3]. Nonetheless, the introduction of the wireless network to control the physical process brought some challenges, such as induced delay, and packet dropout, which harm the stability and reduce the system's performances. The model of the NCS with data packet dropout and network induced delay has been treated in [4–6]. There are also other challenges of using the network as a mean to impart the information between the system components, like malicious intrusions, viruses, and cyber-attackers who always find a way to access the system network and make critical damages. Consequently, reinforcing system security attracts the attention of the specialists. Remarkable papers have studied the security of the network and the cyber-attacks [7].

This article addresses a particular cyber-attack, termed as replay-attack, studied first in [8]. At our best knowledge, the trajectory tracking control (TTC) problem of NCS under a replay-attack based on an accurate mathematical model of the replay-attack is not fully investigated, and this will be the subject of this paper. In our previous work [1, 9, 10], we modeled the NCS under a replay-attack in three different ways. In [10], we used the Markovian jump linear system to model the replay-attack. In [1], we neglected the communication delay, and we assumed that the adversary attacks just the sensor reading, and in [9], we supposed that the adversary attacks the sensor reading and the actuator's inputs simultaneously. This article involves a new contribution compared to the studies mentioned above. In this study we will give a new model of the NCS against a replay-attack, in which we take into account the effect of the delay that exists in the communication channel (sensor-controller).

This extended version deals with the TTC of a DC motor controlled through a wireless network and exposed to a replay-attack. The adversary seeks to destabilize the system by recording secretly the sensor reading and subsequently replayed it to the controller. To protect its anonymity on the network and stay undetectable for the most prolonged period, the adversary appears at different times (randomly), that sounds similar to the Stuxnet cyber-attack [11]. Moreover, we will take into account the packet dropout and the delay in the communication channel (sensor-controller). Nevertheless, the communication channel (controller-actuator) will be supposed perfect, that means 100% of packets have successfully arrived at the actuator from the controller without any delay.

The study will contain four fundamental parts. We will start by

*Corresponding Author Email: reda.elabbadi@edu.uiz.ac.ma

defining the structure of the global system. Then, we will give the model of the network. After that, we will provide our model of the replay-attack. Finally, by using the linear matrix inequalities (LMIs) formulation and the cone complementary linearization (CCL) approach we will calculate the controller gain which helps us to find the TTC gain.

2 Structure of the Global System

The structure of the NCS against a replay-attack is given in Figure1. The sensor measurement $z(k_i)$, which has successfully arrived at the controller (the switch S1 is closed), will be first saved in a buffer. If a measurement $z(k_i)$ is lost during the transmission (the switch S1 is opened), the controller will utilize the measurements that are already stocked in the buffer to calculate the new control law $v(k_i)$. To avoid being detected by the classical detectors, the adversary will apply the attack at various times (randomly). The switch S2 models the replay-attack, if S2 is opened this means there is no attack in the buffer, whereas if S2 is closed this means there is an attack; this switching will harm the stability and the system's performances, which will reflect negatively on the trajectory tracking.

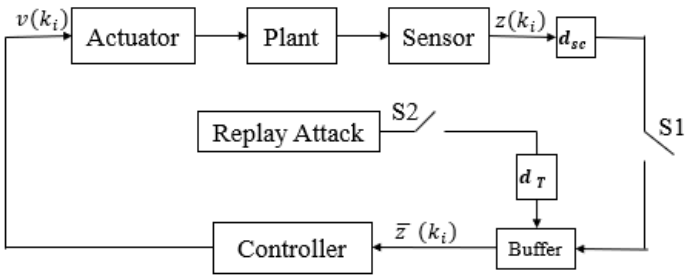


Figure 1: System architecture.

The plant of the system is defined as follows:

$$z(k_{i+1}) = Gz(k_i) + Hv(k_i), \tag{1}$$

where $z(k_i) \in \mathbb{R}^n$ is the system state and $v(k_i) \in \mathbb{R}^m$ is the control law. $G \in \mathbb{R}^{n\times n}$ is the state matrix and $H \in \mathbb{R}^{n\times m}$ is the input matrix. The state feedback controller is:

$$v(k_i) = F_1\bar{z}(k_i), \tag{2}$$

where F_1 is the controller gain, and $\bar{z}(k_i) \in \mathbb{R}^n$ is the controller input which will be defined in the next subsections.

2.1 Wireless network model

The iterative approach described in [12, 13] was adjusted to give a model of the wireless network with the packet dropout and the communication delay d_{sc} in the channel (sensor-controller).

The Figure 2 shows the packets sent from the sensor to the controller. The green packets represent the received packets, whereas the red ones represent the dropouts packets. The notations k_i and $k_i + m$ represent respectively the green packets and the red packets, where , $i \in \mathbb{Z}$ and , $m \in \mathbb{N}^*$.

At first, we will ignore the effect of the replay-attack and we will concentrate on the data packet dropout and the communication delay. Hence, the controller input can be written as:

$$\bar{z}(k_i) = z(k_{i-d_{sc}}). \tag{3}$$

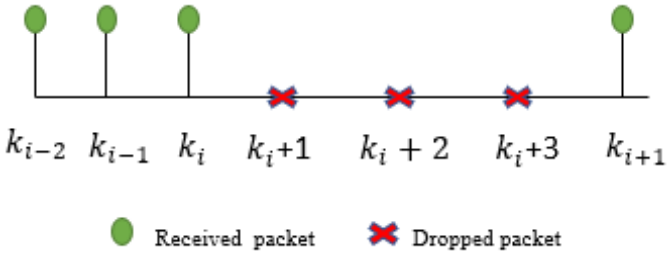


Figure 2: Received and lost packets [12].

As reported by Figure 2, the iterative approach can be defined as:

$$\begin{aligned}
z(k_i+1) &= Gz(k_i) + HF_1z(k_{i-d_{sc}}).\\
z(k_i+2) &= Gz(k_i+1) + HF_1z(k_{i-d_{sc}}),\\
&= G^2z(k_i) + GHF_1z(k_{i-d_{sc}}) + HF_1z(k_{i-d_{sc}}),\\
&= G^2z(k_i) + (GHF_1 + HF_1)z(k_{i-d_{sc}}).\\
z(k_i+3) &= Gz(k_i+2) + HF_1z(k_{i-d_{sc}}),\\
&= G^3z(k_i) + (G^2HF_1 + GHF_1 + HF_1)z(k_{i-d_{sc}}),
\end{aligned}$$

For time instant k_{i+1}, the mathematical model of the system with packet dropout and delay is:

$$z(k_{i+1}) = G^Nz(k_i) + \sum_{j=0}^{N-1} G^jHF_1z(k_{i-d_{sc}}), \tag{4}$$

with N is the number of successive dropped packets.

2.2 Replay-attack model

We assume that an attacker has connected to the buffer can replace the received packet $z(k_i)$ by the previous one $z(k_{i-d_T})$, with d_T is the replay-delay. For example, in the Figure 3 the third packet (102) was exposed to an attack with d_T=2T_e, so it was replaced by the first packet (100). The same for the packet (200) which was replaced by the packet (198).

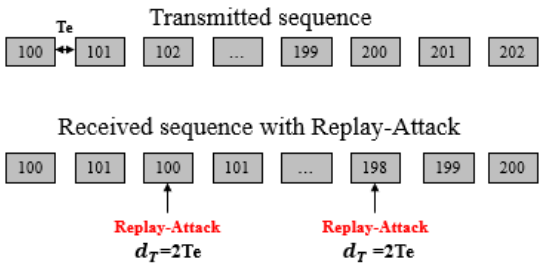


Figure 3: Replay-attack.

As reported by Figure 3, the system under a replay-delay can be defined as:

$$
\begin{aligned}
z(k_i+1) &= Gz(k_i) + HF_1 z(k_{i-d'_T}).\\
z(k_i+2) &= Gz(k_i+1) + HF_1 z(k_{i-d'_T}),\\
&= G^2 z(k_i) + GHF_1 z(k_{i-d'_T}) + HF_1 z(k_{i-d'_T}),\\
&= G^2 z(k_i) + (GHF_1 + HF_1) z(k_{i-d'_T}).\\
z(k_i+3) &= Gz(k_i+2) + HF_1 z(k_{i-d'_T}),\\
&= G^3 z(k_i) + (G^2HF_1 + GHF_1 + HF_1) z(k_{i-d'_T}).
\end{aligned}
$$

For time instant k_{i+1}, the mathematical model of the system against a replay-attack is:

$$z(k_{i+1}) = G^N z(k_i) + \sum_{j=0}^{N-1} G^j HF_1 z(k_{i-d'_T}), \tag{5}$$

where $d'_T = d_T + d_{sc}$.

The overall system will switch between two subsystems. Subsystem 1 if S2 is "off", and subsystem 2 if S2 is "on".
From (4) and (5) the global system becomes:

$$z(k_{i+1}) = G^N z(k_i) + \gamma \sum_{j=0}^{N-1} G^j HF_1 z(k_{i-d_{sc}}) + (1-\gamma) \sum_{j=0}^{N-1} G^j HF_1 z(k_{i-d'_T}), \tag{6}$$

where the variable γ equals one if S2 is "off", and equals zero if S2 is "on".
The augmented state can be written as:

$$\hat{z}(k_i) = [z^T(k_i)\, z^T(k_{i-1}) \cdots\ z^T(k_{i-d_{sc}}) \cdots\ z^T(k_{i-d'_T})]^T. \tag{7}$$

The overall system (6) can be written as:

$$\hat{z}(k_{i+1}) = \Phi_\sigma\, \hat{z}(k_i), \tag{8}$$

in which $\sigma = 1, 2$, and

$$\Phi_\sigma = \begin{bmatrix} G^N & 0 & \cdots & \gamma \sum_{j=0}^{N-1} G^j HF_1 \\ I & 0 & \cdots & 0 \\ 0 & I & 0 & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ & & & \end{bmatrix} \begin{bmatrix} 0 & \cdots & (1-\gamma) \sum_{j=0}^{N-1} G^j HF_1 \\ \vdots & \cdots & 0 \\ \ddots & \ddots & \vdots \\ 0 & I & 0 \end{bmatrix}. \tag{9}$$

Therefore, the overall system can be equivalent to an asynchronous dynamical system expressed in (8),

with

$$\Phi_1 = \begin{bmatrix} G^N & 0 & \cdots & \sum_{j=0}^{N-1} G^j HF_1 \\ I & 0 & \cdots & 0 \\ 0 & I & 0 & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ & & & \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \cdots & 0 \\ \ddots & \ddots & \vdots \\ 0 & I & 0 \end{bmatrix}, \tag{10}$$

and

$$\Phi_2 = \begin{bmatrix} G^N & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ 0 & I & 0 & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ & & & \end{bmatrix} \begin{bmatrix} 0 & \cdots & \sum_{j=0}^{N-1} G^j HF_1 \\ \vdots & \cdots & 0 \\ \ddots & \ddots & \vdots \\ 0 & I & 0 \end{bmatrix}. \tag{11}$$

3 Stability and Control Design

Lemma 1 *[14] The asynchronous dynamical system $z_{k+1} = f_s(z_k), s = 1...N'$, is exponential stable if there exists a Lyapunov function where $\beta_1\|z\|^2 \le V(z) \le \beta_2\|z\|^2$, $\beta_{1,2} > 0$, and given positive scalars α_s satisfying:*

$$V(z_{k+1}) - V(z_k) < (\alpha_s^{-2} - 1)V(z_k), \tag{12}$$

$$\alpha_1^{r_1}\alpha_2^{r_2}...\alpha_s^{r_s} > 0, \tag{13}$$

with r_s, ($s \in \mathbb{N}$) is the occur rate of discrete event satisfying this two conditions, $r_s > 0$ and $\sum_{s=1}^{N'} r_s = 1$.

Theorem 1 *If there exist symmetric matrices $P_1 > 0$, $P_2 > 0$ and given scalars $\alpha_\sigma > 0$, $\sigma = 1, 2$ satisfying:*

$$\alpha_1^r \alpha_2^{1-r} > 0, \tag{14}$$

$$\begin{bmatrix} -P_1 & \Phi_\sigma^T \\ \Phi_\sigma & -\alpha_\sigma^{-2} P_2 \end{bmatrix} < 0, \tag{15}$$

with minimizing the trace ($P_1.P_2$) subject to:

$$\begin{bmatrix} P_1 & I \\ I & P_2 \end{bmatrix} \geqslant 0. \tag{16}$$

Then, the system (8) is exponential stable.

Proof: applying (12) to (8), we have

$$V(\hat{z}_{k_{i+1}}) - V(\hat{z}_{k_i}) < (\alpha_\sigma^{-2} - 1)V(\hat{z}_{k_i}).$$

Hence,

$$V(\hat{z}_{k_{i+1}}) < \alpha_\sigma^{-2} V(\hat{z}_{k_i}).$$

Since $V(\hat{z}_{k_i}) = \hat{z}^T(k_i)P_1^{-1}\hat{z}(k_i)$,

$$\hat{z}^T(k_i)\Phi_\sigma^T P_1^{-1}\Phi_\sigma\hat{z}(k_i) < \alpha_\sigma^{-2}\hat{z}^T(k_i)P_1^{-1}\hat{z}(k_i).$$

Therefore,

$$\Phi_\sigma^T P_1^{-1}\Phi_\sigma - \alpha_\sigma^{-2}P_1^{-1} < 0. \tag{17}$$

To rewrite (17) in a matrix form. We will utilize the Schur compliment. Then, (17) becomes:

$$\begin{bmatrix} -P_1 & \Phi_\sigma^T \\ \Phi_\sigma & -\alpha_\sigma^{-2}P_1^{-1} \end{bmatrix} < 0. \tag{18}$$

Remark 1 *It is clear that* (18) *is not linear because of the existence of* P_1 *and its inverse* P_1^{-1} *in the same matrix. To fix this problem we will make a change of variable* ($P_2 = P_1^{-1}$)*, and to guarantee the convergence of* P_2 *to* P_1^{-1}*, we will use the CCL approach [15]. The inequality matrices* (18) *becomes:*

$$\begin{bmatrix} -P_1 & \Phi_\sigma^T \\ \Phi_\sigma & -\alpha_\sigma^{-2}P_2 \end{bmatrix} < 0, \tag{19}$$

with

$$P_2 = P_1^{-1}. \tag{20}$$

The CCL approach is an algorithm used to guarantee that, P_2 equals P_1^{-1}, by minimizing the trace ($P_1.P_2$) subject to:

$$\begin{bmatrix} P_1 & I \\ I & P_2 \end{bmatrix} \geqslant 0. \tag{21}$$

Solving the LMIs using Yalmip Toolbox, we can calculate the gain F_1 which will help us to find the trajectory tracking gain F_2.

$$F_2^{-1} = C(I - (G + HF_1))^{-1}H. \tag{22}$$

4 Application

4.1 DC motor Model

A DC motor is a machine which converts direct current electrical power into mechanical power. The DC motor has vast applications in many fields including NCS. Owing to this importance, we chose the DC motor as an application system where we will apply our approach.

In this paragraph, the mathematical model of a DC motor will be studied.

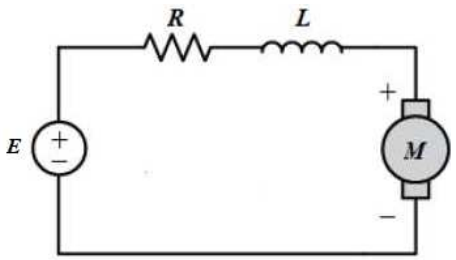


Figure 4: DC motor equivalent circuit model.

As shown in Figure 4, the system input is the voltage source (E), while the system output is the rotational speed $\dot{\theta}$. The physical parameters of the DC motor are given in Table 1:

Table 1: Physical parameters of the Dc motor.

Symbol	Description	Value
J	Moment of inertia	0.02 Kg.m^2
b	Motor viscous friction constant	0.2 N.m.s
K_e	Constant of emf	0.02V.s/rad
K_t	Motor torque constant	0.02N.m/A
R	Resistance	1.5 Ω
L	Inductance	0.5H

The motor torque and the back emf (e) are given in (23) and (24):

$$T_q = K_t.i, \tag{23}$$

$$e = K_e.\dot{\theta}. \tag{24}$$

Let us consider the constant K such that $K = K_t = K_e$. From the Figure 4, and employing the Kirchhoff's voltage law, the electrical equation of the DC motor is described as:

$$J\ddot{\theta} + b\dot{\theta} = Ki, \tag{25}$$

$$L\frac{di}{dt} + Ri = E - K\dot{\theta}. \tag{26}$$

If we choose $[\dot{\theta}, i]^T$ as a state variables, the state space representation will be written as:

$$A = \begin{bmatrix} \frac{-b}{J} & \frac{K}{J} \\ \frac{-K}{L} & \frac{-R}{L} \end{bmatrix}, B = \begin{bmatrix} 0 \\ \frac{1}{L} \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \text{D} = 0.$$

Replacing the parameters by their values, the state space representation becomes:

$$A = \begin{bmatrix} -10 & 1 \\ -0.04 & -3 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \text{D} = 0.$$

The command "c2d" in Matlab is used to passe from the continuous-time to discrete-time, where the sampling time is $T_e = 0.1s$. The discrete-system can be written as follow:

$$G = \begin{bmatrix} 0.3678 & 0.0563 \\ -0.0021 & 0.7407 \end{bmatrix}, H = \begin{bmatrix} 0.0066 \\ 0.1728 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \text{D} = 0.$$

4.2 Simulation and results

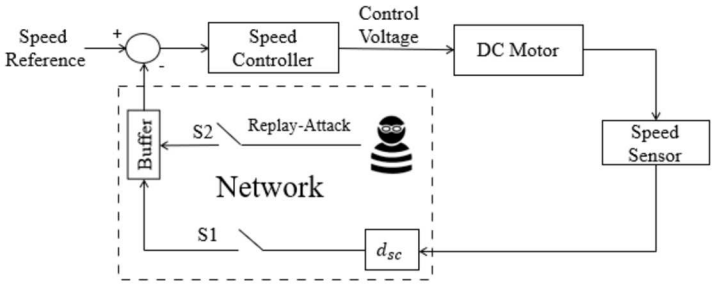


Figure 5: DC motor control under network issues.

The Figure 5 shown the structure of the DC motor under wireless network issues. The initial condition is $z(0) = [0 \;\; 0]^T$, $v_i = 0$, for $i \leq 0$, the maximum number of the successive packets losses during the transmission is N=3 packets, the communication delay equals to 0.1s, and the replay-delay equals to 0.3s, that means $d_{sc} = 1$, and d_T=3. We will study three different situations. In the first situation, the event rate of the switch S2 equals 0.1, in the second situation, the event rate equals 0.5, in the third, the event rate equals 0.9.

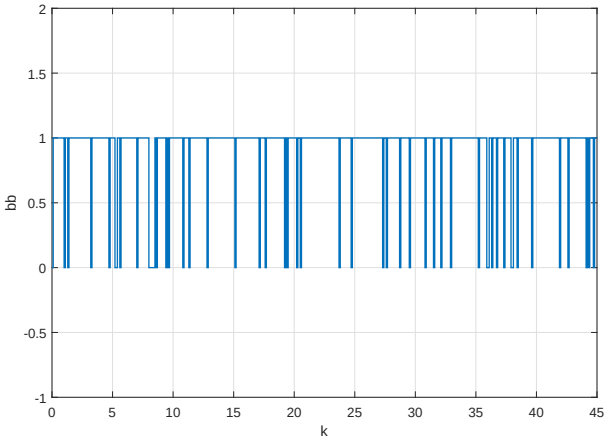


Figure 6: Event rate of S2 is 0.1.

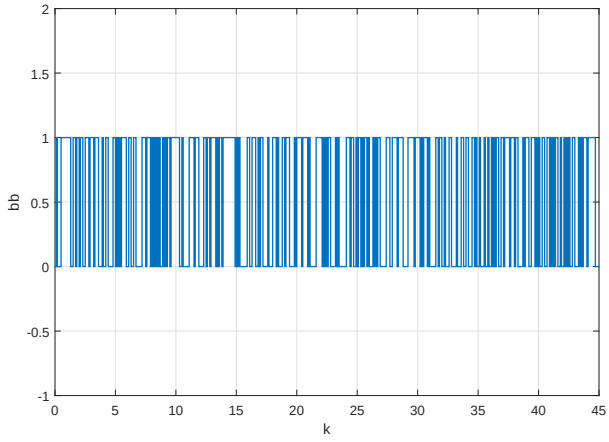


Figure 7: Event rate of S2 is 0.5.

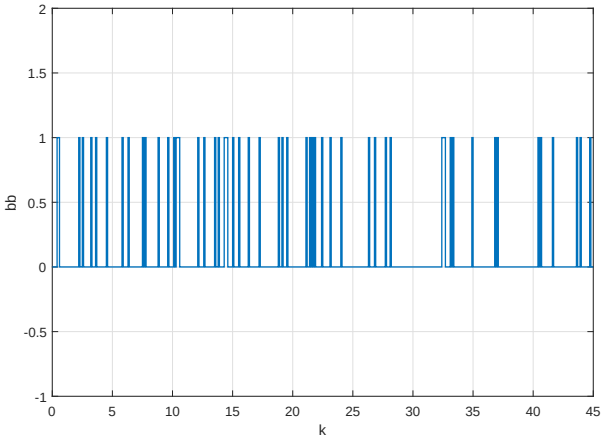


Figure 8: Event rate of S2 is 0.9.

The figures (Figure 6, Figure 7, Figure 8) show the different event rate of the switch S2. 0.1, 0.5 and 0.9 respectively.

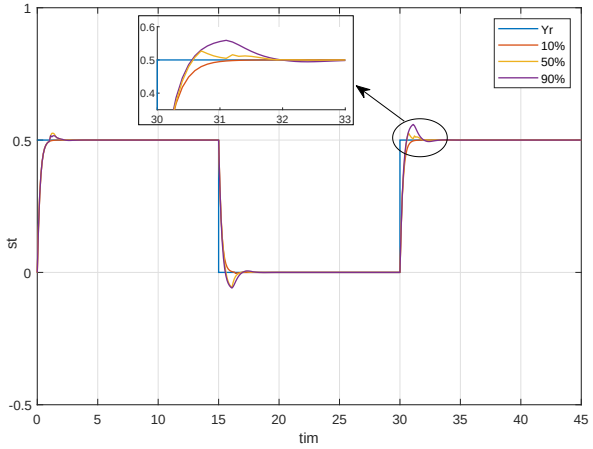


Figure 9: Trajectory tracking.

According to Theorem 1, the LMIs are feasible for $\alpha_1 = 0.1$ and $\alpha_2 = 0.45$. Solving the LMIs in Theorem 1 utilizing the Yalmip toolbox, we can find the controller gain and the trajectory tracking gain of the system as follow:
$F_1 = [0 \;\; 0.0012]$ and $F_2 = 20.9085$.

From Figure 9, we can see that if the event rate of the switch S2 equals 0.1 (the percentage to have an attack is 10%) the output can tracks perfectly the trajectory Y_r, the same thing happens if the chance to have an attack rises to 50% or 90%. But, in these two cases an overshoot appears. However, the results stay acceptable, which reflects the potency of our approach.

5 Conclusion

This paper dealt with the TTC issue of a DC motor controlled through a wireless network. In this extended version, we took into account the packet dropout and the delay in the communication channel (sensor-controller). On the other hand, the communication channel (controller-actuator) was assumed perfect, that means all

data have successfully transferred from the controller to the actuator without any delay. We also considered that the DC motor exposed to a replay-attack, where a cyber-adversary sought to destabilize the system and diminished its performances. A new mathematical model of the NCS under replay-attack was proposed. A sufficient condition for the stability of the resulting asynchronous dynamical system was given in the form of LMIs. The controller gain F_1 and the trajectory tracking gain F_2 were obtained by solving these LMIs employing the CCL approach. Finally, the simulation results proved the effectiveness of our approach. As a perspective of this study, our attention will be oriented towards studying the same problem with communication delay and packets losses in both communication channels.

References

[1] R. El Abbadi and H. Jamouli, "Stabilization of cyber physical system with data packet dropout and replay attack via switching system approach," in *2019 4th Conference on Control and Fault Tolerant Systems (SysTol)*. IEEE, 2019, pp. 325–329, doi: 10.1109/SYSTOL.2019.8864787.

[2] T. C. Yang, "Networked control system: a brief survey," *IEE Proceedings-Control Theory and Applications*, vol. 153, no. 4, pp. 403–412, 2006, doi: 10.1049/ip-cta:20050178.

[3] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, 2007, doi: 10.1109/JPROC.2006.887288.

[4] J. Xiong and J. Lam, "Stabilization of linear systems over networks with bounded packet loss," *Automatica*, vol. 43, no. 1, pp. 80–87, 2007, doi: 10.1016/j.automatica.2006.07.017.

[5] L. Qiu, Q. Luo, S. Li, and B. Xu, "Modeling and output feedback control of networked control systems with both time delays; and packet dropouts," *Mathematical Problems in Engineering*, vol. 2013, 2013, doi: 10.1155/2013/609236.

[6] L. Qiu, Q. Luo, F. Gong, S. Li, and B. Xu, "Stability and stabilization of networked control systems with random time delays and packet dropouts," *Journal of the Franklin Institute*, vol. 350, no. 7, pp. 1886–1907, 2013, doi: 10.1016/j.jfranklin.2013.05.013.

[7] J. M. R. Hernan, "Detection of attacks against cyber-physical industrial systems," Ph.D. dissertation, 2017.

[8] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2009, pp. 911–918, doi: 10.1109/ALLERTON.2009.5394956.

[9] R. EL Abbadi and H. Jamouli, "Stabilization of a cyber physical system with network issues," in *2019 8th International Conference on Systems and Control (ICSC)*. IEEE, 2019, pp. 508–512, doi: 10.1109/ICSC47195.2019.8950544.

[10] R. El Abbadi and H. Jamouli, "Stabilization of cyber physical system exposed to a random replay attack modeled by markov chains," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2019, pp. 528–533, doi: 10.1109/CoDIT.2019.8820311.

[11] P. Shakarian, J. Shakarian, and A. Ruef, "Attacking iranian nuclear facilities: Stuxnet," *Introduction to cyber-warfare: A multidisciplinary approach*, pp. 223–239, 2013, doi: 10.1016/B978-0-12-407814-7.00013-0.

[12] A. Routh, S. Das, and I. Pan, "Stabilization based networked predictive controller design for switched plants," in *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*. IEEE, 2012, pp. 1–6, doi: 10.1109/ICCCNT.2012.6396001.

[13] M. Yu, L. Wang, T. Chu, and G. Xie, "Stabilization of networked control systems with data packet dropout and network delays via switching system approach," in *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, vol. 4. IEEE, 2004, pp. 3539–3544, doi: 10.1109/CDC.2004.1429261.

[14] A. Rabello and A. Bhaya, "Stability of asynchronous dynamical systems with rate constraints and applications," *IEE Proceedings-Control Theory and Applications*, vol. 150, no. 5, p. 546, 2003, doi: 10.1049/ip-cta:20030704.

[15] L. El Ghaoui, F. Oustry, and M. AitRami, "A cone complementarity linearization algorithm for static output-feedback and related problems," *IEEE transactions on automatic control*, vol. 42, no. 8, pp. 1171–1176, 1997, doi: 10.1109/9.618250.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 274-279 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Comprehensive Review of Traditional Video Processing

Helen Kottarathil Joy*, Manjunath Ramachandra Kounte

School of ECE, REVA University, Bengaluru, 560064, India

ARTICLE INFO

Article history:
Received: 08 September, 2020
Accepted: 01 November, 2020
Online: 10 November, 2020

Keywords:
Video compression
MPEG
HEVC
Deep learning

ABSTRACT

Video and its processing are an interesting area as the increase in usage of internet videos, online streaming, CCTV, impact of internet on normal crowd increased. The need to know about video and its processing become an eminent area in research in current era. The paper tries to cover the traditional video processing, the advancement in video codec from the initial year, its origin, features, drawbacks and advancement lead to next stage. It provides an insight to need of video compression, steps involved in it, followed by overall review about video compression in various areas. The detailed explanation with reason of emergence, origin, characteristics are pointed. This information helps to add knowledge about the past and that helps to focus on the advancement and transitions that can be done to the video codecs. It summarizes the advancement in recent video processing using CNN, NN, deep learning too.

1. Introduction

'Video' the meaning has shifted far from a set of moving pictures from a traditional point of view to far extends. But the knowledge of traditional video processing is important to have a clear perspective towards the current video processing era. Video processing have a vast meaning from enhancement of various parameter in video, resolution, restoration of videos, denoising of videos, video compression etc. Each has its own impact and development in each stage of video processing development from video in a digital camera to HDTV to mobile camera to 4K,8K,10K videos. "The development of video compression started from the very started from the very start with a shadow of image compression such as Huffman coding, Golomb coding [1], arithmetic coding [2] etc. Later transform coding was introduced by encoding in spatial frequency including Fourier transform [3], Hadamard transform [4], Discreet Cosine Transform [5] etc followed by JPEG" [6]

The scope of video standardisation focus on video optimisation, allows complexity reduction for implementation and it no guarantee in terms of quality. Only decoder, bit stream and syntax of decoder is standardised. The scope if standardisation lies in the codec part of video processing. The basic knowledge about the fundamental terms in video processing helps to have strong view about any processing or standardisation done in video.

Video processing needs basic overview about the fundamentals like bitrate, display resolution, frame rate, frame type, interlacing, aspect ratio, video quality and compression techniques.

Bit rate is termed as the video data transferred in a frame of time or number of bits in the video transferred in a time durartion.it is related directly to the sharpness of the frame. This is a vital thing when we consider live broadcast into consideration. The term bitrate should be handled intelligently based on the encoder and decoder. Transfer speed also is a matter to be considered here. Display resolution in video is another area to be noted in video processing. Screen resolution /display resolution was not that important part in traditional era of video processing but as time moved and screen came with various sizes colour option, video

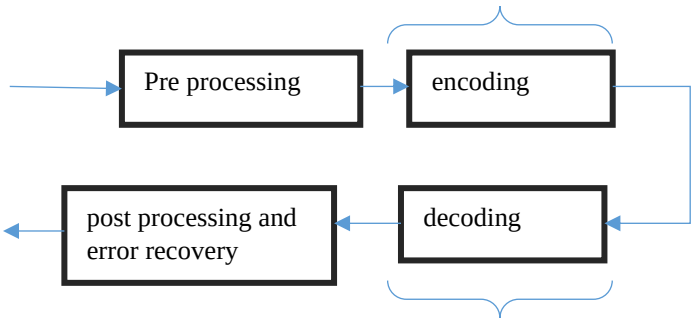


Figure 1: Scope of standardization

*Corresponding author: Helen Kottarathil Joy, helenjoy88@gmail.com

Screen resolution comes with an extension of 'p' and 'i' like 780p,1080p,1080i etc, it represents progressive /interlaced. The CRT monitors and TV followed interlaced scanning were the problem is flickering. Progressive scanning came into market as boom because the recent display screen responds fast. The refresh rate will be 60 normally to have better view effect. There are various frame resolution or frame type as in table

Table 1: Screen resolution Table

Sl.no	Resolution name	Horizontal pixels count	Vertical pixel count	Another Name
1	720p	1280	720	HD
2	1080p	1920	1080	Full HD
3	1440p	2560	1440	Quad HD
4	2160P	3840	2160	4K
5	4320P 8k	7680	4320	8K
6	4320p 10k	10240	4320	10K

Its noted that high resolution video can be viewed in low resolution screen. Video resolution doesn't matter on the screen it watching as it is taken care by the down sampling part in the decoder.

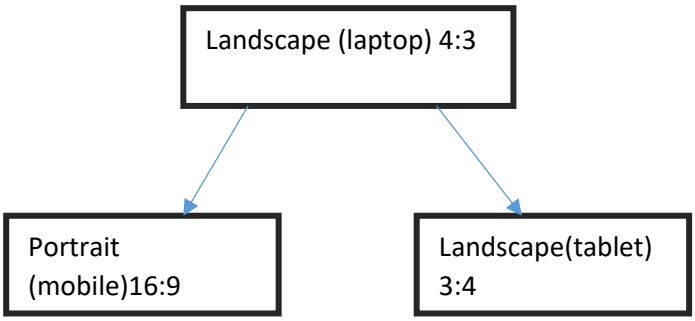


Figure 2: Aspect Ratio ideas

Aspect ratio says how the screen is with respect to the height of the display. In early ages 4:3 was common aspect ratio. In early 2010's by the boom of mobiles 16:9 become so prominent. Display orientation is another thing to be known. Size of the screen resolution should be considered properly for better display.

The paper is divided into three sections. The section 2 give brief idea about the steps in video compression, section 3 give a review of traditional video coding standards its origin, features and developmental stage towards next version. Comparison is done between various compressions followed by a overview about recent video compression with CNN and deep learning [7]. The paper ends with a conclusion with the need of learning traditional video compression for future research.

2. Video compression

2.1. Overview of video compression

Video compression is defined as a data reduction method ie used to encode the video. Video coding process helps that is the reduction of size of video file by making it compatible to store and need only less bandwidth for transfer. Video compression started as a succession of image compression. The compression mainly happens in a better way in video processing as the information carried by the frames are similar in most cases. A better percentage of compression can be achieved by taking this area into consideration. Video compression is essential as storage is taken into consideration. The disk space is important as we can more videos in fact more information can be loaded. Various transforms like DCT, DST etc can also be considered to reduce file size or in other words compression.

2.2. The stages of video compression

The basic steps in video processing are divided into six steps staring with partitioning the picture, predicting the similarity within the frame and between the frame. Majority of compression happens in this stage [8]. The predicted frame and other information's are transformed using DCT and quantized. All these were coded by using entropy coding and send as bit steams with proper bit rate.

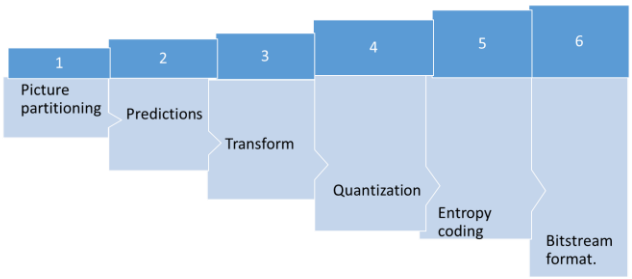


Figure 3: Basic Video Compression Steps

The input frame is transformed using DCT after picture partitioning (Splitting picture into macro blocks as apart of coding tree unit), then quantized. The quantized frame is inversed and compared with the previous frame in motion compensation and estimate is done by subtracting reconstructed frame from input frame and the residual with motion compensation. This is done by intra and inter prediction techniques basically comparing CTU within the frame and between frame to reduce redundancy and to gain maximum compression. The residual frame obtained will be 75 percentage compressed by this stage. The obtained frame 1&2 is transformed and quantized and the reconstructed frame is compared and subtracted. The process repeats and the output is encoded and framed and send in prescribed bit stream format.

3. Traditional video coding standards

Digital video technology covers the area of communication digital video telephony, digital TV, video storage, and a series of applications. This helps in the development of video codec in a faster way. The way of development of video codec are mainly focused on the organization ITU and ISO/IEC, joint venture of IFC and ISO are also available. ITU standards covers video codecs from H.261, H.263, H.263+, and this focus mainly in the video compression in real time communication for example video

conferencing whereas ISO/IEC mainly focus on the internal streaming, video storage etc. The bird view of video compression is shown in figure 4. Based on the purpose and application the compression features vary, upon that the segregation can be as the

one used in web or internet application, in cinemas, and for other medias. The ISO/IEC organization works on MPEG series as ITU/VCEG works on H.26x series. The ancestry of recent video compression followed this way.

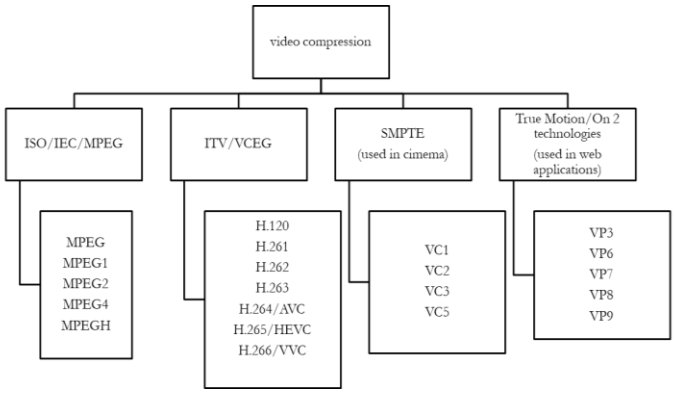


Figure 4: Video Compression Overview

Table 2: Specifications of H.262

1	Bit rate	1.5-2.0 Mbps CBR/VBR constant /variable bit rate
2	Packet/ cell loss rate (CLR)	<1 in 10 (-8)
3	BER	<1 in 10(-10)
4	Packet/ cell delay variation	<500ns

3.1. H.120 Video Compression Standard

The video coding standard starts with H.120 i.e. developed in the year 1984 by CCITT currently known as ITU-T. The primary application was video telephony and teleconferencing over ISDN. The bit rate available was 64kbits /sec.H.120 was not that good enough for real life applications, its spatial domain was good and temporal resolution was poor. The discussion came that encoding should have less than 1 bit as an average in a pixel. This leads to the idea of DCT block-based codec ie H.261.

3.2. H.261 Video Compression Standard

H.261 is the first member in the H.26x family developed in the year 1988-1990 in November by VGCE ie ITU-T9 video coding expert groups. The focus for the development of H.261 was similar to H.120 like video conferencing in ISDN and video telephony as it couldn't satisfy it [9]. The bit rate it was focusing was multiples of 64 kbits /sec. It was used as a backward compatibility mode in H.323 and some video conferencing system still it couldn't be perfect and need refinement that leads to next video codec. But still it remains as a milestone in the history of video codecs.

3.3. H.262/ Moving Picture Expert Group-2

H.262/MPEG-2 was proposed in 1995 by the joint venture of VCEG and MPEG. It is mainly for storage of video like SVCD, digital video, Blue ray, broadcasting, DVD video etc. Its bit rate is suitable for only less than 1Mbps, performs bad for high bit rate. MPEG -2 is a lossy compression method with motion estimation vector, DCT quantization an encoding by Huffman coding, the technique happens in this sequence [10]. The motion vector estimation compares the frame and approximate it into similar set of video frame in a translated form with changes between the frame. This helps to reduce the greater percentage of temporal redundancy in the video frame. This is followed by DCT that helps to convert the spatial information to frequency domain and discard all high frequency information that will not affect the visual experience of human eye. Quantization is applied to the coefficient of DCT. Huffman coding shortens the code and helps in better compression but its lossy with entropy coding.MPEG is good for high quality video with a particular set of transmission limit but not that better for internet transmission as the QoS of MPEG can't be satisfied always with intend.

3.4. H.263 Video Compression Standard

H.263 was designed for low bitrate mainly on video conferencing. This was evolved from the pros and cons of H.261,

MPEG-1, MPEG-2 standard developed in 1996 by ITU -T9. The main application was B-ISDN video on mobile phones (3Gp), video conferencing and video telephony adopted H.324 PSTN. the bit rate followed was 33.6 Kbps. It was developed as their levels that is 1,2,3 as H.263, H.263+, H.263++ with various extra features added on it. H263+ was focusing on improved compression performance and bit stream scalability. H.263++/H.263 2000 namely MPEG 4 part 2, it helps surveillance system its simple profile formal.

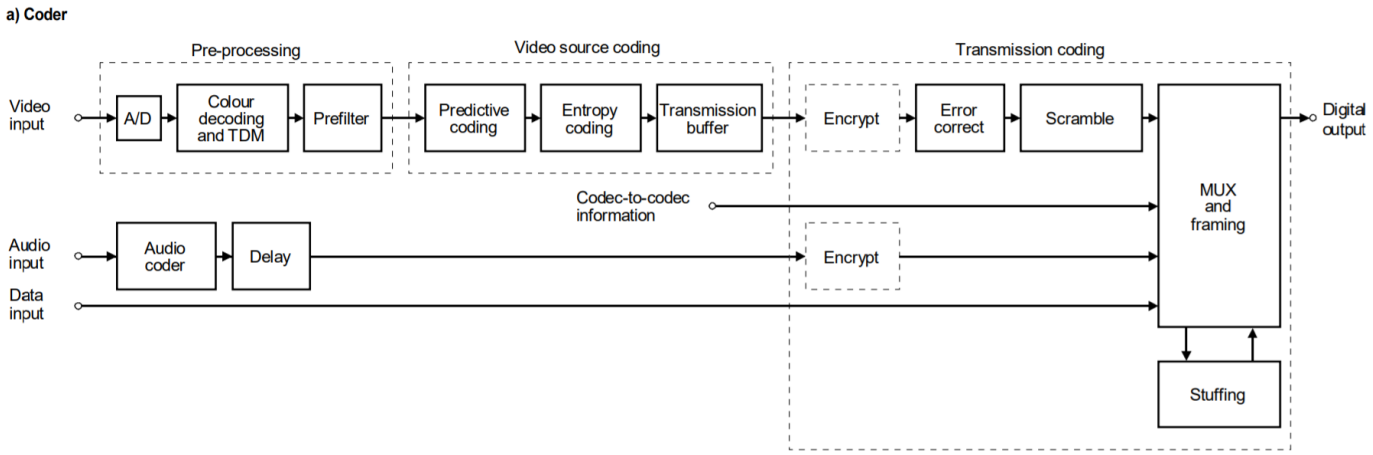


Figure 5: H.261 [7]

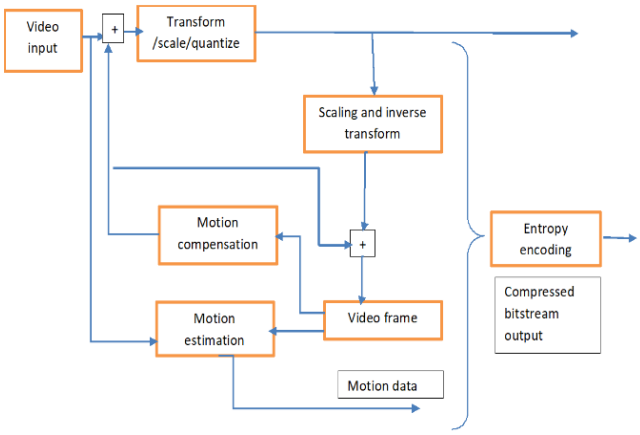


Figure 6: Hybrid video encoder [1]

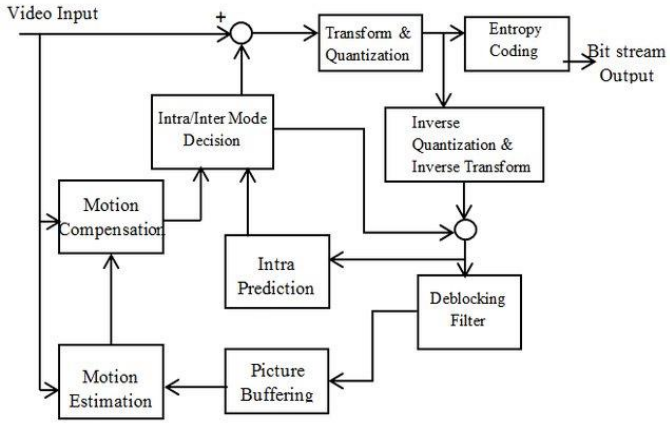


Figure 7: H.264 codec [12]

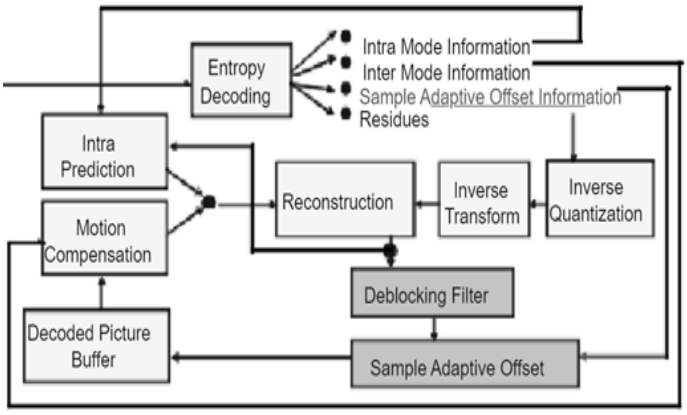


Figure 8: Video Compression with HEVC [14]

3.5. H.264 Video Compression Standard

H. 264/MPEG 4, AVC i.e. [10]an extended version of H.263, MPEG 4 visual released on 30th may 2003 is developed by ISO/IEC and ITU and H .264 and distribution of video content. The resolution it support is 4096X2304 that includes UHD and 4K videos. It uses variable bit rate. Advanced Video Coding /AVC is 25th version of AVC ie MPEG 4 part 10[11]. It is the best-known video coding standard for media storage like Blue ray disc and widely used in streaming in internet source that used variable bit rate. Scaling video coding is a domain feature here added with multi view video coding, 3D-AVC and multi resolution frame compatible (MFC). The application area includes DVBC (digital video broadcast project), HD-DVD, ATSC (advanced television system connector), CCTV (closed circuit TV), DSLR etc.

3.6. H.265 Video Compression Standard

H.265 is devoloped as the successor of AVC with reduced file size, reduced bandwidth[12]. There are CTUs in HEVC than macro blocks in H.264. The HEVC have variable CTU size from 4x4 to 64x64. The main features of HEVC includes flexibility in partitioning size, flexibility in trasform block size and predicting the modes, parallel processing architecture, sophistication for prediction, signalling mode and motion vector[13]. HEVC architecture has encoder and decoder. In encoder picture partioning is done. In the next process is the partioned picture is predicted by intra /inter prediction and the result is subtractred from the original unit[14]. The residue is transformed by DCT and quantised in next step. At the end transformed output, predicted information, mode information, and header are entropy encoded. In decoder the counterpart does the reverse operation to deliver the picture to other end of communication[15].

3.7. Society of Motion Picture and Television Engineers standards

It was introduced in founded in 1916 as the Society of Motion Picture Engineers or SMPE is a global professional association, of engineers, technologists, and executives working in the media and entertainment industry. An internationally recognized standards organization, SMPTE has more than 800 Standards, Recommended Practices, and Engineering Guidelines for broadcast, filmmaking, digital cinema, audio recording, information technology (IT), and medical imaging[16-19]. In addition to development and publication of technical standards documents, SMPTE publishes the SMPTE Motion Imaging Journal, provides networking opportunities for its members, produces academic conferences and exhibitions, and performs other industry-related functions"[20].Compression Systems in SMPTE has standardized five VC standards: VC-1 to VC-5 to provide well-reviewed documentation and enhanced interoperability. The latest of these is the VC-5 standard family that provides documentation and reference software for the video compression used in GoPro systems and workflows. SMPTE also has a new project to document the Apple ProRes codec.

3.8. Video Processing for Internet

On 2 codec designs are popularly known as VP3 then VP7.VP6 was used in flash 8 video codec. It was used in google, skype, YouTube etc. Microsoft and Windows media player version 9 properties similar to Real video, DivX and On 2 tech. AVC improves compression quality, H.264 becomes leading standard and used by many video applications such as in iPod PlayStation Portable, as well as in TV broadcasting standards like DVB-H and DMB[21].AVS audio video standards were initialized by China, the main reason china holds the key factors in development of consumer electronics. Coding efficiency is equivalent to H.264 while computation complexity is less. It was widely used standard [22].

Table 3: Comparison Table of video codecs

Parameter	HEVC	AVC	VP9
Compression Efficiency	Twice compared to AVC	Less than HEVC	Less than HEVC
Resolution &uploading speed	Less compared to AVC	Better compared to both	More
Bandwidth required for broadcast	15mbps	32mbps	More than HEVC
Compatibility	4k TVs, VOD	3Dvideo coding.	Better in chrome, opera Firefox…
Royalty	Not open source	Not open source	Open source
Computation cost	low	high	low
Encoding quality	Good for low bit rate videos	Use lower bit rates better than previous standards	Good for high bit rate videos
FPS	Support up to 300fps	Support up to 59.94fps	Support up to 60fps for video

4. Neural network and Deep learning-based video coding standards.

After the development of H.265/HEVC, the blocks in the codec were explored and improved by adding neural network/ deep learning techniques [22-27] into it. The development of video codec in the following year will be mainly on making a smart video codec by incorporating the deep learning features into it[28-30]. The trial of neural network /deep learning into HEVC standard by the application of trying to make smart modules in HEVC. By using multiple features, the modules are trying to be made smart. The future of codec lies in the hands of deep learning i.e. the future is expecting a smart video codec by absorbing the basic steps in codec and improvising it by adding deep learning techniques into it. Work have been done in HEVC modules intra-prediction, inter-prediction, quantization and encoding and loop filter with neural network methods [31] that can be revised and learned separately.

As an overview the neural network for video coding can be reviewed this way, for intra prediction all the method developed for still image can be used. As we know the applications are many. For neural network based intra-prediction the effective tool considered can be motion, with the better training network it can be effectively done. Some other research happened in this area using CNN are post processing, resolution up conversion sub pixel interpolation, loop filtering, intra-coding, encoding optimization etc.

5. Conclusion

Video and its processing are the area that is focusing on a smart era in the near future with high quality advancement. The paper points the video processing developmental stages from initial days till now. This helps to focus on the recent research in video processing and thus paves way for new research knowing about the origin, features and properties of early codecs, earns us to work on the future of video processing. The early era of video compression started with basic coding like Huffman coding, arithmetic coding etc and transforms like Fourier transform, DCT etc followed by the hybrid video processing era with many compression techniques from Motion JPEG to H 265. The detailed study of this helps to identify that in each stage the was trying to improve the compression ratio, reduce the complexity of circuit, and with high uploading speed. The research and advancement in this area can be done by having a strong base about the existing video processing. The drawbacks of the system can be reviewed and that helps in the research to make it smarter one. The future is focusing on smart video processing and deep learning, NN, CNN are the strong pillars for intelligent video processing.

Conflict of Interest

The authors declare no conflict of interest

References

[1] D.A. Huffman, “A Method for the Construction of Minimum-Redundancy Codes,” Proceedings of the IRE, ,**40**(9) , 1098-1101,1952,doi:10.1109/JRPROC.1952.273898.

[2] S.W. GOLOMB, “Run-Length Encodings,” IEEE Transactions on Information Theory, **12**(3), 399-401, 1966, doi:10.1109/TIT.1966.1053907

[3] I.H. Witten, R.M. Neal, J.G. Cleary, “Arithmetic coding for data compression,” Communications of the ACM, **30**(6), 520-540, 1987, doi:10.1145/214762.214771.

[4] H.C. Andrews , W.Pratt ,”Fourier Transform Applications”, Proc. Hawaii Int. Conf. System Sciences , 677-679, 1968, doi:10.5772/2658.

[5] W.K. Pratt, H.C. Andrews, J. Kane, “Hadamard Transform Image Coding,” Proceedings of the IEEE, **57**(1), 58-68, 1969, doi:10.1109/PROC.1969.6869.

[6] H. K. Joy and M. R. Kounte, "An Overview of Traditional and Recent Trends in Video Processing," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 848-851, 2019, doi: 10.1109/ICSSIT46314.2019.8987896.

[7] A.S. Coyner, J.P. Campbell, M.F. Chiang, Demystifying the Jargon: The Bridge between Ophthalmology and Artificial Intelligence, Ophthalmology Retina, 2019, doi:10.1016/j.oret.2018.12.008.

[8] A.N. Netravali, J.D. Robbins, “Motion-Compensated Television Coding: Part I,” Bell System Technical Journal, **58**(7), 1703-1718,1979, doi:10.1002/j.1538-7305.1979.tb02238.x.

[9] N. Ahmed, T. Natarajan, K.R. Rao, “Discrete Cosine Transform,” IEEE Transactions on Computers, **100**(1), 90-93, 1974, doi:10.1109/T-C.1974.223784.

[10] D.K. Kwon, M. Budagavi, V. Sze, W.S. Kim, Video compression, 2017, doi:10.1201/b12494.

[11] T. Wiegand, G.J. Sullivan, G. Bjøntegaard, A. Luthra, “Overview of the H.264/AVC video coding standard,” IEEE Transactions on Circuits and Systems for Video Technology, **13**(7), 560-576, 2003, doi:10.1109/TCSVT.2003.815165.

[12] Wen Gao, "AVS standard - Audio Video Coding Standard Workgroup of China," 14th Annual International Conference on Wireless and Optical Communications, 2005. WOCC 2005, Newark, NJ, USA, 125–166, 2005, doi: 10.1109/WOCC.2005.1553738.

[13] M.A. Ansari, I.U. Khan, “Performance analysis and evaluation of proposed algorithm for advance options of H.263 and H.264 video codec,” in 2015 International Conference on Recent Developments in Control, Automation

and Power Engineering, RDCAPE 2015, 371-376, 2015, doi:10.1109/RDCAPE.2015.7281427.
[14] G.J. Sullivan, J.R. Ohm, W.J. Han, T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," IEEE Transactions on Circuits and Systems for Video Technology, **22**(12), 1649–1668, 2012, doi:10.1109/TCSVT.2012.2221191.
[15] H. Lv, R. Wang, X. Xie, H. Jia, W. Gao, "A comparison of fractional-pel interpolation filters in HEVC and H.264/AVC," in 2012 IEEE Visual Communications and Image Processing, VCIP 2012, 1-6, 2012, doi:10.1109/VCIP.2012.6410767.
[16] K.M. Khan, J. Arshad, M.M. Khan, "Secure digital voting system based on blockchain technology," International Journal of Electronic Government Research, **14**(1), 53–62, 2018, doi:10.4018/IJEGR.2018010103.
[17] V.M. Harshini, S. Danai, H.R. Usha, M.R. Kounte, "Health record management through blockchain technology," in Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019, 1411-1415, 2019, doi:10.1109/icoei.2019.8862594.
[18] H.K. Joy, S.L. Das, "A novel approach for biomedical web image super resolution," in Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2013, 876–879, 2013, doi:10.1109/ICCPCT.2013.6528858.
[19] S.J. Kamble, M.R. Kounte, "Machine Learning Approach on Traffic Congestion Monitoring System in Internet of Vehicles," in Procedia Computer Science, **171**(1), 2235–2241, 2020, doi:10.1016/j.procs.2020.04.241.
[20] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, S. Wang, "Image and Video Compression with Neural Networks: A Review", IEEE Transactions on Circuits and Systems for Video Technology, **30**(6), 1683-1698, 2020, doi:10.1109/TCSVT.2019.2910119.
[21] S. Matsumoto, T. Yoshihisa, T. Kawakami, Y. Ishi and Y. Teranishi, "A Distributed Video Processing System for Internet Live Broadcasting Services," 2016 19th International Conference on Network-Based Information Systems (NBiS), Ostrava, 2016, 311-316, doi: 10.1109/NBiS.2016.28.
[22] M.R. Kounte, P.K. Tripathy, P. Pramod, H. Bajpai, "Implementation of Brain Machine Interface using Mind wave Sensor," in Procedia Computer Science, 244–252, 2020, doi:10.1016/j.procs.2020.04.026.
[23] B. Tian, L. Li, Y. Qu, L. Yan, "Video Object Detection for Tractability with Deep Learning Method," in Proceedings - 5th International Conference on Advanced Cloud and Big Data, CBD 2017, 397-401, 2017, doi:10.1109/CBD.2017.75.
[24] H. Wu, X. Zhang, B. Story, D. Rajan, "Accurate Vehicle Detection Using Multi-camera Data Fusion and Machine Learning," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 3767-3771, 2019, doi:10.1109/ICASSP.2019.8683350.
[25] M.R. Kounte, P.K. Tripathy, P. Pramod, H. Bajpai, "Analysis of Intelligent Machines using Deep learning and Natural Language Processing," in Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020, 956-960 ,2020, doi:10.1109/ICOEI48184.2020.9142886.
[26] W. Zhang, D. Zhao, L. Xu, Z. Li, W. Gong, J. Zhou, "Distributed embedded deep learning based real-time video processing," in 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings, 001945-001950, 2017, doi:10.1109/SMC.2016.7844524.
[27] M.R. Kounte, P.K. Tripathy, P. Pramod, H. Bajpai, "Implementation of Brain Machine Interface using Mind wave Sensor," Procedia Computer Science, **171**, 244–252, 2020, doi:10.1016/j.procs.2020.04.026.
[28] J. Lainema, F. Bossen, W.J. Han, J. Min, K. Ugur, "Intra coding of the HEVC standard," IEEE Transactions on Circuits and Systems for Video Technology, **22**(12) ,1792-1801, 2012, doi:10.1109/TCSVT.2012.2221525.
[29] J.L. Lin, Y.W. Chen, Y.W. Huang, S.M. Lei, "Motion vector coding in the HEVC Standard," IEEE Journal on Selected Topics in Signal Processing, **7**(6), 957-968 , 2013, doi:10.1109/JSTSP.2013.2271975.
[30] X. Zhang, S. Wang, Y. Zhang, W. Lin, S. Ma, W. Gao, "High-Efficiency Image Coding via Near-Optimal Filtering," IEEE Signal Processing Letters, 2017, doi:10.1109/LSP.2017.2732680.
[31] L. Vinet, A. Zhedanov, "A 'missing' family of classical orthogonal polynomials," in Journal of Physics A: Mathematical and Theoretical, **44**(8), 085002, 2011, doi:10.1088/1751-8113/44/8/085201.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 280-289 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Overmind: A Collaborative Decentralized Machine Learning Framework

Puttakul Sakul-Ung[1,*], Amornvit Vatcharaphrueksadee[2], Pitiporn Ruchanawet[3], Kanin Kearpimy[1], Hathairat Ketmaneechairat[1], Maleerat Maliyaem[1]

[1]*King Mongkut's University of Technology North Bangkok Bangkok, 10800, Thailand*

[2]*North Bangkok University Bangkok, 12130, Thailand*

[3]*King Mongkut's University of Technology Thonburi Bangkok, 10140, Thailand*

ARTICLE INFO

Article history:
Received: 29 August, 2020
Accepted: 01 November, 2020
Online: 10 November, 2020

Keywords:
Decentralization
Machine Learning
Distributed Artificial Intelligence
Feature Transfer
PM2.5 Prediction
Collaborative Machine Learning

ABSTRACT

This paper is an extension of work originally presented in PM2.5 Prediction-based Weather Forecast Information and Missingness Challenges: A Case Study Industrial and Metropolis Areas, which focused on imputation algorithm to solve missingness challenge and demonstrated a basic prediction system to prove the proposed algorithm, II-MPA. Distributed and decentralized systems, recently, have been proven for their effectiveness in multiple perspectives. This paper introduces "Overmind", the solution that governs and builds the network of decentralized machine learning as a prediction framework named after its functionality: it aims to discover a set of data and associated attributes for assigning machine learnings in the collaborative decentralized manner. Overmind also empowers feature transfer learning with data preservation. It demonstrates how discovered features are transferred and shared among synergic agents in the network. This model is tested and evaluated the accuracy against the traditional single machine learning prediction model in the original work. The results are satisfactory in both prediction performance and transfer learning.

1. Introduction

This paper is an extension of work in [1]. Traditionally, a single machine learning is built to process upon a dataset; this can be served as a recommendation, prediction, classification, etc. Unfortunately, the drawback of the traditional model is realized where machine learning problems and algorithms are far enough that all scenarios can be resolved using a single model [2]. A traditional approach relies on the training process over a whole dataset; data exploration, perhaps, consumes time and resources. Also, the experience in data analytics does not guarantee the effectiveness of the built machine learning systems to specific business scenarios. Moreover, discovering new attributes in real-world scenarios may require a brand-new dataset and retraining process. As business prefers to preserve data as much as possible, this is supported by system design which deals with a new data pattern and the least trade-off and penalty in performance.

This paper innovates a framework under the "Data Blind" concept – it indicates the knowledge over dataset and scenarios, which are unknown and undiscovered. Hence, the proposed framework performs data analytics for building machine learning models that serve multiple purposes. Overmind, a system conceptual model in distributed artificial intelligence, is named after its theoretical design. Overmind aims for managing dataset, discovering data pattern, and creating collaborative agents, as a complete framework. This paper experiments a demonstration of how it works, and steps, which are taken to gather results and contributions. An interpretation of behaviours and outcomes will be studied and presented in future works.

2. Previous works

The concept of Overmind has been forged base on previous machine learning algorithms and the decentralized systems. We also explore related works on feature transfer and selection. Thus, this section illustrates in two aspects, which are Multi-agent and Decentralized System, and Feature Transfer.

*Corresponding Author: Puttakul Sakul-Ung, King Mongkut's University of Technology North Bangkok Bangkok, Thailand, puttakul.s@gmail.com

2.1. Multi-Agent and Decentralized System

A term of Distributed Artificial Intelligence (DAI) has been defined based on the concept of distributed intelligent agents. Many projects and methodologies have been reported and summarized into categories proposed in [3] as system conceptual and agent conceptual models. The fine examples of DAI contributions were presented in [4–6] in the area of Intelligent Manufacturing System (IMS) since 1990. Then, a decentralized and distributed machine learning works were published in 1993 [7, 8]. The Cooperative Distributed Problem Solving (CDPS) that involves active agents sharing cooperative behaviours and protocols was invented in 1995 [9]. However, the cooperation is working based on broadcasting negotiation between a system and active agents by receiving the message to perform solving collaboratively.

Undeniably, this paper is highly influenced by the concept proposed in years ago. Although, machine learning has been continuously improved its capability and intelligence to the level that can resolve complex problems in the real-world scenario. Today, a decentralized reinforcement learning [10], which is again inspired by the multiagent system, describes problem statements that require large-scale real-world machine learning whereas the completed work focuses on two dimensions - Networked Evolution Strategies (NES), and Bayesian Optimality in the Wisdom of the Crowd (BOWC). NES describes network structure of communication between parallel-processing agents, that speed up learning process. This network structure of communication inspires Overmind's concept to integrate network analysis components and allow communication among decentralized machine learnings. Lastly, BOWC draws attention to the modelling on how humans learn from the influence of each other; this also inspires the Overmind concept for incorporating the knowledge management framework to mimic the collaboration in humans.

Decentralized Clustering in Large Multi-Agent Systems [11] draws a concept of using multi-agents in decentralized clustering, which is one of the major machine learning objectives. To overcome data clustering real-world problems such as data is widely distributed, data sets are volatile, and data instances also cannot be compactly processed, decentralization of multi-agent has proven its better performance against the traditional k-mean algorithm. This reviewed work hints decentralized clustering of a single dataset, perhaps, gives significant contributions in pre-data processing.

Recently in 2019, Feature Distributed Machine Learning (FDML) is studied collaboratively called FDML: A Collaborative Machine Learning Framework for Distributed Features [12]. This work inspires privacy preservation in feature engineering when one machine learning model can acquire prediction power by using additional features from other machine learning without sharing actual features. This works fully introduces the collaborative machine learning framework with high capability in transfer learning. The experiments are conducted using a different algorithm, and the results yield significance in real-world application development.

Nowadays, a concept of distributed and decentralized artificial intelligence has been widely reported, which shows upcoming era in machine learning integrated fields of research such as artificial intelligence in distributed ledger technology [13], artificial intelligence as IoT [14, 15], distributed security solution [16, 17], etc. Surprisingly, these integrations are all based on a distributed and decentralized concept.

2.2. Transfer Learning

In this sense, transfer learning refers to a process where discovered knowledge is fed to existing machine learning for reinforcement and adaptation over modelling that intents to re-train for better performance. Furthermore, the communication among models in a network is established for transferring particular knowledge from one machine learning to others [18]

Similarly, work in [19] introduces the transferring knowledge over different classes. This work also emphasizes transfer learning in feature selection for machine learning models using a proposed framework. The results of this work influents possibility of transfer learning in real-world dataset across classes and domains.

In sum, this paper draws conceptual designs based on previous works in two dimensions, which are distributed and collaborative systems, and transfer learning. However, the algorithm and process behind our work are different from the origins due to its' concepts and purposes. Ultimate research questions under our contributions are following: (1) how to establish an end-to-end platform for demonstrating collaborative machine learning models in a dynamic network in order to power up the prediction and (2) how to create a flexible system for transfer learning between models in a network that reflects the reinforcement.

3. Methodology

This work follows a quantitative research method. Firstly, the literature review is performed by limiting the scope of previous works and using keyword-searching. The results of the review are compounded into the conceptual design of the proposed framework, Overmind, which will be described in the next section of this paper. The reviewing of the proposed framework is conducted by researchers and authors. Experiment and testing are conducted on our previous work dataset, Thailand Air Quality Index (AQI) and weather forecast information. The result of applying this proposed framework is studied by comparison against the traditional framework on the same dataset, as presented in sections 5 and 6 of this paper.

4. Overmind: A Framework for Decentralized Machine Learnings

Overmind, a collaborative decentralized machine learning framework, is proposed considering dealing with anonymous dataset before building a machine learning model to serve purposes such as prediction, recommendation and classification. Figure 1 shows the main conceptual design of deploying multiple collaborative machine learning models (“agents”) each of which is responsible for a specific cluster of data. Collaboration means the agents can communicate and exchange knowledge, and share the workload for data processing.

This paper proposes Overmind as a collection of functional processes which are:

1. Determine Data Clustering

2. Assign Agent
3. Establish Network
4. Transfer Learning

In the following section, the aforementioned functional processes are explained in the respective order.

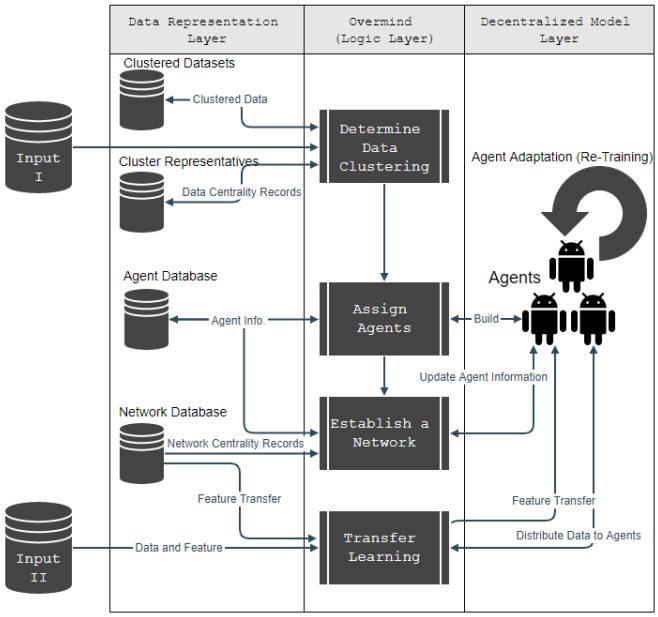


Figure 1: Overmind Architecture

4.1. Determine Data Clustering

Firstly, just the traditional approach, Overmind mimics the data cleansing process by using unsupervised learning algorithms to divide the dataset into clusters. This can be demonstrated as follows:

1. Let D be the training dataset where $D = \{d_1, d_2, d_3, \dots, d_n\}$, given that n is the number of instances.
2. Let F be the feature set, where $F = \{f_1, f_2, f_3, \dots, f_z\}$, given that z is the number of features in D.
3. Overmind deploys unsupervised learning method to divide D into clusters, given C is a set of clusters where $C = \{c_1, c_2, c_3, \dots, c_k\}$ and members in D are allocated and distributed over members of C, and k is the number of clusters.
4. The framework proposes s as centrality node of similarity graph of data under c_i. Hence, each cluster contains centrality node of data which is represented as s_i, where $i \leq k$

Figure 2 illustrates the output of determining data clustering and assigning agent.

4.2. Assign Agent

Agents are the supervised machine learning models trained using data in each cluster. Overmind determines the number of agents and builds a supervised machine learning algorithm to create agents, which can be explained as follows:

1. Let A be a set of agents trained by using C, given that a_i is trained by using data under c_i where $i \leq k$
2. Hence, a_i will be agent based on z features, given each feature contains weight/coefficient for particular a_i. Let W_i be a set of weight/coefficient of features for particular agent a_i

4.3. Establish Network

To build a collaborative framework, there shall be linkages between agents and rules for cooperation. This paper introduces a similarity network of agents based on the graph as below:

1. Let G be a graph of pairs $(V(G), E(G))$ where $V(G) = A$.
2. Similarity of agents in A is measured by W_i of particular a_i, given that y is a pre-determined threshold of similarity.
3. $E(G)$ will be created and linked between agent in $V(G)$ where the similarity measurement of a pair of agents satisfies y. This results in a network of agents.
4. Denote the centrality of G by H_G. If the centrality of G corresponds to multiple agents, Overmind selects one.

Eventually, Overmind creates decentralized machine learning as a collection of nodes in a network by using similarity measurement. Each node represents a supervised machine learning model.

assessed in term of similarity, which can be advanced for future works when one would like to measure affinity among trained models.

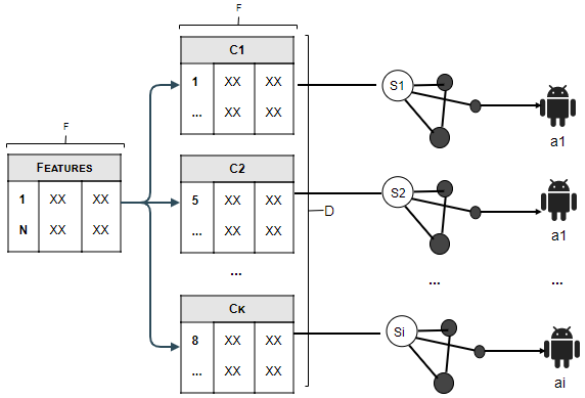


Figure 2: Illustration of Determining Data Clustering and Assigning Agent

4.4. Transfer Learning

This section describes two major transfer learnings of Overmind –data and feature transfers.

4.4.1 Data Transfer

Data transfer is involved when new data arrive for processing. Assumption on testing scenario when d_{n+1} is discovered and fed to framework, Overmind determines a suitable a_i for d_{n+1} by using s_i. This process implies similarity measurement on the centrality of clusters and new data instances. The steps are taken as below:

1. Let D' be a test dataset of this framework. Hence, $\{d'_1, d'_2, d'_3, \dots, d'_q\} = D'$.
2. Overmind determines each d'_x in D' and measures the similarity of d'_x to s_i.
3. Overmind assigns d'_x to a_i which has the highest similarity measured between s_i and d'_x. This step can be considered as

a batch process where D' is divided into subsets for feeding to particular a_i.

4. The accuracy of framework is measured as average accuracy of d'_x or average accuracy of subsets in D'.

4.4.2 Feature Transfer

Features and Feature Transfer are challenging components in designing the machine learning model. A traditional approach may require retraining of model when dealing with newly discovered features. A retaining process takes resources and time consuming for a big dataset. Yet, in the real-world scenario, some business may obtain a set of new information, both data and features, whether it contributes significances or not, but they have no willing to dispose of an existing dataset or pre-trained model.

In this sense, feature transfer refers to the ability that the existing trained model determines and deploy the discovered new features without retraining model; this means the transfer process will occur in every agent in the network. To elaborate flexibility in a dynamic machine learning model which accepts changes in both dataset and features, Overmind is conceptually designed to overcome the aforementioned challenges by described steps below:

1. When f_{z+1} is discovered within a new dataset D_{new}, Overmind predetermines a need of f_{z+1} by mimicking the supervised machine learning model of H_G and process upon D_{new} along with f_{z+1}. However, to avoid data volume bias, Overmind will proportionally random data in each c_i to form a dataset as equal as D_{new}.
2. Let accuracy of forming dataset is denoted by acc_{ext} and accuracy of model upon D_{new} is denoted by acc_{new}.
3. If $acc_{ext} > acc_{new}$, Overmind drops f_{z+1} from D_{new} and takes as data transfer process. This indicates that f_{z+1} cannot improve performance of existing system.
4. If $acc_{ext} \leq acc_{new}$, Overmind considers f_{z+1} as significant feature and process further steps for feature transfer process which is described below:
   4.1. Where the existing dataset lacks of f_{z+1} true values, Overmind creates imputation model by changing target label to f_{z+1} using forming dataset. This imputation model is denoted by a_{imp}.
   4.2. Overmind inserts f_{z+1} for each c_i in C, denoted by c'_i. This may result in extreme missing values in f_{z+1}. Hence, a_{imp} is used for imputation of missingness.
   4.3. Starting from H_G, agent a_i retrains its model along with f_{z+1}, the retrained agents are denoted by a'_i
   4.4. Retrained agents a'_i perform cross-validation and compare a result against a_i. There are two possible outcomes
      - a'_i gives a better performance, then f_{z+1} is accepted. a'_i adjusts W_i, denoted by W'_i, or else
      - a_i gives a better performance, then f_{z+1} is rejected. a'_i flags weight/coefficient of f_{z+1} as zero in W'_i
5. Overmind reshapes a network of agents, this creates $G' = (V(G), E(G'))$. Perhaps, centrality of network may be different from G, Overmind updates the new centrality, denoted by $H_{G'}$

Figure 3 shows the establishing of network of agents and transferring learning that impacts changes over a network.

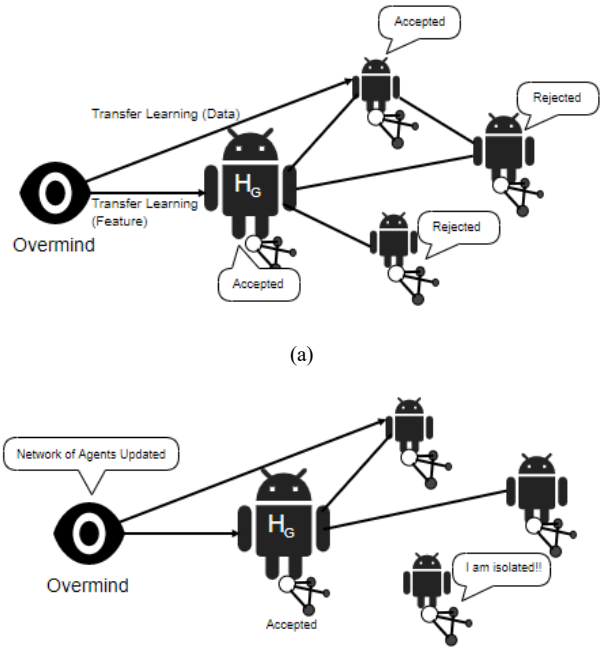


Figure 3: (a) Illustrating an established network with transfer learnings for both data and feature (b) Demonstrating a reshaped network resulted from transfer learning process.

Until now, a proposed framework creates a dynamic network which illustrates the transfer process among decentralized agents. Hence, this is called the collaborative decentralized machine learning framework.

5. Performance evaluation: a case study Thailand PM2.5 and weather forecast information database

To evaluate framework performance, this paper simulates the framework over the testing dataset, which is Thailand PM2.5 and weather forecast dataset presented in [1]. This dataset contains 2,383 instances with 21 features (including target). Overmind's simulation takes place in two functions, Framework Performance and Demonstration on Transfer Learning for Feature Transfer, to predict PM2.5 values from a given set of features.

5.1. Framework Performance

This paper deploys 80% of dataset for training and 20% for testing, for both traditional machining learning model –a single model –and Overmind.

In the training process, five uncorrelated features are removed from the dataset. Two algorithms, Gradient Boosted Trees (GBT) and Neural Net (NN), are used in this paper because the previous work has described Thailand PM.5 prediction using these algorithms. The set of tuning parameters is left by default using RapidMiner 9.6. For GBT, number of trees = 50 with learning rate = 0.01. For NN, training cycle = 200 with learning rate = 0.01. In addition, 10 folds cross-validation with automatic sampling type is used.

Throughout this paper, gradient color codes are used to indicate how good the results are. Green-colored fields indicate the higher accuracy and correlated values and red-colored fields indicate the opposite.

The clustering and agents training results show in Table 1

Table 1: Distributed Agents in Training Process

Agents	Train Dataset (0.8, 1906, F=15)		
	Data Instances (N)	GBT (RMSE)	NN (RMSE)
$D: a_s$	1906	13.50	7.21
$c_0: a_0$	1162	6.56	5.25
$c_1: a_1$	256	17.14	9.62
$c_2: a_2$	488	12.28	7.00
s: single traditional model $C = \{c_0, c_1, c_2\}$ $Distributed\ Agents, A = \{a_0, a_1, a_2\}$			

Table 1 proves a better performance, measured by Root Mean Square Error (RMSE), on agents, a_0 and a_2 than on a traditional single model, a_s, in both algorithms. It may indicate the abnormal data pattern in c_1 which causes variation of performance for a traditional model and assumption of removing these instances are not acceptable by business scenario. In other words, 13.43% of the instances drop performance of agent for from RMSE 6.56 to 13.50, and it seems to be unacceptable tradeoffs.

Establishing a network of agents, we explore the weights of features based on GBT variable importance percentage as shown in Table 2

Table 2: Variable Importance in agents based GBT

Variable Importance	$D: a_s$	$c_0: a_0$	$c_1: a_1$	$c_2: a_2$
AQI	0.762	0.689	0.123	0.282
CO	0.114	0.004	0.683	0.202
PM10	0.075	0.139	0.069	0.112
O3	0.017	0.003	0.041	0.018
PRESSURE	0.009	0.095	0.035	0.002
PRECIP	0.007	0.000	0.013	0.000
HUMIDITY	0.006	0.020	0.019	0.008
CLOUD	0.005	0.018	0.007	0.008
SO2	0.005	0.011	0.000	0.357
NO2	0.000	0.001	0.003	0.004
WIND	0.000	0.011	0.001	0.001
GUST	0.000	0.001	0.001	0.003
TEMP	0.000	0.006	0.005	0.003
TEMP_FEEL	0.000	0.002	0.000	0.000
TIME_POINT	0.000	0.000	0.000	0.000

By examining variable importance in Table 2, both of three highest importance features and lowest importance features were examined giving a clue that the criteria as feature weight on the prediction of PM2.5 differ for each agent. The data in each cluster should require a specifically trained agent.

Recall, an establishing network of agents based GBT algorithm, Overmind determines the similarity measurement of agents using variable importance. A Euclidean Distance Measurement is used to find the distance among agents. In Table 3, the similarity is measured by using the equation:

$$Similarity(s, d) = 1 - Distance(s, d) \quad (1)$$

where s is source node, d is destination node.

Table 3: Agent Similarity Measurement

Source	Destination	Distance	Similarity
$D: a_s$	$c_0: a_0$	0.17	0.83
$D: a_s$	$c_1: a_1$	0.86	0.14
$D: a_s$	$c_2: a_2$	0.60	0.40
$c_0: a_0$	$c_1: a_1$	0.89	0.11
$c_0: a_0$	$c_2: a_2$	0.58	0.42
$c_1: a_1$	$c_2: a_2$	0.62	0.38

Setting a threshold $y = 0.3$ means the edges will be created when the similarity between nodes is higher or equal to 0.3. However, this threshold is an adjustable parameter, the lower the threshold the higher chance for connected agents, the similarity of agents affects and shapes the transfer process. Where connection exists, a pair of the agent can communicate and transfer both data and feature, in another hand, isolated agent requires specially and specifically consideration. In this study, 0.3 is represented for the demonstration that creates linkages between agents unless the transfer process cannot be illustrated. Considering a tradition model, an agent in a network will results in Figure 4, however, this network will be examined again in the feature transfer section.

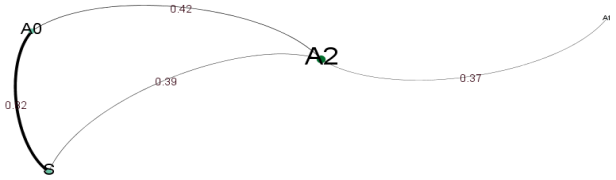


Figure 4: Network of Agents (|F|=15, y =0.3)

In the testing process, 477 instances of the dataset were distributed to specific agents based on the data similarity measurement. Moreover, the prediction assessments on both Gradient Boost Tree (GBT) and Neural Network (NN) were shown in Table 4.

The results show high accuracy when distributed agents perform a prediction over the associated clustered dataset. Undeniably, the traditional model is flexible and stable for input data. Perhaps, it should be considered as a tradeoff. We calculate average performance for distributed agents as:

$$Acc = \frac{\sum_0^i RMSE(a_i, c_i)}{i} \quad (2)$$

where $RMSE(a_i, c_i)$ is an accuracy of prediction by a_i on dataset c_i

Table 4: Test Results

Agent	Trained By	Test Dataset (0.2, n=477)					
		GBT (RMSE)			NN (RMSE)		
		c_0	c_1	c_2	c_0	c_1	c_2
		294	55	128	294	55	128
a_s	477	8.28	22.87	11.25	4.29	8.69	6.99
a_0	294	5.80	36.46	18.51	3.99	22.41	12.34
a_1	55	38.64	17.04	21.40	12.30	7.80	12.04
a_2	128	19.32	22.34	11.00	13.31	15.36	7.23

The GBT has achieved average performance RMSE of 11.28 whereas the traditional model has obtained RMSE of 14.13. On the other hand, the NN achieve the RMSE of 6.34, while the traditional model achieves at 6.65. These performances described in Figure 5.

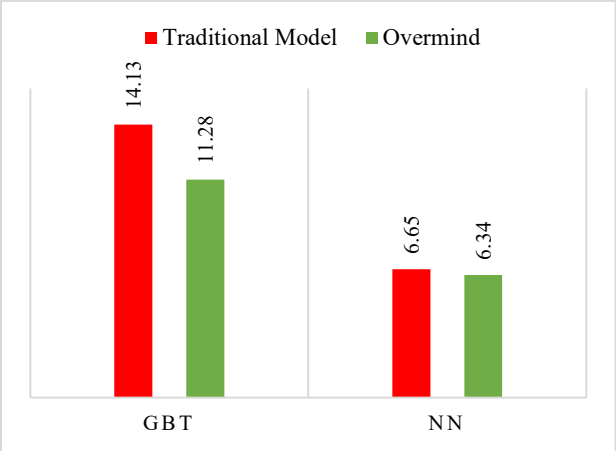


Figure 5: Average Performance, measured in RMSE.

5.2. Demonstration on Transfer Learning for Feature Transfer

In this demonstration, one correlated feature from a training dataset is removed, but a testing dataset remains with full features. Based on variable importance in Table 2, AQI is assumedly removed from the training dataset. The results of newly trained agents are shown in Table 5.

Table 5: Distributed Agents in Training Process, |F|=14

Agents	Train Dataset (0.8, 1906, |F|=14)		
	Data Instances (n)	GBT (RMSE)	NN (RMSE)
D: a_s	1906	14.28	9.97
c_0: a_0	553	15.72	11.96
c_1: a_1	234	16.20	12.45
c_2: a_2	1119	8.17	7.05

The performance of two minor clustered (n =553, and n = 234) significantly drops. However, it is not yet comparable results where |F|=15 because the clustering process may yield a different output. Base on this experiment, a same threshold value, $y = 0.3$ will result isolated a_2 which indicates that the data in c_0 and c_1 should be removed from the demonstration and only one agent required in this scenario, it is a_2; the demonstration will be then bias, hence, we lower a threshold to $y = 0.2$ and the result of networks of agents shows in Figure 6.

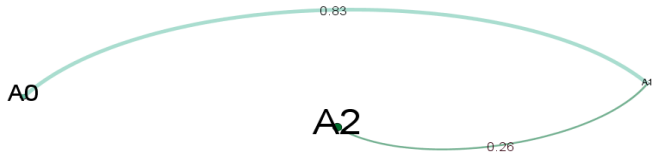


Figure 6: Network of Agents (|F|=14, y =0.2)

Figure 6 shows that a centrality of network is a_1, this means centrality of the network is determined neither by agent performance nor numbers of instances. The centrality of the network is responsible for sharing and transferring the discovered both data and feature.

Demonstration on a discovered new dataset and feature (n=477, |F|=14+1), assumedly, AQI is a new feature. Overmind firstly determines the need for AQI data if it can improve the performance of the overall agent or not. Overmind takes sampling proportional data from clusters in a network, then building a model to evaluate performance against a new dataset.

Table 6: Comparing Results for a Transfer Learning

Agent	No. of instance (n)	No. of feature (|F|)	GBT (RMSE)	NN (RMSE)
Existing Dataset: acc_{ext}	477	14	14.56	10.12
New Dataset: acc_{new}	477	14+1	12.43	7.90

Table 6 shows a better performance with a new feature, hence, a transfer learning should be accepted for the new feature, AQI.

An existing dataset in a network contains no value of AQI, and the business scenario does preserve existing dataset. Hence, an imputation model is built by using a new dataset for training process (n=477, |F|=14+1). Eventually, an existing dataset is imputed for a new feature, AQI; it is deployed to the centrality of the network, all agents accept AQI as a new feature because the comparison identifies AQI can improve performance as shown in Table 7.

Table 7: Distributed Agents in Training Process, |F|=14+1 (imputed AQI)

Agents	Train Dataset (0.8, 2383, |F|=14+1 imputed AQI)		
	No. of instances (n)	GBT (RMSE)	NN (RMSE)
D': a'_s	1906	13.39	6.85
c'_0: a'_0	553	14.19	7.10
c'_1: a'_1	234	14.55	9.55
c'_2: a'_2	1119	7.03	4.71

Overmind update a network of agents based on changing in similarity measurement among agents.

A re-connected nodes between a'_2 and a'_0 is formed, this results in a fully connected network, and the centrality of the network is shifted to a'_2 because it has the highest performance and controlling a majority of data in the network (N=1119). Overmind then continues data transfer process as described in (A). The data prediction results are shown in Figure 7.

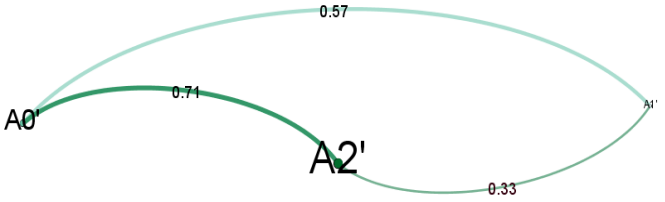


Figure 7: Network of Agents (F=14+1, Y=0.2)

Table 8: Test Result after feature transfer process

Agent	Test Dataset (0.2, 477, |F|=14+1)						
	No. of instance	GBT (RMSE)			NN (RMSE)		
		c_0	c_1	c_2	c_0	c_1	c_2
a'_s	477	14.44	17.02	11.92	7.11	13.11	6.36
a'_0	162	13.25	18.16	17.23	7.07	16.9	11.79
a'_1	37	19.73	13.61	25.73	19.81	10.38	15.64
a'_2	278	18.39	21.18	13.22	11.64	16.21	8.62

The comparison shows average performance in RMSE using GBT with 13.36 in this testing dataset, while a traditional model would give an accuracy, 14.46. For NN, the average performance in RMSE is 8.69, and 8.86 for a traditional model. This comparison is described as shown in Figure 8.

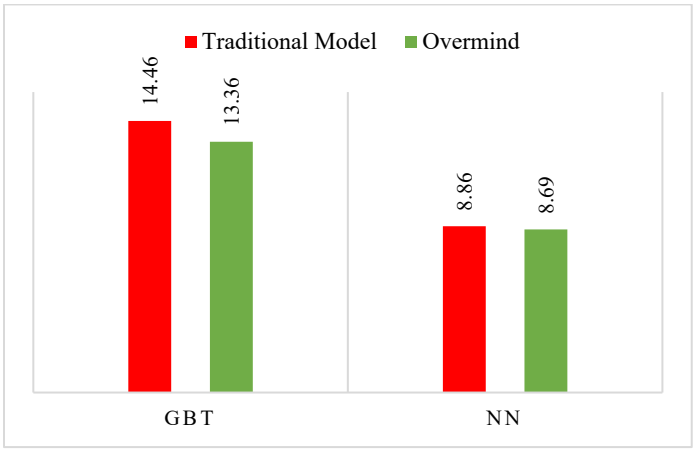


Figure 8: Average Performance, measured in RMSE after feature transfer process

Figure 8 shows penalty in transfer process caused by the imputation model that increases the RMSE after transfer process as shown in Figure 9. However, it should be reminded that demonstrating feature transfer is taken place on the important variable which is the AQI.

To verify the consistency of this approach, the progression of the framework functional processes, Determine Data Clustering, Assign Agent, and Establish Network has been performed. However, in the second round, the dataset is slightly different because (1) 1,906 instances (80%) of training dataset have imputed AQI values, and (2) 477 instances (20%) of the testing dataset have true values of AQI. To avoid confusing, the agents in the second execution is denoted by $a_{x:2}$ where x is a number of clusters produced by this execution.

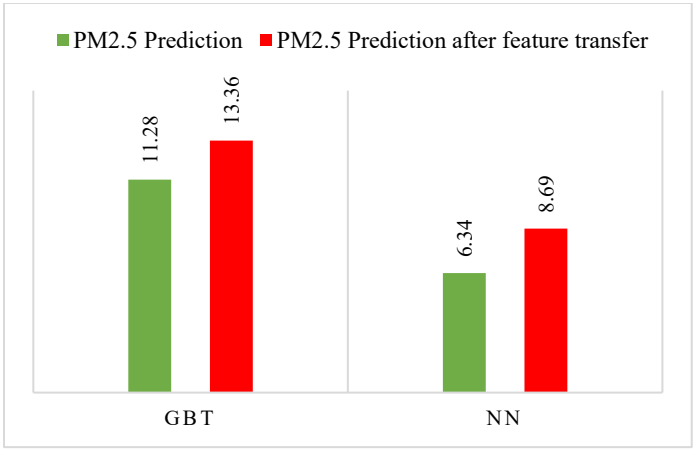


Figure 9: Penalty after feature transfer process

Table 9: Distributed Agents in Training Process by Second Execution

Agents	Train Dataset (1, 2383, |F|=15)		
	No. of instances (n)	GBT (RMSE)	NN (RMSE)
$D: a_{s:2}$	2383	13.05	6.95
$c_0: a_{0:2}$	1458	6.24	4.22
$c_1: a_{1:2}$	316	17.11	11.27
$c_2: a_{2:2}$	609	11.90	6.44
s: single traditional model $C = \{c_0, c_1, c_2\}$ $Distributed\ Agents, A = \{a_{0:2}, a_{1:2}, a_{2:2}\}$			

The result in Table 9 shows consistency over the clustering process which still produce three clusters with similar proportions. Moreover, the overall performance of agents is consistency and even slightly improved. Examining agent similarity using variable importance between the first and second executions are shown in Table 10 and Figure 10. The comparison in Figure 10 shows changes in variable importance, but the agents are keeping their characteristics over the same set of features.

Eventually, a network of agents is re-established as shown in Figure 11; a network of agents also preserves its structure where the centrality of network remains unchanged. However, if the similarity threshold is recalled in the very first demonstration of establishing a network where $y = 0.3$, the isolated agent $(a_{1:2})$ may occur which is assigned to the minority of data (n=316) that contribute poor performance.

Table 10: Variable Importance in agents based GBT by Second Execution

Variable Importance	$D: a_{s:2}$	$c_0: a_{0:2}$	$c_1: a_{1:2}$	$c_2: a_{2:2}$
AQI	0.762	0.689	0.123	0.282
CO	0.114	0.004	0.683	0.202
PM10	0.075	0.139	0.069	0.112
O3	0.017	0.003	0.041	0.018
PRESSURE	0.009	0.095	0.035	0.002
PRECIP	0.007	0.000	0.013	0.000
HUMIDITY	0.006	0.020	0.019	0.008
CLOUD	0.005	0.018	0.007	0.008

Variable Importance	$D: a_{s:2}$	$c_0: a_{0:2}$	$c_1: a_{1:2}$	$c_2: a_{2:2}$
SO2	0.005	0.011	0.000	0.357
NO2	0.000	0.001	0.003	0.004
WIND	0.000	0.011	0.001	0.001
GUST	0.000	0.001	0.001	0.003
TEMP	0.000	0.006	0.005	0.003
TEMP_FEEL	0.000	0.002	0.000	0.000
TIME_POINT	0.000	0.000	0.000	0.000

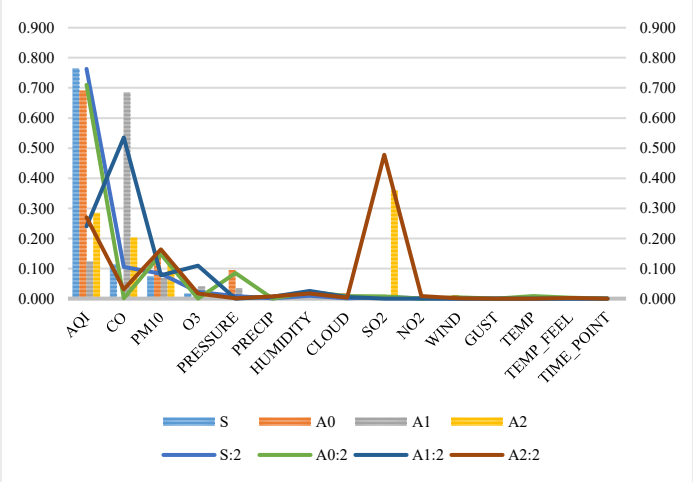


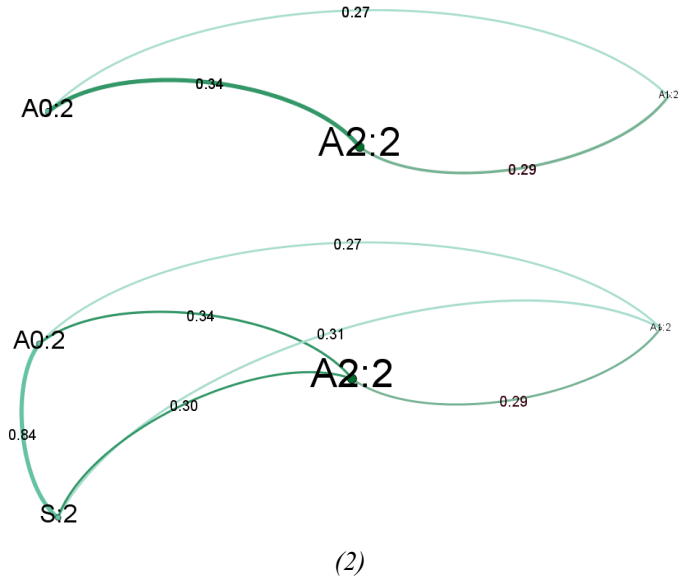
Figure 10: Comparison of Agent Variable Importance based GBT by first and second executions

This scenario implies that the business may consider, (1) a separate studying of agents in order to build a specific algorithm for prediction over the associated dataset, (2) removing associated data instances out of dataset in order to reduce inconsistency and improve prediction performance, (3) reducing a threshold (y) in order to keep particular agent connected to a network because this agent can be a specific dataset where a business foresees this data will be increasing over time in soon future.

6. Results and Discussion

Till now, demonstrations of both framework performance, and the transfer process are explained. The summary of the performance of Overmind with four comparing scenarios in cross-validation agents' average performance can be concluded as follows:

1. Scenario I: Overmind assigns distributed agents on a fully completed dataset.
2. Scenario II: Overmind assigns distributed agents on a dataset where importance attributes is removed.
3. Scenario III: Overmind assigns distributed agents to demonstrate a situation where a new feature is discovered and transferred, and
4. Scenario IV: Overmind assigns retrained distributed agents and restates its network



(2)

Figure 11: (1) Network of Agents in Second Execution (F=15, Y=0.2) without Traditional Model, (2) Network of Agents in Second Execution (F=15, Y=0.2) with Traditional Model,

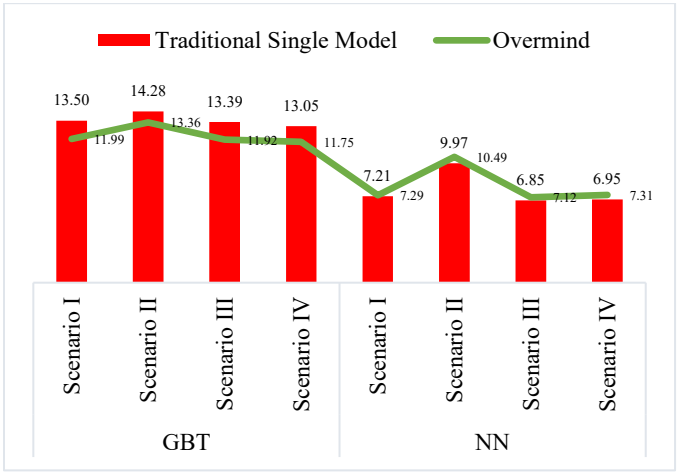


Figure 12: Comparison of Average Performance in Cross-Validation Agent Generated Models measured in RMSE

Figure 12 explains, for all scenarios, the cross-validation performance in Overmind generated models is impressive compared to a traditional single model under GBT algorithm. In the other side, the performance is not satisfied when compared to the traditional model under NN algorithm.

To summarize testing performance in testing data, we divide into two scenarios, (1) Scenario I: Overmind assigns distributed agents on a fully completed dataset, (2) Scenario II: Overmind assigns distributed agents to demonstrate a situation where a new feature is discovered and transferred. The results are shown in Figure 13.

Figure 13 indicates a better testing performance in all scenarios, but the performance is slightly better in the NN algorithm. In sum, Overmind can achieve its primary intended design and concept, this achievement can be demonstrated through our previous dataset for PM2.5 prediction in Thailand.

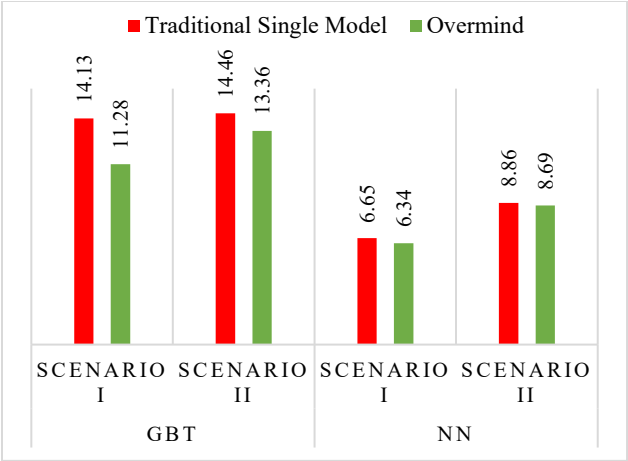


Figure 13: Summary and Comparison on Testing Performance measured in RMSE

7. Conclusion

Overmind is a framework aiming to demonstrate and facilitate a decentralized machine learning-based network analysis. This paper proves the concept and its performance through the mixed dataset between Thailand PM2.5 dataset and weather forecast dataset.

In data prediction, this framework can significantly reduce the data bias scenario, where input data is uncontrollable, variant, and clustered. Moreover, Overmind has proven its performance by comparing an average accuracy measured in RMSE on both traditional models (single machine learning) and distributed agents (multiple machine learning, agents). Overmind performs better performance than the traditional method in selected dataset using the GBT algorithm, reducing the RMSE from 14.13 to 11.28 (20.16% decrease). Furthermore, the RMSE is slightly reduced from 6.65 to 6.34 (4.6% decrease) in the NN algorithm.

A feature transfer process is demonstrated as a conceptual design for data preservation. However, the evaluated performance yields penalty due to imputation method.

In sum, Overmind has successfully formed a collaborative decentralized machine learning framework serving as an alternative in the design of a machine learning system. Possible contributions of Overmind are:

7.1. Higher Performance and Unaffected by Data Bias.

Overmind proves its effectiveness as a prediction system in the selected testing dataset. The design of Overmind is not affected by data bias because distributed agents are designed to function over associated sub-dataset.

7.2. Distributed Resource Utilization

Overmind mimics the concept of a distributed system so that it inherits the advantage where resources are allocated in a decentralized manner. Each agent utilizes computational resources for processing associated input instead of processing the whole dataset. This contribution also involves parallel processing and queue system where particular agents may transfer their queued inputs, as a bottleneck, to similar connected agents in a network for parallel processing where penalty inaccuracy can be traded off. This parallel processing for Overmind is already considered and planned for future work.

7.3. Dynamic Models in Collaborative Manner

Overmind demonstrates dynamic machine learning models for feature discovery and transferring across collaborative network analysis with data preservation. The concept of similarity measurement and network of agents has innovated methods in a dynamic framework. There are still opportunities in exploring the aforementioned methods to study optimization parameters and techniques.

8. Future Work

A study on optimization is recommended on Overmind's parameters. An interpretation over a network of agents formed by Overmind is an interesting topic to be explored and described how it contributes significances to business viewpoints.

In addition, the enterprise system architecture of Overmind can be improved and enhanced for higher performance and resilience such as queueing system and redundancy in agent level For example, when one agent is in abnormal stage, data can be continuously processed by neighbors or connected agents in the network; this will reduce the impact of disruption and error in artificial intelligence system.

Acknowledgment

We would like to thank the members of Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, for their guidance and suggestions. Importantly, we would also like to thank Bangkok Air Quality for supporting data.

References

[1] P. Sakul-Ung, P. Ruchanawet, N. Thammabunwarit, A. Vatcharaphrueksadee, C. Triperm, M. Sodanil, "PM2.5 Prediction based Weather Forecast Information and Missingness Challenges: A Case Study Industrial and Metropolis Areas," in 2019 Research, Invention, and Innovation Congress (RI2C), 1–5, 2019,doi:10.1109/RI2C48728.2019.8999941.

[2] K.F. Li, "Engineering Distributed Intelligence: Issues and Challenges," in 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems, 404–408, 2014, doi:10.1109/CISIS.2014.56.

[3] L. Eduardo, C. Hern, "On distributed artificial intelligence," The Knowledge Engineering Review, **3**(1), 21–57, 1988, doi:10.1017/S0269888900004367.

[4] G.M.P. O'Hare, A.W.J. Chisholm, "Distributed Artificial Intelligence: An Invaluable Technique for the Development of Intelligent Manufacturing Systems," CIRP Annals, **39**(1), 485–488, 1990, doi:https://doi.org/10.1016/S0007-8506(07)61102-0.

[5] B. Chaib-draa, "Industrial applications of distributed AI," **38**(11 %J Commun. ACM), 49–53, 1995, doi:10.1145/219717.219761.

[6] S. Kirn, "Organisational intelligence and distributed AI," No. 40. Arbeitsberichte des Instituts für Wirtschaftsinformatik, Westfälische Wilhelms-Universität Münster, 1995.

[7] Y. Feng-Chao, H. Yu-Kuen, "Design of a distributed artificial intelligence system shell," in Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation, 192–195 vol.1, 1993, doi:10.1109/TENCON.1993.319961.

[8] S. Sen, "Multiagent systems: Milestones and new horizons," Trends in Cognitive Sciences, **1**(9), 334–340, 1997, doi:10.1016/S1364-6613(97)01100-5.

[9] G. Weiß, L. Steels, "Distributed Reinforcement Learning," Springer Berlin Heidelberg, Berlin, Heidelberg: 415–428, 1995.

[10] D. Adjodah, Decentralized Reinforcement Learning Inspired by Multiagent Systems, Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, 1729–1730, 2018.

[11] E. Ogston, B. Overeinder, M. van Steen, F. Brazier, "A method for decentralized clustering in large multi-agent systems," in Proceedings of the second international joint conference on Autonomous agents and multiagent systems - AAMAS '03, ACM Press, New York, New York, USA: 789, 2003, doi:10.1145/860575.860702.

[12] Y. Hu, D. Niu, J. Yang, S. Zhou, FDML: A Collaborative Machine Learning Framework for Distributed Features, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2232–2240, 2019, doi:10.1145/3292500.3330765.

[13] K.D. Pandl, S. Thiebes, M. Schmidt-Kraepelin, A. Sunyaev, "On the Convergence of Artificial Intelligence and Distributed Ledger Technology: A Scoping Review and Future Research Agenda," IEEE Access, **8**, 57075–57095, 2020, doi:10.1109/ACCESS.2020.2981447.

[14] Y. Lin, Y. Lin, C. Liu, "AItalk: a tutorial to implement AI as IoT devices," IET Networks, **8**(3), 195–202, 2019, doi:10.1049/iet-net.2018.5182.

[15] X. Lin, J. Li, J. Wu, H. Liang, W. Yang, "Making Knowledge Tradable in Edge-AI Enabled IoT: A Consortium Blockchain-Based Efficient and Incentive Approach," IEEE Transactions on Industrial Informatics, **15**(12), 6367–6378, 2019, doi:10.1109/TII.2019.2917307.

[16] H.F. Rashvand, K. Salah, J.M.A. Calero, L. Harn, "Distributed security for multi-agent systems - review and applications," IET Information Security, **4**(4), 188–201, 2010, doi:10.1049/iet-ifs.2010.0041.

[17] X. Zhao, R. Jiang, "Distributed Machine Learning Oriented Data Integrity Verification Scheme in Cloud Computing Environment," IEEE Access, **8**, 26372–26384, 2020, doi:10.1109/ACCESS.2020.2971519.

[18] S.J. Pan, Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, **22**(10), 1345–1359, 2010, doi:10.1109/TKDE.2009.191.

[19] Z. Wang, X. Ye, C. Wang, P.S. Yu, "Feature Selection via Transferring Knowledge Across Different Classes %J ACM Trans. Knowl. Discov. Data," **13**(2), 1–29, 2019, doi:10.1145/3314202.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 291-298 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Based on Reconfiguring the Supercomputers Runtime Environment New Security Methods

Andrey Molyakov*

Institute of information technologies and cybersecurity, Russian State University for the Humanities, Moscow, 117534, Russia

ARTICLE INFO

Article history:
Received: 09 April, 2020
Accepted: 18 May, 2020
Online: 29 May, 2020

Keywords:
Kripke structures
Sandbox
Stationary and on-board
Supercomputer systems
CTL logic

ABSTRACT

This paper is an extension of work originally presented in 2019 Third World Conference on Smart Trends in Systems Security and Sustainability (WorldS4) [1]. Author describes two new methods: reactive protection method (without delay after detecting an attack), which consists in virtualizing the execution environment of supercomputers processes if the calculated state descriptor falls into the "risk" zone and based on monitoring requests for allocation of resources in accordance with the rules of the security policy in the form of temporal modal structures CTL logic and method for reconfiguring the runtime environment of the supercomputers taking into account the mobility requirements (built-in computations) based on the application of the trajectories of computing state security descriptors on Kripke structures. The methods develop a number of provisions of the theory of information security, based on the new concept of Information Security of stationary and onboard supercomputer computing systems as a calculated convolution of the states of the execution environment (hardware or virtual) and system software.

1. Introduction

Widely used computers of various classes are essentially systems with von Neumann architecture, and the program execution model is a universal Turing machine with a tape of calculations, left and right shifts.

The performance criterion in the classical Popek - Goldberg theory for Turing machines is ambiguous. In the original work it was formulated twice [1]:

a) "a statistically prevailing subset of virtual processor instructions must be executed directly by the physical processor, without the intervention of a virtual machine monitor";
b) "all harmless instructions are executed directly by the physical processor, without the intervention of a virtual machine monitor".

These conditions are not identical. They are identical only if most of the instructions executed by the virtual machine are harmless, which cannot be guaranteed in the general case. In proving sufficient conditions for constructing a virtual machine monitor, Popek and Goldberg use the second definition.

Since the 70-80s, electronic technology has undergone significant changes. With the advent and development of hardware virtualization successfully implemented [2, 3]:

• *SMEP (Supervisor Mode Execution Prevention*). Prevention of code execution in supervisor mode) is a technology developed by Intel to protect your computer from hacker attacks and other threats that use the so-called "supervisor mode".

• *SMAP (Supervisor Mode Access Prevention)*. It prevents writing to memory and reading code from it that unauthorized uses supervisor mode, while SMEP only prevents the execution of this code.

Supervisor mode is the preferred processor mode used by the kernel of the operating system. This mode is also called kernel mode. The opposite is the user mode in which user applications work [4, 5].

For SMEP to work, in addition to the corresponding processor, a suitable operating system is required. However, SMEP, although significantly complicating the task of hacking the system, still does not guarantee its complete protection. Therefore, later (in Broadwell architecture processors), in order to increase security and protect against vulnerabilities that were not resolved by SMEP, SMAP technology was additionally introduced [6, 7].

Direct execution of guest instructions whose encodings correspond to the privileged instructions of the physical processor on the host is impossible, since they certainly cause a trap to be thrown when executed in the user mode in which virtual machines are running. For example, the encodings of new secure instructions for Intel-64 architecture usually correspond to invalid encodings on processors

*Andrey Molyakov, Moscow, 117534, Kirovogradskaya street, 25/2, Russia. Tel: 8-495-388-0888 & andreimolyakov@mail.ru

of previous generations of this architecture. That is, new instructions are usually privileged on processors of previous generations, because they cause exceptions [8, 9].

Many hardware processors and various research groups are interested in supporting hardware transactional memory [10]. IBM was the first to develop a system with transactional memory hardware support in IBM Blue Gene [11] supercomputers and z-Enterprise EC12 servers [12]. AMD announced the development of the Advanced Synchronization Facility (ASF) extension [13], which is a variant of the hardware transactional memory for the Intel 64 architecture. Sun Microsystems has developed a processor code-named ROCK [14], but this project was closed. The Intel® Transactional Synchronization Extensions (Intel® TSX) instruction set [15], consisting of Intel® Hardware Lock Elision (Intel® HLE) and Intel® Restricted Transactional Memory (Intel® RTM) extensions, has been added to the sixth generation Intel Core processors [16].

There are also many software implementations of transactional memory ideas that do not require special support from the hardware. However, the study of existing solutions shows their extremely low productivity [17].

Moreover, with the advent of hardware transactional memory, the Intel RTM extension allows you to perform nested transactions, but at the same time, all internal XBEGIN and XEND pairs should not create new save points, that is, rollback of a nested transaction leads to a rollback of the entire transaction chain. The introduction of transactional memory extensions implies that the semantics of all instructions working with memory, as well as instructions affecting the state of the processor that are not stored in savepoints, must be changed in order to support the new execution mode.

In order to avoid critical errors, it is necessary to implement a multi-level "sandbox" in which the agents of the hypervisor function and monitor the execution of processor processing instructions at all levels of the hierarchy.

2. New actual requirements for information security

The Based on the Theorem on a protected hypervisor OS, formulated and proved by Zegzhda, two requirements can be formulated [18]:

a) *Requirement 1 for virtualization of protected resources*: in a supercomputer computing system, all operations of application OS applications on protected resources must be carried out by virtualizing the resource in a virtual application environment;

b) *Access control requirement 2*: in a supercomputer computing system, all operations of application OS applications on protected resources must be controlled by hypervisor protection tools.

Further developing the idea, one can formulate theoretical premises for the solvability of the supervenience problem: all possible operating modes of the processor should be considered - the real hardware mode, the protected hardware mode, the real virtual machine mode, the protected virtual machine mode. This means that the multi-domain protected hypervisor must ensure the correct execution (software using the interpreter or hardware) of all the instructions of the guest system, as well as guarantee isolation of virtual machines. At the same time, all safe instructions common to the guest and host systems are executed directly by the physical processor (privileged modes).

The virtual machine monitor operates in mode 0 (in a separate domain), therefore, any interpretation errors, memory leaks, incorrect pointers, addressing errors of data code segments, program codes, and page addressing errors do not lead to an abnormal termination.

The solution to the problem of providing a modified instruction mnemonics for processors of different generations in order to support hardware transactional memory is to prevent the direct execution of these instructions in a virtual environment - they must be simulated programmatically on virtual processors (unprivileged modes).

Taking into account the development of the element design base and hardware technologies, we can formulate ***new requirements for information protection*** [18]:

a) Protective equipment should control all informational interactions without exception;

b) Security features should be developed independently of application programs and rely on an abstract representation of information interactions;

c) Protective equipment should control information interactions based on clearly defined rules that make up the formal model;

d) A mechanism should be provided to assess the safety of both the current state of the system and predict the safety of future states.

The principle of homogeneity of memory is the main postulate of von Neumann architecture. This means that for supercomputer systems in which record high performance is important, a fundamentally new architecture is required. Sharing the bus for program memory and data memory leads to a bottleneck in von Neumann's architecture, namely limiting the bandwidth between the processor and memory compared to the amount of memory. Due to the fact that program memory and data memory cannot be accessed at the same time, the processor-memory channel bandwidth and memory speed significantly limit the processor speed. Storage of data and commands in different places solves the problem of the "memory wall" on highly loaded computational problems [1].

3. Fundamental scientific problem

Over the past 25 years, OS security tools (for example, SMEP, SMAP, PaX, ExecShield, ASLR, DEP, Flusk, Patchguard and etc.) have come a long way, but failed to provide protection against current threats. The fundamental problem for this field of knowledge (computer science and computer technology) is known as the ***problem of supervenience***: taking into account the high asynchrony and multi-connectedness of processes performed with supercomputer systems, the huge amount of processed data (Petabytes of data) there is no logical correspondence between changes in the programs of the runtime environment of supercomputers and changes hardware components in terms of implementing an isolated environment and monitoring security policy rules. For the chosen class of speakers, the solution to the fundamental problem is the development of the theoretical, scientific, practical and organizational and technical foundations

of reactive protection of stationary and onboard supercomputers, which consists in virtualizing the execution environment of the processes of applied and system software, in the interest of increasing their security and reliability, as well as substantiation of the principles for constructing mechanisms for implementing such protection for next generation automated systems.

The main contradiction for this class of computing systems is the ratio of energy intensity, speed and security. With the introduction of additional functional elements using the example of a hypervisor and a transactional memory controller, the characteristics of performance and energy consumption change. Concurrently executed queries conflict when they read and modify a certain database element, and the resulting conflict can lead to an erroneous result that could not be obtained if these queries were executed sequentially. Transactional memory provides a lightweight transaction mechanism for control flows running in a shared address space. It guarantees atomicity and isolation of parallel tasks [10, 11].

Atomicity ensures that changes in the state of a program made by code that is executed in a transaction are invisible from the point of view of other transactions executed in parallel [12].

Isolation ensures that concurrent tasks do not affect the outcome of the transaction, so that the transaction produces the same result as if no other task was being performed. Transactions provide the basis for constructing parallel abstractions, which are building blocks that can be combined without knowing their internal details, much like procedures and objects provide suitable abstractions for composing sequential code [13].

Integration with hardware transactional memory requires solving 5 main problems of computer science: limited application, debugging complexity, process synchronization and exclusion of access, resource control in conditions of parallelism and high asynchronous processes, emulation of different types of processors [14]. For multi-domain protected systems, trust is ensured by the following factors: *"Transparency"* – invisibility for the application OS, working directly with equipment, the amount of hypervisor code is small compared to the OS and applications, therefore it is easier to ensure that there are no vulnerabilities, all actions and events inside the virtualized systems are reversible - attacks can be quickly neutralized by rollback or reset.

Monitoring the exchange of a virtualized system with the external environment allows you to abandon a thorough study of the virtualized system itself.

Limitations associated with loss of productivity are extremely low, due to the record high performance characteristics of supercomputers.

4. Research methodology: new concepts and definitions

Information security is based on access control to objects of managing and guest OS, these objects can be attributed to different levels of protection. Examples of objects are directories, files, network sockets, registers, and interrupt handlers and some special-use memory areas by operating systems. The traditional access control approach involves the use of access attributes (rights) in requests to these objects for performing certain operations on them. If the verification of such attributes is successful, then access to the object at its security level is allowed, then the requested operation is performed on it. With this approach, it is technically possible to intercept a request and use its access rights in a substitute request aimed at malicious impact.

New concepts and definitions are introduced in the work: supersecurity, information security of stationary and onboard supercomputer computing systems, a descriptor for assessing the security of system states.

Supersecurity is a property of a supercomputer system based on self-configuration and self-control. By using the high-performance reserves of these supercomputers, you can create a hypervisor that provides reliable protection against attacks associated with high asynchronous processes of the supercomputers and the ability to perform false transactions as a result of destabilizing external influences and accidental hardware failures.

The information security of stationary and onboard supercomputer computing systems is formalized as "a calculated convolution of the states of the execution environment (hardware or virtual) and system software".

The descriptor for assessing the security of system states is a convolution that is "calculated" based on the attributes of ongoing processes that are implemented at the level of microprocessor cores and the interworking environment. All processes have their own descriptors calculated and methods for their processing — get, set, delete and etc. are specified.

5. New Safe Operations Model

The safe operations model includes 3 components: commands, data and timestamps. The space of operations in calculating the descriptors is a variety of Kripke's "worlds". Each world of Kripke is assigned its own digital double and form a knowledge base for machine learning. The class of operations that are characteristic of supercomputers is the operations of multiplication and division in the form of successive shifts in the tagged structures of packets of requests from clients, guest and control OSs.

From the point of view of category theory, we work with the Group object - the category group. Objects are a group in the form of a residue ring, morphisms are mappings preserving the group structure. Category theory studies concepts through how these concepts interact with each other. We forget how these concepts are implemented, and we only look at the properties of connections, abstracting from the type of processor architecture (scalar, vector, MIPS, classic x86_64, tile architecture like Tilera, mass-multithread, hybrid and etc.), processor capacity (32-bit, 64-bit, 128-bit and etc.).

A type refers to a class type when a type provides certain operations with a specific expected behavior. For example, the tau type may belong to the Functor class if it has a specific behavior similar to a collection:

a) The type **tau** is parameterized over another type, which you should consider as the type of the collection element. The type of the complete collection is then similar to Scheme_type = {Int, String, Bool and etc.};
b) If you contain integers, strings, or Booleans, respectively. If the element type is unknown, it is written as a parameter of type a. Examples include lists (zero or more elements),

type "Unknown" (zero or one element of type a), sets of elements of type a, arrays of elements of type a, all kinds of search trees containing values of type a, etc.

2. Another property that tau must satisfy is that if you have a function like a –> b (function on elements), then you should be able to use this function and product for a related function over collections. You do this with the fmap operator, which is shared by each type of Functor class. The operator is actually overloaded, so if you have an even function with type Int -> Bool, then

Definition 1. A functor is a kind of collection (a set of configurations for recursively computing a hash), for which, if you are provided with a function on elements, fmap will return the function in the collections.

Definition 2. The context of the operation is a set of tuples consisting of logical variables. A collection of collections is a conjunction of atomic predicates defined on many tuples.

I appeal to the theory of functional programming, we will define a functor for comparing is_equal of two hash values with support for different class type templates :

```
class is_equal

{ private: scheme_type v; public:

is_equal(scheme_type value) : v(value) {}

bool operator () (scheme_type x)

 { return x == this –>v; } };

scheme_type count_zero(const std::vector< scheme_type>& data)

{ return std::count_if(data.begin(), data.end(), is_equal(0)); },
```

scheme_type – data type (int, uint, float, double and etc.).

It is not possible to implement a complete hash function that is valid for all types. You cannot just convert an object to raw memory and hash bytes.

In addition, this idea fails due to padding technology when creating an index for each record. Because of this, it is necessary to take into account the context of a particular operation. To implement a universal convolution calculation algorithm, it is necessary to take into account different processor operating modes and the principle of type conversion and alignment of orders. We need to normalize the presentation of metadata operations in terms of aligning the boundaries of digital structures.

The basic principles of the algebraic structure on which the whole theory and methodology are based: The principle of Soft power, Self-organization, Supersecurity. The descriptors of the state safety assessment function are calculated on the set, which is a multiplicative group of the residue ring modulo 8.

Classical propositional logic is a "black and white" model; utterances are static, unchanged in time. In the ordinary propositional logic, sentences that do not explicitly or implicitly contain properties whose truth changes with time do not adequately formalize. We want to study and verify systems that evolve over time.

The proposed approach to the description of operations is based on the classification of risks of information security breaches and analysis of the context of the implementation of outgoing directives, through which data can be transmitted bypassing the requirements of the adopted security policy, which leads to a violation of the security of computing nodes resources.

The carriers of the analyzed operations are the sets of objects and access subjects to which various security levels (labels) are assigned. To control the security level of operations generating new entities, for example, operations, we will use the sign of immutability of the object generating the access entity.

6. Theoretical calculations of the reactive protection method

6.1. Algebra of operations with objects processed on supercomputers

For supercomputers the sign of immutability cannot be a constant indefinitely: simultaneously, a huge number of processes are launched. Temporal logic is needed to describe the states of subjects and access objects. We need a model of threats to the integrity of the execution environment of IC processes, in which instead of the classic link "subject, object, predicate" a new paradigm of writing security policy rules is implemented and new entities are defined - "subject", "object" and "descriptor for assessing state security": for each i-th threat, the value of the state security assessment function is calculated, the arguments of which are given in the form of a conjunction of predicates of eight logical variables, the subjects are the guest and control OS, and the objects are the components of the hype Sizer and transactional memory controller.

Temporal logic of branching time consider possible calculations (paths on a tree) - trajectories on a scan of the Kripke structures.

The Kripke structure is a transition system with labeled states and unlabeled transitions. Sweep defines infinite chains of states - possible calculations. Each state can have not one, but many chains - continuations, and is the root of its tree of stories (calculations).

The structure of Kripke M is the five M = (S, S0, R, L, AF).

In our case [19, 20], AF = {context_id | Dom_id | S | Ord | Context_type | TCU | TR}. Let an arbitrary formula Fi of CTL logic and a Kripke structure M. be given. For each subformula ψi of formula Fi, the marking algorithm performs the following steps:

a) We proceed to the construction of the numbering of the programs. Each operator is uniquely characterized by a pair - type (name) and a list of parameters, including operator labels, variables, functional and logical expressions. We only need to encode each of the possible values of the operator parameters. Labels of operators do not need special coding, since they are natural numbers;

b) Since each program calculates a certain function, the introduced Gödel numbering generates some numbering

of the functions. On the one hand, this numbering is not one-to-one, since we code the program syntax, and any syntactically different programs have different codes. On the other hand, every computable function is computed by an infinite class of programs.

However, such numbering has fundamental and practically useful properties. Let A be the countable class of functions for assessing the safety of supercomputers states: A = {f1, f2 ..., fn}.

Even though each of the fi functions is computable by some algorithm, this does not guarantee the existence of a single algorithm for computing all functions. We call a class A uniformly enumerable if there exists a two-place computable function F such that the class A consists exactly of functions of the form F (n, x) for some n from the set N.

We call the function a universal function of class A. nF (n, x) is the uniform numbering of class A.

Our main goal is to show that the numbering of the computable functions that we have determined is uniform with respect to other numbers and that by the number of the computable function in the given uniform numbering we can effectively find its number. Using the function for assessing the safety of SC states, we can subsequently accurately calculate the numbering states on the whole variety of Kripke structures, and then identify the classes of safe and dangerous operations.

Lemma 1. If A is an effective set, then for any effective set B: AB is effective, as well as any Cartesian product A1, A2 ... An of effective sets is effective, the set A *, A * = An, of all finite sequences of elements of A is effectively countable.

The proof of the statement is trivial and follows from the main theorem of arithmetic. We formulate and prove a theorem on the enumerability of the class of operations.

Theorem 1. Class A operations on the residue ring modulo m are uniformly enumerable, and the hash value calculation function F is a universal function of class A only if the number of hierarchy (nesting) levels is 8.

Evidence. In fact, operations are performed on data tuples, which are a set of i-th elements of an allergic structure (bit, atomic predicate, predicate conjunction, any set of predicative and functional symbols, vector variable, scalar, real number) of any dimension m (regardless of encodings and positional number system). We work with matrices, where the number of rows is equal to the number of nesting (hierarchies) of the computation space N, and the i-th column is an element of a structure of dimension m in the residue ring Z (m), i = 1 ... n. Using the Kornfeld formula to assess the confidence probability of security, each factor PΛi is a geometric decreasing progression [21].The most suitable values are 0.1 and 0. (1). Criteria of temporal logic (lack of new patterns over time) satisfies only the solution N = 8.

Moreover, nF (n, x) is an effective numbering of class A, since with a change in the dimension of the parameter n the regular relations between the elements of the set do not change, new properties of objects of our algebraic structure do not appear.

The main result of theoretical calculations is that we have proved universality and uniformity for given classes of operations of functional transformations. Regardless of the dimension of the input parameters, the nature of the relationships between the types, the type of operands, the type of processor architecture, the operation scheme and implementation of the algorithms, we have found such a number of hierarchy levels (N = 8) that there is an invariant in the algebraic system. If we evaluate the time parameters, this regularity property is preserved indefinitely (from 0 to + ∞).

6.2. Integration of the algebraic structure with architectural modifications of supercomputers

We proceed directly to the construction of a weighted multigraph-model for performing operations of supercomputers, which is a tree. The starting point is the top of the tree structure, describes a transactional memory controller that interacts directly with the hypervisor verifier module. The controller in conjunction with the verifier implements security mechanisms: an isolated multi-domain address space is represented as non-overlapping memory areas, each of which corresponds to a vertex of the graph lying in the second level of the root structure. The vertices of the third level correspond to the subsets of the components that make up the hypervisor. Since we have 8 levels of the request processing hierarchy, the block is divided into eight subsets of level details of possible states indicated by S8, ... , S1.

The last level of the root structure of a multigraph is represented by agents of the hypervisor (graph leaves) through which communication with the external environment and all possible attacks on the supercomputers occur. Advances to each next level are a higher level of abstraction of the description of request processing, a transition from the lower level of the specification to the upper. Each state can have not one, but many chains - continuations, and is the root of its tree of stories (calculations).

The reactive protection method should include not only deductive algorithms for calculating descriptors for assessing the security of states and identifying threats based on marking the states of fulfillment of supercomputers requests, but also implementing inductive learning based on self-diagnostics and explanatory decision-making mechanisms based on the concept of machine learning. We calculate such impacts that ensure the safe functioning of supercomputers, and block dangerous and suspicious ones.

7. Method for reconfiguring the runtime environment of the supercomputers

The hash is calculated recursively for all processes (subprocesses) and an effective value is obtained for the regular measurement of the query path (based on past experience, pre-training). In the process, the system is being trained. Each new configuration is classified and recognized in the future based on the characteristic set of markers. Coding of identification features - a set or conjunction of simple predicates on which the formula Fi is given [20]. The subformula ψ is a recursive call of the same algorithmic procedure, only with a different set of characteristic features and properties.

6 configurations of "colors" of the tree of computation histories. We select two colors: Blue - the predicate of the presence of the transient process, Red - the predicate of changing the type of the context of the operations performed.

a) configuration allows you to track changes in local states at the same hierarchy level and identify the launch of child processes as part of the base process with the markers of interest to us;
b) configuration analyzes the entire state tree based on the selected marker (the mode of a full run of calculations in case of restarting the task). The story tree is being rebuilt;
c) analyzes the change in two parameters in the near branches of the tree (local scale, in one domain);
d) configuration to track a single (irregular) parameter change in the far branches of the computation tree;
e) configuration regularly repeated changes of the selected identification parameter along the entire trajectory of movement;
f) configuration studies the branches of the tree; over time, in the future, changes in two parameters may appear on different trajectories. At the same time, processes can be started in different domains.

Sets of configurations adapted to different contexts of operations execution, in the form of matrices of access rules, are stored by agents of the hypervisor monitoring information security events in the memory of the transactional memory controller. In the process of training, the rules of safety rules are updated and adjusted. The system dynamically evolves and modifies sets of modal rules based on CTL temporal logic grammars for responding to signaling events. Identification criteria for threats in the semantic interpretation of a set of atomic predicates.

Thus, the hypervisor in conjunction with the transactional memory controller is a multi-agent system with machine learning.

There are deductive checks that correspond to the principles of classical logic, but we are using inductive algorithms with metadata about the operation of system components at different levels of the hierarchy, where, along with checking the integrity in the form of changing the values of the calculated hash functions for each process, access control is implemented in the form Labeling allowed transients when switching between security domains and thread migrations. The history of various system configurations is kept, new events are recognized due to a variant new set of predicates and the values of their trajectories on Kripke structures.

8. Discussion

8.1. Optimization, normalization of descriptor calculation

For optimization and normalization, it is necessary to specify an exactly safe data volume. The ***axiomatic of our algebraic structure*** are as follows:

a) Multiplication and division operations are specified (cyclic shifts to the right or left);
b) A lot of descriptor calculations is a residue ring modulo m = 8;
c) The ring must be symmetrical with respect to the performance of the operations of multiplication and division;
d) A request for any tap of a processor with a classic von Neumann architecture is a combined tagged structure in which a data segment and a program code segment are stored in the same RAM sections. In this case, the L-operand for the recursive calculation of ψi is stored in the data section, the R-operand for the recursive calculation of ψi is stored in the code section.

The principle of supervenience is the absence of differences of one kind in the absence of differences of another type: the absence of differences in the set of process descriptors in the absence of changes in the configuration of software and hardware agents. There is a one-to-one logical correspondence between a change in the set of a vector variable stored in the hypervisor generative tables, a map of transactional memory states (hardware level), and a hash function value (software level). In the process of passing the request, the key line does not change normally. If it changes, then it means that the route for passing the request by a third-party agent is being modified.

A-property – the identity of the values of the security assessment descriptors for the i-th process at the n-level of the hierarchy (running programs at the OS level, each process is identified by a hash value, formalized as a hash function value).

B-property – configuration of software and hardware components (formalized as a set of matrices in the form of 8 16-bit key lines, based on which the value of the function is calculated).

8.2. The solvability condition for the supervenience problem for supercomputer systems

In this case, one must take into account the upper or lower case of addresses (a segment of a data code or a command code, a combined format for representing the structure in RAM). Then the amount of precisely safe calculations doubles and you need to introduce an additional 2 * 4 = 8 register fields in the form of significant bits of the n-dimensional vector Ψ (n) = 8 to control the addressing and offset boundaries in the pipeline of processed requests and prevent the execution of "false transactions" ". Thus, the dimension of the control parameter k is 8. In this case, the bits (upper and lower registers) of the pointer to the RAM memory cells are not mixed.

If this condition is not fulfilled for the comparison functor, then the boundaries of the data segment and code are violated, the descriptor is recursively calculated: for N < 8, empty bits remain (you can write zeros or junk data in them), and normalize the processed data for N > 8 (BDC - format and type conversion of operands) is not possible.

To get this type of representation of a data set in the form of tuples of n-logical variables, the minimum set (basis) is 8. When the number of iterations is a multiple of 8, we increase the amount of data of recursive calculations, but the effects and patterns do not change. With a decrease in the number of iterations (less than 8 levels of the query processing hierarchy), automatic alignment of the processed data in the BDC format leads to the fact that

individual bits or groups of bits remain uninitialized, which allows dangerous shifts in the supercomputers.

The optimization of descriptor calculations is due to eight interval constraints in the processing cycle to obtain a productive sample, using only eight significant bits (rather than infinitely large numbers) for storing and processing the results. Based on the foregoing, we formulated the solvability condition for the fundamental problem of establishing logical correspondence (supervenience) of the program execution process on the supercomputer, by changing the state of the hypervisor components and supercomputers hardware: the number of levels of the request processing hierarchy by the hypervisor N in interaction with the transactional memory controller should be eight.

Solvability condition. The number of levels of the request processing hierarchy by the hypervisor N in interaction with the transactional memory controller should be eight.

A change in the states of processes at the upper level leads to a coordinated change in states at the lower level; there is a similarity relation. When using two rings of protection (N = 2) and in the absence of physically separated storage of programs and data, the process can gain access to the "alien" segment and increase privilege levels.

9. Results

Development, taking into accounts the specifics of supercomputers, of a fundamentally new technological solution creating an isolated program execution environment in the form of an 8-level sandbox with the implementation of control mechanisms both at the level of hypervisors and at the level of transactional memory controllers. Thus, an important scientific and technical problem has been solved in the field of creating information security tools for a new class of systems - stationary and on-board supercomputers. All the means of protecting information that existed today were the hardware and software components of the protected system itself, without integration with hardware transactional memory and the introduction of multi-level control, it was impossible to solve the problem of detecting and identifying various types of threats at all levels of the hierarchy of query execution. In the course of experimental studies, new scientific results were obtained that confirm the efficiency and minimal performance loss of applying hardware virtualization technology in the form of a multi-level "sandbox" for promising supercomputers compared to using traditional clusters.

Since it is impossible to control the operation of all equipment, but only to monitor the execution of requests at the level of the components of the hypervisor and the controller of transactional memory, during the research, the maximum level of functioning of information protection agents was found - S8. With this configuration, when the number of hierarchy levels is N = 8, the execution of context-sensitive operations becomes quasi-determined with a confidence probability of approximately 0.9.

The threshold for performance loss when using the verification complex is less than 6-7%. The effectiveness of the developed security system was assessed based on the use of various security tools (used to protect clusters and mainframes) and analysis of the number of successful recognitions and errors.

The following results of experimental studies are obtained:

a) with the number of training samples n > = 86, stable detection with critically minimal errors of the 1st and 2nd kind is ensured;
b) only 8 clusters are enough for effective detection of malicious code, which significantly reduces the cost of implementing a software and hardware solution. Each cluster is bound to a security domain. Their number is also 8.

Thus, an unambiguous correspondence has been achieved between the number of hierarchy levels and the number of protection domains and the number of countable clusters tied to a certain "sandbox" level. To ensure the requirements of the international standard ISO 5725 for the convergence and reproducibility of measurement data, the experiments were performed in a series of 5 repeated experiments for 14 days until the observed data were stabilized**.**

10. Conclusion

a) The theorem on the uniform enumerability of functions for assessing the safety of states is proved. In this case, the means of protection of stationary and onboard supercomputers are considered as an object of research for the first time;
b) The solvability condition is formulated, as a consequence of the theorem on the uniform enumerability of functions, the problems of establishing the logical correspondence (supervenience) of the program execution process on a supercomputer, by changing the states of the components of the hypervisor and equipment;
c) A reactive protection method has been developed (without delay after detecting an attack), which consists in virtualizing the execution environment of supercomputers processes if the calculated state descriptor falls into the "risk" zone and based on monitoring requests for allocation of resources in accordance with the rules of the security policy in the form of temporal modal structures CTL logic;
d) A method has been developed for reconfiguring the runtime environment of supercomputers taking into account the mobility requirements (built-in computations) based on the application of the trajectories of computing state security descriptors on Kripke structures;
e) It is proposed to use a model of threats to the integrity of the execution environment of supercomputers processes, in which instead of the classic link "subject, object, predicate" a new paradigm of writing security policy rules is implemented and new entities are defined - "subject", "object" and "descriptor for assessing state security" : for each i-th threat the value of the state security assessment function is calculated, the arguments of which are given in the form of a conjunction of predicates of eight logical variables, the subjects are the guest and operating systems, the objects are components of the hypervisor and transactional memory controller;

f) A safe operation model has been developed that considers, from the standpoint of "integrity" of the runtime environment, both supercomputers hardware at the microprocessor level and system software, describing its processes in the form of decomposition into an 8-level hierarchical structure: each process privilege level uniquely corresponds to a domain number protection and operation mode of the microprocessor.

References

[1] Molyakov A. S., "New security descriptor computing algorithm of Supercomputers" / 2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4), IEEE Xplorer Digital Library. https://doi.org/10.1109/WorldS4.2019.8903965

[2] Popek G. J., Goldberg R. P., "Formal Requirements for Virtualizable Third Generation Architectures" / Communications of the ACM, 1974.

[3] F. Leung, G. Neiger, D. Rodgers et al, "Intel® Virtualization Technology" : Intel Technology Journal., 2006, 10.

[4] AMD Corporation. — AMD64 Architecture Programmer's Manual Volume 2: System Programming, 2013.

[5] Stephan Diestelhorst, Martin Pohlack, Michael Hohmuth et al., "Implementing AMD's Advanced Synchronization Facility in an out-of-order x86 core" / 5th ACMSIGPLAN Workshop on Transactional Computing, 2010.

[6] Intel® Software Development Emulator, 2012. https://software.intel.com/en-us/articles/intel-software-development-emulator .

[7] Rechistov Grigory, Plotkin Arnold, "Implementation of Intel Restricted Transactional Memory ISA Extension in Simics" / Procedia Computer Science, 2013, 18 , 1804 – 1813.

[8] Herlihy Maurice, Moss J. Eliot B., "Transactional memory: architectural support for lock-free data structures" / Proceedings of the 20th annual international symposium on computer architecture, ISCA '93,New York,, USA : ACM, 1993, 289–300.

[9] Amy Wang, Matthew Gaudet, Peng Wu et al., "Evaluation of Blue Gene/Q Hardware Support for Transactional Memories" / Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques, PACT '12, 2012, 127–136.

[10] Jacobi Christian, Slegel Timothy, Greiner Dan, "Transactional Memory Architecture and Implementation for IBM System Z" / Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture, 2012., 12, 25–36.

[11] Jaewoong Chung, Stephan Diestelhorst, Martin Pohlack et al., "ASF: AMD64 Extension for Lock-free Data Structures and Transactional Memory" / Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture, 2010.

[12] Moir Mark, Moore Kevin, Nussbaum Dan, "The adaptive transactional memory test platform: a tool for experimenting with transactional code for ROCK" / Proceedings of the twentieth annual symposium on Parallelism in algoithms and architectures, SPAA '08, New York, USA : ACM, 2008, 362–362.

[13] Maurice Herlihy, Victor Luchangco, Mark Moir, William N. Scherer , "Software Transactional Memory for Dynamic-sized Data Structures" / Proceedings of the Twenty-second Annual Symposium on Principles of Distributed Computing, PODC '03, New York, USA : ACM, 2003, 92–101.

[14] Dice Dave, Shalev Ori, Shavit Nir, "Transactional Locking II" / Proceedings of the 20th International Conference on Distributed Computing, DISC'06, Berlin, Heidelberg : Springer-Verlag, 2006, 194–208.

[15] Milo M. K. Martin, Daniel J. Sorin, Bradford M. Beckmann et al., "Multifacet's general execution-driven multiprocessor simulator (GEMS) toolset" / SIGARCH Comput. Archit. News, 2005, Vol. 33, no. 4, 92–99.

[16] Yi Liu, Yangming Su, Cui Zhang et al..,Efficient Transaction Nesting in Hardware Transactional Memory / Architecture of Computing Systems - ARCS 2010, Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, 138–149.

[17] Moore Kevin E, "LogTM: Log-Based Transactional Memory" / In Proceedings of the Twelfth IEEE Symposium on High-Performance Computer Architecture, 2006, 258–269.

[18] .Zegzhda D..P., "Building secure operating systems based on virtualization technology" [presentation], St. Petersburg, 2018, 33 p. https://docplayer.ru/storage/77/76383820/1589898846/qi9gw3tca9toeeJG7VgqWQ/76383820.pdf

[19] Molyakov A. S., "A Prototype Computer with Non-von Neumann Architecture Based on Strategic Domestic J7 Microprocessor" / Automatic Control and Computer Sciences., 2016, 50(8), 682 -686.

[20] Molyakov A. S., "Token Scanning as a New Scientific Approach in the Creation of Protected Systems: A New Generation OS MICROTEK" / Automatic Control and Computer Sciences, 2016, 50(8), 687-692.

[21] Molyakov A. S., "Threat model and theoretical foundations of the reactive protection method of supercomputers" / Natural and technical sciences, Company Sputnik +, 2019, 7, 197–201.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 296-313 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Cluster Centroid-Based Energy Efficient Routing Protocol for WSN-Assisted IoT

Nalluri Prophess Raj Kumar*, Josemin Bala Gnanadhas

Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, 641114, India

ARTICLE INFO

Article history:
Received: 31 May, 2020
Accepted: 19 July, 2020
Online: 28 July, 2020

Keywords:
Internet of Things
Wireless sensor networks
Energy management
Basestation
Clustering
Energy Centroid

ABSTRACT

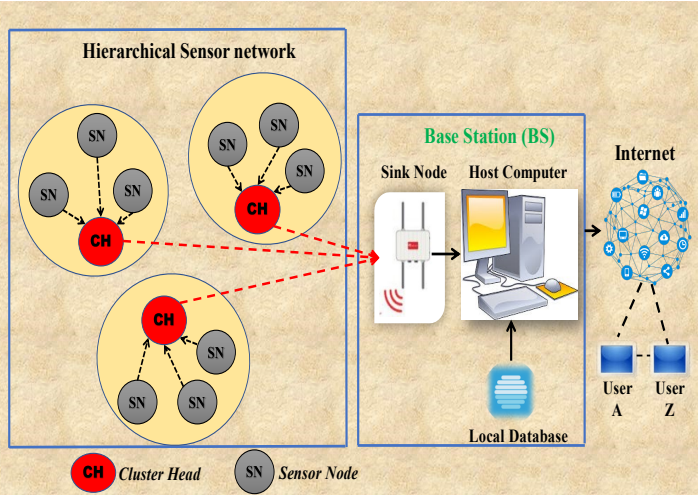
Wireless sensor network is highly resource constrained, where energy efficiency and network lifetime plays a major role for its sustenance. As the sensor nodes are battery operated and deployed in hostile environments, either recharging or replacement of batteries in sensor nodes is not possible after its deployment in inaccessible areas. In such condition, energy is the vital factor for the survival of sensor node in the sensing field. In order to increase the network lifetime and balance the energy consumption, robust routing protocols are required. The proposed network routing has three phases: 1. Network initiation phase to create a zone which enables the communication among local nodes 2. Zone co-ordinator selection phase algorithm to form zone cluster and its re-election procedure and 3. Zone head selection with its replacement phase based on energy centroid positional information and distance to the basestation to distribute load equally among zone co-ordinators, local sensor nodes. The data path between zone heads and basestation is distance centric and is optimized at one hop and dual hop levels to avoid data packet loss at zoneheads. Each zone is designed to own atmost ¼ rth of deployed sensor node count through uniform random deployment. Simulations results when basestation is placed inside sensing field indicates that the proposed network algorithm outperforms when benchmarked against similar protocols like conventional LEACH, Traditional PEGASIS, existing PRRP, ES3 protocols in terms of performance metrics like Network energy consumption, Average energy consumed by sensor node, Packet delivery ratio, Packet loss percentage and Network throughput.

1. Introduction

Internet of Things (IoT) is a concept to interlink the mechanism of the object with the internet and communicate with them to get identified by others. IoT defines the world where a physical object can be connected and communicated in an intelligent way. IoT can build the network comprising physical objects and provide them the capability to gather and share their own information with others. Wireless sensor network is a subset of IoT as it has the technology that can be utilized within an IoT system to meet the user demands. So, a sensor network assisting IoT, in an application using wireless network may have an advantage of lesser cost, friendly deployment of sensor nodes and better scalability. But a main drawback is that, its energy resources are hard to replace as they operate in rugged environment. This makes the energy management a key metric in wireless sensor network assisted IoT applications. A simple architecture of WSN- assisted IoT is shown in Figure 1.

Here a sensor node (SN) gathers information from environmental properties like temperature, pressure or moisture of air and convert it into digital form to send them wirelessly via *cluster* heads to the *base station* (BS). A basestation in contrast to sensor node possesses more computational power, large memory and connected to best source of energy unlike sensor nodes. In an IoT application, a basestation deployed can store, analyze and visualize the *sensed data* collected from cluster heads (CH). The basestation provides *graphical user interface* to interact with users directly or forward sensed data to a *remote server* via *Internet*. Then these *remote servers* will relay sensed data to the authorized users. Also, this sensed data can be saved as *web pages*, so that it can be accessed worldwide via *Internet*.

* Corresponding Author: Nalluri Prophess Raj Kumar, Karunya Institute of Technology and Sciences, +91-8309329926 & prophessms2k17@gmail.com, nalluriprophess@karunya.edu.in



In order to simplify the network management, the *cluster/group/ zone* concept is proposed by various researchers. CH nodes are the managers of grouped sensors nodes. CH has the responsibility to organize sensor nodes in its cluster, framing a routing table, collecting, aggregating and retransmitting the sensed data of the cluster. CH in sensor network will fast deplete their energy due to load imposed on them. To reduce the energy dissipation of CH nodes, the communication distance between CH and the basestation should be low or else information needs to be multi-hop forwarded from CH to basestation with the help of intermediate cluster heads.

Aiming at higher energy efficiency for the sensor network proposed, an extension of work originally presented in an IEEE conference [1] is modified as Cluster Centroid based Energy Efficient Routing (CEER) to balance energy consumption in cluster based WSN-assisted IoT. The main contributions of this research work are as follows:

- Producing a zone-clustering protocol which operates based on energy centroid position. The residual energy of sensor nodes and the communication distance to the basestation is taken into consideration here.
- To optimize the path between the zone head and the basestation, a dual hop or single hop communication routing algorithm is proposed. Zone heads count and distance metric are taken into consideration in a network area of *A* sq.m
- To optimize the path between the local sensor nodes in the zone and the basestation, zone co-ordinators are selected in the zone clusters. One among them acts as the zone head to forward sensed data to avoid local packet loss and long-distance communication among sensor nodes inside the zone.
- The user is provided with an option to opt for either 2 X 2 CEER (mostly suitable for dense environments) or 4 X 4 CEER (mostly suitable for sparse environments) to analyze the results based on their own preferred node deployment type. Provided the basestation should be placed inside the sensing field of the user defined area.

The research contributions reduced the average energy consumption of sensor nodes inside the zone clusters which eventually increased the lifetime of sensor nodes in the network when simulated under a simulation time t = 200 Sec. Moreover, the network lifespan is increased, as the sensor nodes balance the energy consumption among themselves (zone co-ordinator and zone head replacement strategy) to live for a longer period of time and contribute to the network throughput.

The rest of the paper is organized as follows: In section 2, the related work is discussed. The proposed CEER protocol and its phases along with its radio energy dissipation model is explained in section 3. The simulation results and discussions of the proposed protocol in comparison to the existing works is shown in section 4. Section 5 concludes the paper with further proceedings.

2. Related Work

Based on the network architecture, hierarchical routing protocols are classified into three types [2]: 1. Cluster-based; 2. Chain-based; 3. Tree-based protocols.

In *cluster-based* routing, network area is divided into clusters with the assistance of BS or sometimes through grid clustering and so on. A cluster head is selected for each cluster to transfer information collected from all the sensor nodes to the basestation either directly or with the assistance of other cluster heads or sensor nodes outside the cluster. In *chain-based* routing, the nodes are arranged in a chain like structure and only one node will act as the chain head to transmit the entire information of the chain to the basestation directly in the entire network area. In *tree-based* routing, all the sensed information by the leaf node is carried to the parent node (sink node) and at last, the information is sent to the root node i.e., BS.

2.1. Conventional Clustering approach

The first and foremost hierarchical routing protocol designed for wireless sensor networks is the LEACH protocol. In the open system interconnections (OSI) reference model, cluster routing protocols works on the network layer which is connected to both the data link layer and the transport layer. The primary function of the network layer is to permit different networks to be interconnected. It also translates logical network addresses into physical address. So, a network layer forms the major part in forwarding and routing the information across the network.

The LEACH [3-6] protocol is explained in detail in section 4.1.1. Its main characteristic is the local cluster generation, dynamic CH node rotation with data fusion inside the cluster. By generating a random number which is between 0 or 1, all sensor nodes are given provision to be the CH and the threshold set for selection of node is calculated as in Eq. (1)

$$T(n) = \begin{cases} \frac{P}{1-P[r*mod\left(\frac{1}{P}\right)]} & n \in N \\ 0 & Otherwise \end{cases} \quad (1)$$

The generated random number and $T(n)$ are compared for each sensor node. If the random number generated by a sensor node is less than the calculated threshold value $T(n)$, then that particular sensor node is selected as a cluster head for the current round. In Eq. (1) P is the desired percentage of sensor node inside the local cluster to be selected as cluster head, r is the round number, n is

node count and *T(n)* is the threshold. Number of cluster head nodes are not constant in LEACH due to its random selection of cluster head nature. The drawback of LEACH [7] is that, the different cluster count noted in each round will make number of sensor nodes in each cluster different. Uneven sensor node count in clusters dissipate uneven energy in each round. During CH selection, residual energy of sensor node is not taken into consideration in LEACH protocol.

2.2. Energy based clustering approach

An Energy Based Clustering – Self organizing map (EBC-S) to bring the effectiveness in cluster-based routing algorithms is proposed in [8] for topological clustering and incorporating a topological energy-based clustering technique to achieve extended lifetime and network coverage. The assumption made on BS is that it has no constraints on energy resources. A hierarchical and distance-based clustering technique is proposed in [9] which utilizes a new rank-order distance measure for agglomerative hierarchical clustering. Here, authors generate a rank order list by sorting all other sensor nodes in near neighbors by absolute distance. The distance based on rank order of two sensor nodes is computed using their rank order. The algorithm designed by them by grouping sensor nodes into small number of sub-clusters is similar to CEER networking like the formation of zone clusters inside the zone in CEER. The word *centroid* for wireless sensor networks is used by authors in [10-12] for *k*-means clustering algorithm. This algorithm is mainly based on Euclidean distance of nodes and CH selection based on residual energy of nodes. In this method, sink node/BS collects information about the *identifier, position* and *residual energy* of all nodes and store that information.

The steps in *k*-means clustering algorithm for WSNs are as follows:

- To form *k* clusters of sensor nodes, *k* centroids are to be selected initially at different locations inside the network area *A* sq.m.
- Euclidean distance from each sensor node to the selected centroids is calculated and information of a sensor node is saved in its nearest centroid. Thus, *k* initial clusters are formed in network area.
- After initial round, the position of centroid in each cluster is calculated again to check for position change from the previous one. If it is so, again calculation of Euclidean distance step repeats, else clusters are finalized.

2.3. Mobile sink node clustering approach

Data generated by an individual sensor may not appear significant, but overall data generated by sensor nodes of network area in dense environments is big. So, utilizing sink node's mobility is the concept proposed by authors in [13] to enable data gathering. The proposed technique by authors may reduce energy consumption of sensor nodes, but create additional challenges like determining sink node's trajectory and cluster formation prior to data aggregation. The mobile sink data collection process, cluster head selection problem and *mobile sink* path optimization is discussed in [14]. The mobile sink path optimization is formulated as shortest path finding problem. So, artificial bee colony algorithm is used to find optimal solution and the shortest path of mobile sink to improve data collection efficiency in network area.

2.4. Grid based Clustering approach

Grid clustering for energy optimization of sensor nodes is proposed in [15,16], where size of grid is directly related to the transmission coverage range of the node. So, as the grid size is reduced, the transmission range for sensor node also decrease and thus conserves the energy. But it may lead to more control overhead and consumes extra bandwidth. Moreover, if the grid size is too small, CH has to dissipate more energy to transmit its information to basestation. The similar simulation is done in proposed CEER protocol in the form of 2 X 2 CEER and 4 X 4 CEER protocols. But depending on concentration of network nodes in network area the concerned protocol is implemented in proposed research.

2.5. Centroid (Midpoint) based routing approach

An energy efficient clustering protocol based on K-means algorithm named Energy Efficient Clustering Protocol (EECPK-means) has been proposed by authors in [17] for WSNs where midpoint algorithm is used to improve initial centroid selection procedure. It considers residual energy and Euclidean distance as the parameter metrics for CH selection. A clustering algorithm using spatial correlation is proposed in [18], which groups sensor nodes with similar readings into one cluster and reporting the same, as the reading of the entire group. One node is selected as CH using centroid method here. The sensor node which has minimum distance to cluster centroid point is chosen to be the cluster head of the similar reading sensor group. Similarly, an energy efficient clustering protocol to prevent unbalanced clusters based on firefly and midpoint algorithms is proposed in [19] which uses residual energy and Euclidean distance as the performance metrics. It produced balanced clusters to balance CHs load and increase network lifetime. Similarly, the term *gateway node* is used in [20] to reduce the data load on cluster heads and forward the data to BS. But their assumption that every gateway node should be in the range of its neighbor gateway node in their proposed topology proves high algorithm complexity and limits user defined random sensor node deployment process. The term *energy centroid* for sensor networks is used in [21] where each cluster is designed to own 25% of sensor nodes using *distance centroid* algorithm. Here CH selection is based on *energy centroid* while the communication between CHs and the basestation is distance centric. An *energy centroid* algorithm for WSN-Assisted IoT is proposed by authors in [22] with the EECRP (Energy Efficient Centroid based Routing Protocol) algorithm when the BS is placed inside the network. This algorithm produced better results than conventional LEACH routing algorithm but no proper routing phase is described for the

communication between the CHs and the BS. Moreover, the EECRP algorithm finds the CHs, and transfer the information of cluster to the BS directly in single hop. But the proposed finds the help of other ZHs to relay the information of ZH which is far from BS and thus has the provision to opt for dual hop communication to the BS. Most of the existing centroid routing approaches has following drawbacks when choosing the initial centroids randomly for CHs.

- An empty cluster.
- Residual nodes
- User desired clusters as input to algorithm.
- Unbalanced workload on the CHs
- Non-selection of optimal CHs count.
- Unsuitable for Multi-hop routing.

3. Cluster Centroid based Energy Efficient Routing (CEER) protocol

In this section, deployed sensor node network either densely or sparsely distributed in uniform random fashion is utilized to obtain optimal path from source node to basestation/sink using cluster centroid-based routing. Subsequently the radio energy dissipation model is introduced. Moreover, few assumptions are presented for better understanding of our proposed work.

3.1 Assumptions

Wireless sensor network (WSN) in CEER consists of Sink node/Basestation along with other sensor nodes. The communication established between sensor nodes and basestation is multihop based. The routing structure is hierarchy. The sensor network is grouped to zone clusters (Interzonal communication) and each cluster has four zonal quadrants (Intrazonal communication). The *cluster / zonehead* collects information from four *zone co-ordinators* which is collected from each of four *zone-quadrants* or *zone-clusters*. The cluster/zone head communicates with the basestation either in single hop (directly) or through multihop (*relaying zone heads*) communication. In the proposed work the following assumptions are made.

- Distribution of sensor nodes in network is uniform random.
- There is only one sink node located inside the sensing field. (Placed at centre of sensing field).
- Both sink and sensor nodes position is static/fixed i.e., the position of sensor nodes will not change once the sensor network arrangement is done.
- All nodes can adjust their transmission power according to distance of transmission by using free space model or multipath fading model.
- All sensor nodes deployed are isomorphic i.e., equipped with equal energy levels when deployed and has the same processing and communication capabilities.
- Sensor nodes are aware of their locations through some localization techniques i.e., the locational co-ordinates are already fed into them.
- Sensor nodes are capable of transceiving information.
- Energy cost for zone formation is on BS and no control packets for sensor nodes.
- BS has complete knowledge about the energy level and locational co-ordinates of sensor nodes in network area of *A* sq.m.

Once the sensor network formation is done, position of sensor node is not changed. The location co-ordinates of sensor nodes are fed into the node already during network deployment. We assume that every sensor node knows the basestation co-ordinates after initial broadcasting as well as their residual energy at any time. The shape of sensing area in which sensor nodes are distributed is in rectangle. Network region is constructed using cartesian co-ordinate system (xlocation, ylocation). The notations and its explanation are provided in Table 1.

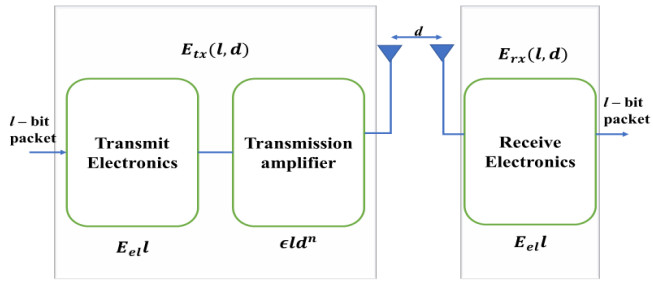
Table 1: Notations in CEER protocol and its explanation

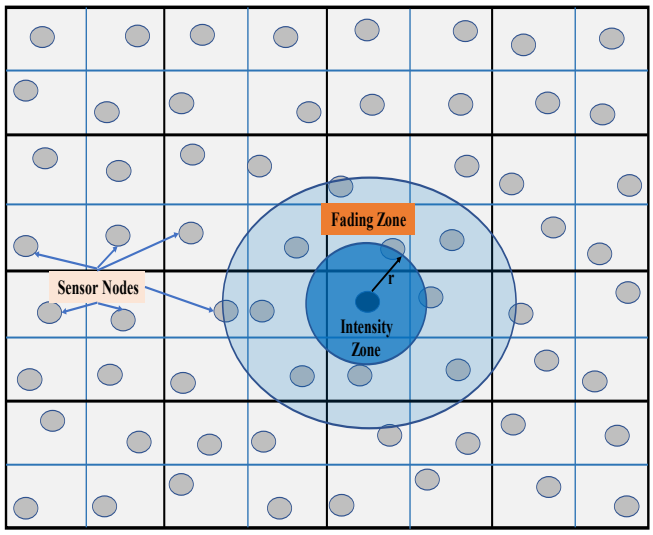
Notation	Explanation
l	*data packet*
d	*distance*
A	*Area of sensing field in Sq.m*
d_{th}	*distance threshold*
$\mathbf{E_{tx}}$	*Energy of transmitter*
$\mathbf{E_{rx}}$	*Energy of receiver*
$\mathbf{E_{el}}$	*Energy dissipated at tx or rx*
$\epsilon_{\mathbf{fsm}}$	*Free space model* (d^2 *power loss*)
$\epsilon_{\mathbf{mpf}}$	*Multi-path Fading* (d^4 *power loss*)
BS	*Base Station*
ZH	*Zone Head*
ZC	*Zone Co-ordinator*
E_{simR}	*Energy consumed in a Simulation Round*
E_{ZH}	*Energy consumed by Zone Head*
E_{SN}	*Energy consumed by a Sensor Node*
K_i	*I Zone Count*
$E_{ZH\ to\ BS}$	*Energy consumed in transferring data from ZH to BS*
$E_{Reception}$	*Energy consumed by ZH while receiving data from ZC*
$E_{Aggregation}$	*Energy consumed by ZH in transferring l bit data packet to BS*
$d_{ZH\ to\ BS}$	*distance between ZH and BS*
$dist_{MAX}$	*Maximum broadcast distance of BS in Sensor Network of Area A sq.m*
$L_{x,y}$	*Cartesian locational co-ordinates of Sensor Node*
$d_{S\ to\ D}$	*distance from source node to destination node*
NS	*Total network size in Sq.m*
Z_S	*Zone Size in Sq.m*
Z_q	*Zone quadrant/Zone cluster size in Sq.m*
$(\overline{X_C}, \overline{Y_C})$	*Mathematical centroid co-ordinates for sensor field*
$(\overline{X_{ec}}, \overline{Y_{ec}})$	*Energy centroid co-ordinates of CEER network sensor field*

3.2 Radio Energy Dissipation Model of CEER Network

The clustering in CEER is dependent on the energy model implemented below in figure 2.

From the radio energy model shown in figure 2, the required energy to transmit *l* bit data to distance *d* can be formulated as:





$$E_{tx}(l,d) = \begin{cases} l.E_{el} + l.\epsilon_{fsm}.d^2 \text{ if } d < d_{th} \\ l.E_{el} + l.\epsilon_{mpf}.d^4 \text{ if } d \geq d_{th} \end{cases} \quad (2)$$

The energy used in receiving l bit packet at the receiver is formulated as Eq. (3)

$$E_{rx}(l,d) = l.E_{el} \quad (3)$$

where E_{el} is per bit dissipated energy at the transmitter or receiver, ϵ_{fsm} and ϵ_{mpf} reflect "*free-space model*" (d^2 *power loss*) and "*Multi-path fading model*" (d^4 *power loss*) conditions. As shown in Figure 3, the data communication process of sensor nodes to the *intensity zones* uses "*free-space model*" and "*Multi-path fading model* to its *fading zones*. d_{th} is the distance threshold set for both models and d represents the distance between source sensor to destination sensor. From Eq. (2) distance threshold may be expressed as follows in Eq. (4)

$$d_{th} = \sqrt{\frac{\epsilon_{mpf}}{\epsilon_{fsm}}} \quad (4)$$

Now consider A sq.m sensor network area with N sensors deployed in uniform random fashion and divided into K zones or clusters. In hierarchical approach the total utilized energy within a simulation round E_{SimR} is calculated as follows Eq. (5)

$$E_{simR} = \sum_{i=1}^{K} E_{ZH_i} + \sum_{j=1}^{K} E_{SN_j} \quad (5)$$

where E_{ZH_i} is the utilized energy by Zone Head (ZH) when it receives information from Zone Co-Ordinators and relaying information on behalf of other Zone Heads to base station along with its own zonal information to basestation. E_{SN_j} is the energy consumed by sensor nodes in zone quadrants of a zone including zone co-ordinators.

Energy consumed by zoneheads in each zone/cluster is defined as Eq. (6)

$$E_{ZH_i} = E_{(ZH\ to\ BS)_i} + E_{Reception_i} + E_{Aggregation_i} \quad (6)$$

where $E_{(ZH\ to\ BS)_i}$ is the Energy utilized when the ZH of zone i transfer information to basestation, $E_{Reception_i}$ is the utilized energy while receiving information from the zone co-ordinators within the zone cluster i, $E_{Aggregation_i}$ is the energy consumed to process l bit packet to the basestation by ZH. Now $E_{(ZH\ to\ BS)_i}$ is calculated as Eq. (7)

$$E_{(ZH\ to\ BS)_i} = \begin{cases} l.E_{el} + l.\epsilon_{fsm}.d_{to\ BS_i}^2 \text{ if } d < d_{th} \\ l.E_{el} + l.\epsilon_{mpf}.d_{to\ BS_i}^4 \text{ if } d \geq d_{th} \end{cases} \quad (7)$$

$$E_{Reception_i} = |N_i|.l.E_{el} \quad (8)$$

$$E_{Aggregation_i} = |N_i|.l.Z_{pc} \quad (9)$$

where $|N_i|$ in Eq. (8) is the node count in zone i and Z_{pc} in Eq. (9) is the zone path cost of a bit reporting the basestation.

From the assumptions made above, some zone heads operate in free space mode and others in amplification mode to reach basestation with the zonal information. Here zone heads using amplification mode to transmit their information uses relaying zone heads to reach basestation. Let total ZH count in both modes be m. Then $\sum_{i=1}^{K} E_{(ZH\ to\ BS)_i}$ is formulated as Eq. (10)

$$\sum_{i=1}^{K} E_{(ZH\ to\ BS)_i} = l\,(K.E_{el} + \epsilon_{fsm}.\sum_{i=1}^{K-m} d_{to\ BS_i}^2 + \epsilon_{mpf}.\sum_{i=1}^{m} d_{to\ BS_i}^4 \quad (10)$$

So, total consumed energy by all ZH's in sensing area A x A m is given by Eq. (11)

$$\sum_{i=1}^{K} E_{ZH_i} = l\left((K+N).E_{el} + N.Z_{pc} + \epsilon_{fsm}.\sum_{i=1}^{K-m} d_{to\ BS_i}^2 + \epsilon_{mpf}.\sum_{i=1}^{m} d_{to\ BS_i}^4\right) \quad (11)$$

Now the Energy consumed by all sensor nodes (N-K) other than K zone heads is given by $\sum_{j=1}^{K} E_{SN_j}$ in Eq. (13)

Energy consumed by single sensor node inside a zone is given by Eq. (12)

$$E_{(SN)_j} = \begin{cases} l.E_{el} + l.\epsilon_{fsm}.d_{to\ BS_j}^2 \text{ if } d < d_{th} \\ l.E_{el} + l.\epsilon_{mpf}.d_{to\ BS_j}^4 \text{ if } d \geq d_{th} \end{cases} \quad (12)$$

Similar to Eq. (11) let n be the node count operating in both free-space and amplification modes. So,

$$\sum_{j=1}^{N-K} E_{SN_j} = l\left((N-K)E_{el} + \epsilon_{fsm}\cdot\sum_{j=1}^{N-K-n} d^2_{to\ ZH_j} + \epsilon_{mpf}\cdot\sum_{j=1}^{n} d^4_{to\ ZH_j}\right) \quad (13)$$

From Eq. (11) and (13), the total energy consumed in a simulation round of t sec in proposed hierarchical approach routing is concluded in Eq. (14)

$$E_{simR} = l\left(2.N.E_{el} + N.Z_{pc} + \epsilon_{fsm}\left(\sum_{i=1}^{K-m} d^2_{to\ BS_i} + \sum_{j=1}^{N-K-n} d^2_{to\ ZH_j}\right) + \epsilon_{mpf}\left(\sum_{i=1}^{m} d^4_{to\ BS_i} + \sum_{j=1}^{n} d^4_{to\ ZH_j}\right)\right) \quad (14)$$

3.3 Zone or cluster creation and Node deployment

To implement CEER protocol zone formation plays a major role, as it reduces overhead and energy consumption in data transmission within the zone/cluster. By further dividing the zone into 4 quadrants each the energy consumption of zoneheads/cluster heads is reduced. Zone co-ordinators are created within a quadrant for collecting the entire information of quadrant and sending it to zone head. From Figure 4 one can see the formation of a zone, out of K zones with four equal quadrants (Q_n, $where\ n = 1,2,3,4$) displayed inside the sensor network. So, inside a 2 X 2 CEER network i.e., a sensor network formed with 2 zones in a row and 2 columns in a sensor area of A x A m, will give better results when CEER is implemented with a greater number of sensors mostly suitable for dense environment. If a CEER has to be implemented in a sensor network with lesser number of nodes then zone formation of 4 X 4 or 8 X 8 finds more suitable for application in sparse environments. Let the intersection point of all the quadrants within a zone be ZQI (Zone Quadrant Intersection) located at co-ordinates (x, y). Now zone quadrant formation algorithm is coded as Algorithm 1

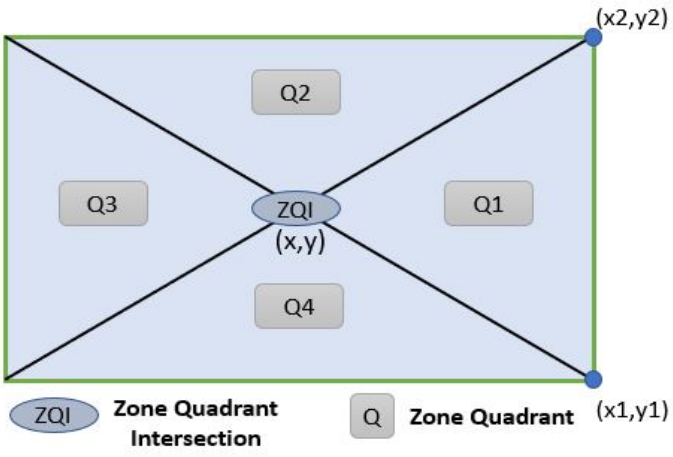


Figure 4: Zone formation in CEER Network

Algorithm 1: Zone Quadrant Formation

Step 1: theta1 = **compute** *angle* (x, y, x1, y1);

theta2 = **compute** *angle* (x, y, x2, y1);

Step 2: compute *angle* (x1, y1, x2, y2); x = x1 - x2; y = y1 - y2;

Step 3: *angle* = atan2(y, x);

Step 4: *angle* = *angle**180/pi;

Step 5: while (*angle* < 0) *angle* += 360;

Step 6: while (*angle* > 360) *angle* -= 360;

Step 7: return *angle*;

Assuming that the sensor nodes are static throughout the simulation round, they are distributed uniform randomly within the zones. In the sensor network, the locational co-ordinates of sensor nodes are obtained using the following Eqs. (15) , (16)

X-coordinate $X_c = sensingField_{Length} * random$ (15)

Y-coordinate $Y_c = sensingField_{breadth} * random$ (16)

Where $sensingField_{Length}$ and $sensingField_{breadth}$ are the network size locational paramters and *random* denotes positional deployment of node in random fashion inside sensing field and is the number between 0 and 1. Since the simulation is done through Network simulator 2 (NS2 version 2.32), nodes are deployed using *scen*(N) command where N is the Node count in Network area *A* sq.m

The distribution of energy in entire network is done equally among all sensor nodes inside sensing area. Each sensor computes the distance ($d_{S\ to\ D}$) between its own location $L_{x,y}$ to destination sensor or the Base station B_S. The distance ($d_{S\ to\ D}$) is estimated using the Pythagorean Theorem. So, finding the distance between sensors by using the coordinating points of source sensor (X_S, Y_S), and the destination sensor (X_D, Y_D) as shown in Figure 5 and then computing the distance between them is as follows in Eq. (17)

$$(d_{S\ to\ D}) = \sqrt{|X_S - X_D|^2 + |Y_S - Y_D|^2} \quad (17)$$

In R X C CEER network formed with R row zones and C column zones inside A sq.m network field, let total network size be NS = A^2m as represented in fig.5. Consider Z_S as one zone size in sq.m which is split into further equal quadrants of each size Z_q $sq.$m. Now NS is formulated from Eq. (18), (19)

$$Z_S = (4 * Z_q)\ \text{sq.m} \quad (18)$$

$$\text{NS} = (R * C * Z_S)\ sq.m \quad (19)$$

Suppose if the 2 X 2 CEER network is implemented in 100 sq.m network field. Z_S is 4 * Z_q = 25 sq.m. Then total network size 2 * 2* 25 = 100 sq.m. from Eq. (19)

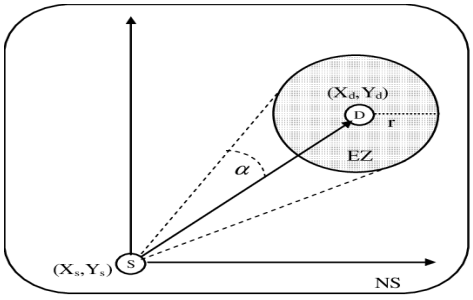


Figure 5: Expected Zone (EZ) formed in radius "r" for the destination sensor w.r.t source sensor locational co-ordinates at an angle α

When the CEER network of network size 2 X 2 is considered i.e., with 2 rows and 2 columns totally comprising of 4 zones. Each zone Z_s will occupy an equal area of 25 sq.m and so total network size NS would be 25 * 2 (Rows of zones formed) * 2 (Columns of zones formed) = 100 Sq.m

3.4 Zone Clustering Algorithm Scheme

The proposed CEER algorithm find the most appropriate zone head (ZH) node for the zone based on energy centroid. The algorithm has 3 phases: 1. Network initiation phase. 2. Zone co-ordinator selection phase and 3. Zone head selection and its replacement phase

3.4.1 Network initiation phase

The aim of this phase is to create mutual message exchange path between sensor nodes and basestation. These messages will have nodes positional, energy level information along with the average energy of deployed network. Also, information of the zone co-ordinators (ZC's) is chosen by the basestation in its first round and the longest transmission distance too. Firstly, sensor nodes will send their *position* message to the *basestation (BS)*. The packet format of *position* message is shown in Figure 6.

msg type	*Sensor ID*	*X-coordinate*	*Y-coordinate*	*Energy Level*

Figure 6: Position message packet format

The *msg type* header shows that the packet has a sensor node locational information. The *Sensor ID* is given to identify the sensor, sending its location update to base station. The position of sensor node in the sensing field is given in cartesian co-ordinate (X, Y). Here *X-coordinate* and *Y-coordinate* of sensor is sent. *Energy level* in the packet format shows the residual energy of sensor node at particular time *t*.

When this location message packet is received from all sensor nodes in the sensing field within the time limit set by timer, basestation will start to estimate the distance from itself to all the sensor nodes which have updated their location message. Consequently, the zone formation will be done with its quadrants, inside the sensing field area. Now the basestation starts identifying the *zone clusters* based on the estimated distances calculated previously. Subsequently, it updates the *node table* with the position information of node and its energy level. Then, the basestation broadcast the *ACK* (Acknowledgement) message specifically to the sensor nodes in one zone. The *ACK* packet format is shown in Figure 7

msg type	$dist_{MAX}$	*ZC ID*	*Average Energy*

Figure 7: Pack format of ACK message

msg type here is used to send the intended information to particular zone cluster nodes from *basestation / sink*. $dist_{MAX}$ field provides the information on maximum broadcast range to each node in particular *zone cluster* and is calculated by basestation and set it as a *communication threshold* for CEER network. Note that, $dist_{MAX}$ is calculated by taking the *Average Energy* from Eq. (2) *ZC ID* is needed here, so that the basestation can send the above-mentioned information to the particular *zone co-ordinator*. The *Average Energy* indicates the CEER network average energy intended for each node. The *ACK* message receiving node will update its routing table according to *ACK* packet content. At the end of this phase, the mutually exchanged information is updated in routing tables of both basestation and sensor nodes memory. Also, the routing table information is updated in real time, as the CEER network enhances to further phases.

3.4.2 Zone co-ordinators selection phase

The main aim of this phase is to select zone *co-ordinator* for every *zone cluster*. After receiving the *position* message and delivering *ACK* message, the basestation selects the node which has the *Energy Level* greater than the *Average Energy*. In the initial round, ZC selection is mostly random as the sensor nodes are equipped with equal energy levels at the deployment stage. In the consecutive rounds as the basestation updates the broadcast information, it is saved by all sensor nodes. After receiving the updated broadcasting information from basestation which has the information on node that can act as the ZC, the sensor node will check its own ID whether it is the ZC. So, if the ID matches then that node claims its election as the ZC for that transmission round. It also acts as a transceiver by receiving the information of all sensor nodes in a *zone cluster* and transmitting it to the ZH from then. Zone Head is elected later based on energy centroid in routing phase as shown in Figure 8. If suppose a sensor node finds its own ID different from BS broadcasted information, it saves the energy by activating its receiving antenna and closing the transmitter antenna to wait for further updates from the basestation through ZC.

3.4.3 Zone head selection and its replacement phase

After selecting the zone co-ordinators, the ZCs' broadcasts the schedule message with their IDs and the positional information to neighbour ZC nodes. Based on ZC ID in *ACK* message and schedule message received by neighbour nodes to determine whether they belong to that zone quadrant/cluster, zone clustering phase is completed. Now the sensor nodes inside each zone cluster will send their positional and energy level information to their ZC. Now zone coordinators which fall under zone K_i (where i=1, 2, …., (R*C) as shown in Figure 8) has to find the energy centroid location of the zone K_i. The ZC node which is nearest to that energy centroid will act as ZH for that zone. So, ZH will be one among the four zone coordinators of a selected zone.

By electing the ZH in the above-mentioned way, the CEER network can balance the energy consumption of networking sensor nodes inside the sensing field area. Now the CEER network satisfies the four aspects of our protocol framework. Firstly, in default simulation round, the temporary ZCs' are chosen by basestation itself, thereby gaining an overall picture of CEER network. When the simulation is processing, official ZC is selected in the zone cluster, which shows that CEER network is self-organizing. Also, one among the selected ZCs' of a zone closer to the energy centroid can act as zone head (ZH), which improves coverage of CEER network. The position of energy centroid calculation is described in next section which is done purely based on positional and residual energy information of sensor nodes inside CEER network. So, for a sensor node to become a ZH it should reach the ZC stage first and its energy level must prove worthy to become ZH after considering the closest distance parameter.

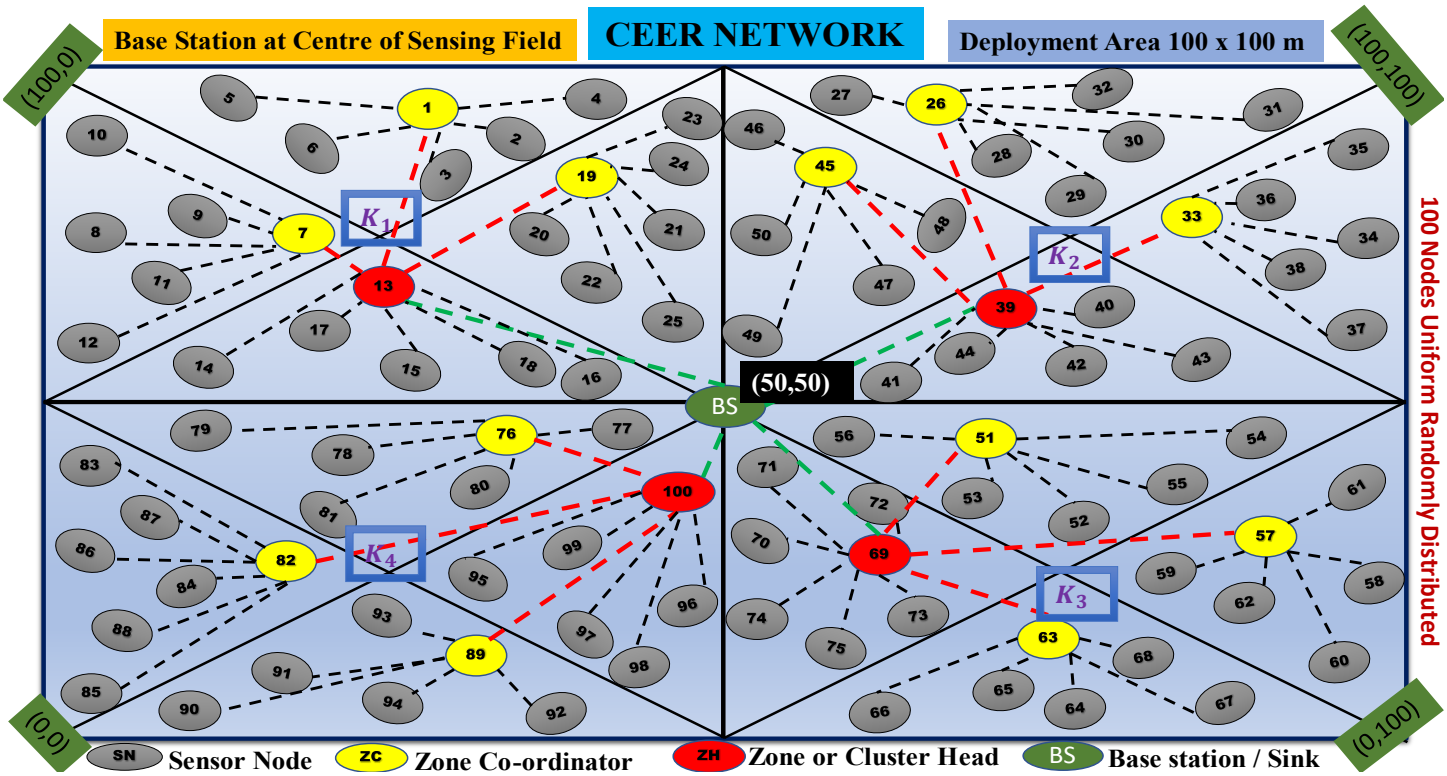


Figure 8: Zone formation and routing process in the CEER network when the basestation (BS) is placed at centre co-ordinates of Network area *A* sq.m

3.5 Zone Energy Centroid Selection Algorithm

In a concept of mathematics, centroid is the point of concurrency. Indirectly, it is the intersection point of all the medians in a geometric figure. Median is the value separating the higher half from the lower half of a data sample i.e., a population or a probability distribution. Now consider the ZC as a median of all the sensor nodes in a zone cluster. These zone coordinators from each zone quadrant are the set of medians. Now all these zone coordinators fulfil their duty to find the centroid position and which would be an imaginary point of mass concentration. Then the ZC near to that point will be selected as ZH for one simulation round. Here the concept is introduced as *Zone Energy Centroid* slightly named different from traditional mathematical centroid for few reasons. Firstly, median in terms of sensor nodes doesn't make sense. Secondly, finding the intersection point of the medians in terms of sensor operation to find zone head is meaningless. Finally, residual energy of sensor nodes is the only metric which changes in the sensor network operation. So, energy centroid in CEER network operation is to display the distribution of remaining energy of entire network.

Here, to calculate the Energy Centroid, the residual energy level and location of zone co-ordinators *i* is taken into consideration. In the zone cluster, if the weight of ZC is known then position of centroid is calculated using Eq. (20) and (21). In the field of mathematics, calculation for centroid in cartesian co-ordinate system $(\overline{X_C}, \overline{Y_C})$ is calculated as

$$\overline{X_C} = \frac{dM_y}{M} = \frac{\iint_{Az} x.D\ dw}{\iint_{Az} D\ dw} \tag{20}$$

$$\overline{Y_C} = \frac{dM_x}{M} = \frac{\iint_{Az} y.D\ dw}{\iint_{Az} D\ dw} \tag{21}$$

where $\overline{X_C}$ and $\overline{Y_C}$ are the results of centroid in mathematical prospective for CEER network. Here *Az* is the zone geometrical area. *dw* is the centroid weight differential. D is the density of nodes weight. dM_y and dM_x are the static moments of *x* and *y* axis respectively.

In CEER network, when the weight of ZC is unknown i.e., as the term "weight" makes no sense in ZH formation, ZC locational and energy level information is gathered to obtain position of *Energy Centroid*. When its position is obtained, ZC nearer to that *Energy Centroid* can take the responsibility of particular zone by acting as a ZH. Once ZH is selected for that particular simulation round, its locational information is updated in the routing tables of ZH itself, the basestation and the other ZH's in the network area of A sq.m by basestation. The position of centroid is given as $(\overline{X_{ec}}, \overline{Y_{ec}})$ in zone K_i as obtained from Eq. (22) and (23)

$$\overline{X_{ec}} = \frac{\sum_{p=1}^{4} \frac{RE_{ZC_p}}{E_0} .X}{N_j(p)} \tag{22}$$

where RE_{ZC_p} is the Residual Energy of Zone Co-Ordinator p in zone K_i. X and Y are the location co-ordinates of Zone Co-Ordinator p. $N_j(p)$ is the j node count in the zone cluster with zone coordinator p.

When the Zone Heads of *K zones* are ready to send packets to basestation, they estimate their own distance $d_{ZH\ to\ BS}$ to basestation placed at center of network area with $dist_{MAX}$ sent by the basestation in *ACK* packet. If $d_{ZH\ to\ BS} < dist_{MAX}$ then

ZH's will start sending the packets to base station, else ZH's will find their nearest neighbor ZH from pool of $(ZH_i)\ where\ i = 1,2, \ldots (K-1)$ in terms of distance, to transmit its information in maximum dual hop to the basestation. Then the condition to be checked is whether $d_{ZH_i\ to\ BS} < dist_{MAX}$ or not. If fulfilled then proceed to transmit the information to the nearest ZH and it will send that information to basestation on its behalf, else store the packets in ZH buffer queue and wait for the next round. This process of storing packets in ZH and waiting for consecutive round will be the least case, and if it has to happen there might be packet loss in rare scenario. So, a trading exists in CEER network between energy consumption and packet loss to avoid the long-distance communication of ZH nodes which may lead to more packet loss. So, the energy consumed in a simulation round of *t* sec is calculated through Eq. (14) after updating the Zone path cost Z_{pc}.

Algorithm 2: CEER Network Routing Process

Step 1: Sensor Nodes $SN_j \xrightarrow{\textbf{position}\ PACKET}$ Basestation *BS*

Step 2: Calculate *Average Energy* ***and*** $dist_{MAX}$

Step 3: $BS \xrightarrow{\textbf{ACK}\ PACKET} SN_J$

Step 4: ***if*** (Zone Head *ZH* prevail)

Step 5: ***then Begin*** *Zone Clustering* $\xrightarrow{Result}$ *Zone co-ordinator ZC*

Step 6: *ZCs'* $\xrightarrow{Schedule\ Message}$ neighbour *ZC* nodes,

Step 7: **Recalculate** *Energy Centroid* $\xrightarrow{Result}$ **updated** ZH ID

Step 8: **timer** is **set,**

Step 9: ***and if*** **(**Node count $N \geq$ *Zone Cluster* **count)**

Step 10: ***and if*** (pre-set **timer** time not reached ***then***

Step 11: $SN_J \xrightarrow{sensed\ data}$ *ZC (p)*, where *p = 1,2,3,4* in a **zone**

Step 12: *ZC (p)* $\xrightarrow{Energy\ Centroid\ information}$ **forward** data to *ZH* K_i

Step 13: ***if*** $(d_{ZH\ to\ BS} < dist_{MAX})$

Step 14: K_i (where *i* = 1, 2, …., (R*C)) $\xrightarrow{forward\ Data}$ *BS directly*

Step 15: ***elseif (***$d_{ZH_i\ to\ BS} < dist_{MAX}$);

Step 16: ***then*** $\xrightarrow{forward\ Data}$ **nearest** ZH_i where *i* = (1,2, …., (*K -1*)

Step 17: ***return***

Step 18: ***end if***

Step 19: ***return***

Step 20: ***end if***

Step 21: ***return***

Step 22: ***end if***

Step 23: **Recalculate** *Average Energy* ***and*** $dist_{MAX}$

Step 24: $BS \xrightarrow{\textbf{ACK}\ PACKET} SN_J$

Step 25: **Select** Zone Head *ZH* nodes

Step 26: ***return***

Step 27: ***end if***

Step 28: ***return***

Step 29: ***end if***

As shown in Eq. (14), the energy consumed by CEER network in the initial stage is for the basic information exchange i.e., for sending *position* message and *ACK* message. But when the network enters the stable stage of transferring the sensed information by sensors, the energy consumed by the above messages can be ignored. As it would be very less when the routing proceeds further rounds. We know that hierarchical sensor networks are good at achieving better results in local zone clusters (in our case). So, local sensor nodes i.e., nodes inside a *zone cluster* can only communicate with their ZCs'. Also, the average transmission distance of the local sensor is purely dependent on the location of ZC inside a *zone cluster*. In most of the conventional-clustering protocols like LEACH (Low Energy Adaptive Clustering Hierarchy), the location of sensor node is not considered while choosing the cluster head (CH). As a result, when the CH is chosen it may be located at edges of the cluster and it makes the sensor node inside the cluster to consume most of its energy for communicating with the cluster head. Contrarily, the proposed CEER chooses the ZC and ZH in an entirely different way. Also, the ZC of the next round is chosen by ZC of current round based on metrics encoded in the script. The ZH is chosen as the centre of energy network to optimize energy consumption parameter.

4 Simulation Results and Discussions

The simulation tool used for simulating CEER network is NS-2 version 2.32. Network Simulator (*ns*) is a discrete event simulator targeted at networking research. Based on few sources in NS-2, few assumptions are made for the CEER simulation. The simulation is based on fact that there may be no loss of sensed data from environment by sensor nodes, since the transmission medium used is reliable by default in NS2. Since the omnidirectional antenna is used in simulation, the radio coverage will be in circular direction. The configured node uses free space model /multipath fading model as its propagation *type*.

Here the physical layer is used as wireless channel transmission medium, where the node transmits and receive information. Interface Queue type can be either Drop Tail /Priority Queue based on Routing. If the channel stays busy/channel unavailable i.e., the basestation is far reachable for ZH when $d_{ZH\ to\ BS} > dist_{MAX}$ and even neighbour Zoneheads $d_{ZH_i\ to\ BS} > dist_{MAX}$ in our case, the packet will be saved to the queue. Here the number of packets that can be hold in queue is set to 50 packets. Here the topology loaded is of Flat Grid Type. Two sensor nodes can mutually interchange their data with each other if they lie within their sensing range. But as the topology changes there may be a change

in the nodes position so that the node may move out of its neighbours sensing range. Then (GOD) General Operation Director) will monitor Sensor Nodes information in simulation of Wireless Networks. But GOD needs to be re-created for every instance of simulation. GOD creation requires node count to be predefined, so that memory could be reserved for those nodes to capture information. Simulation area of CEER network is 100 x 100 m. The node configuration used for NS-2 simulation of CEER network is shown in Figure 9.

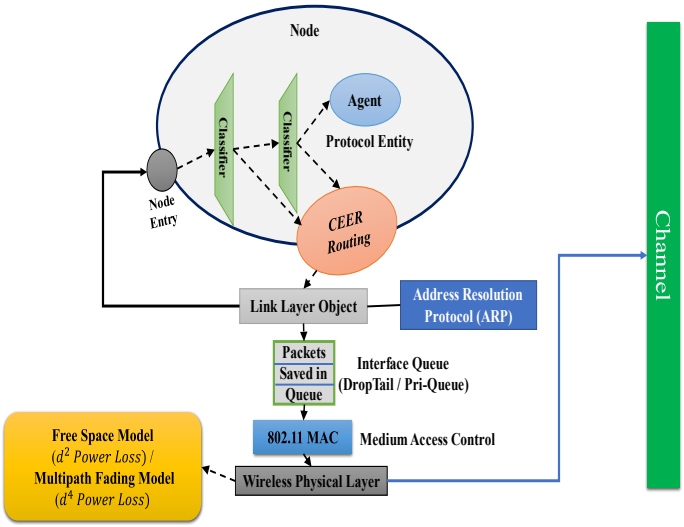


Figure 9: Node configuration of CEER network in NS2 simulator

The "*free space model*" assumes the ideal propagation condition with only one LOS (Line-of-sight) path between transmitter and receiver. It basically represents the communication range as a circle around the transmitter. The receiver receives all packets, if its location is within the circle. This model is best suitable for short distance communication. Otherwise, packets are lost. So, according to *free space model* the received signal power in free space at distance d from the transmitter is given as Eq. (24) [23]

$$\epsilon_{fsm} = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L} \quad (24)$$

where P_t is transmitted signal power. G_t and G_r are the transmitter and receiver antenna gain respectively. L ($L \geq 1$) is the system loss and λ is the wavelength. It is common to select $G_t = G_r = 1$ and $L = 1$ in *ns* simulations.

In "*multipath fading model*" both the LOS i.e., direct path and the obstacle reflection path are considered. This model gives more accurate prediction for long distance communication than the *free space model*. In this model the received power at distance d is given as Eq. (25) [24]

$$\epsilon_{mpf} = \frac{P_t G_r G_t {h_t}^2 {h_r}^2}{d^4 L} \quad (25)$$

where L = 1 in case of *multipath fading model* model.

P_t is the signal transmitted power, G_r is the receiver antenna gain,G_t is the transmitter antenna gain, h_t and h_r are the heights of transmit and receive antennas respectively.

Since *multipath fading model* doesn't provide good result for short distance communication due to oscillation caused by constructive and destructive combination of two rays, *free space model* best suits the situation when d is small. From Eq. (25) as the distance d increases the sensor node experience faster power loss. So, to utilize the benefits of both models, we calculate the distance threshold d_{th} from Eq. (4). When $d < d_{th}$ Eq. (24) is used and when $d \geq d_{th}$ Eq. (25) is used. So, distance threshold is calculated as Eq. (26)

4.1 Simulated protocols for comparison with CEER network

The simulation results of proposed CEER protocol is compared with conventional LEACH (Low-Energy Adaptive Clustering Hierarchy), Traditional PEGASIS (Power Efficient Gathering in Sensor Information Systems), Existing PRRP (Position Responsive Routing Protocol) and Pre-proposed ES3 (Energy Efficient Sink Selection Scheme). The results of CEER protocol are taken from two scenarios when R X C is 2 X 2 and 4 X 4 when the BS is placed at center of the sensing field. The evaluated performance metrics in the simulation are Total Energy Consumption of network, Average Energy Consumed by a sensor node, Packet Delivery Ratio (PDR), Dropping ratio of packet and the Network Throughput.

4.1.1 LEACH protocol (Low Energy Adaptive Clustering Hierarchy) [3-6]

Aim: To Improve the lifespan of Wireless Sensor Networks by reducing Energy Consumption in the Network.

- LEACH protocol is a TDMA based MAC Protocol.
- It is a *clustering-based* proactive routing Protocol.

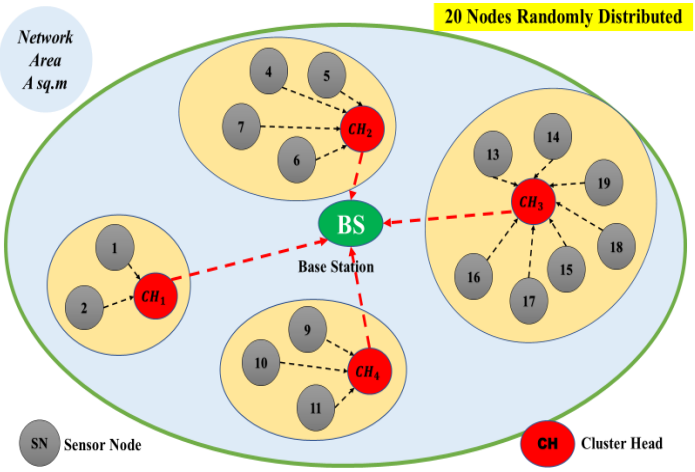


Figure 10: Routing in LEACH Protocol

Protocol Description:

- LEACH is self-organizing and adaptive clustering Protocol.
- In Figure 10 sensor nodes (SN) are organized into local clusters of their choice, where one of them will be a cluster head.
- This randomized approach is used to reduce the energy consumption of one particular sensor node in relaying the information collected from the sensor nodes to the sink.

- The responsibility of cluster head is not only to collect the information from its cluster, but to aggregate the data to be sent to base Station (BS) to reduce the quantity of messages, resulting in less consumption of energy which increases the network lifetime.
- LEACH uses randomized rotation of cluster heads to distribute the energy load among the sensor nodes deployed in the sensing field.

Advantages:

- Cluster Head (CH) aggregates the entire information which reduces the traffic load of the network.
- Since this protocol uses routing with single hop between sensor nodes to cluster heads it saves energy.
- Thus, increasing the lifetime of sensor network.
- Also, geographical location of networking node used to create a cluster is not required.
- LEACH uses no control information from BS and does not require any global knowledge of network.

Drawbacks:

- The protocol does not provide the information on the cluster head count in the network to the Basestation
- Also, if cluster head dies due to any reason the cluster become useless as there is no possibility of that particular cluster info reach the base station.
- Division of clusters is random and results in uneven distribution of clusters.
- E.g. Some clusters have more nodes and some clusters have lesser nodes.
- Some clusters heads are placed at the center of the cluster while some are in the border of cluster.
- This results in an increase in energy consumption and reduces the performance of the network.

4.1.2 PEGASIS protocol (Power Efficient Gathering in Sensor Information Systems)

Aim: To gather the information or data received from sensor nodes and transmit the same to the close neighbors in distance and iterating the leader nodes for transmitting received data to the basestation.

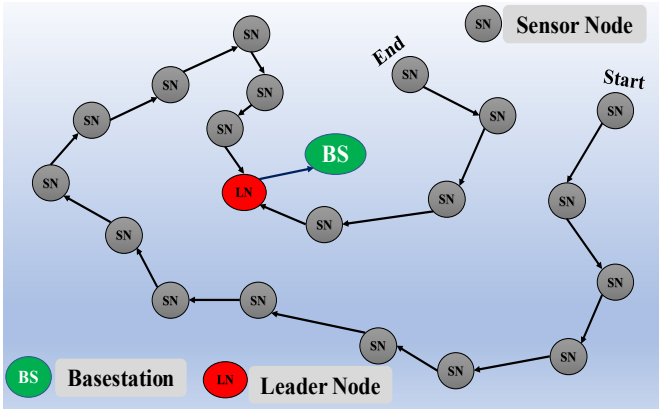


Figure 11: Routing in PEGASIS Protocol

Protocol Description:

- PEGASIS [25,26] is a near optimal chain-based routing protocol as shown in Figure 11. which is an improvement over LEACH.
- It facilitates distribution of energy load evenly among sensor nodes.
- The deployment of sensor nodes in sensing field is random and those sensor nodes arrange themselves into a chain using greedy algorithm.
- Also, the computation of this chain is done by basestation and broadcasts it to all sensor nodes.
- In PEGASIS, every sensor node receives and transmit atleast one packet in every round and be a leader in atleast once in *n* rounds and *n* is also the node count.

Advantages: (Over LEACH)

- Secondly, the leader node can receive atmost two messages from its neighbor nodes which is not the same in the case of LEACH protocol.
- For E.g. In the network of 100 nodes, cluster head (CH) in LEACH receives 20 messages at a time if the 20 nodes transfer their information at a time to the cluster head inside a cluster.
- For every round of communication one node i.e., leader node takes the entire responsibility to transmit the message to the basestation.
- In PEGASIS, data fusion helps in reducing the quantity of data being transmitted between sensor nodes and the basestation.
- Data fusion combines one or more data packets from different sensor measurements to produce a single packet for better inference.
- PEGASIS protocol is used mostly to characterize and monitor the quality of environment.

Drawbacks:

- The distance between the basestation and the leader node is not predefined and so the transmission distance between the leader node and basestation can be far.
- Energy level of the leader node is not considered for its selection.
- Since only one leader node can exist for one instance of communication, it may be the bottleneck for the network in causing delay.
- Redundant transmission (same data or information received more than once) exists as only one node takes the responsibility of transmitting entire chain network information to the basestation.

4.1.3 PRRP protocol (Position Responsive Routing Protocol)

Aim: PRRP protocol [27,28] is mainly designed to reduce the energy consumed by each node by minimizing the time the sensor

node is in idle listening state and by reducing the average communication distance over the network.

Protocol Description:

Sink/Basestation builds the tree-based sensor network by broadcasting the control message containing six fields *sender, position, type, level, parent and energy*. The BS creates network tree and each node receives tree information. The protocol finds logically, how many nodes lies in a particular tier T_n as shown in Figure 12.

Energy threshold E_{th}in PRRP is calculated using Eq. (27).

$$E_{th} = lE_{el} + lE_{amp}d^2 + 8lE_{el} \quad (27)$$

Where lE_{el} is the dissipated energy in processing l bits, E_{amp} is the energy of the amplifier electronics used in simulation. Energy consumption of transmitter E_{tx} and receiver E_{rx} are given by Eq. (28) and Eq. (29) respectively.

$$E_{tx} = lE_{el} + lE_{amp}d^2 \quad (28)$$

$$E_{rx} = lE_{el} \quad (29)$$

Where lE_{el} is the dissipated energy in processing l bits, E_{amp} is the Energy of the amplifier electronics used in simulation.

The lifespan of cluster head (CH) in PRRP is calculated in terms of sampling interval number N_s, as in Eq. (30)

$$N_s = \frac{E_{in} - E_{th}}{E_t + lE_r + (T_s - (l+1)T_t)P_s} \quad (30)$$

Where E_{in} denote initial energy of sensor field. E_t is energy consumed in transmitting one data bit. T_s and T_t are sampling time and transit time respectively and P_s is sensor node power.

The sensor nodes are deployed in a random fashion across sensing field as shown in Figure 11 and are assumed that they are aware of its location using GPS or by any other location means. The gateways for the base station are the nodes within tier T_0 . Each tier is having a vertical cross-sectional area on either side of Base station. Each vertical cross-section is provided with *tier-head* which acts as a sink for all the nodes in that vertical cross-sectional tier. *Tier-heads* are selected based on node energy level, its distance from the base station.

Advantage:

- It uses TDMA (Time division multiple access technique) to transfer the data from *tier heads* to the Basestation.
- No delay constraints in data transmission.

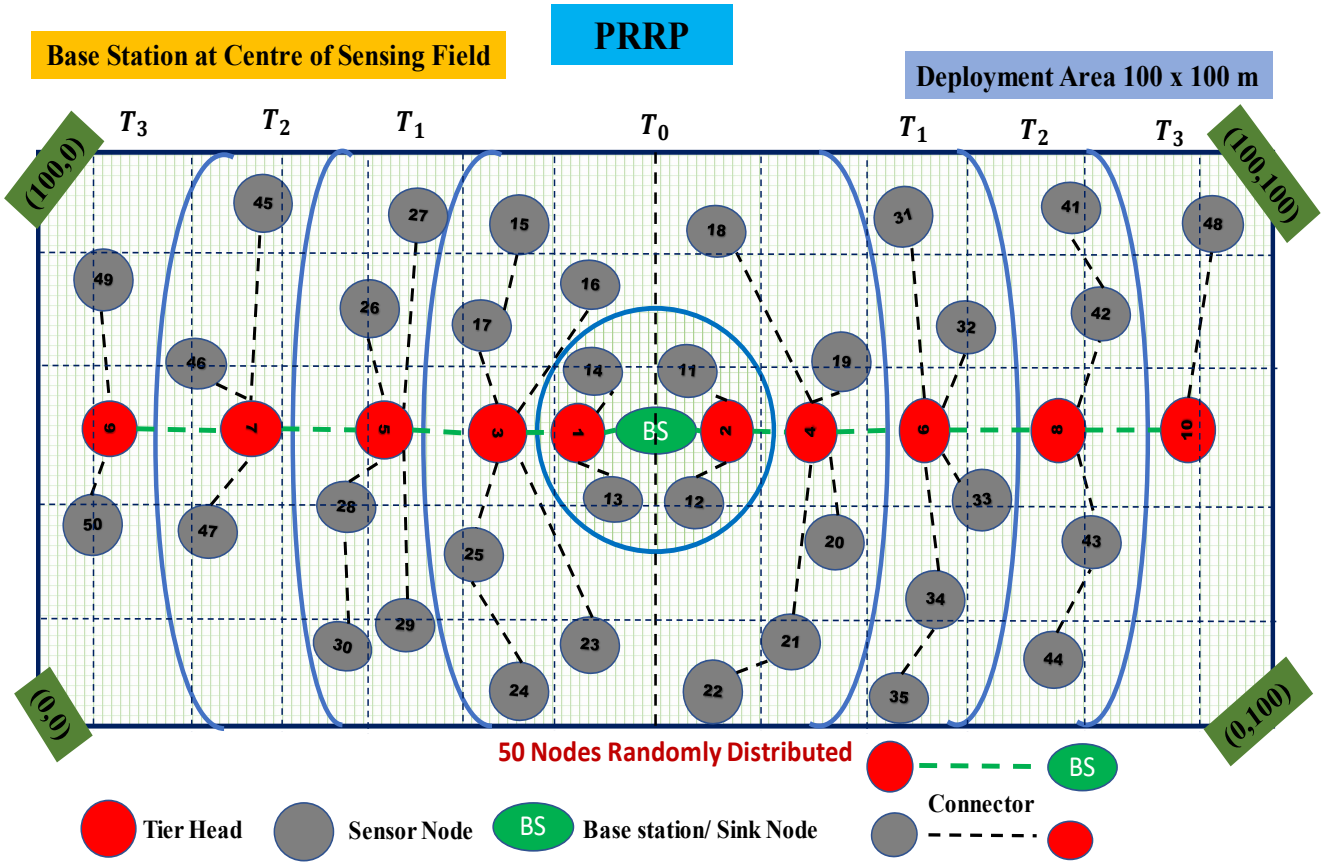


Figure 12: Tier formation and routing process in PRRP (Position Responsive Routing Protocol) when the basestation is placed at the center of the sensing field

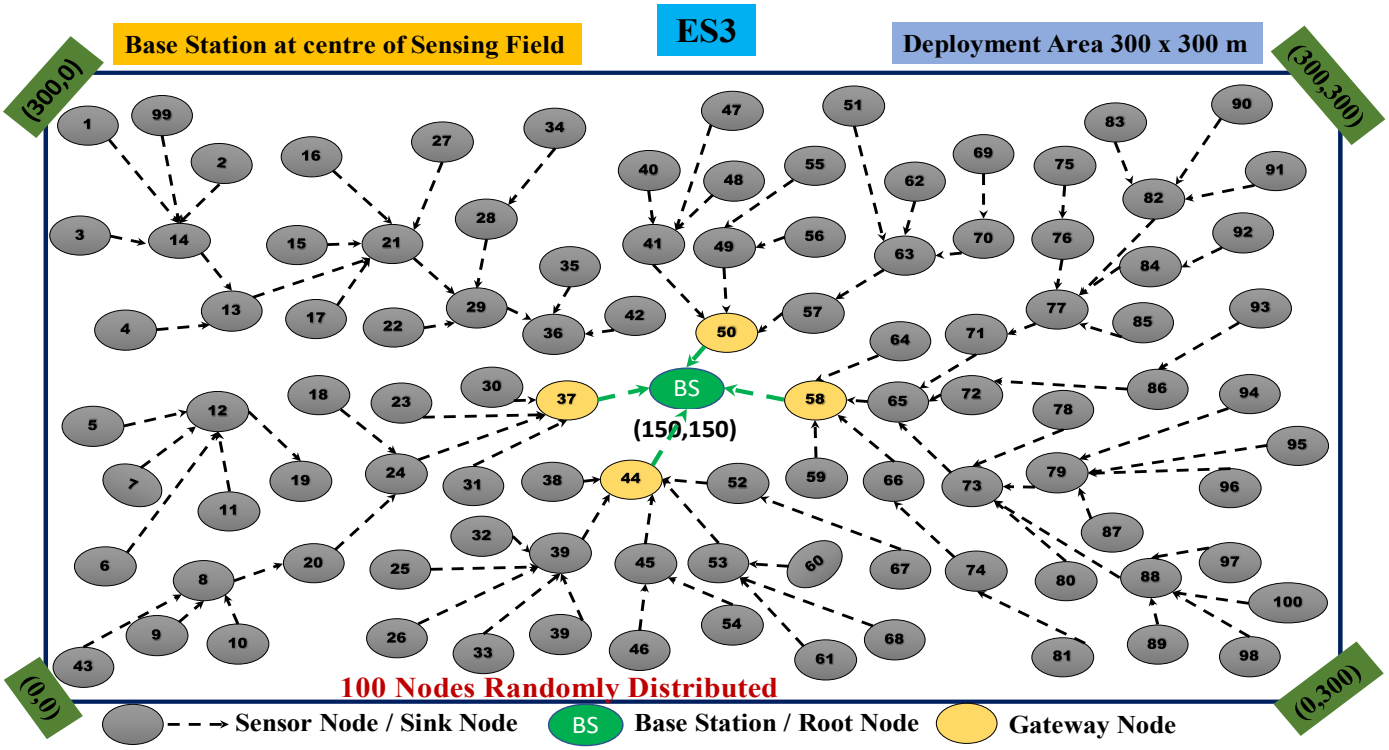


Figure 13: ES3 (Energy Efficient Sink Selection Scheme) protocol routing process when the BS is placed inside the sensing field

Drawback:

- It uses only *free-space model* (d^2 *power loss*) for propagation of sensor information.
- The energy consumption of tier heads located in tier T_3 which is far from BS will deplete its energy faster than normal and so it leads to reselection of tier heads frequently in the tiers far from BS.

4.1.4 ES3 protocol (Efficient Sink Selection Scheme)

Aim: To increase network performance and reduce average energy consumed by the node in large-scale sensor networks.

Protocol Description:

ES3 [29,30] uses *tree-based* routing. Here the nodes in ES3 network are classified into three types: Base station / Root Node, sink node/sensor node and Gateway node. Root node is the first node existed in the network with its hop zero. This node search for child nodes by sending broadcast packets. Here every sensor node can act as a sink node for relaying information of cluster of sensor nodes. Gateway nodes are utilized to handle the communication load of BS. The routing process in ES3 when BS is placed at center of sensing field is shown in Figure 13.

The energy model used in ES3 is same as in CEER network as explained from Eq. (2) and Eq. (3). The weight of sensor node when acting as a sink, differs from its calculation of CEER network as shown in Eq. (31) [31]

$$W_{S_n} = \frac{\rho}{dist_m} + \frac{\vartheta}{C_m+1} + \sigma R_m + \frac{\tau}{HoP_m+1} \quad (31)$$

Here, W_{S_n} is the weight of sink node S_n. C_m is the current count of child nodes connected to sink node S_n. R_m is sink node S_n left over energy. $dist_m$ is the distance between current node and Sink node S_n . We define the root node hop to be 0 and gate way nodes hop by default is set to 1. HoP_m is the hop of sink node S_n. ρ, ϑ, σ, and τ are standardized parameters of four variables $dist_m$, C_m , R_m and HoP_m and are set at values 15, 6, $1/20$, 11 respectively during the simulation.

Advantages:

- Suitable for Large-scale sensor networks
- Provides balance between various network performance metrics like Energy consumption, communication distance, hop count and child node count.

Table 2: Comparison between the PEGASIS, PRRP, ES3 and CEER protocols

	PEGASIS	PRRP	ES3	CEER
Scalability (Increasing number of nodes after network establishment)	very low	Low	Medium	High
Transmission delays	Very high, since it uses a simple control token passing approach	Medium, as only the tier heads of the extreme tiers have to use the intermediate tier heads to relay the information to BS	Quite high, as no proper clustering structing is maintained during the data transmission phase	Quite low, as the local quadrant information is routed directly to zone co-ordinators and one among them will take the responsibility to relay the information to basestation
Distribution of sensor Nodes in sensing field	Randomly Distributed	Uniform Random distribution	Randomly Distributed	Uniform Random Distribution
Control Message Overhead	Medium	Low as the network area is divided into tiers and overhead is added to data when the information is relayed through neighbour tier heads	Medium	Very low as the overhead is added to the data only when an entire zonal information has to be relayed through neighbour zonal heads in single or dual hop to reach the basestation
Uniform Distribution of Energy	High, when implemented will lesser node count as it has cain network architecture	Medium, due to its tier-based network architecture as the extreme right or extreme left tier heads faces few problems	Quite High, as it approach is tree based hierarchical approach	Quite High, as the network area is uniformly divided into equal zones and further into equal quadrants
Energy Efficiency	Low	Low as the sensor nodes in extreme tier heads has to dissipate more energy.	Very high, as it balances distance, hop, number of child nodes and residual energy	Very high, as it uses centroid clustering approach to find the best zonal head of a particular zone
Inter cluster/ quadrant structure	1-Hop	1-Hop	Multihop	1-Hop
Algorithm Complexity	High	Medium	High	Medium

Table 3: Simulation Parameter Table of ES3, PRRP and CEER networking protocols

Parameter	**Value**		
Protocol	**PRRP**	**ES3**	**CEER**
Network area (m)	100 × 100 m	100 × 100 m	100 × 100 m
Number of nodes	100	100	100
Sensor deployment	Uniform Random	Random	Uniform Random
Location of the sink	Centre of the Sensing Field	Centre of the Sensing Field	Centre of the Sensing Field
Channel Access Mechanism	MAC IEEE 802.11	MAC IEEE 802.11	MAC IEEE 802.11
Propagation Model	Free space Model	Free space/Multipath fading	Free space/Multipath fading
Total Simulation Time (sec)	200 Sec	200 sec	200 Sec
Initial Energy of each Node (J)	2 Joule	1.5 Joule	2 Joule
Packet Size (l)	64 Bytes	64 Bytes	64 Bytes
Transmitting/Reception Energy	400nJ/byte	$0.4\mu J/byte$	$50nJ/bit$
ϵ_{fsm} (d^2 power loss) (Amplifier Energy) E_{amp}	10pJ/bit/m^2	$0.01nJ/bit/m^2$	$10pJ/bit/m^2$
ϵ_{mpf} (d^4 power loss)	NA	$1.3nJ/bit/m^2$	$1.3nJ/bit/m^2$
Energy Threshold (initial) E_{th}	1.5J	$0.75J$	$1J$

E_{th} (For one simulation Round)	0.4 mJ	0.4 m*J*	0.4 m*J*
Bandwidth	0.2 Mbps	0.2 Mbps	0.2 Mbps
Beam Forming Energy (*nJ/bit*)	5	5	5
Capture Threshold	10	10	10
Carrier Sense Threshold	3.652e-8 i.e., 250 m	22.427e-8 i.e., 50 m	3.652e-8 i.e., 250 m
Receiver Threshold	3.652e-8 i.e., 250 m	22.427e-8 i.e., 50 m	3.652e-8 i.e., 250 m
Interface Queue Length for Node	50 Packets	2 Packets	50 Packets
Receiving Power of Node	0.1 W	0.395 W	0.1 W
Transmitting Power of Node	0.2 W	0.66 W	0.2 W
Route Timeout	10 sec	10 sec	10 sec
Active Route Timeout	10 sec	10 sec	10 sec
Reverse Route Life	6 sec	6 sec	6 sec
Broadcast ID save	Up to 6 Times	Up to 6 Times	Up to 6 Times
Route Request Retries	3 Times	3 Times	3 Times
Maximum Route Request Timeout	10 Sec	10 Sec	10 Sec
Node Traversal Time	0.03 Sec	0.03 Sec	0.03 Sec
Local Repair Wait Time	0.15 Sec	0.15 Sec	0.15 Sec
Network Diameter	30 Hops	10 Hops	30 Hops
Route Reply Wait Time	1 Sec	1 Sec	1 Sec
Address Resolution Protocol Delay	0.01 Sec	0.01 Sec	0.01 Sec
Hello Interval	5 Sec	5 Sec	5 Sec
Allowed Hello Loss	3 Sec	3 Sec	3 Sec
Bad Link Lifetime	3 Sec	3 Sec	3 Sec
Max Hello Interval	(1.25 * Hello Interval)	(1.25 * Hello Interval)	(1.25 * Hello Interval)
Min Hello Interval	(0.75 * Hello Interval)	(0.75 * Hello Interval)	(0.75 * Hello Interval)
RTQ Maximum Length	64 Packets	64 Packets	64 Packets
RTQ Timeout	30 Sec	30 Sec	30 Sec
Max Errors Allowed	100	100	100

Drawback:

- Delay in data transmission due to more control overheads and normalized overheads added to sensed data transfer process.

4.2 Simulation performance metrics with results

The simulation parameters and their values used in NS-2 simulation of protocols PRRP, ES3 and CEER are displayed in Table 3 and their comparison results with conventional LEACH and traditional PEGASIS are shown in graphs under respective sections below.

4.2.1 Comparison in terms of Total Energy Consumption of Sensor Network

Augmenting the energy balance in sensor nodes minimizes the energy dissipation during network communications. Total energy consumption of sensor network is the energy consumed to perform transmission, reception and data aggregation.

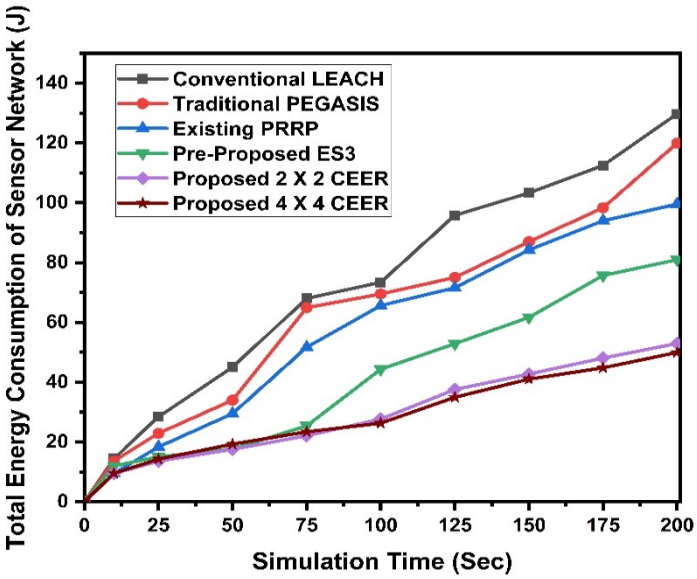


Figure 14: Total energy consumed by proposed and pre-existing sensor networks

in a simulation round of t = 200 Sec when BS is placed at centre of sensing area with co-ordinates (50,50)

When the protocols LEACH, PEGASIS, PRRP, ES3 and CEER are simulated under the same platform of NS-2 version 2.32 for a simulation round of t = 200 Sec, CEER network proved to consume less energy than the other network as shown in Figure 14. The energy saving table of CEER networks when compared to others is shown in Table 4.

Table 4: Outperformance by CEER network in terms of Energy at the end of simulation round

	Comparison with ES3	Comparison with PRRP	Comparison with PEGASIS	Comparison with LEACH
2 X 2 CEER	Saves 28 J	Saves 46.627 J	Saves 66.9947 J	Saves 76.80 J
4 X 4 CEER	Saves 31 J	Saves 49.632 J	Saves 70 J	Saves 79.80 J

4.2.2 *Comparison in terms of Average Energy Consumption of Sensor Node in Sensing Field* [32,33]

The performance metric "Average Energy consumption of sensor node" is vital in CEER as it determines its network lifetime. In any other protocol this metric has its importance as it shows how efficient is the proposed or implemented algorithm to compete with the existing works. In our scenario, when this metric result is compared with existing works as shown in Figure 15, CEER network saves energy consumed by a sensor node as its algorithm finds energy effective sink for every sensor node deployed in the network area and transmit the data through it to reach the basestation. Indirectly, the energy load is distributed equally and effectively among all the sensor nodes in the network.

From Figure 15, CEER network shows that a sensor node deployed inside its network can have longer lifespan i.e., around 62% than conventional LEACH, 56% than traditional PEGASIS, 38% than existing PRRP and 22% than our pre-proposed ES3.

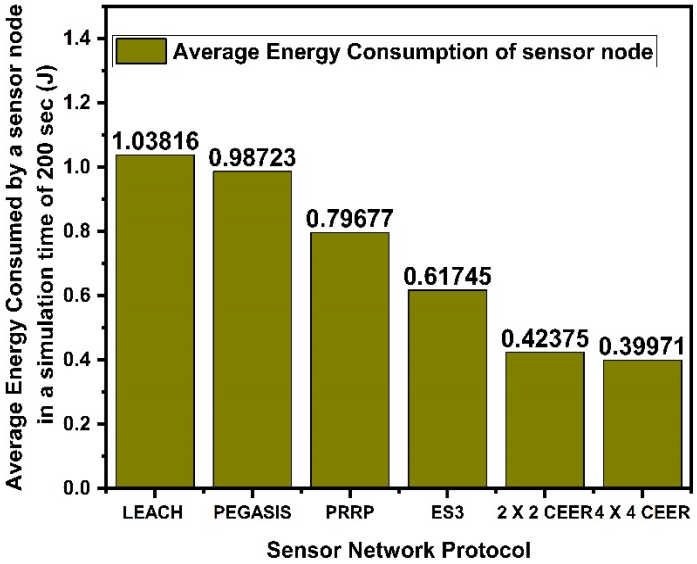


Figure 15: Average Energy consumed by a sensor node in simulated network protocols at the end of simulation time of t = 200 sec when the basestation is placed at co-ordinates (50,50) in the network area.

4.2.3 *Comparison in terms of packet delivery ratio (PDR) of Sensor Network*

Packet Delivery Ratio (PDR) is defined as the percentage ratio of packets received by the basestation and the generated packets by the source sensors. In *ns* terms it is defined as the ratio between the client of packages originated by the "application layer" CBR (Constant Bit Rate) sources and the package count received by CBR sink (BS) at the final destination.

$$Packet\ Delivery\ Ratio\ (PDR) = \frac{(\sum Total\ packets\ received\ by\ BS\ at\ time\ t) * 100}{\sum Total\ packets\ sensed\ by\ sensor\ nodes\ in\ network\ at\ time\ t} \quad (32)$$

PDR in the CEER network is high as this hierarchical protocol is constructed with stable routing operational procedure. As shown in Figure 16, at the beginning of the network initialization and initial data transfer phase, the PDR of CEER network might be low, but as the network enters the stable operation phase as mentioned in section 3, the PDR is quite good and more as compared to other simulated works under simulation time t and BS at the centre of the sensing field.

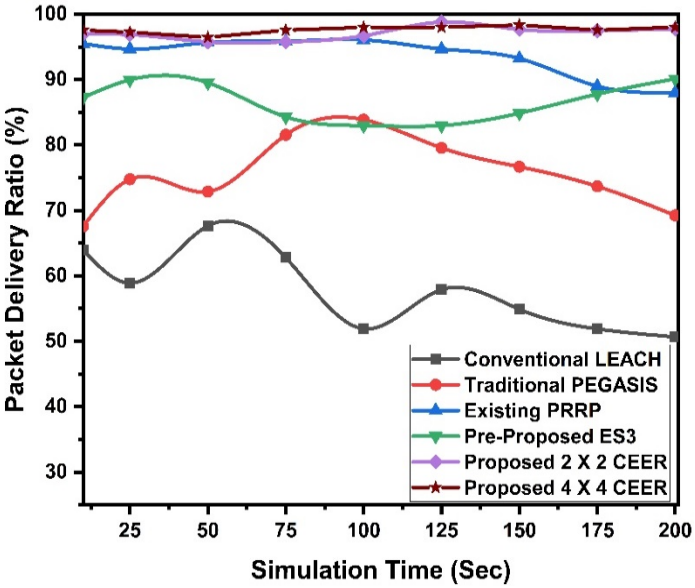


Figure 16: Packet delivery ratio (PDR) of simulated sensor networks in a simulation time t = 200 Sec, when the BS is placed at centre of sensing field at location co-ordinates (50,50)

4.2.4 *Comparison in terms of packet loss percentage in Sensing Field*

Packet loss rate is calculated in wireless sensor networks to check the reliability of a communication network path. This packet loss percentage is equal to the unreceived packets by the BS to the total packets sensed by sensor nodes in the network. Packet loss occurs due to failure of sensor node that carries data across the sensing area of the network.

$$Packet\ loss\ (P.L) = \begin{pmatrix} Total\ packets\ sensed\ by \\ sensor\ nodes\ in\ network \\ in\ simulation\ time\ t \end{pmatrix} - \begin{pmatrix} Total\ Packets\ received \\ by\ BS \\ in\ simulation\ time\ t \end{pmatrix} \quad (33)$$

Packet Loss Percentage =

$$\frac{(P.L*100)}{\text{Packets sensed by sensor nodes in network in simulation time } t} \quad (34)$$

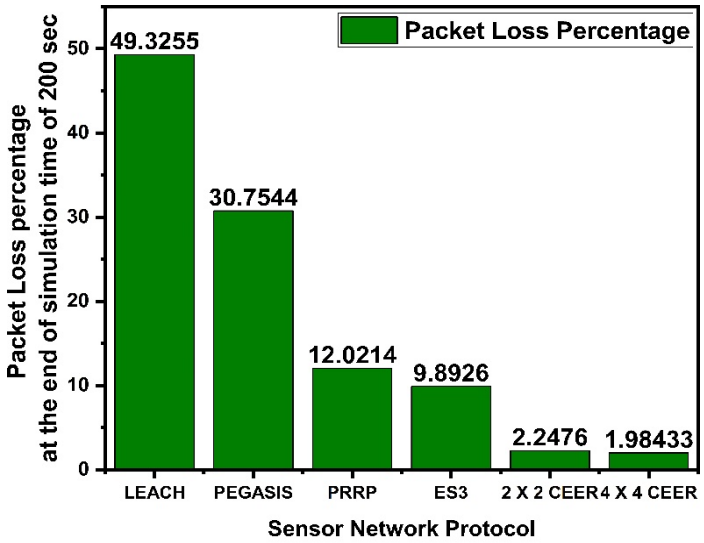


Figure 17: Packet Loss Percentage of sensor networks simulated in time t = 200 sec

Results from Figure 17 shows acceptable packet loss for CEER networks. Reliable protocols react to packet loss automatically and so as the CEER network as its nodes save the excess packets received in one transmission round in the queue memory to avoid further packet loss. Moreover, when compared to conventional LEACH and traditional PEGASIS protocols, CEER network packet loss is more than 47%, 28% lesser respectively. Similarly, when compared to existing PRRP protocol, it is around 10% lesser.

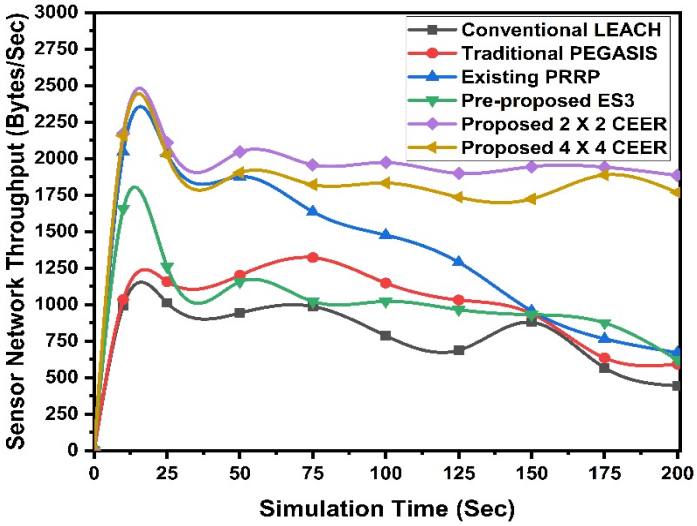


Figure 18: Sensor network throughput measured in bytes/sec of the protocols simulated in a simulation time of t = *200* sec when the BS is placed at locational co-ordinates (50,50)

4.2.5 Comparison in terms of Sensor Network Throughput

Throughput is the rate at which information is sent through the network. Network throughput in CEER networks is an amount of data transferred successfully from source sensors to base station at time t sec and its result graph is shown in Figure 18. High ratio of unsuccessful packet delivery leads to lower throughput and reduces the network lifespan. Throughput capacity is usually measured in *bits/sec*.

*Throughput (bits/sec) = (Packet count successfully received at BS * packet size)/Total time t spent in delivering those packets from source sensor to base station* (35)

Packet loss, latency and jitter are all related to reduce the sensor network throughput. All these three metrics are proved to be lesser in CEER network than the compared protocols when simulated based on the assumptions made in section 3.

5 Conclusions and Further proceedings

In this research work, we propose a clustering centroid-based energy efficient routing (CEER) protocol to solve the problem of forming zones and its clusters to transmit information energy efficiently to the basestation or sink node. An optimized path is achieved through CEER networks in the zones formed, by selecting energy efficient zone co-ordinators and zone heads. Energy centroid routing is proposed in this paper by considering the residual energy of sensor nodes and communication distance to the sink/basestation. The data path between zone heads and basestation is optimized at both one-hop level and two-hop levels by relaying the information to basestation via intermediate zone heads to avoid packet loss at the zone heads and sometimes directly to basestation itself if the transmission distance between the zone head and the basestation is less. From the simulation results, when the basestation is placed at center of the sensing field, CEER networks can transmit an ample information by consuming less amount of energy. Its simulation results proved that the average energy consumption of a sensor node inside CEER network consumes energy i.e., atleast 60% and 55% less than the conventional LEACH and traditional PEGASIS protocols respectively and atmost 35% than the existing PRRP protocol.

In future work, we have planned to optimize the path between the Zone Heads (ZH) and the Basestation (BS) by introducing multi-hop communication between them unlike the current dual hop and to implement it in IoTN scenario when the BS is placed outside the sensing field at the co-ordinates (50,125). In case if the basestation is still far reachable that the ZH has to dissipate most of its energy for data transfer, we planned to optimize the path using ANT routing strategy to avoid the packet loss.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] N. P. R. Kumar and G. J. Bala, "An Energy Efficient Quadrant Based Position Responsive Routing Protocol," 2019 2nd International Conference on Signal Processing and Communication (ICSPC), Coimbatore, India, 5-10, 2019. https://doi.org/10.1109/ICSPC46172.2019.8976781.

[2] H. Ali et al., "Clustering methods for cluster-based routing protocols in wireless sensor networks: Comparative study", Comparative study. International Journal of Applied Engineering Research, **12**(21), 11350-11360, 2017. http://www.ripublication.com

[3] W. Xinhua and W. Sheng, "Performance Comparison of LEACH and LEACH-C Protocols by NS2," 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Hong Kong, 254-258, 2010. https://doi.org/10.1109/DCABES.2010.58.

[4] N. Wang and H. Zhu, "An Energy Efficient Algorithm Based on LEACH Protocol," 2012 International Conference on Computer Science and Electronics Engineering, Hangzhou, 339-342, 2012. https://doi.org/10.1109/ICCSEE.2012.150.

[5] L. Li and C. Liu, "An Improved Algorithm of LEACH Routing Protocol in Wireless Sensor Networks," 2014 8th International Conference on Future Generation Communication and Networking, Haikou, 45-48, 2014. https://doi.org/10.1109/FGCN.2014.18.

[6] S. Gupta and N. Marriwala, "Improved distance energy based LEACH protocol for cluster head election in wireless sensor networks," 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, 91-96, 2017. https://doi.org/10.1109/ISPCC.2017.8269656.

[7] R. Dasondhi, M. Singh, D. Kulhare, "An Algorithm for Balanced Cost Cluster-Heads Selection for Wireless Sensor Network" International Journal of Engineering Research & Technology, **01**(10), 2012. IJERTV1IS10526

[8] N. Enami, R. A. Moghadam and K. D. Ahmadi, "A new neural network based energy efficient clustering protocol for Wireless Sensor Networks," 5th International Conference on Computer Sciences and Convergence Information Technology, Seoul, 40-45, 2010. https://doi.org/10.1109/ICCIT.2010.5711026.

[9] S. Mostafavi, V. Hakami, "A new rank-order clustering algorithm for prolonging the lifetime of wireless sensor networks" Int J Commun Syst., 33(e4313), 2020, https://doi.org/10.1002/dac.4313

[10] P. Sasikumar and S. Khara, "K-Means Clustering in Wireless Sensor Networks," *2012* Fourth International Conference on Computational Intelligence and Communication Networks, Mathura, 140-144, 2012. https://doi.org/10.1109/CICN.2012.136.

[11] W. Fakhet, S. E. Khediri, A. Dallali and A. Kachouri, "New K-means algorithm for clustering in wireless sensor networks," 2017 International Conference on Internet of Things, Embedded Systems and Communications (IINTEC), Gafsa, 67-71, 2017. https://doi.org/10.1109/IINTEC.2017.8325915.

[12] Jorio A. Sanaa El Fkihi, Brahim Elbhiri, and Driss Aboutajdine "An Energy-Efficient Clustering Routing Algorithm Based on Geographic Position and Residual Energy for Wireless Sensor Network" Journal of Computer Networks and Communications, 2015, https://doi.org/11.10.1155/2015/170138.

[13] D. Takaishi, H. Nishiyama, N. Kato and R. Miura, "Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks" IEEE Transactions on Emerging Topics in Computing, **2**(3), 388-397, 2014. https://doi.org/10.1109/TETC.2014.2318177.

[14] R. Vijayashree and C. Suresh Ghana Dhas., "Energy efficient data collection with multiple mobile sink using artificial bee colony algorithm in large-scale WSN", Automatika, **60**(5), 555-563, 2019, https://doi.org/10.1080/00051144.2019.1666548

[15] G. Kumar and J. Singh, "Energy efficient clustering scheme based on grid optimization using genetic algorithm for wireless sensor networks," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 1-5, 2013. https://doi.org/10.1109/ICCCNT.2013.6726634.

[16] N. Thakur and R. K. Chauhan, "Conservation of energy by using grid clustering in wireless sensor networks," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, 591-596, 2016. https://doi.org/10.1109/PDGC.2016.7913192.

[17] A. Ray and D. De, "Energy efficient clustering protocol based on K-means (EECPK-means)-midpoint algorithm for enhanced network lifetime in wireless sensor network," IET Wireless Sensor Systems, **6**(6), 181-191, 2016. https://doi.org/10.1049/iet-wss.2015.0087.

[18] A. A. Shaikh and D. J. Pete, "Spatial Correlation and Centroid Based Clustering in Wireless Sensor Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 1-5, 2018. https://doi.org/10.1109/ICCUBEA.2018.8697416.

[19] R. Daniel and K. N. Rao, "EEC-FM: Energy Efficient Clustering based on Firefly and Midpoint Algorithms in Wireless Sensor Network," KSII Transactions on Internet and Information Systems, **12**(8), 3683-3703, 2018. https://doi.org/10.3837/tiis.2018.08.008

[20] Q. Kashif et al., "Optimized Cluster-Based Dynamic Energy-Aware Routing Protocol for Wireless Sensor Networks in Agriculture Precision", Journal of Sensors. 1-19, 2020. https://doi.org/10.1155/2020/9040395.

[21] S. Loganathan, J. Arumugam, "Energy centroid clustering algorithm to enhance the network lifetime of wireless sensor networks"., Multidim Syst Sign Process 2019. https://doi.org/10.1007/s11045-019-00687-y

[22] J. Shen, A. Wang, C. Wang, P. C. K. Hung and C. Lai, "An Efficient Centroid-Based Routing Protocol for Energy Management in WSN-Assisted IoT," IEEE Access, 5, 18469-18479, 2017. https://doi.org/10.1109/ACCESS.2017.2749606.

[23] H. T. Friis, A note on a simple transmission formula.Proc. IRE, 34, 1946.

[24] T. S. Rappaport, Wireless communications, principles and practice.Prentice Hall, 1996.

[25] R. Hetal, V. Sangeeta, A. Mohammad, "Comparative Study of PEGASIS Protocols in Wireless Sensor Network. IOSR Journal of Computer Engineering," 16, 25-30, 2014. https://doi.org/10.9790/0661-16512530

[26] R. K. Yadav and A. Singh, "Comparative study of PEGASIS based protocols in wireless sensor networks," 1st India International Conference on Information Processing (IICIP), Delhi, 1-5, 2016. https://doi.org/10.1109/IICIP.2016.7975320.

[27] N. Zaman, A.B. Abdullah, "Energy Optimization through Position Responsive Routing Protocol (PRRP) in Wireless Sensor Network," International Journal of Information and Electronics Engineering, **2**(5), 748-751, 2012. https://doi.org/10.7763/IJIEE.2012.V2.199

[28] N. Zaman and A. B. Abdullah, "Position Responsive Routing Protocol (PRRP)," 13th International Conference on Advanced Communication Technology (ICACT2011), Seoul, 644-648, 2011.

[29] T. Qiu, X. Liu, L. Feng, Y. Zhou and K. Zheng, "An Efficient Tree-Based Self-Organizing Protocol for Internet of Things," IEEE Access, 4, 3535-3546, 2016. https://doi.org/10.1109/ACCESS.2016.2578298.

[30] Nalluri, Raj Kumar & Bala, G. "An Efficient Energy Saving Sink Selection Scheme with the Best Base Station Placement Strategy Using Tree Based Self Organizing Protocol for IoT", Wireless Personal Communications **109**(2), 869-895, 2019. https://doi.org/10.1007/s11277-019-06595-5.

[31] H. Zhang, P. Chen and S. Gong, "Weighted spanning tree clustering routing algorithm based on LEACH," 2010 2nd International Conference on Future Computer and Communication, Wuha, 223-227, 2010. https://doi.org/10.1109/ICFCC.2010.5497381.

[32] S. Krit & L. Elmaimouni, "Energy consumption in wireless sensor network: simulation and comparative study of flat and hierarchical routing protocols"., IADIS International Journal on Computer Science and Information Systems. 12, 109-125, 2017.

[33] H. Oudani, J. Laassiri, S. Krit and L. El Maimouni, "Comparative study and simulation of flat and hierarchical routing protocols for wireless sensor network," 2016 International Conference on Engineering & MIS (ICEMIS), Agadir, 1-9, 2016. https://doi.org/10.1109/ICEMIS.2016.7745357.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 299-306 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Complex Order $PI^{\alpha+j\beta}D^{\gamma+j\theta}$ Design for Surface Roughness Control in Machining CNT Al-Mg Hybrid Composites

Ravi Sekhar [*,1], Tejinder Paul Singh[2], Pritesh Shah[1]

[1]*Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University) (SIU), Pune, India*

[2]*Department of Mechanical Engineering, Thapar Institute of Engineering & Technology, Patiala, India*

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 01 November, 2020
Online: 08 November, 2020

Keywords:
Complex order controller
Fractional calculus
Fractional order controller
Nano composites
Carbon nanotubes
Fractional modeling
Bode diagram

ABSTRACT

Accurate machining control is indispensable for the smart factories of tomorrow. Variations in controller responses may cause unacceptable process deviations during machining leading to productivity losses and possible damage. In the present work, a complex order $PI^{\alpha+j\beta}D^{\gamma+j\theta}$ (COPID) controller was designed to effectively control surface roughness generation while machining CNT Al-Mg hybrid composites. Performance of the designed complex order controller was compared against the conventional PID and fractional order PID (FOPID) controllers for the machined surface roughness system. Output signal responses indicate that the complex order controller attains the desired surface roughness set point with zero percent overshoot in almost same settling time (46 sec) as PID (41 sec) and FOPID (46 sec) controllers. The PID and FOPID output signals registered overshoots of 96.8 % and 36.7 % respectively. Similarly, in case of control signals (feed rate) the COPID controller successfully minimised peak overshoot to 4.6 %; as compared to 64.2 % and 96.5 % in case of the PID and FOPID controllers respectively. The COPID controller was also effective in reducing its peak time response metric (2.001 % for control signal and no peak time for output signal due to zero overshoot). In comparison, the PID and FOPID controllers recorded higher response peak times (7 sec / 26 sec for the PID ouput/control and 5 sec / 7.84 sec for FOPID output/control signal responses). Overshoot elimination in output signal (surface roughness) is crucial for consistency of the machined surface quality. Similarly, overshoot minimisation in control signal (feed rate) is critical because excessive feed rate can damage the cutting tool, work piece, machinery and is a potential safety hazard for the machine operator as well. Hence, the COPID controller can be safely and extensively applied in smart industrial control systems of the future.

1 Introduction

This paper is an extension of the work originally reported in the 2019 IEEE International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE) [1]. In that work, firstly the multi input single output system of surface roughness generation during machining of the CNT Al-Mg composites was identified using the ARX and ARMAX model structures. Thereafter, the fractional order PID (FOPID) controller was designed to control the identified system at the desired set point (surface roughness). The authors reported that the ARMAX model based FOPID control exhibited better time domain characteristics as compared to the ARX model based FOPID controller design. The current work extends this work further to include complex order controller design to improve the time domain performance even more. PID controller has also been implemented for better comparisons. Moreover, the output signal time domain specifications corresponding to the machining variables not discussed in the previous work have also been included in the present study.

Metal matrix composites are new age materials finding new and varied applications due to their superior mechanical properties over conventional alloys [2]. Carbon nanotubes (CNT) in particular have attracted attention of researchers worldwide owing to the superior strengthening characteristics imparted by these nano sized particles to the base metal alloys [3]. Aluminum based alloy systems have almost a universal presence across industry sectors and products [4]. Combination of carbon nantubes and aluminum-magnesium alloys results in composites with superior mechanical characteristics.

[*]Corresponding Author: Ravi Sekhar, Symbiosis Institute of Technology, Pune, India Contact No. +91 20 28116400 & Email ravi.sekhar@sitpune.edu.in

Modeling and designing effective controllers is an important step towards automating the machining of these modern materials.

Generally, PID controllers find extensive applications due to their ease of implementation and maintenance in industrial systems. However, specialized machinery may require more sophisticated controllers having better design characteristics. Fractional order controllers offer a better design alternative because they have two more parameters to tune as compared to the traditional PID controllers [5, 6]. These controllers are based on fractional calculus, a branch of integral calculus that enables fractional orders of the model structure parameters. Fractional order PID controllers (FOPID) have been demonstrated to exhibit better performance characteristics in various application areas such as control systems, signal processing, medical research, material science and many more [7]–[10]. FOPID controllers prove to be more robust against system variations against their PID counterparts. Therefore, FOPID controllers are being increasingly preferred in more dynamic systems across industries [11]–[14]. Machining of complex materials like the metal matrix composites is one such example of a dynamic system; wherein the material composition is anisotropic and calls for a more robust controller architecture [6].

Complex order PID (COPID) controllers are the next step in the controller evolution [15]. They are an extension of the FOPID controllers just as the FOPID are extended versions of the PID controllers. The COPID have further two parameters to tune over the FOPID controllers. In the recent years, COPID controllers have been gradually attracting increasing investigations into their design and application aspects [16]–[18]. Khandani et al [19] designed a complex order controller for DC motors. Pinto and Carvalho [20] derived a complex order model for HIV infection drug resistance. Machado [18] optimized complex order controller for linear and non linear systems using genetic algorithm. Ayadi and Amairi [21] applied numerical optimization to tune complex order controller for a second order time delay resonant system. Hanif et al [22] implemented genetic algorithm for tuning of complex order controller design. Silva et al [23] applied complex order structure to model hexapod robot locomotion based on foot-ground transfer function. Jacob et al [15] reviewed frequency domain, time domain and stability aspects in complex order modeling of linear systems. Tare et al [24] compared integral, fractional and complex order controller performances for fractional order systems. Adams et al [25] determined the transfer function solution to the complex order differential equation. Shahiri et al [26] investigated robust control of a non linear fuel cell system using complex order architecture. From the literature review, it appears that the application of COPID controller in machining systems needs attention. The current work aims to bridge this gap.

The CNT Al-Mg composite experimental details may be referred to in the previous part [1] of the current work. It involved variation of the input machining parameters (cutting speed, feed and depth of cut) and data collection of the corresponding values of the machined surface roughness. Machining experiments were performed on a CNC turning center (Simple Turn 5057 Haas Automation) using chemical vapour deposition (CVD) coated tool inserts (Seco CNMG120408). Surface roughness was measured by a Handy Surf E-DTS570 tester. The best performing model structure (ARMAX) describing this machined surface roughness system in the previous work was adopted in the current study for the COPID and PID controller design. The FOPID controller time domain characteristics from the previous work [1] were referred to in the current work for comparisons with the COPID and PID controller responses. In the previous work, high peak overshoots were detected in output / control signal responses of the FOPID controller. The current study aimed to minimise / eliminate peak overshoots by employing COPID controller design. The following section details upon the design methodologies followed for the implementation of the PID, FOPID and COPID controllers in composite material machined surface roughness system.

2 Controller Design Methodology

This section describes the design methodologies of the PID, FOPID and COPID controllers for the machined surface roughness system. All controllers were designed to attain output set point of 1 micron surface roughness. In machining, cutting speed is inversely related to surface roughness whereas depth of cut directly increases machined surface roughness levels. However, feed rate has the most significant effect on surface roughness generation [2]. Therefore, feed rate parameter was selected as the manipulated variable in the current study whereas cutting speed and depth of cut were maintained at their mid levels as per the experimental design [1]. Thus, all controllers were designed to control this multi input (cutting speed, feed, depth of cut) and single output (surface roughness) system as shwon in Figure 1.

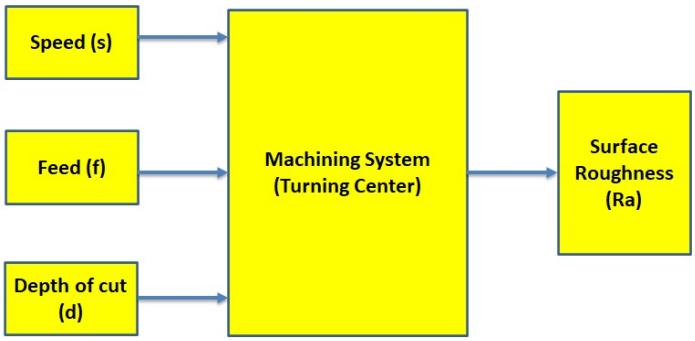


Figure 1: Multi input single output system schematic for the nano composite machined surface roughness system

2.1 PID controller

The PID controller has been widely implemented in industries since many years because of its simple structure and compatibility with the industrial automation software [27]. In the current work, Simulink block was utilised to implement PID controller for the machined surface roughness plant. The PID controller structure is given as follows:

$$C(s) = K_P + K_I\left(\frac{1}{s}\right) + K_D\left(\frac{N}{1 + N\frac{1}{s}}\right) \quad (1)$$

where, K_P is the proportional gain, K_I is the integral gain, K_D is derivative gain and N is the filter coefficient of derivative. The PID block was auto tuned in Simulink using the time domain approach.

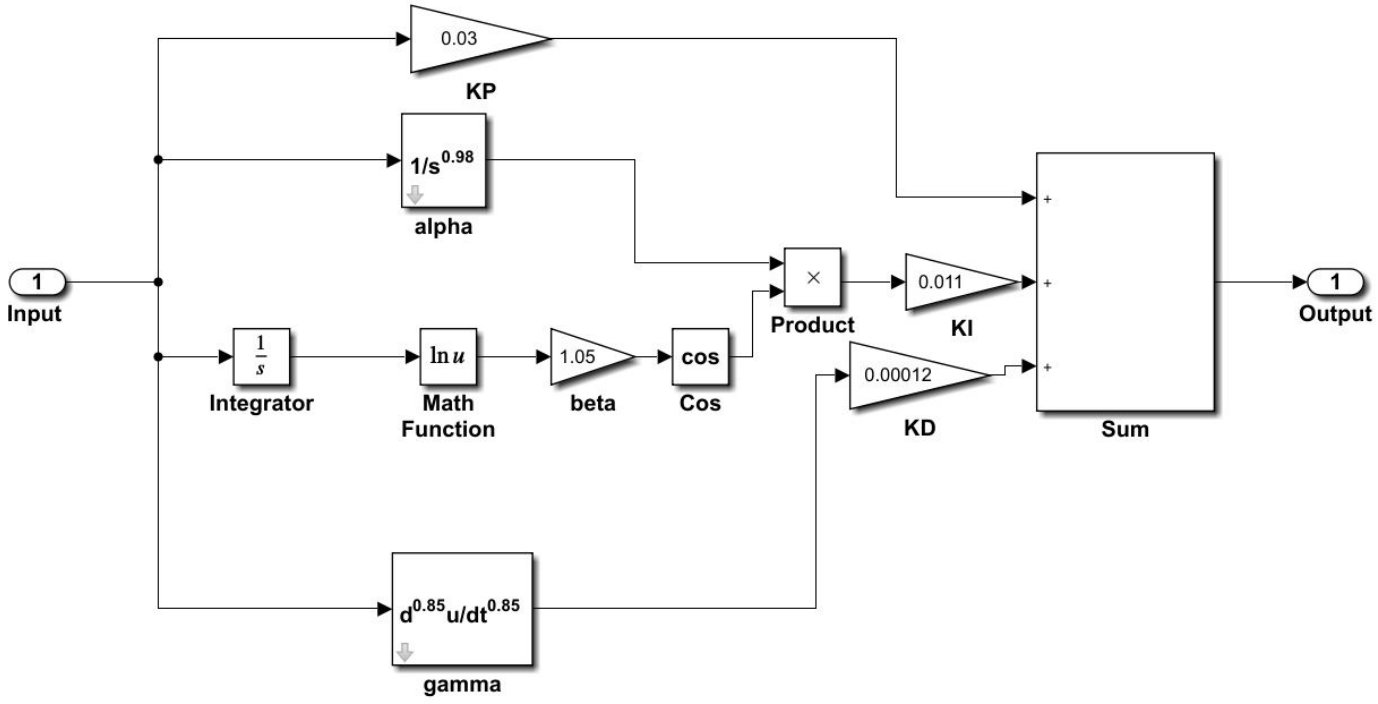


Figure 2: Complex $PI^{\alpha+j\beta}D^{\gamma+j\theta}$ Controller Structure in Simulink

2.2 FOPID controller

Fractional PID (FOPID) controller is an extension of the classical PID controller and is designed based on fractional calculus [28]. In this controller, the orders of integration and derivation are real numbers [29, 14]. The FOPID controller has the following controller structure:

$$C(s) = K_P + K_I\left(\frac{1}{s}\right)^{\lambda} + K_D s^{\mu} \tag{2}$$

where μ is the order of derivative and λ is the order of integration. There are five parameters in the fractional PID controller. This controller has the property of isodamping making it more robust against variations in the system parameters. The values of λ and μ were taken between 0 to 2 for the closed loop system to be stable.

Following equations were employed to synthesise the fractional calculus for the fractional PID controller application:

$$\omega_u = \sqrt{\omega_h \omega_b}$$

$$\omega_0^{'} = \alpha^{-0.5}\omega_u; \omega_0 = \alpha^{0.5}\omega_u;$$

$$\frac{\omega_{k+1}^{'}}{\omega_k^{'}} = \frac{\omega_{k+1}}{\omega_k} = \alpha\eta > 1$$

$$\frac{\omega_{k+1}^{'}}{\omega_k} = \eta > 0; \frac{\omega_k}{\omega_k^{'}} = \alpha > 0$$

$$N = \frac{\log(\omega_N/\omega_0)}{\log(\alpha\eta)}$$

where, N is the order of approximation, ω_h and ω_b are the frequency bounds for approximation.

In the present work, FOMCON toolbox was used for implementation of FOPID and COPID controllers [12]. In this toolbox, Simulink blocks are available for various applications of fractional calculus. It can be used for designing controllers in continuous/discrete time scales, fractional order state space and many more. In FOPID design, the derivative action was considered zero ($K_D = 0$ and $\mu = 0$). The other gain parameters are taken from PID controller settings. Value of λ was selected based on the approach given by Shah and Agashe [29].

2.3 COPID controller

Complex order PID (COPID) controller idea originated from the 3rd generation of CRONE controller [30, 11]. COPID controller is similar to FOPID controller except that the orders of integration and derivation are complex instead of real numbers. The COPID controller has the following structure [16, 26].

$$C(s) = K_P + K_I\left(\frac{1}{s}\right)^{\alpha+j\beta} + K_D\,(s)^{\gamma+j\theta} \tag{3}$$

where, K_P is the proportional gain, K_I is the integral gain, K_D is derivative gain and $\alpha, \beta, \gamma, \theta$ are the orders of the complex order controller.

To implement the above equation, a simplification procedure was carried out in two steps. Firstly, the complex integration was written as:

$$K_I\left(\frac{1}{s}\right)^{\alpha+j\beta} = K_I\left(\frac{1}{s}\right)^{\alpha}\left(\frac{1}{s}\right)^{j\beta}$$
$$K_I\left(\frac{1}{s}\right)^{\alpha+j\beta} = K_I\left(\frac{1}{s}\right)^{\alpha} * e^{\ln\left(\frac{1}{s}\right)^{j\beta}}$$
$$K_I\left(\frac{1}{s}\right)^{\alpha+j\beta} = K_I\left(\frac{1}{s}\right)^{\alpha} * e^{j\beta\ln\left(\frac{1}{s}\right)}$$
$$K_I\left(\frac{1}{s}\right)^{\alpha+j\beta} = K_I\left(\frac{1}{s}\right)^{\alpha} * \left[\cos\left(\beta\ln\left(\frac{1}{s}\right)\right) + j\sin\left(\beta\ln\left(\frac{1}{s}\right)\right)\right]$$

The imaginary part of the above equation cannot be synthesised for time domain implementation of the system. Hence, for implementing the COPID controller in Simulink, imaginary part of the above equation was omitted [16].

$$K_I\left(\frac{1}{s}\right)^{\alpha+j\beta} = K_I\left(\frac{1}{s}\right)^{\alpha} * \left[\cos\left(\beta\ln\left(\frac{1}{s}\right)\right)\right] \quad (4)$$

Similarly, derivative component of the Eq. (3) was simplified for implementation as follows -

$$K_D * s^{\gamma+j\theta} = K_D * s^{\gamma} * s^{j\theta}$$
$$K_D * s^{\gamma+j\theta} = K_D * s^{\gamma} * e^{\ln s^{j\theta}}$$
$$K_D * s^{\gamma+j\theta} = K_D * s^{\gamma} * e^{j\theta\ln s}$$
$$K_D * s^{\gamma+j\theta} = K_D * s^{\gamma} * [\cos(\theta\ln s) + j\sin(\theta\ln s)]$$

The imaginary part of the above equation cannot be synthesised for time domain implementation of the system. Hence, for implementing the COPID controller in Simulink, imaginary part of the above equation was omitted [16].

$$K_D * s^{\gamma+j\theta} = K_D * s^{\gamma} * [\cos(\theta\ln(s))] \quad (5)$$

Combining Eq. (4) and (5), the COPID controller structure may be rewritten as

$$C(s) = K_P + K_I\left(\frac{1}{s}\right)^{\alpha} * \left[\cos\left(\beta\ln\left(\frac{1}{s}\right)\right)\right] + K_D * s^{\gamma} * [\cos(\theta\ln(s))] \quad (6)$$

The COPID controller given in Eq. (6) was initially tuned by incorporating the reference values of FOPID controller and was further fine tuned based on the principles given in literature [31, 29]. The fine tuned COPID controller was implemented in Simulink (Matlab) as shown in figure 2. The initial value of integer order integrator was assumed to be a small number during the controller implementation.

3 Results and Discussions

This section describes the different controller structures, output / control signal responses, time domain characteristics and Bode plot for the COPID controller. The discrete-time ARMAX model (order 3331) with 1 sec sample time from the previous work [1] was obtained as follows:

$$A(z)y(t) = B(z)u(t) + C(z)e(t)$$
$$A(z) = 1 - 0.3675z^{-1} + 0.4876z^{-2} - 0.3772z^{-3}$$
$$B_1(z) = -0.001019z^{-1} - 0.003656z^{-2} + 0.01446z^{-3}$$
$$B_2(z) = 0.6741z^{-1} + 0.2097z^{-2} + 13.82z^{-3}$$
$$B_3(z) = -1.569z^{-1} + 0.4574z^{-2} - 1.532z^{-3}$$
$$C(z) = 1 + 0.4142z^{-1} + 0.554z^{-2} - 0.4004z^{-3}$$

This model depicts the CNT Al-Mg composite machined surface roughness system considered in the current study. The PID controller applied on this model structure has the following structure:

$$C(s) = 0.0407 + 0.0143\frac{1}{s} - 0.0633\frac{0.4367}{1 + 0.4367\frac{1}{s}} \quad (7)$$

The output and control signal responses of this PID controller for a set point of 1 micron surface roughness are shown in Fig. 3 and Fig. 4 respectively.

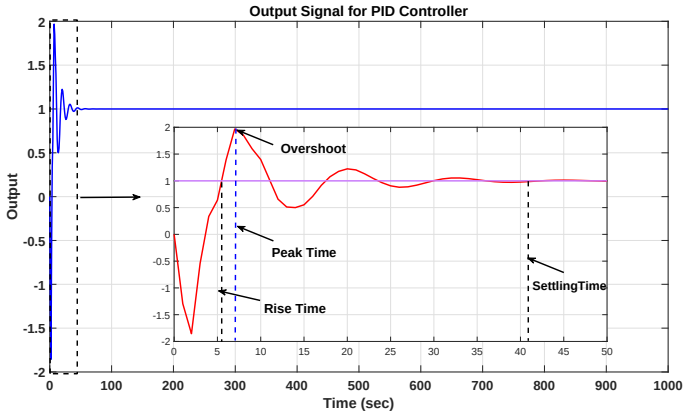


Figure 3: PID controller output (surface roughness, μ) for the machined surface roughness system

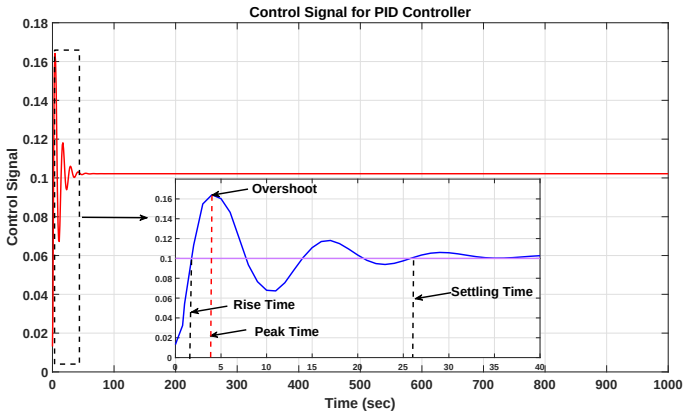


Figure 4: PID control signal (feed rate, mm / rev) for the machined surface roughness system

Following is the structure of the FOPID controller applied to the

composite machined surface roughness system -

$$C(s) = 0.0407 + 0.0143\frac{1}{s^{0.952}} \tag{8}$$

Fig. 5 and Fig. 6 depict the output and control signal plots for the FOPID controller set at 1 micron surface roughness output.

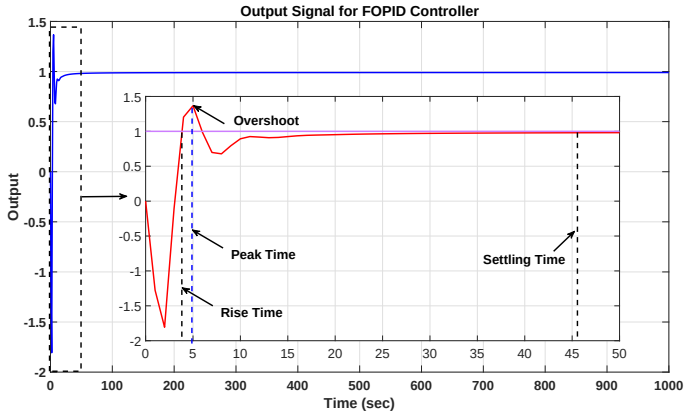


Figure 5: Fractional PID controller output (surface roughness, μ) for the machined surface roughness system

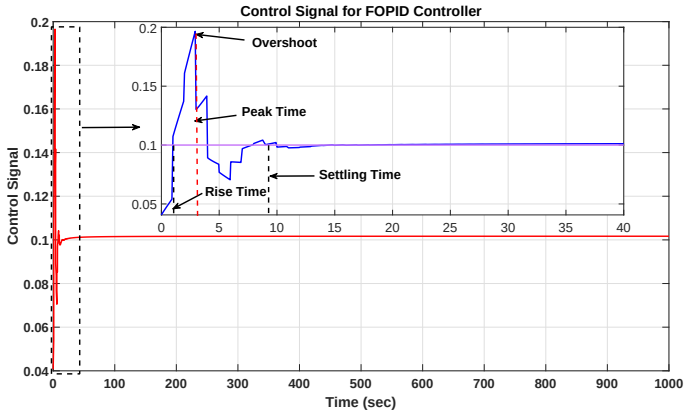


Figure 6: Fractional PID control signal (feed rate, mm / rev) for the machined surface roughness system

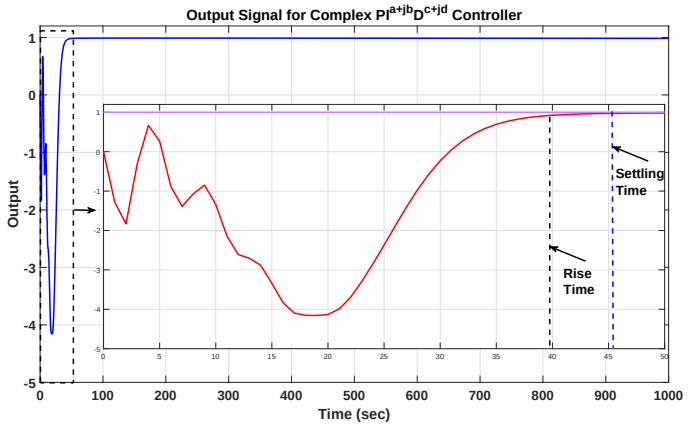


Figure 7: Complex $\mathrm{PI}^{\alpha+j\beta}\mathrm{D}^{\gamma+j\theta}$ controller output (surface roughness, μ) for the machined surface roughness system

The complex order controller structure for the same system is shown below -

$$C(s) = 0.03 + 0.011\frac{1}{s^{0.98+j1.05}} + 0.00012s^{0.85+j0} \tag{9}$$

The output and control signal trends for the COPID controller are shown in Fig. 7 and Fig. 8 respectively.

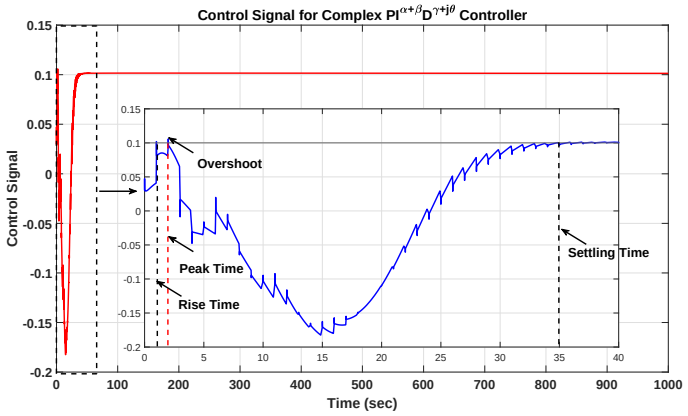


Figure 8: Complex $\mathrm{PI}^{\alpha+j\beta}\mathrm{D}^{\gamma+j\theta}$ control signal (feed rate, mm / rev) for the machined surface roughness system

Table 1: Time domain specifications of output signals

	PID	FOPID	COPID
Rise Time (sec)	5	3.8	39
Peak Time (sec)	7	5	NA
Settling Time (sec)	41	46	46
Overshoot (%)	96.8	36.7	0

Table 2: Time domain specifications of control signals

	PID	FOPID	COPID
Rise Time (sec)	2	1	1
Peak Time (sec)	4	2.924	2.001
Settling Time (sec)	26	7.844	34.38
Overshoot (%)	64.2	96.5	4.6

Table 1 shows the output signal (surface roughness) time domain specifications of the PID, FOPID and COPID controllers for the machined composite surface roughness system. The PID controller scores slightly better than the FOPID and COPID controllers in case of response settling time. PID also scores significantly better than the COPID controller in terms of the rise time. FOPID controller attains the least rise time. However, the COPID controller positively removes peak overshoot from the output signal. The other two controllers have high overshoots in their respective output signal responses. High overshoot in the surface roughness output is undesirable from quality perspective. Generally, critical industrial components require tight dimensional and surface roughness tolerances. Hence, for such cases of precision manufacturing, complex

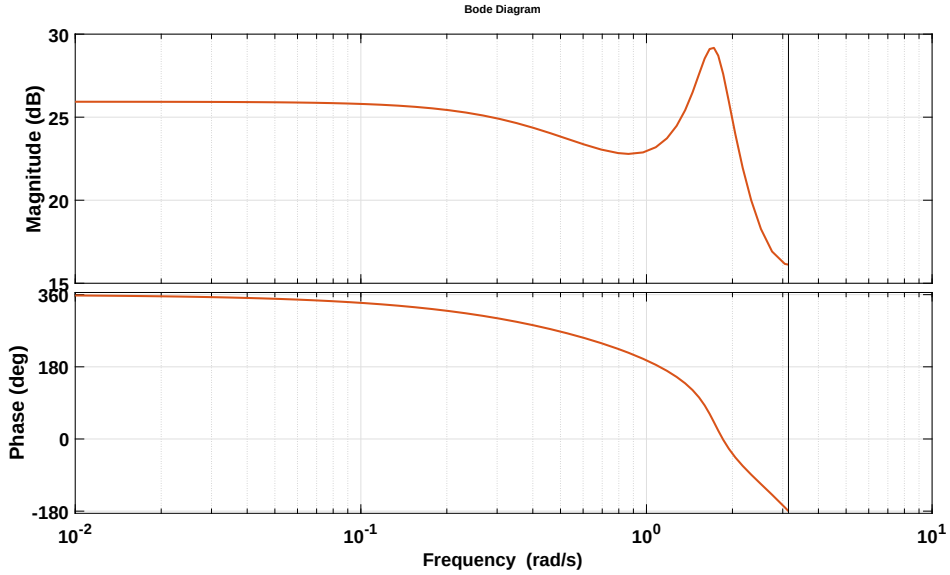


Figure 9: Bode plot for complex order controller

order controller promises to be a better alternative over PID and FOPID controllers.

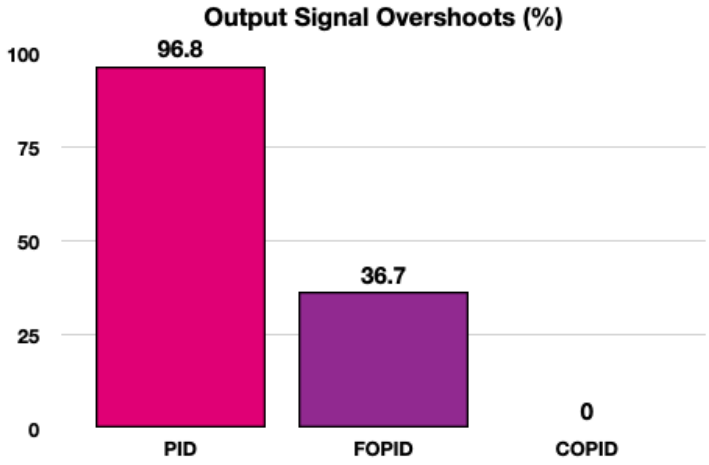


Figure 10: Output Signal Overshoots (%)

Table 2 shows the time domain characteristics of the control signal (feed rate) responses. In this case, the FOPID controller performs well in terms of the rise and settling times. The COPID outperforms the rest for the peak time and overshoot percentages. FOPID has the worst overshoot percentage, whereas the PID controller has the worst peak time characteristic. However, PID controller does perform better than the COPID for settling time. In the current study, control signals correspond to feed rate manipulation in the actual machining systems. High overshoots in feed rate settings may lead to cutting tool damage / breakage, machine tool chatter and possible safety hazard for the machinist. Therefore, minimisation of peak overshoots in feed rate manipulation is a critical machining requirement being fulfilled by the complex order PID controller. Fig. 9 and 10 give visual representations of the relative differences among the output and control signal overshoot percentages attained by the three controllers considered in the current study. Thus, the superiority of the COPID controller over others may be appreciated in light of the importance of the peak overshoot minimisation. Fig. 11 shows the Bode diagram for the complex order controller. This plot indicates that the gain and phase margin for the COPID controller is ∞, implying that the machined composite surface roughness system is robust against variations in the system parameters. This result is crucial from the point of view of the anisoptropic structure of composite materials; that lead to system variations during machining.

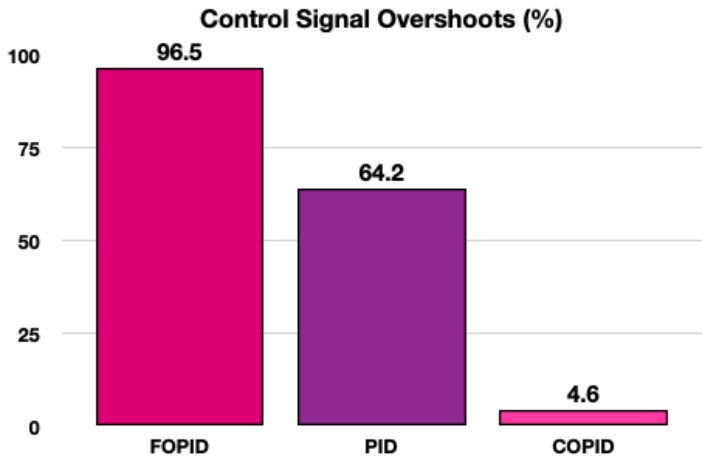


Figure 11: Control Signal Overshoots (%)

4 Conclusions

The current work is an extension of the preliminary study reported in the 2019 IEEE International Conference on Mechatronics, Robotics

and Systems Engineering (MoRSE) [1]. The preliminary study involved system identification and FOPID control of surface roughness evolution in machining of CNT Al-Mg composite materials. The ARMAX model structure controlled by the FOPID controller produced best output signal responses as reported in the maiden paper [1]. The current study expanded that work by including PID and COPID controller designs on the best performing ARMAX model structure determined in the previous work. All three controllers' responses were compared in terms of time domain specifications of the respective output and control signal responses. The current work also included analysis of the output signal trends for the said controllers (which was not considered in the preliminary study). The COPID controller conclusively demonstrated its ability to minimise and eliminate peak overshoot percentages in the control signal (feed rate) and output signal (surface roughness) responses respectively. The other two controllers (PID and FOPID) proved effective in reducing other time domain metrics viz. rise, peak and settling times. However these controllers were unable to stem the peak overshoots to minuscule levels. It is absolutely important to minimise and possibly eliminate the control/output signal overshoots to avoid the adverse affects of such overshoots on the machining process and product quality. This work establishes that the complex order PID controllers can be safely and widely adopted towards effective control of nano composite manufacturing systems for high productivity.

References

[1] R. Sekhar, T. Singh, P. Shah, "ARX/ARMAX Modeling and Fractional Order Control of Surface Roughness in Turning Nano-Composites," in 2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE), 97–102, IEEE, 2019, doi:10.1109/MoRSE48060.2019.8998654.

[2] R. Sekhar, T. Singh, "Mechanisms in turning of metal matrix composites: a review," Journal of Materials Research and Technology, **4**(2), 197–207, 2015, doi:https://doi.org/10.1016/j.jmrt.2014.10.013.

[3] R. Sekhar, T. Singh, "Determination of Johnson Cook Parameters in Turning of Micro and Nano Reinforced Aluminum Composites using Trust Region Reflective Algorithm," International Journal of Innovative Technology and Exploring Engineering, **8**(12), 1712–1716, doi:10.35940/ijitee.L3183.1081219.

[4] V. S. Jatti, R. Sekhar, R. Patil, "Study of ball nose end milling of LM6 al alloy: Surface roughness optimisation using genetic algorithm," Int. J. Eng. Technol, **5**, 2859–65, 2013, doi:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.411.1573&rep=rep1&type=pdf.

[5] I. Podlubny, "Fractional-order systems and fractional-order controllers," Institute of Experimental Physics, Slovak Academy of Sciences, Kosice, **12**(3), 1–18, 1994, doi:http://people.tuke.sk/igor.podlubny/pspdf/uef0394.pdf.

[6] R. Sekhar, T. Singh, P. Shah, "Micro and Nano Particle Composite Machining: Fractional Order Control of Surface Roughness," in Third International Conference on Powder, Granule and Bulk Solids: Innovations and Applications PGBSIA 2020 February 26-28, 2020, 35, 2020, doi:https://www.pgbsia.com/wp-content/uploads/Conference-Proceedings-PGBSIA-2020.pdf#page=40.

[7] P. Shah, R. Sekhar, S. Agashe, "Application of Fractional PID Controller to Single and Multi-Variable Non-Minimum Phase Systems," International Journal of Recent Technology and Engineering, **8**(2), 2801–2811, 2019, doi:10.35940/ijrte.b2805.078219.

[8] P. Shah, R. Sekhar, "Closed Loop System Identification of a DC Motor using Fractional Order Model," in 2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE), 69–74, IEEE, 2019, doi:https://doi.org/10.1109/MoRSE48060.2019.8998744.

[9] P. Shah, S. Agashe, "Design and optimization of fractional PID controller for higher order control system," in International conference of IEEE ICART, 588–592, 2013, doi:https://www.semanticscholar.org/paper/Design-and-Optimization-of-Fractional-PID-for-Order-Shah-Agashe/6e32b76ffe9609e0423e062b8aedd87361dedd4c.

[10] R. Bhimte, K. Bhole, P. Shah, "Fractional Order Fuzzy PID Controller for a Rotary Servo System," in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 538–542, IEEE, 2018, doi:https://doi.org/10.1109/ICOEI.2018.8553867.

[11] S. Das, Functional fractional calculus, Springer Science & Business Media, 2011, doi:https://doi.org/10.1007/978-3-642-20545-3.

[12] A. Tepljakov, "FOMCON: fractional-order modeling and control toolbox," in Fractional-order Modeling and Control of Dynamic Systems, 107–129, Springer, 2017, doi:https://doi.org/10.1007/978-3-319-52950-9_6.

[13] P. Shah, S. Agashe, "Design of controller for a higher order system without using model reduction methods," Progress in Fractional Differentation and Applications, **3**, 289–3004, 2017, doi:http://dx.doi.org/10.18576/pfda/030405.

[14] S. Pritesh, S. AGASHE, A. J. KULKARNI, "Design of a fractional $PI^{\lambda}D^{\mu}$ controller using the cohort intelligence method," Frontiers of Information Technology & Electronic Engineering, **19**(3), 437–445, 2018, doi:https://doi.org/10.1631/FITEE.1601495.

[15] J. A. Jacob, A. V. Tare, V. A. Vyawahare, V. N. Pande, "A review of time domain, frequency domain and stability analysis of linear complex-order systems," in 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 164–169, IEEE, 2016, doi:https://doi.org/10.1109/WIECON-ECE.2016.8009110.

[16] M. Shahiri, A. Ranjbar, M. R. Karami, R. Ghaderi, "New tuning design schemes of fractional complex-order PI controller," Nonlinear Dynamics, **84**(3), 1813–1835, 2016, doi:https://doi.org/10.1007/s11071-016-2608-5.

[17] C. Li, X. Dao, P. Guo, "Fractional derivatives in complex planes," Nonlinear Analysis: Theory, Methods & Applications, **71**(5-6), 1857–1869, 2009, doi:https://doi.org/10.1016/j.na.2009.01.021.

[18] J. T. Machado, "Optimal controllers with complex order derivatives," Journal of Optimization Theory and Applications, **156**(1), 2–12, 2013, doi:https://doi.org/10.1007/s10957-012-0169-4.

[19] K. Khandani, A. A. Jalali, M. R. R. Mehdiabadi, "Robust complex order controller design for DC motors," in 20th Iranian Conference on Electrical Engineering (ICEE2012), 900–903, IEEE, 2012, doi:https://doi.org/10.1109/IranianCEE.2012.6292481.

[20] C. M. Pinto, A. R. Carvalho, "Effect of drug-resistance in a fractional complex-order model for HIV infection," IFAC-PapersOnLine, **48**(1), 188–189, 2015, doi:https://doi.org/10.1016/j.ifacol.2015.05.162.

[21] A. Guefrachi, S. Najar, M. Amairi, M. Aoun, "Tuning of fractional complex order PID controller," IFAC-PapersOnLine, **50**(1), 14563–14568, 2017, doi:https://doi.org/10.1016/j.ifacol.2017.08.2093.

[22] O. Hanif, G. B. Babu, S. Sharma, "Performance Improvement of PI^{x+iy} D Fractional Complex Order Controller using Genetic Algorithm," in 2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 1–5, IEEE, 2018, doi:https://doi.org/10.1109/AEEICB.2018.8480981.

[23] M. F. Silva, J. T. Machado, R. S. Barbosa, "Complex-order dynamics in hexapod locomotion," Signal processing, **86**(10), 2785–2793, 2006, doi:https://doi.org/10.1016/j.sigpro.2006.02.024.

[24] A. V. Tare, J. A. Jacob, V. A. Vyawahare, V. N. Pande, "Design of novel optimal complex-order controllers for systems with fractional-order dynamics," International Journal of Dynamics and Control, **7**(1), 355–367, 2019, doi:https://doi.org/10.1007/s40435-018-0448-5.

[25] J. L. Adams, T. T. Hartley, L. I. Adams, "A solution to the fundamental linear complex-order differential equation," Advances in Engineering Software, **41**(1), 70–74, 2010, doi:https://doi.org/10.1016/j.advengsoft.2008.12.014.

[26] M. Shahiri, A. Ranjbar, M. R. Karami, R. Ghaderi, "Robust control of nonlinear PEMFC against uncertainty using fractional complex order control," Nonlinear Dynamics, **80**(4), 1785–1800, 2015, doi:https://doi.org/10.1007/s11071-014-1718-1.

[27] K. J. Åström, P. Eykhoff, "System identification—a survey," Automatica, **7**(2), 123–162, 1971, doi:https://doi.org/10.1016/0005-1098(71)90059-8.

[28] P. Shah, S. Agashe, "Review of fractional PID controller," Mechatronics, **38**, 29–41, 2016, doi:https://doi.org/10.1016/j.mechatronics.2016.06.005.

[29] P. Shah, S. Agashe, "Experimental Analysis of Fractional PID Controller Parameters on Time Domain Specifications," Progress in Fractional Differentiation and Applications, **3**, 141–154, 2017, doi:http://dx.doi.org/10.18576/pfda/030205.

[30] P. Lanusse, A. Oustaloup, B. Mathieu, "Third generation CRONE control," in Proceedings of IEEE Systems Man and Cybernetics Conference-SMC, volume 2, 149–155, IEEE, 1993, doi:https://doi.org/10.1109/ICSMC.1993.384864.

[31] K. H. Ang, G. Chong, Y. Li, "PID control system analysis, design, and technology," IEEE transactions on control systems technology, **13**(4), 559–576, 2005, doi:https://doi.org/10.1109/TCST.2005.847331.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 308-313 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Economic and Environmental Analysis of Life Expectancy in China and India: A Data Driven Approach

Nittaya Kerdprasop*,1, Kittisak Kerdprasop[1], Paradee Chuaybamroong[2]

[1]*School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand*

[2]*Department of Environmental Science, Thammasat University, Rangsit Campus 12120, Thailand*

ARTICLE INFO

Article history:
Received: 08 June, 2020
Accepted: 13 September, 2020
Online: 17 September, 2020

Keywords:
Data analysis
Life expectancy
Environment and economic
Web graph
Classification and regression tree

ABSTRACT

A data analytic approach presented in this work covers both data descriptive and predictive modeling with two main objectives: (1) discovering factors related to longevity of populations in the two most populated nations, China and India, and (2) generating life expectancy predictive models for both countries. Descriptive modeling methods to explore major environmental and economic factors anticipating to affect longevity patterns of people are web graph analysis and chi-squared automatic interaction detection (CHAID) techniques. Web graph analysis has been applied for the ease of visualization and CHAID is for discovering factors leading to longevity. From the analysis results, particulate emission including ozone pollution and PM2.5 concentrations are the most important factor threatening life of populations in both China and India. To predict number of years an individual is expected to live based on the available environmental and economic factors, several statistical and machine learning techniques are applied and it turns out that a linear regression model yields the most accurate prediction result.

1. Introduction

Longevity, education and income are the three main indicators that the United Nations Development Program (UNDP) has adopted for computing the human development index (HDI) to assess development level of each country [1]. The HDI captures three main essential aspects: a long and healthy life, an ability to have sufficient knowledge, and a decent standard of living. The health aspect is measured from life expectancy at birth, which is number of years a new-born baby is expected to live, averaging from the cohort. The knowledge aspect is measured from years of schooling. The standard of living aspect is measured from the gross national income per capita. The HDI is the geometric mean of all three aspects that have gone through the normalized process. This paper focuses on the longevity indicator through the life expectancy at birth measure, as it is considered [2], [3] a reflect for the good health of population.

The analysis of longevity trends within and across nations is of interest among many groups of researchers [4]-[7]. The number of years an individual is expected to live is also important to actuaries for making an optimal and economical insurance and pension plans [8]-[10]. Both positive and negative factors affecting longevity have been investigated by several researchers. For instance, Chinese researchers had explored the factors relating to energy consumption in daily life of Chinese people through the use of coal and electricity [11], [12]. The results are that coal usage relates to shorten life, whereas domestic electricity usage shows positive correlation to longevity. However, the choice of energy sources depends on the household income. Such economic and socio-economic factors had been proven by many researchers [13], [14], [15] that they can affect the long-life of population.

Exposure to polluted environment is also another important factor studied by various researchers [16], [17] to assess its impact against longevity. Environment factors reported to have negative impact toward shorten longevity includes carbon dioxide (CO2) emission [18], [19], particulate matter (PM10) and sulfur dioxide (SO2) concentrations [20], [21]. A poor climate condition also shows negative impact toward longevity of population [22], [23].

From the literature review, it can be noticed that most researchers study longevity by building a predictive model to forecast number of years the populations are expected to live using various methods including regression [24], autoregressive integrated moving average [25]-[27], and neural network algorithms [28], [29]. Some researchers [30] apply an ensemble method to make forecasting through a number of models and use the averaging scheme to predict the years of living.

*Corresponding Author Nittaya Kerdprasop, nittaya@sut.ac.th

In this work, we make our contribution to the demography as well as the environmental fields of research by proposing a different longevity analysis approach in which both numeric and categorical modeling are applied, instead of the sole numeric computation. Our data analytic method generates descriptive model through categorical computation to reveal life-threatening factors and also produces predictive model to make a numeric prediction toward number of years an individual is expected to live based on the available economic and environmental factors. The data source and our analysis methodology are explained in Section 2. Results from the data analytic approach are presented in Section 3. Performance evaluation of the predictive models is shown in Section 4. The conclusion is in Section 5.

2. Materials and Methods

The main purpose of our research is to present descriptive and predictive modeling methods to automatically discover major factors influencing good health of people living in the highly populated countries. We choose China and India to be our case study because the two countries are the most populated ones in the world (China = 1.43 billion, India = 1.36 billion [31]) and the trends of human development index (HDI) of the two countries are almost similar. The geographic location of China and India and the trend in HDI improvement from the years 1990 to 2018 are shown in Figures 1 and 2, respectively.

We focus the HDI analysis along the dimension of long and healthy life through the discovery of major factors affecting life expectancy at birth among populations in China and India. The life expectancy at birth together with other 16 economic and environmental indicators expecting to play some role on lifespan during 1960-2017 have been accessed from the World Bank database [32]. The summary of all attributes is listed in Table 1.

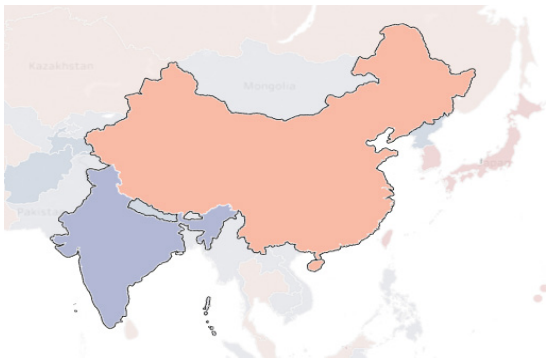


Figure 1: Location of China (above -- in orange) and India (below -- in grey) (image source: https://data.worldbank.org/country/)

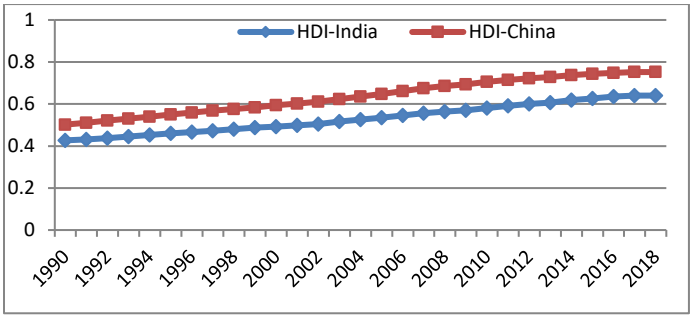


Figure 2: HDI improvement of China and India from the years 1990 to 2018 (data source: http://hdr.undp.org/en/data)

Table 1: Selected attributes for life threatening analysis

Attribute	Description	Unit
National_ income	Net national income per capita	annual % growth
Education_ expense	Education expenditure -- excluding investments in buildings and equipment	% of GNI
Forest_ depletion	Net forest depletion -- if growth exceeds harvest, this figure is zero	% of GNI
Particulate_ emis	Particulate emission damage -- calculated as foregone labor income due to premature death due to exposure to ozone pollution and indoor concentrations of $PM_{2.5}$ in households cooking with solid fuels	% of GNI
Agri_met_ emis	Agricultural methane emissions -- from animals, animal waste, rice production, agricultural waste burning	% of total
CO2_ emission	CO_2 emissions -- carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring	metric tons per capita
Electric_ power	Electric power consumption	kWh per capita
Energy_use	Energy use	kg of oil equivalent per capita
Exports	Exports of goods and services	% of GDP
Forest_area	Forest area -- excluding tree stands in agricultural production systems and trees in parks and gardens	% of land area
GDP_growth	GDP growth	annual %
GNI	GNI per capita growth	annual %
High-tech_exports	High-technology exports -- products with high R&D intensity, such as in aerospace, computers, pharmaceuticals, scientific instruments, and electrical machinery	% of manufactured exports
Imports	Imports of goods and services	% of GDP
Industry	Industry, value added -- industries correspond to ISIC divisions 10-45 and 15-37	% of GDP
Manufacturing	Manufacturing, value added -- industries belonging to ISIC divisions 15-37	% of GDP
Life	Life expectancy at birth -- the number of years a newborn infant would live if patterns of mortality at the time of its birth were the same throughout its life	years

Among all 17 numeric data attributes, life expectancy at birth is the target of our analysis. The main steps in the data-driven analytical approach are illustrated in Figure 3.

The first step is data extraction including the selection of data attributes from the World Bank database and the data preparation to be in a suitable format for further analysis steps. The next step is data exploration, which is the analysis of correlation among data attributes. The third step is the discovery of factors threatening life of population in China and India. This step needs the transformation from numeric to be categorical through the binning approach. The descriptive model to reveal important factors affecting lifespan is derived by the algorithm chi-squared automatic interaction detection, or CHAID [33]. This algorithm has been adopted because its model represented as a tree structure

has many advantages such as efficiency, interpretability and successful adoption to solve a wide range of problems [34],[35]. To visualize factor association, we adopt the web graph method.

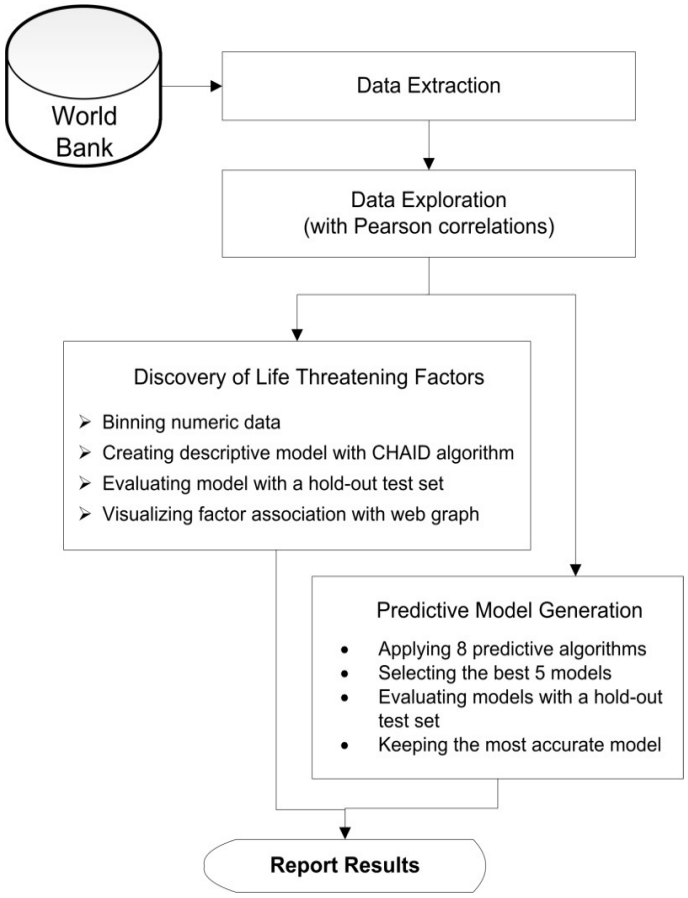


Figure 3: The main steps for data-driven analytical approach

The fourth and also the last step of our analytical method is the predictive model generation. We adopt 8 algorithms in this modeling step including regression, generalized linear model (GLM), k-nearest neighbors (kNN), support vector machine (SVM), classification and regression tree (CART), chi-squared automatic interaction detection (CHAID), artificial neural network (ANN), and linear.

The obtained models are evaluated for their performances observed during the training process. The best 5 models are then selected for further testing with the hold-out method in which the separated test dataset has been applied. In the training and testing steps, the total 48 records of data during the years 1960-2017 had been split as two separate subsets: training and testing. The training dataset comprises of 45 records, whereas the testing dataset contains 13 records.

After evaluating prediction performance of the best 5 models using the same set of test data, the most accurate model is kept as the final model. This model is to be used for forecasting the number of years an individual is expected to live. It is worth noting that forecasting model for one country may be different from another because the modeling is a data driven approach such that different set of training data may yield different modeling result, even though set of data attributes is the same.

3. Analytical Results

According to the design of our data analytical approach, there are three main steps of data analysis: data exploration, descriptive model creation, and predictive model generation. We thus present the results from these three main steps sequentially in the following three subsections.

3.1. Data Exploration Results

The results of Pearson correlation analysis is shown in Table 2. It can be noticed similarity between the two countries that source of energy and amount of energy usage have strong positive influence toward longevity. The values in terms of import, export and industry are also among the top-5 positive factors associating with longevity.

In case of correlation analysis to reveal negative factors to lifespan of population, we can notice that particulate emission is among the top-five factors associating to life shortening in both China and India. However, other factors such as forest depletion, forest area, the export of high-technology product, the expense in education and the national income per capita are also appeared as having negative influence toward longer life of population in the two countries. These results are preliminary data exploration. The next step is the in-depth analysis discovering only prominent factors relating to longevity of population.

Table 2: Pearson correlation between life expectancy at birth and economic and environmental factors

	Top-5 Positive Factors (correlation)	**Top-5 Negative Factors** (correlation)
China	Energy_use (0.856) Electric_power (0.837) Imports (0.800) Exports (0.767) Industry (0.734)	Forest_depletion (-0.862) Education_expense (-0.862) Particulate_emis (-0.709) High-tech_exports (-0.689) Forest_area (-0.672)
India	Electric_power (0.954) Energy_use (0.940) Industry (0.902) Exports (0.875) Imports (0.838)	High-tech_exports (-0.853) Particulate_emis (-0.848) Forest_area (-0.798) National_income (-0.739) Forest_depletion (-0.717)

3.2. Descriptive Modeling Results

Descriptive analytics refer to the process of applying statistical and other intelligent techniques to provide insight into the historical data to gain some understanding about the important factors, hidden patterns or concealed behavior. To understand characteristics of longevity pattern, we apply CHAID algorithm to reveal prominent factors affecting lifespan and display it as a tree structure as shown in Figure 4 (root of a tree is on the left hand side). It can be seen from the patterns that particulate emission damage is the most important environmental factor shortening lifespan of populations in both China and India.

A web graph to display association between economic and environmental factors and life expectancy at birth for each population group is also shown in Figure 5. The thickness of line linking each node in a graph represents strength of association. The thicker is the stronger. The web graph also shows the strong association between a short lifespan and a high level of particulate emission damage for both Chinese and Indian populations.

China

Particulate_emis in [0.27-0.34] => Lifespan in [74.29-76.41]

Particulate_emis in [0.35-0.75] => Lifespan in [69.93-74.28]

Particulate_emis > 0.75 => Lifespan in [67.61-69.92]

Particulate_emis is unknown

Exports in [2.49-8.32] => Lifespan in [59.08-67.60]

Exports in [8.33-17.91] => Lifespan in [67.61-69.92]

Exports in [17.92-36.03] => Lifespan in [59.08-67.60]

(Accuracy on test data = 100%)

India

Particulate_emis in [0.78-1.02] => Lifespan in [64.96-68.80]

Particulate_emis in [1.03-2.21] => Lifespan in [59.92-64.95]

Particulate_emis in > 2.22 => Lifespan in [54.71-59.91]

Particulate_emis is unknown

Industry in [21.41-26.16] => Lifespan in [47.72-54.70]

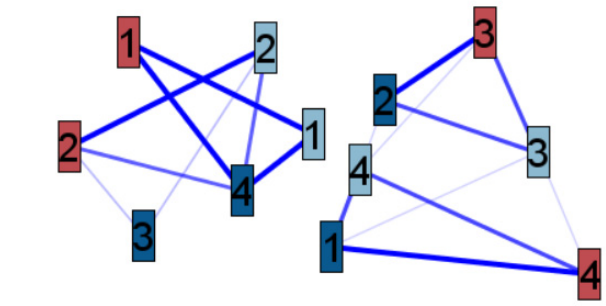
Industry in [26.17-27.84] => Lifespan in [54.71-59.91]

Industry in [27.85-31.13] => Lifespan in [47.72-54.70]

(Accuracy on test data = 84.62%)

Figure 4: Factors affecting lifespan of populations in China and India discovered by the CHAID algorithm

China



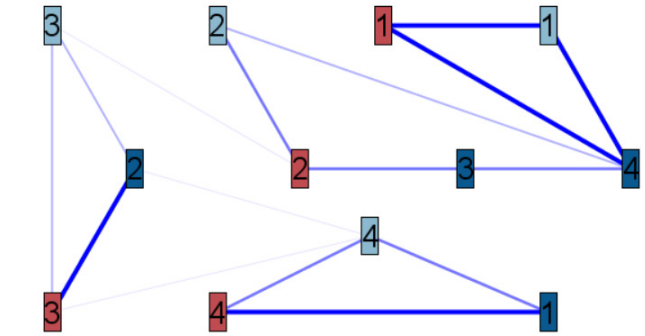
Life expectancy at birth (LE)

Particulate emission damage (PE)

Exports (EX)

LE: 1=[59.08-67.60], 2=[67.61-69.92],
3=[69.93-74.28], 4=[74.29-76.41]
PE: 1=[0.27-0.34], 2=[0.35-0.75],
3 = >0.75, 4 = unknown
EX: 1=[2.49-8.32], 2=[8.33-17.91],
3=[17.92-23.58], 4=[23.59-36.03]

India



Life expectancy at birth (LE)

Particulate emission damage (PE)

Industry (IN)

LE: 1=[47.72-54.70], 2=[54.71-59.91],
3=[59.92-64.95], 4=[64.96-68.80]
PE: 1=[0.78-1.02], 2=[1.03-2.21],
3 = >2.22, 4 = unknown
IN: 1=[21.41-26.16], 2=[26.17-26.72],
3=[26.73-27.84], 4=[27.85-31.13]

Figure 5: Factors contributing to longevity of populations displayed as a web graph

3.3. Predictive Modeling Results

Predictive modeling is the data analytical approach to generate a model with the main aim of using that model to forecast future event. We apply both statistical and machine learning methods to build a predictive model from the training dataset. From the model performance comparison, linear regression is a method yielding the most accurate result for both China and India cases. The linear regression models to predict lifespan of population in each country are presented in Figure 6. It can be noticed that education expense is among the first two factors appeared in the models and it shows positive influence toward long life of population in both countries.

China: *Life expectancy at birth*

= Education_expense * 6.195 +
Agri_methane_emission * 0.00005864 +
CO2_emission * -4.16 +
Electric_power_consumption * -0.008426 +
Energy_use * 0.03816 +
Exports * -0.07075 +
Forest_area * 0.0001171 +
GDP_growth * 0.01842 +
GNI * -0.00001071 +
Hi-tech-exports * -0.0002215 +
Imports * 0.3694 +

Industry * -0.2848 +
Manufacturing * 0.000447 + 48.7

India: *Life expectancy at birth*

= National_income * -0.02832 +
Education_expense * 2.28 +
Forest_depletion * -0.02815 +
Agri_methane_emission * -0.000006419 +
CO2_emission * 8.185 +
Electric_power_consumption * 0.02312 +
Energy_use * -0.04361 +
Exports * 0.2372 +
Forest_area * 0.00003288 +
GDP_growth * -0.2307 +
GNI * 0.3371 +
Hi-tech-exports * 0.00004354 +
Imports * -0.1426 +
Industry * 0.7558 +
Manufacturing * -0.173 + 35.8

Figure 6: Factors contributing to longevity of populations in China and India

4. Model Evaluation

We test performance of descriptive and predictive models with a separate set of test data. The descriptive model derived by CHAID algorithm to reveal longevity pattern of population in China with 100% of accuracy, whilst the accuracy of CHAID model drops to 84.62% when the algorithm tries to fit model with data from Indian training cases. The decrease in accuracy may occur from the low level of homogeneity in the training data.

To build a predictive model, we apply eight algorithms and then select the best five models to test performance on a hold-out dataset. The evaluation results for the case of China population are illustrated in Table 3, whereas the results of India population modeling assessment are shown in Table 4. The best predictive model from both population groups with the least mean absolute error is the one derived from the linear regression algorithm.

Table 3: Predictive model evaluation results for China

Algorithm	Correlation	Number of Factors	Mean Absolute Error
Regression	**0.977**	13	**0.602**
CHAID	0.972	7	0.769
CART	0.971	12	0.916
GLM	0.874	16	1.597
SVM	1.0	16	3.212

Table 4: Predictive model evaluation results for India

Algorithm	Correlation	Number of Factors	Mean Absolute Error
Regression	**0.996**	15	**0.544**
CHAID	**0.996**	16	0.562
CART	0.988	7	1.071
GLM	0.926	16	1.507
SVM	0.895	15	1.888

5. Conclusion

We present data analytical approach to create descriptive and predictive model with the two main aims: to discover threatening factors that may shorten lifespan and to predict the number of years on average an individual is expected to live based on the economic and environmental variables. The analytical approach is data driven in the sense that model is to be generated from the training data. Therefore, using different datasets can result in obtaining different models, even though the same data variables and the same learning algorithm have been applied.

To derive a descriptive model, we adopt the CHAID algorithm. The training data are World development indicators accessed from the World Bank database with two specific countries, China and India, and the target of analysis is life expectancy at birth which is one aspect that the UNDP applied for computing the human development index of a country. The models of China and India show one common finding that particulate emission damage due to the exposure to ambient ozone pollution and PM2.5 is the most important factor affecting shorter longevity. We also apply several learning algorithms to derive predictive models to forecast number of years an individual is expected to live based on the available economic and environment factors. The most accurate predictive model is the one built from the linear regression method.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been supported by grant from Suranaree University of Technology through the funding of Data and Knowledge Engineering Research Unit in which the first and second authors are principal researchers.

References

[1] UNDP, "Human Development Indices and Indicators: 2018 Statistical Update," http://hdr.undp.org/sites/default/files/hdr2018_technical_ notes.
[2] D. Dicker et al., "Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: A systematic analysis for the Global Burden of Disease Study 2017," The Lancet, **392**(10159), 1684–1735, 2018.
[3] M. Kim, Y.H. Khang, "Why do Japan and South Korea record very low levels of perceived health despite having very high life expectancies?," SSRN Electronic Journal, article 3276420, 2018.
[4] L. Wang, Y. Li, H. Li, J. Holdaway, Z. Hao, W. Wang, T. Kraft, "Regional aging and longevity characteristics in China," Archives of Gerontology and Geriatrics, **67**, 153-159, 2016.
[5] S. Wang, K. Luo, R. Ni, Y. Tian, X. Gao, "Assessment of elemental background values and their relation with lifespan indicators: A comparative

study of Jining in Shandong Province and Guanzhong area in Shaanxi Province, northern China," Science of The Total Environment, **595**, 315–324, 2017.
[6] L.A. Johnston, "The economic demography transition: is china's 'not rich, first old' circumstance a barrier to growth?," Australian Economic Review, Jul., 1–21, 2019.
[7] R. Bai, J. Wei, R. An, Y. Li, L. Collett, S. Dang, W. Dong, D. Wang, Z. Fang, Y. Zhao, Y. Wang, "Trends in life expectancy and its association with economic factors in the belt and road countries—evidence from 2000–2014," International Journal of Environmental Research and Public Health, **15**(12), article 2890, 2018.
[8] E. Kurtbegu, "Replicating intergenerational longevity risk sharing in collective defined contribution pension plans using financial markets," Insurance: Mathematics and Economics, **78**, 286–300, 2018.
[9] E. Debonneuil, S. Loisel, F. Planchet, "Do actuaries believe in longevity deceleration?," Insurance: Mathematics and Economics, **78**, 325–338, 2018.
[10] L. Mayhew, D. Smith, D. Wright, "The effect of longevity drift and investment volatility on income sufficiency in retirement," Insurance: Mathematics and Economics, **78**, 201–211, 2018.
[11] S. Wang, K. Luo, "Life expectancy impacts due to heating energy utilization in China: Distribution, relations, and policy implications," Science of The Total Environment, **610–611**, 1047–1056, 2018.
[12] S. Wang, Y. Liu, C. Zhao, H. Pu, "Residential energy consumption and its linkages with life expectancy in mainland China: A geographically weighted regression approach and energy-ladder-based perspective," Energy, **177**, 347–357, 2019.
[13] W.C. Cockerham, Y. Yamori, "Okinawa: an exception to the social gradient of life expectancy in Japan," Asia Pacific Journal of Clinical Nutrition, **10**(2), 154–158, 2001.
[14] J. Jiang, L. Luo, P. Xu, P. Wang, "How does social development influence life expectancy? A geographically weighted regression analysis in China," Public Health, **163**, 95–104, 2018.
[15] C. Lee, M. Kim, "The relationship between internet environment and life expectancy in Asia," Review of Integrative Business & Economics, **8**(2), 70–80, 2019.
[16] K. Hassan, R. Salim, "Population ageing, income growth and CO2 emission: empirical evidence from high income OECD countries," Journal of Economic Studies, **42**(1), 54–67, 2015.
[17] J.C. Yeh, C.H. Liao, "Impact of population and economic growth on carbon emissions in Taiwan using an analytic tool STIRPAT," Sustainable Environment Research, **27**(1), 41–48, 2017.
[18] N. Kerdprasop, K. Kerdprasop, "Regression tree analysis of CO2 emissions and environmental factors to the survival rate of population in Thailand and China," in the International MultiConference of Engineers and Computer Scientists (IMECS), 286–290, 2016.
[19] N. Kerdprasop, K. Kerdprasop, "Association of economic and environmental factors to life expectancy of people in the Mekong basin," in the 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), 1984–1989, 2017.
[20] L. Wang, B. Wei, Y. Li, H. Li, F. Zhang, M. Rosengerg, L. Yang, J. Huang, T. Kraft, W. Wang, "A study of air pollutants influencing life expectancy and longevity from spatial perspective in China," Science of The Total Environment, **487**, 57–64, 2014.
[21] W. Song, Y. Li, Z. Hao, H. Li, W. Wang, "Public health in China: An environmental and socio-economic perspective," Atmospheric Environment, **129**, 9–17, 2016.
[22] J. Lv, W. Wang, Y. Li, "Effects of environmental factors on the longevous people in China," Archives of Gerontology and Geriatrics, **53**(2), 200–205, 2011.
[23] J. Robine, F.R. Herrmann, Y. Arai, D.C. Willcox, Y. Gondo, N. Hirose, M. Suzuki, Y. Saito, "Exploring the impact of climate on human longevity," Experimental Gerontology, **47**(9), 660–671, 2012.
[24] G. Gulis, "Life expectancy as an indicator of environmental health," European Journal of Epidemiology, **16**, 161–165, 2000.
[25] T. Torri, J.W. Vaupel, "Forecasting life expectancy in an international context," International Journal of Forecasting, **28**(2), 519–531, 2012.
[26] K.J. Foreman et al., "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and territories," The Lancet, **392**(10159), 2052–2090, 2018.
[27] M.D. Pascariu, V. Canudas-Romo, J.W. Vaupel, "The double-gap life expectancy forecasting model," Insurance: Mathematics and Economics, **78**, 339–350, 2018.
[28] M. Kanevski, R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, V. Demyanov, S. Canu, "Environmental data mining and modelling based on machine learning algorithms and geostatistics," Environment Modelling & Software, **19**(9), 845–855, 2004.
[29] M. Leuenberger, M. Kanevski, "Extreme Learning Machines for spatial environmental data," Computers & Geosciences, **85**, 64–73, 2015.
[30] V. Kontis, J.E. Bennett, C.D. Mathers, G. Li, K. Foreman, M. Ezzati, "Future life expectancy in 35 industrialised countries: Projections with a Bayesian model ensemble," The Lancet, **389**(10076), 1323–1335, 2017.
[31] United Nations, Department of Economic and Social Affairs, Population Division, "World Population Prospects 2019: Data Booklet," https://population.un.org/wpp/Publications/
[32] The World Bank, "World Development Indicator," https://databank.worldbank.org/source/world-development-indicators
[33] G. Kass, "An exploratory technique for investigating large quantities of categorical data," Applied Statistics, **29**(2), 119–127, 1980.
[34] S. Jang, H. Choi, Y. Jung, E. Moon, T. Yoon, "A comparison of H1N1 and H3N2 viruses using decision tree and Apriori algorithm," International Journal of Machine Learning and Computing, **6**(1), 76–79, 2016.
[35] S. Boonamnuay, N. Kerdprasop, K. Kerdprasop, "Classification and regression tree with resampling for classifying imbalanced data," International Journal of Machine Learning and Computing, **8**(4), 336–340, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 314-322 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Deep Learning Model for A Driver Assistance System to Increase Visibility on A Foggy Road

Samir Allach*, Mohamed Ben Ahmed, Anouar Abdelhakim Boudhir

LIST Laboratory, FSTT, UAE University, Tangier 90000, Morocco

A R T I C L E I N F O

Article history:
Received: 22 May, 2020
Accepted: 13 July, 2020
Online: 28 July, 2020

Keywords:
ADAS
Deep learning
Image dehazing
Foggy road
Object detection

A B S T R A C T

For many years, a lot of researches have been made to develop Advanced Driver Assistance Systems (ADAS) that are based on integrated systems. The main objective is to help drivers. Hence, keeping them safe under different driving conditions. Visibility for drivers remains the biggest problem faced on the road in an atmosphere of fog. In this paper, we examine a system that can be employed to substantially enhance visibility through using deep neural networks. Researches done recently- which are based on deep learning for eliminating image fog- have made clear that an end-to-end proposed system is such an effective model. However, it becomes a must to extend the idea to end-to-end real-time video deshazing. In this paper, we introduce a model of image dehazing. It is based on Convolutional Neural Networks (CNN) as a basis for developing the video dehazing model. As in addition, we concatenate our model with the faster RCNN for detecting objects on the road in real time. The experimental results on our image datasets shows the performance of our model with regard to Peak Signal to Noise Ratio (PSNR=19.823) and Structural Similarity (SSIM =0.8501). On the dataset of the synthesized videos, our model achieved a performance of PSNR = 21.4032 and SSIM = 0.9354. Moreover, with the concatenation of our dehazing model with Faster R-CNN (regions with convolutional neural networks), our proposed system displays desirable visual quality and a remarkable progress of the object detection achievement on blurred images with mean Average Precision (mAP) equal to 0.933 during the day and 0.804 at night.

1. Introduction

Advanced Driver Assistance Systems (ADAS) are mainly designed to offer vehicle drivers help, thereby minimizing a potential threat to their safety. The majority of these systems are based on image processing algorithms, such as those allowing detecting of nearby vehicles and pedestrians and the recognition of signs. Because of this, quite a lot of systems are fixed into vehicles. Though such systems are widely employed to draw drivers' attention once a potential threat appears for vehicle drivers, they perform less effectively under certain adverse weather conditions where the vision weakened. This happens most notably in the presence of fog.

Eliminating or reducing the fog of an image captured by an ADAS system seems to be difficult and somehow ill-posed phenomenon. It is worth mentioning here that significant developments that took place in deep learning mainstream paved the way to get considerable results in improving vision degraded by fog [1]. Therefore, the elimination of fog requires the estimation of the depth map. What is more, previous assumptions are necessary to estimate the depth map for systems using single images as input.

Very lately, a lot of algorithms have been suggested for detecting objects and eliminating fog [1, 2]. It is worth mentioning here that traditional algorithms need two crucial elements: gathered facts on the environment and developed learning as well. In addition, most object detection and fog elimination algorithms are not suited for real-time uses on account of they consist of considerable arithmetic time.

In this way, authors in this area of study have introduced deep learning approaches for degraded vision so that images can be restored and reconstructed [1]-[3]. Yet, these methods cannot directly be applied for removing the fog from the image. The authors F. Hussain, and J. Jeong, came up with an approach, using Deep Neural Network (DNN).They also assumed that an anonymous complex function could model mathematically the fog in an image [4]. Li Chongyi et al. proposed a cascading CNN

*Corresponding Author: Samir Allach, LIST-FSTT Bp 416 Tangier-Morocco, Tel: +212615571400 & Email: allach.samir@gmail.com

model composed of three components: a component of the shared invisible layers that extract the common features, the work of the global estimate of the atmospheric light and that of the subnet of the average transmission [5]. The researchers Li Boyi and others introduced a model for dehaze of image constructed with a convolutional neurons network (CNN), named "all-in-one dehazing network" and "An All-in-One Network for Dehazing and beyond" (AOD-Net) [1, 6]. It is designed on the basis of a reformulated atmospheric diffusion model. The researchers B.Cai et al. found an end-to-end formable model an excellent alternative for calculating the average transmission, called DehazeNet [2], it takes a hazy image at the input and leaves its transmission matrix, and then image dehazed is recovered by the atmospheric diffusion model. W. Ren et al. used a multi-scale CNN (MSCNN) that created and improved a coarse scale transmission matrix [7].

With regard to the video dehazing, most approaches depend crucially on the phase of improving temporal inconsistencies. The authors Kim et al. suggested inserting a temporal coherence in the cost function, with a clock filter to accelerate the processing [8]. The authors Cai et al. conceived a spatio-temporal optimization for de-hazing video in real time [2].

In recent years, an interest has begun to grow in video modeling using CNN for a huge number of tasks. And here are three examples: super-resolution (SR) [9], blur [10] and classification [11, 12] and style transfer [13]. In [9], the author studied various structure configurations for SR video. In [11, 12], the authors made similar attempts by exploring different connectivity options for video classification. The researchers Liu et al. introduced such adjustable system. In this context, they placed a spatial alignment network between the images [14]. Others introduced an end-to-end CNN to learn how to accumulate information on several hazy images / video [10]. With reference to video style transfer, Chen et al. integrated short- and long-term coherence. This is not to mention they indicated the superiority of multi-image approaches over single-image approaches [13].

In this paper, we present a video dehazing system based on an image-dehazing model. We use convolutional neural networks (CNN). Then, we concatenate our model with the faster RCNN to detect objects on the foggy road in real time. Moreover, we deploy our system to validate the quantitative and visual results obtained.

2. Network Design

Our proposed work is made of two important steps. Firstly, we worked hard to develop an end-to-end CNN model [15] that explicitly learns the mapping relationships between the raw images and their associated transmission maps to recover fog-free images. And then we integrated the Faster R-CNN proposed model [16] for objects detection in video. The algorithm presents the general steps of our model for dehazing and detecting objects in a video:

Algorithm: Video dehazing and object detection

Result: video dehazing and object detection
Let's *i* : counter;
n : number of images after splitting video to images;
V_{in} : input video , *M*: Resulting of dehazed image( );
Dehz( ): Dehaze Function;
Detect( ): Object Detection Function;
Save( ): Saving resulting image;
Concat( ): Image concatenation;
V_{split}*( ):* Splitting video to images;
V_{out}*( ):* Video generating function;
Image1( ): Output image of Vsplit( );
Image2( ): Output image of Detect( );

```
begin
    Insert Vin ;
    Splitting Video to n images;
        Vsplit( Vin );
    Dehazing images and detecting objects using Faster R-CNN;
        i ←0 ;
        while ( i<=n )
            M ← Dehz( Image1 ( i ) ) ;
            Image2 ( i ) ← Detect( M ) ;
            Save( Image2( i ) ) ;
              if i > 0 then
                  Concat( Image2( i ) , Image2( i - 1 ) ) ;
              end
            i ← i + 1;
        end
    Vout( image2( i );
end
```

2.1. Atmospheric diffusion model

The model of atmospheric diffusion is the usual description of the hazy image production process, proposed by McCartney [17] and developed later by Nayar and Narasimhan [18]. The model is written as follows [1,2,19]:

$$I(x) = J(x)\,t(x) + A\,(1 - t(x)) \tag{1}$$

$$t(x) = e^{-\beta d(x)} \tag{2}$$

where I (x) is an observed hazy image, J (x) is the image that recovered after the estimation of A and t(x), A is the global atmospheric light and t (x) is the transmission map.

In equation (2), β is the atmosphere diffusion coefficient and d (x) is used to refer to the distance between the image and the camera.

Equation (2) indicates that when d (x) moves to infinity, t (x) comes near to zero. With equation (1), we have:

$$A = I(x), d(x) \rightarrow inf \tag{3}$$

In reality, d (x) cannot be infinite, but can be a long distance, which gives a very weak transmission t_0. Then, the global atmospheric light (A) is estimated by the following formula

$$A = \max_{y \in \{x | t(x) \leq t_0\}} I(y) \tag{4}$$

2.2. Construction of the proposed CNN model

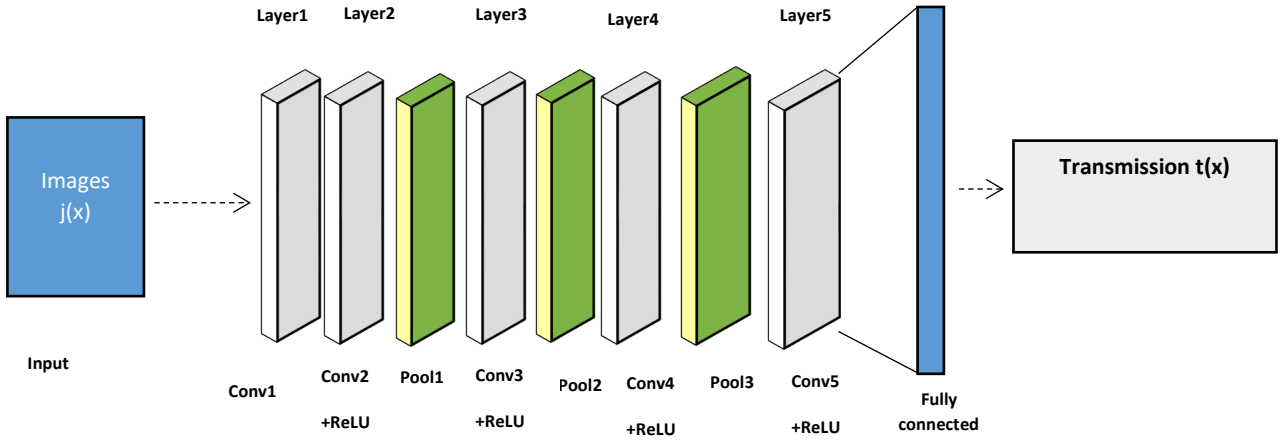
The proposed model for the image frame consists of an estimation module. This module uses convolutional layers to

estimate transmission map t (x) (see Figure 1), followed by a clean image generation module that consists of a multiplication layer of several element by element and the addition layers for generating the recovery image (see Figure 2).

The estimation module is the essential component, responsible for estimating the depth and level of relative disorder. As Figure 1 shows, we use five convolutional layers by merging filters of various sizes (see Table 1).

After comparing the results obtained by different architectures of CNN (see table 1), we concluded that the CNN model that consists of five convolutional layers with the number of 3x3 filters is the most efficient, with a PSNR of 19.8231 and a SSIM 0.8501.

The authors B. Cai et al. used parallel convolutions with filter sizes varied [2]. W. Ren et al. concatenated the distinctive features of the large- scale network with an intermediate layer of the fine-scale network [7]. Influenced by them, we concatenated in our model the layer "pool1" entities with the layers "conv1" and "conv2". The same thing for "pool2" with those of "conv2" and "conv3"; and "pool3" with those of "conv1", "conv2", "conv3" and "conv4". This model captures the characteristics of images at different scales and at the intermediate connections, compensating for the loss of information during convolutions. Notably, in our proposed model each convolutional layer uses only three filters. Consequently, our model is lighter and performs well in terms of PSNR and SSIM values in comparison with other existing deep methods.



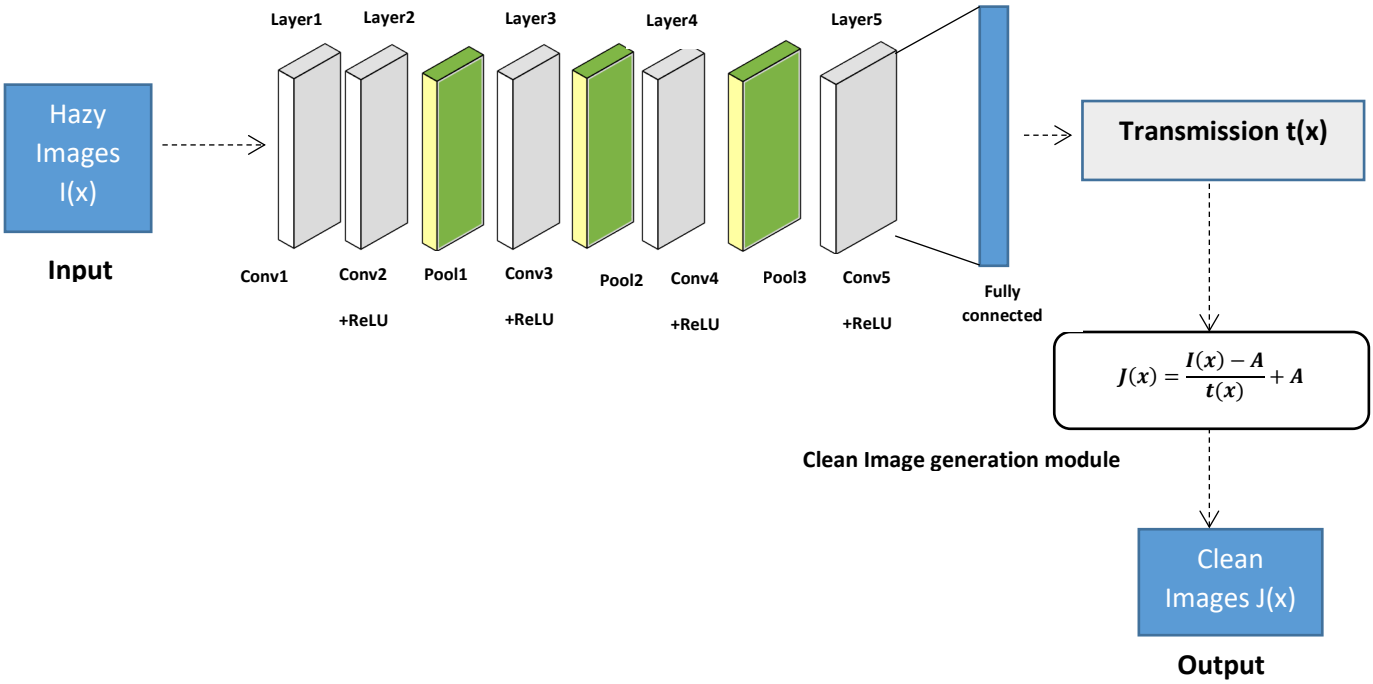


Table 1: Average of PSNR and SSIM results for different architectures of CNN model

Number of convolution layers	Number of filters	PSNR	SSIM
2	3 x 3	-	-
	5 x 5	14.0321	0.6801
	7 x 7	14.1321	0.7032
3	3 x 3	15.6706	0.7791
	5 x 5	16.7653	0.7356
	7 x 7	16.0362	0.7452
4	3 x 3	16.8521	0.7937
	5 x 5	17.7653	0.7745
	7 x 7	18.2381	0.8012
5	**3 x 3**	**19.8231**	**0.8501**
	5 x 5	19.6364	0.8125
	7 x 7	19.5364	0.8029
6	3 x 3	18.5381	0.8101
	5 x 5	18.2381	0.7998
	7 x 7	18.3381	0.8111

2.3. Model of Object detection

Faster R-CNN [16] is proposed to detect objects on images accurately. With ResNet, Faster R-CNN achieved a mean average Precision (mAP) of 76.4% on the PASCAL VOC dataset 2007 and 2012 (see Table 2). By combining region and classifier propositions in a large network, it becomes possible to automatically learn good representations of the features for the task.

In this paper, we examine the detection and recognition of objects in the presence of fog with a view to improve high-level vision tasks to combine with the model for dehazing of video. We opted for the Faster R-CNN model as the basic algorithm for detecting robust objects in real time (see Table 2), approved on synthetic and natural blurred images.

We modified the Faster R-CNN Caffe source code and used ResNet [20] as the Faster-RCNN convolution backend. We came to the conclusion that using ResNet provides a substantial improvement over other architectures. The time reweighting of the chain gives further improvements.

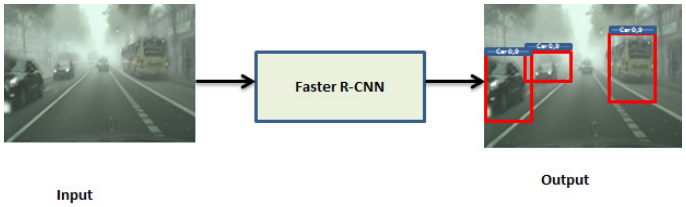


Figure 3: Object detection with Faster R-CNN model

Table 2: Comparison of detection methods on the 2007 and 2012 PASCAL VOC dataset

Method	mAP	FPS	Batch size
Fast R-CNN [21]	0.70	0.5	1
YOLO [20]	0.634	45	1
SSD300 [21]	0.743	46	8
Faster R-CNN VGG-16 [21]	0.732	7	1
Faster R-CNN-ResNet [20]	0.764	5	1

Table 2 shows comparison measurements of mAP and the FPS (frames per second) speed with the batch size, using the dataset PASCAL VOC between Fast R-CNN, YOLO, SSD300, Faster R-CNN VGG-16 and R- Faster CNN-ResNetet. Our chosen detection method (Faster R-CNN-ResNetet) surpasses all methods in terms of mAP. Though a set of certain methods may work at higher speeds, they have lower accuracies (mAP). The Faster R-CNN-ResNetet remains the best method in real time to reach more than 76.4% of mAP.

Outside the dehazing part, the temporal coherence should be taken into consideration for object detection so that we can reach satisfactory results. With our proposed model, we can hereby-advancing Faster R- model CNN ideally suited for video [22]. The first two convolutional layers of Faster R-CNN model to an image, was divided into three equivalent sections to insert the earlier, present and upcoming images, in this order. The last three images are concatenated in the next of the second convolutional layer and go through the resting layers for predicting the bounding boxes of the objects for the present image.

3. Results and discussion

3.1. Dehazing an image

First, we created foggy images synthesized from equations (1) and (2), using the ground truth images with the NYU2 inner depth database.

We put different atmospheric lights (A), selecting them consistently each channel between [0.5, 1.0] and selecting $\beta \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6\}$.

The input data is RGB images with a resolution of 640 x 480 and the output was a depth map with a resolution of 320 x 240.

In our model, about 12 epoch are enough for it to converge, and it works well enough after these times. During the learning phase, the weights of our network are initialized in a random way. We used the ReLU function as a more efficient neuron in our specific context. We opted for the function of Mean Square Error loss (MSE), and that it stimulates the PSNR (Peak Signal to Noise Ratio) and the SSIM (Structural Similarity) as well as the visual quality.

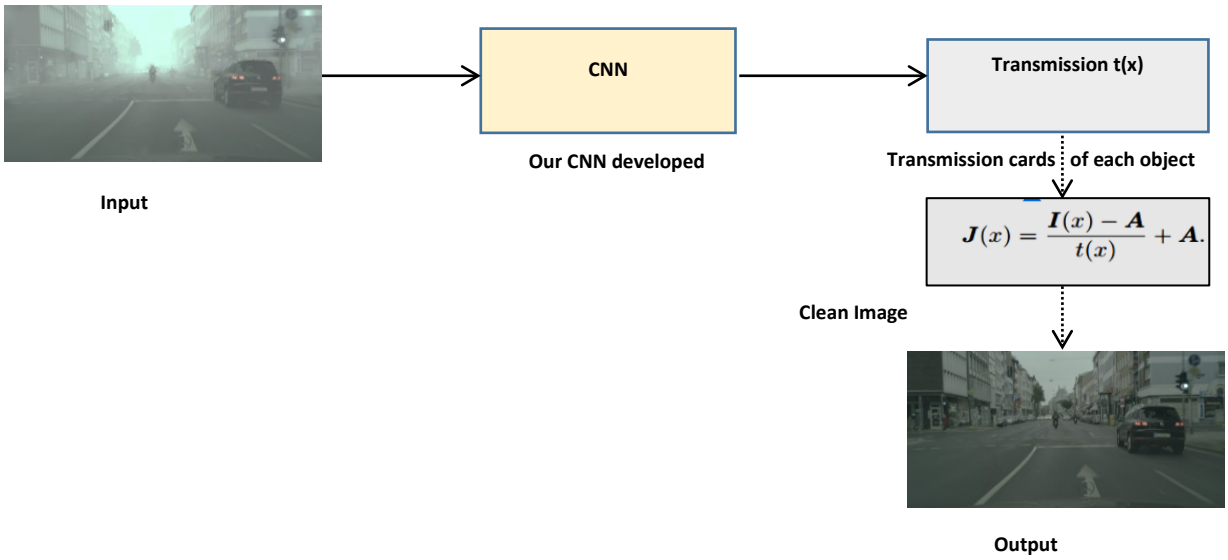


Figure 4: Our model of image dehazing

Table 3: Average PSNR and SSIM results on Test1

	ATM	BCCR	NLD	FVR	DCP	DehazeNet	CAP	MSCNN	AOD-Net	**OUR MODEL**
PSNR	14.1475	15.7606	16.7653	16.0362	18.5385	18.9613	19.6364	19.1116	19.6954	**19.8231**
SSIM	0.7141	0.7711	0.7356	0.7452	0.8337	0.7753	0.8374	0.8295	0.8478	**0.8501**

We compared our basic model proposed with several dehazing methods, in particular: Automatic atmospheric light Recovery (ATM) [24], Regularization of the context constrained to limits (BCCR) [25], No-local Image Dehazing (NLD) [26], Fast visibility restoration (FVR) [27], Dark channel priority (DCP) [28], MSCNN [7], DehazeNet [2], Color Attenuation Prior (CAP) [29] and AOD-Net [1]].

Our synthesized hazy images come along with ground truth images, paving the way for us assess PSNR and SSIM and check out if the results stay accurate.

As our model is optimized from start to finish in case of MSE loss, one should not be amazed to see its PSNR performance superior to other methods. Further, even if SSIM is not directly called the optimization criteria, our model obtains SSIM advantages superior to the other models compared.

It is well known that SSIM more accurately reflects human perception, as SSIM measures apart from pixel level errors. We become faithful to our model with such a consistent SSIM improvements we achieved.

Table 3 shows the promising performance of our model compared to the others, in terms of PSNR and SSIM.

Our method has an advantage greater than 0.2 dB in the PSNR and 0.03 in the SSIM when we compare it to approach AOD-Net.

3.2. Video sequences dehazing

First of all, we brought into existence a dataset of foggy synthetic video from on equation(1), employing 20 chosen videos in the TUM RGB-D dataset [30] that consists of different video sequences. Depth information is refined by means of the filling model of Silber man et al. [31]. After that, we broke our dataset into a learning set made up of 12 videos with 120,000 images and a set of non-overlapping tests called Test2, consisting of 8 short video sequences containing a total of 450 images. Finally, we collected a set of natural hazy video sequences to validate and evaluate the performance of our model.

During the training of our model, we adopted the loss of the mean squared error (MSE), which is aligned with the SSIM, the PSNR and the visual quality.

Table 4: PSNR average and SSIM results on Test2.

	STMRF	EVD-Net	OUR MODEL
PSNR	18.9956	20.9908	21.4032
SSIM	0.8707	0.9087	0.9354

Concerning video-based methods, we compared our model with EVD-Net [19] and STMRF [32] methods (not based on CNN). Then, we got a higher advantage over the two previous approaches. Furthermore, we observed that our advanced system performs quite remarkably showing a difference of 0.5 dB in the PSNR and 0.03 in the SSIM.

3.3. Object detection model

We reformed our adaptive Faster-RCNN model on a set of training data provided by Foggy Cityscapes [33] and on a set of synthesized personal data. Foggy Cityscapes is a foggy synthetic data set. It causes fog to operate on real scenes. The images are rendered utilizing the Cityscapes images and depth maps [8]. It contains 2,975 images in the learning set and 500 images in the validation set. In this experiment, we reported our findings on categories such as person, car, truck, bus, motorcycle and bicycle. We opted for the average score of the channel separately for each class based on validation performance.

On the validation set, we achieve a mAP of 0.777 (see Table 5 for a full breakdown between classes).

Table 5: Average Precision (AP) and Mean Average Precision (mAP) between original Faster R-CNN and adaptive Faster R-CNN with ResNet.

Metrics	Original Faster R-CNN	Adaptive Faster R-CNN with ResNet
Person (AP)	0.731	0.779
Car (AP)	0.809	0.812
Truck (AP)	0.750	0.803
Bus (AP)	0.692	0.792
Motorbike (AP)	0.663	0.723
Bike (AP)	0.621	0.756
mAP	0.711	0.777

3.4. The proposed model Architecture for object detection in hazy video

The merging of our model for video dehazing and the adapted Faster R-CNN model (Figure 5)has given birth to our general model, which naturally displays an interesting tree structure that is locally linked and is subjected to joint and more crucial optimization.

On the validation set of our proposed model, we achieved a mAP of 0.929 (see Table 6).

Table 6: Average Precision (AP) and mean Average Precision (mAP) results with adaptive Faster R-CNN before and after the dehazing operation.

Metrics	Adaptive Faster R-CNN with ResNet	
	Before the dehazing	After the dehazing
Person (AP)	0.779	0.902
Car (AP)	0.812	0.911
Truck (AP)	0.803	0.909
Bus (AP)	0.792	0.970
Motorbike (AP)	0.723	0.981
Bike (AP)	0.756	0.901
mAP	0.777	0.929

4. Deployment of our system

As an embedded platform, we used a Raspberry Pi 3 Model B +, for our experiments we made predictions on new video sequences using the raspberry Pi with camera designated v2.1, based on the Sony IMX219 CMOS type sensor with a resolution of 8 MP (3280 × 2464 pixels) and we downloaded our pre-trained model on our pi. We needed Raspbian Stretch 9, because TensorFlow 1.9 officially supports Raspberry Pi if you use Raspbian 9.

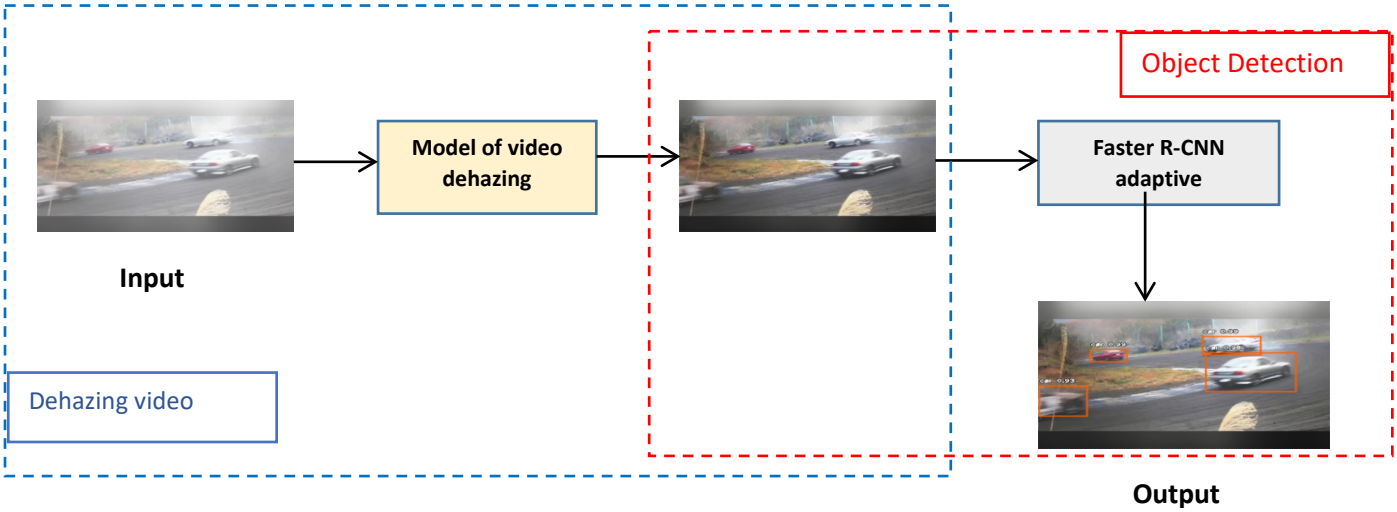


Figure 5: The proposed general model

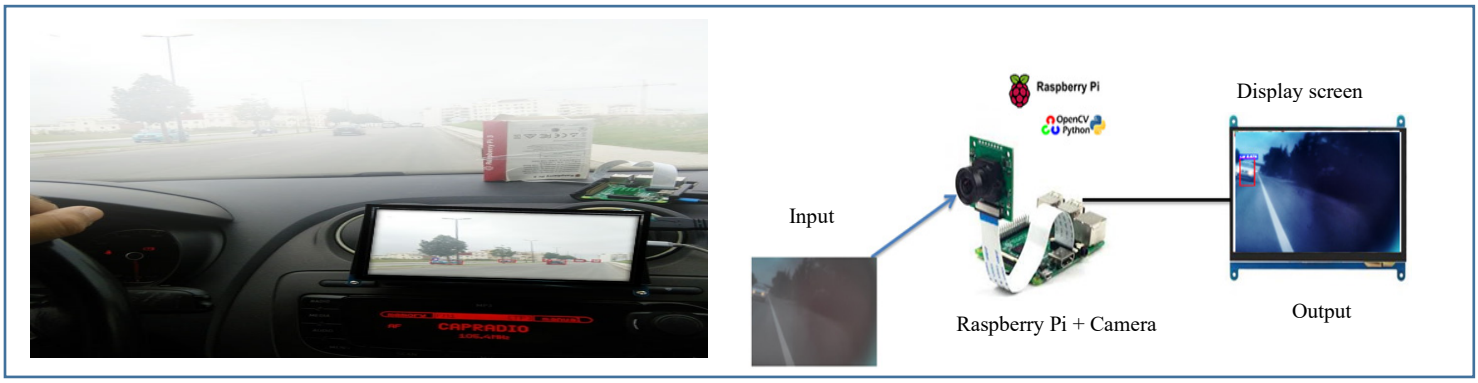
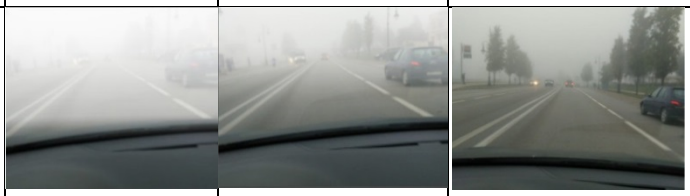
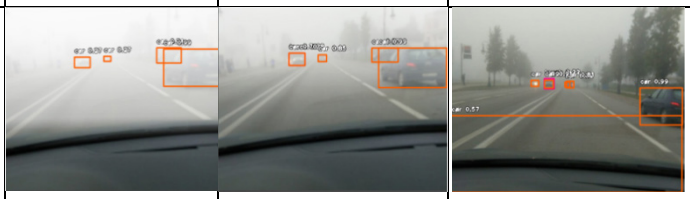


Figure 6: structure of system deployment

Table 7: Results of mAP for 5 states: Heavy haze (day), Medium haze (day), Light haze (day), Medium haze (night), Light haze (night), before and after dehazing.

	During the day			At night	
	Heavy haze	Medium haze	light haze	Medium haze	light haze
Faster R-CNN (adaptive)	mAP = 0.501	mAP = 0.663	mAP = 0.815	mAP = 0.531	mAP = 0.678
Dehazing + Faster R-CNN (adaptive)	mAP = 0.773	mAP = 0.891	mAP = 0.933	mAP = 0.723	mAP = 0.804

Table 8: the visual results of our model on images taken by the raspberry camera

	During the day			At night	
	Heavy haze	Medium haze	Light haze	Medium haze	Light haze
Hazy image					
Objet detection in hazy images					
Objet detection in Image dehazing					

We performed tests on images taken during the day and at night using the Faster R-CNN model adaptive at first, and then the proposed general model (image dehazing + Faster R-CNN adaptive). Table 7 and 8 show the results obtained.

4.1. The Quantitative results

For quantitative results, we calculated the mean Average Precision (mAP) across all images, using both the adaptive Faster R-CNN model and the dehazing model concatenated with the adaptive Faster R-CNN. (See Table 7). During the day, the heavy haze degrades mAP to around 0.27. Adding the dehazing model, mAP is improved by 0.12 for the detection of objects under a light veil conditions, 0.23 in medium fog and 0.27 in thick fog. During the night, with the dehazing model the mAP improves by 0.28 for detecting objects in light haze conditions and 0.21 in the middle fog (see table 7).

4.2. The results visualized

For the visualized results, Table 8 makes a visual comparison of the results when the object is detected on hazy images. And here we are illustrating five cases:

- **During the day:**

- Detection of objects in the Heavy Haze image with adaptive Faster R-CNN before and after dehazing.

- Detection of objects in the Medium haze image with adaptive Faster R-CNN before and after dehazing.

- Detection of objects in the Light haze image with adaptive Faster R-CNN before and after dehazing.

- **At night:**

- Detection of objects in the Medium haze image with adaptive Faster R-CNN before and after dehazing.

- Detection of objects in the Light haze image with adaptive Faster R-CNN before and after dehazing.

Experiments we conducted lead us to this conclusion: as soon as the haze becomes heavier at night, detecting objects becomes less dependable. Most of all, under whichever fog conditions- light or medium or high- our highly advanced system can regularly ameliorate the process of detecting objects.

5. Conclusion

In this paper, we present an intelligent system with a view to enhance visibility quality in atmospheric fog, allowing dehaze and object detection. The proposed system is based on the deep learning model using a light CNN model of five convolutional layers works well in terms of PSNR and SSIM compared to other existing deep methods.

We modified the Faster R-CNN model, using ResNet as its convolution backend. We concluded that using ResNet offers a substantial improvement over other architectures. Further, we concatenated the CNN video dehazing model with the faster RCNN as a robust model compared to the others. Our objective was detecting objects on the road in real time with significant performance in terms of mAP.

Based on the qualitative and visual results, our system demonstrated both efficiency and superiority over other existing systems in terms of PSNR, SSIM, mAP and visual quality.

It should be noted that, though the proposed system was much more applicable in enhancing visibility for road drivers when there is fog, there were some limitations. First of all, we did not test our proposed system in case of a rainy foggy environment. Secondly, our system is not applicable in case of very dense fog. Lastly, it is not also applicable when detecting so many objects. Consequently, in our future work we shall definitely spare no effort to enhance the performance of our system by an automated Chabot. This Chabot interprets the captured image and announces audio warnings in the event of critical situations, a project that mainly aims at developing a system in the situation where the presence of fog and rain come at the same time.

References

[1] B.Li, X.Peng, Z.Wang, J.Xu, D. Feng, “Aod-net: All-in-one dehazing network.” In Proceedings of the IEEE International Conference on Computer Vision, 4770-4778, 2017. DOI: 10.1109/ICCV.2017.511

[2] B.Cai, X.Xu, K.Jia, C.Qing, D. Tao, “Dehazenet: An end-to-end system for single image haze removal.” IEEE Transactions on Image Processing, 25(11), 5187-5198, 2016. DOI: 10.1109/TIP.2016.2598681

[3] C.Dong, C. C.Loy, K.He, X.Tang, “Image super-resolution using deep convolutional networks.” IEEE transactions on pattern analysis and machine intelligence, 38(2), 295-307, 2015. DOI: 10.1109/TPAMI.2015.2439281

[4] F.Hussain, J. Jeong, “Visibility enhancement of scene images degraded by foggy weather conditions with deep neural networks.” Journal of Sensors, 2016. https://doi.org/10.1155/2016/3894832.

[5] C.Li, J.Guo, F.Porikli, H.Fu, Y.Pang, “A cascaded convolutional neural network for single image dehazing.” IEEE Access, 6, 24877-24887,2018. DOI: 10.1109/ACCESS.2018.2818882

[6] B.Li, X.Peng, Z.Wang, J.Xu, D.Feng, “An all-in-one network for dehazing and beyond.” arXiv preprint, 2017. https://arxiv.org/abs/1707.06543

[7] Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., & Yang, M. H. “Single image dehazing via multi-scale convolutional neural networks.” In 2016Springer/Cham, European conference on computer vision, 154-169,2016. https://doi.org/10.1007/978-3-319-46475-6_10

[8] J. H.Kim, W. D.Jang, J. Y.Sim, C. S. Kim, “Optimized contrast enhancement for real-time image and video dehazing.” Journal of Visual Communication and Image Representation, 24(3), 410-425, 2013. http://dx.doi.org/10.1016/j.jvcir.2013.02.004

[9] A.Kappeler, S.Yoo, Q.Dai, A. K. Katsaggelos, “Video super-resolution with convolutional neural networks.” IEEE Transactions on Computational Imaging, 2(2), 109-122, 2016. DOI: 10.1109/TCI.2016.2532323

[10] S.Su, M.Delbracio, J.Wang, G.Sapiro, W.Heidrich, O.Wang, “Deep video deblurring for hand-held cameras.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1279-1288, 2017. DOI: 10.1109/CVPR.2017.33

[11] A.Karpathy, G.Toderici, S.Shetty, T.Leung, R.Sukthankar, L. Fei-Fei, “Large-scale video classification with convolutional neural networks.” In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 1725-1732, 2014. https://doi.org/10.1109/CVPR.2014.223

[12] H.Shen, S.Han, M.Philipose, A.Krishnamurthy, “Fast video classification via adaptive cascading of deep models.” In Proceedings of the IEEE conference on computer vision and pattern recognition, 3646-3654, 2017. DOI: 10.1109/CVPR.2017.236

[13] D.Chen, J.Liao, L.Yuan, N.Yu, G.Hua, “Coherent online video style transfer.” In Proceedings of the IEEE International Conference on Computer Vision, 1105-1114, 2017. DOI: 10.1109/ICCV.2017.126

[14] D.Liu, Z.Wang, Y.Fan, X.Liu, Z.Wang, S.Chang, T. Huang, “Robust video super-resolution with learned temporal dynamics.” In Proceedings of the IEEE International Conference on Computer Vision, 2507-2515, 2017. DOI: 10.1109/ICCV.2017.274

[15] S.Albawi, T. A.Mohammed, S.Al-Zawi, "Understanding of a convolutional neural network." In 2017IEEE International Conference on Engineering and Technology (ICET), 1-6, 2017. DOI: 10.1109/ICEngTechnol.2017.8308186

[16] Y.Xu, G.Yu, Y.Wang, X.Wu, Y. Ma, "Car detection from low-altitude UAV imagery with the faster R-CNN." Journal of Advanced Transportation, 2017. https://doi.org/10.1155/2017/2823617

[17] E. J. McCartney, "Optics of the atmosphere: scattering by molecules and particles." New York, John Wiley and Sons, Inc., 421, 1976. https://doi.org/10.1063/1.3037551

[18] S. G.Narasimhan, S. K. Nayar, "Chromatic framework for vision in bad weather." In 2000 Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 598-605, 2000. DOI: 10.1109/CVPR.2000.855874

[19] B.Li, X.Peng, Z. Wang, J.Xu, D. Feng, "End-to-end united video dehazing and detection." In 2018 AAAI Thirty-Second Conference on Artificial Intelligence, 2018. https://arxiv.org/abs/1709.03919

[20] K.He, X.Zhang, S.Ren, J. Sun, "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778, 2016. DOI: 10.1109/CVPR.2016.90

[21] W.Liu, D.Anguelov, D.Erhan, C.Szegedy, S.Reed, C. Y.Fu, A. C. Berg, "Ssd: Single shot multibox detector." In 2016 Springer/Cham, European conference on computer vision, 21-37, 2016. https://doi.org/10.1007/978-3-319-46448-0_2

[22] S.Ren, K.He, R.Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, 91-99, 2015. DOI: 10.1109/TPAMI.2016.2577031

[23] Z.Li, P.Tan, R. T.Tan, D.Zou, S.Zhiying Zhou, L. F. Cheong, "Simultaneous video defogging and stereo reconstruction." In Proceedings of the IEEE conference on computer vision and pattern recognition, 4988-4997, 2015. DOI: 10.1109/CVPR.2015.7299133

[24] M.Sulami, I.Glatzer, R.Fattal, M.Werman, "Automatic recovery of the atmospheric light in hazy images." In 2014 IEEE International Conference on Computational Photography (ICCP), 1-11, 2014. DOI: 10.1109/ICCPHOT.2014.6831817

[25] G.Meng, Y.Wang, J.Duan, S.Xiang, C.Pan, "Efficient image dehazing with boundary constraint and contextual regularization." In Proceedings of the IEEE international conference on computer, 617-624, 2013. DOI:10.1109/ICCV.2013.82

[26] D.Berman, T.Treibitz, S.Avidan, "Air-light estimation using haze-lines." In 2017 IEEE International Conference on Computational Photography (ICCP), 1-9, 2017. DOI: 10.1109/ICCPHOT.2017.7951489

[27] J. P. Tarel, N. Hautiere, "Fast visibility restoration from a single color or gray level image." In 2009 IEEE 12th International Conference on Computer Vision, 2201–2208, 2009. DOI: 10.1109/ICCV.2009.5459251

[28] K.He, J.Sun, X. Tang, "Single image haze removal using dark channel prior." IEEE transactions on pattern analysis and machine intelligence, 33(12), 2341-2353, 2010. DOI: 10.1109/TPAMI.2010.168

[29] Q.Zhu, J.Mai, L.Shao, "A fast single image haze removal algorithm using color attenuation prior." IEEE transactions on image processing, 24(11), 3522-3533, 2015. DOI: 10.1109/TIP.2015.2446191

[30] J.Sturm, N.Engelhard, F.Endres, W.Burgard, D.Cremers, "A benchmark for the evaluation of RGB-D SLAM systems." In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 573-580, 2012. DOI: 10.1109/IROS.2012.6385773

[31] F. Mal, S. Karaman "Sparse-to-dense: Depth prediction from sparse depth samples and a single image." In 2018 IEEE International Conference on Robotics and Automation (ICRA), 1-8, 2018. DOI: 10.1109/ICRA.2018.8460184

[32] B.Cai, X.Xu, D. Tao, "Real-time video dehazing based on spatio-temporal mrf." In 2016 Springer/Cham, Pacific Rim Conference on Multimedia, 315-325,2016. DOI: 10.1007/978-3-319-48896-7_31

[33] C.Sakaridis, D.Dai, L. Van Gool, "Semantic foggy scene understanding with synthetic data." International Journal of Computer Vision, 126(9), 973-992, 2018. https://doi.org/10.1007/s11263-018-1072-8

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 314-331 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

ISR Data Processing in Military Operations

Ladislav Burita*, Ales Novak

Department of Informatics and Cyber Operations, University of Defence, Brno, 66210, Czech Republic

ARTICLE INFO

Article history:
Received: 17 June, 2020
Accepted: 31 August, 2020
Online: 17 September, 2020

Keywords:
ISR data processing
Business Intelligence
Data Mining
Big Data

ABSTRACT

This paper provides an overview of Intelligence, Surveillance, and Reconnaissance (ISR) data with respect on NATO standards and recommendations; further presents methods, tools, and experiences in ISR data processing in military operations. The steps of the Intelligence cycle and disciplines Business Intelligence (BI), Data Warehousing, Data Mining, and Big Data are presented in the introduction. The literature review is oriented to the analysis of intersections between ISR and BI methods. The next chapter describes the ISR data processing in detail; there are listed structures, formats, standards, and data from the operational point of view. The ISR operational picture is explained, and steps of the ISR data mart is completed. The last part is oriented to Big Data processing; NoSQL, in-memory and streaming databases. The last two chapters are focused on the description of research results in the processing of ISR data. The ISR data mart experiment processes the radio transmission data that consists of detected radio signals. Results are visualized in RapidMiner Studio. The Big Data experiment is realized in Apache Hadoop.

1. Introduction

This review paper is an extension of work originally presented in the 7th International Conference on Military Technologies [1]. Paper summarizes methods, tools, and experiences in ISR (Intelligence, Surveillance, and Reconnaissance) data and information processing in military operations.

The important focus in the paper is oriented on the NATO ISR data standards. The ISR data format includes in authors point of view: the structured data, unstructured (text) data and Big Data.

The ISR data processing mainly uses Business Intelligence (BI) techniques, which include data and text analysis, Data Mining (DM), Data Warehousing (DW) On Line Analytical Processing (OLAP), and Decision Support Systems (DSS).

1.1. The ISR Processes and Data

The Command and Control (C2) of the military units is mostly mentioned as OODA (Observe-Orient-Decide-Act) management cycle [2]. Each phase of the cycle can be supported by appropriate information technologies (IT). The most important phase is orientation, which affects both the creation of the right alternatives for decision and the method of observation and that functions are supported by IT. The Intelligence Cycle (IC) supports the orientation phase. The IC is defined as "The sequence of activities whereby information is obtained, assembled, converted into intelligence and made available for users." The IC goal is to solve commanders' problems related to the operational situation. It consists of four phases [3]: Direction, Collection, Processing and Dissemination.

The ISR process "A coordination through which intelligence collection disciplines, collection capabilities and exploitation activities provide data, information and single-source intelligence to address an information or intelligence requirement, in a deliberate, ad hoc or dynamic time frame in support of operations planning and execution" [4] is a part of data collection in IC.

The ISR activities consist of five phases: Task, Collect, Process, Exploit and Disseminate (TCPED). In the collection phase, by the NATO documents, data are obtained from heterogeneous sources and thus covers multiple intelligence disciplines:

- Human Intelligence (HUMINT) - information that is collected and provided by human sources.
- Acoustic Intelligence (ACINT) - the detection and utilization of acoustic signals or emissions.
- Imagery Intelligence (IMINT) - collection, processing and exploitation of image sequences.
- Measurement and Signature Intelligence (MASINT) – the source of information in scientific and technical area.

*Corresponding Author: Ladislav Burita, ladislav.burita@unob.cz

- Open Source Intelligence (OSINT) – information gathered from publicly available sources.
- Signals Intelligence (SIGINT) - collecting and exploiting electromagnetic signals.

The approach of NATO to ISR area is rapidly changing because of disrupting possibilities in technology and science. In the last years, the community of innovation in NATO has been in cooperation with the NATO Support and Procurement Agency (NSPA), prepared conditions for extensive changes in the AWACS (Airborne Warning and Control System) aerial survey and the method of exchanging information.

NATO is planning the next version of AFSC (NATO Alliance Future Surveillance and Control) as an ISR distributed network systems as opposed to a single platform. This is meaning as combination of technical and human resources that would provide more enhanced surveillance capability opposed a single complex and expensive platforms. [5]

1.2. Business Intelligence

Business Intelligence (BI) is including tools, technologies, and practices data collection, aggregation or integration, processing, analysis, and presentation of in the form of information to support decision-making tasks.

All sources of data including information systems (IS) for decision preparing contain the huge volume of data. It is not possible to obtain the awaiting information for the decision process directly from the data sources and it is unnecessarily time consuming and expensive.

It must be prepared many questions that lead to the desired result. Data are spread in many sources in a various format, and therefore obtaining information is very difficult.

Commanders or directors are able to prepare decision strategy quicker and better by BI; are able to simply prepare or modify company strategies or processes, which leads to competitive advantage in a military area.

BI is an approach that has an impact on the effective decision, and thus increases the quality of the organization and leads to an advantage in warfare.

BI infrastructure includes components that collects, stores, summarizes, aggregate, integrate and analyzes the data produced by an organization or military unit. It is a broad term that encompasses data mining, data warehousing, online analytical processing (OAP), performance benchmarking, descriptive analytics, business process reengineering (BPR) and reporting, see Figure 1.

1.3. The data Warehousing and OLAP

The data warehouse (DW) is prepared as a trusted and integrated data collection for analyze processes with outputs that are directed to the decision process in a company or a military unit. It helps users solve problems and preparation decision in business and military missions.



Figure 1: Business Intelligence [6]

Sources for DW are IS in business, Enterprise Resource Planning (ERP) systems, Customer Resource Management (CRM) data sources, and other data files, see Figure 2. The DW is preparing from data sources thru ETL (Extract – Transform – Load) processes.

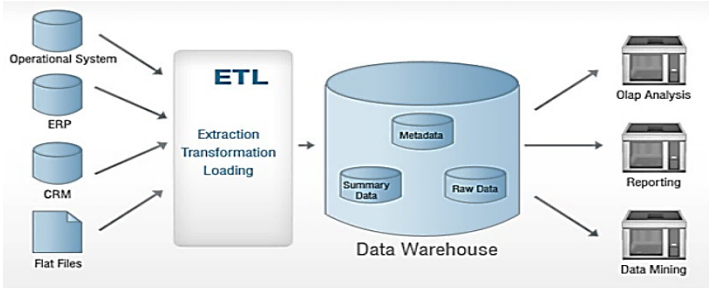


Figure 2: Data Warehousing [7]

The main reason of DW constitution is the possibilities of the IS of other data sources for decision-making support:

- Information systems have not a satisfactory performance in analysis, it takes too long, it is expensive and the historical data are usually missing.
- Data are spread in more inconsistent applications (production, commerce, human, finance, logistics, and other) resources.
- Database structure of the DW data sources are not optimal for decision-making elaboration.

OLAP (or OLAP Cube) is abbreviation of On Line Analytical Processing. It is a multidimensional memory whose dimensions form required view of data structure and many variants of detail view in information. The OLAP cube in gasoline industry consists of dimensions: Time, Product, and Measure, see Figure 3.

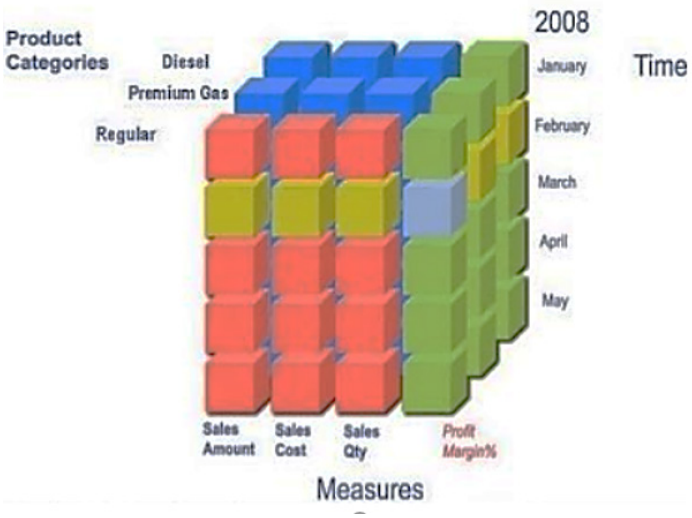


Figure 3: OLAP – the multidimensional memory [8]

Each node of the OLAP consists of product category, time, and sales results (amount, cost, quantity, and profit margin in %). In the prepared data structure, the results can be by managers good monitored from various directions.

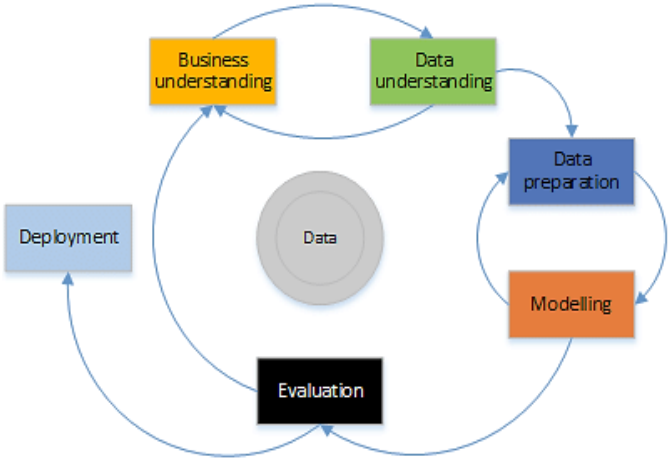


Figure 4: CRISP-DM methodology phases [9]

1.4. Data Mining, Methods and Tools

The mission of Data Mining (DM) is to find information, not known before, from databases. There are used modelling approach and statistical tools to gain knowledge from data. The usually used DM methodology is Cross Industry Standard Process (CRISP) that include processes of data acquisition, processing and transformation by application of DM method, evaluation, visualization, and presentation; see Figure 4.

The effectiveness and quality of the decision process depends on the reliable data; mistakes may lead to bad results. Mistaken data is a result of practical processing and data can be incomplete, inconsistent or should contain problems.

Set of actions in the data processing:

- Cleaning Data - completion the missing values, identify and remove boundary values, eliminate data inconsistency.
- Data Integration – various sources of data should be integrated.
- Data Transformation – aggregation, normalization and reduction makes data useful for analysis.
- Descriptive summarization of data – understanding of structure and properties of the data.
- Discretization - conversion of analogue format into digital form.

The often-used DM methods include decision trees (DT), neural networks (NN), statistical methods (SM), and genetic algorithm (GA). The popular method of DM is statistics: discrimination and regression analysis, and pivot tables.

In DT method is data divided into groups which represent the same class. The goal is to find the optimal tree by the principle that differentiate chosen attributes the best.

Method NN is based on set computational neurons with weighted connections. NN functions are set up by testing data.

GA is used if it is a problem to find the quick solution, and method is likened from the natural environment.

The most popular BI SW in the academic environment are products SAS Institute and IBM SPSS.

SAS is the statistical SW for data management, various kinds of analysis (multivariate, predictive, or statistical) for BI, a policy investigation. SW is able to process data from many sources, perform analysis on them, and makes available a graphical user interface for common user and developed solution using the special language for technical users.

The most extensive SAS solution offers for customer intelligence. It includes parts for marketing, social media and web analytics, is able to profile customers, predict their business steps, and provides optimal communication.

The SAS SW for enterprises and organizations is preparing for risk and compliance analysis, includes risk simulation, scenario building to reduce the risk and development of corporate policies.

The SAS supply-chain SW prepares forecasting for company development, solution of products distribution, inventory and optimal pricing. Exists also the SAS SW for environmental sustainability, economic and social consequences in causal relations between production and its impact on the eco-environment. The SAS has solution for government, industry, trade, and communications. [10]

The SPSS SW suite solves various kind of analysis (statistical, text, predictive), includes solution for artificial intelligence (machine learning algorithms).

SPSS offers open-source connections, integration with Big Data tools, and simply deployment into own applications.

The SPSS SW is prepared as flexible and ease to use for all categories of users (from common to advance).

SW is offered for all sizes and complexity of projects and can improve company or organization efficiency, minimize risk, and find opportunities [11].

1.5. Big Data

Big Data is a term that includes primarily theme with amount of data, its format, generation and changing of data, and processing. The first characteristics of big data is large *volume*. Simple example is possible to take from social media that millions users generate every second data that volume is constantly growing.

The next characteristics is *velocity* of generating and changing data. Simple example is possible to take from Internet of Things (IoT), which sensors generate data about your condition. Sometimes they are data streams, their processing cannot be performed by conventional analytical methods.

The third characteristics is dealing with *variety* of formats. Data can be standardized or unstandardized, formalized or structured (databases, files), unstructured or text (documents), web pages, pictures, schemas, graphs, etc.

The last two characteristics refers to the *veracity* refers to the credibility and reliability of the data, and the *value* is the importance given by company or military organization to use data. In a military environment, the reliability of data, for example when assessing the enemy's situation, is crucial.

The summary of Big Data definition concludes the five "V" specification: volume, velocity, variety, veracity and value. There are special means and techniques for big data processing, an overview of which is at Figure 5. They are divided into three groups, and 15 sub-groups:

1. Computing tools.
2. Storage tools.
3. Support technologies.

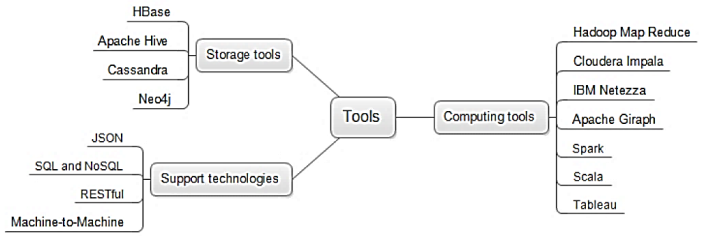


Figure 5: Mind map representation of Big Data tools [12]

As the variety, volume, veracity, value, and velocity of data is undergoing a rapid transformation and aggravating the cognitive burden on the Warfighter, armed forces require technologies to store, process, verify, and disseminate actionable intelligence at pace.

It is widely accepted that decision-making in the future information-dominated battlespace will depend on big data processing using artificial intelligence (AI) methods. It is important to address ways to gain and retain the information advantage through disruptive innovation and big data analytics, addressing cloud computing, the acquisition of data from disparate sources, data validation, visualization, accelerated and autonomous decision-making, and network resilience through cyber defense mechanisms. AI, big data technologies and ML will transform ISR and command and control (C2). [13]

2. The Literature Review

The aim of literary review is to map the interest of authors of articles, indexed on Scopus, in the topic published in the article. Information from the selected articles is oriented about the knowledge of the ISR area in relation to information technologies (IT). Overview of the selected relevant articles is listed in Table 1. The goal of individual queries was to find the relevant papers only in one area of interest. The queries were directed to all fields of the papers (the first version of the literature review, on March 26, 2020). The IT query found papers that correlate the query (BI and DW and DM and BigD) and in complex query was the result connected to ISR area. Only one paper [14] has met a complex selection requirement. The paper deals with the application of big data analytics in meteorological research.

Table 1: Number of relevant papers to the query - the first version

Query	ISR	BI	DW	DM	BigD
Individual	3 981	30 478	36 840	552 791	228 493
IT		619			
Complex	1				
ISR and DM	8 (open access, citated>0)				
ISR and BigD	9 (open access, citated>0)				

After the first version of the literature review was discovered that more than one-half of the found papers have keyword "ISR" matching only in references. Therefore, a second literature review was carried out (on April 2, 2020), where the queries were directed only to fields Title, Abstract, and Key-words of the papers. The overview of the selected relevant articles is listed in Table 2.

Table 2: Number of relevant papers to the query – the second version

Query	ISR	BI	DW	DM	BigD
Individual	1 697	7 503	12 849	169 113	81 292
IT		22			
Complex	0				
ISR and DM	15 (5, citated>0)				
ISR and BigD	12 (5, citated>0)				

The ten selected articles (see Table 3) were further analyzed in detail; see Table 3 that shows the areas of interest. Each paper was described by three themes – characteristics of the paper interest. Among the most, numerous characteristics of the analyzed papers are "military" and "data or information fusion", these themes integrate orientation and interest of most analyzed papers with connection to the theme "ISR data processing". This first conclusion of the analysis correlates completely with the focus and aim of the article.

Let us now get focus more closely on the analyzed articles. First, how data or information fusion is assumed in chosen articles and with what requirements is this operation connected? In addition, the second focus: Which artificial intelligence (AI) techniques, agent based, and knowledge approaches were used.

In [15], the author explains information superiority than can be achieved by application of the four RIGHT: information, time,

format, and people. However, it is only right when skilled commanders are able getting information use in praxis. Information superiority is possible only by intensive application of information and communication technologies that spread information in almost real time by remote access regardless of time and space. The condition of good cooperation on the battlefield is information interoperability and security in broaden contexts. That is why NATO is intensively developing standardization in the field of data exchange, which includes the concept Coalition Shared Data (CSD). It supports operational processes in NATO Joint ISR and by using the Intelligence Cycle.

Table 3: Areas of interest in the analyzed papers

Area of interest	Paper										Σ
	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]	
Agent-based system		1			1						2
AI-Machine learning							1				1
AI-Neural network									1		1
Air Force						1					1
Cybersecurity										1	1
Data or information fusion	1	1	1	1					1		5
DNN architecture				1							1
GDPR								1			1
HPEC architecture						1					1
Information management	1										1
ISR mission support						1					1
Knowledge approach				1							1
Military	1	1	1		1	1	1		1	1	8
Navy					1					1	2
OSINT and BDA								1			1
Personal security								1			1
QuEST strategy							1				1
STACK model			1								1

Explanation to Table 3:

- AI-Artificial Intelligence
- BDA-Big Data Analytics
- DNN-Dual Node Network
- GDPR-General Data Protection Regulation
- HPEC-High Performance Embedded Computing
- OSINT-Open Source Intelligence
- QuEST-Qualia Exploitation of Sensor Technology
- STACT-Spatial, Temporal, Algorithm, and Cognition

In [16], the author presented elements to achieve integration of information (information fusion). The solution includes a distributed surveillance system and intelligence databases. Data sources comes from electronically situational awareness, controlled by military and civilian security operators and commanders in principle of ISR. The surveillance system works automatically; integrates data from aircraft systems, sensors, and background information into a military situational picture. Components for information fusion are agent-based mechanisms for integration and drawing using different reasoning techniques, knowledge modelling, information extraction and transformations.

The data fusion strategies for ISR missions, similar to previous paper, is presented in [17]. Extraction of sematic information from various set of sensors and its processing is based on new AI algorithms and computational power. Content of the project is bringing challenges for information fusion algorithms for moving targets, combination of image data with different temporal frequency, getting the contextual information by monitoring of mobile phones. Name of solution is STACK- Spatial, Temporal, Algorithm, and Cognition.

In [18], the author described the distributed fusion systems for integration. Solution includes adaptive methods that improve usefulness data sources. |The data processing cycle supports to find needed data, to develop models (situation modelling), and evaluate predictive characteristics of model (impact of prediction to success in mission). The goal of system is to improve availability, quality, variety, and actuality of data for information fusion. Machine learning techniques learn to find typical symptoms for integration data with fusion system and ontology. Extension of the Data Fusion & Resource Management Dual Node Network (DNN) architecture was applicate. The data mining methods are used to find models in data sources with help of automatically learn component behaviors by fusion procedure.

Project Albert is oriented to key areas that traditional modeling and simulation techniques often do not capture satisfactorily and uses two data management concepts, data farming and data mining, to assist in identifying areas of interest. The actual solution in Project includes four agent-based models that allow agents to interact with each other and produce emergent behaviors [19].

The Air Force Research Laboratory is developing new computing architecture, designed to provide high performance embedded computing solution to meet operational and tactical real-time processing intelligence surveillance and reconnaissance missions [20].

The achieved results in use of big data and relevant technologies, introduction of artificial intelligence (machine learning) learning are specified in [21]. The situation understanding is solved in collaboration human and machine using the Qualia Exploitation of Sensor Technology (QuEST). The QuEST is, in comparison to the limits of traditional big data processing, a completely progress and offer a higher-level system in situational awareness preparation.

Analyses, focused to open source intelligence and big data analytics with the emphasis on reconnaissance and how personal security is part of that entity. The main question is how privacy manifests itself as part of both mentioned data sources [22].

The big data from military sources are extensively multimodal and multiple intelligence, where numerous sensors along with human intelligence generate structured sensor and unstructured audio, video and textual ISR data. Data fusion at all levels remains a challenging task. While there are algorithmic stove-piped systems that work well on individual modalities. There is no system that can seamlessly integrate and correlate multi-source data; as textual, hyperspectral, and video content. It is not effective to solve integration of individual data sources gradually; better results can be achieved by integration at a higher level of the model by deep fusion of all data sources, while achieving accuracy of integration and obtaining the basic properties of signals [23].

US Navy is proposing emerging maritime strategy of command, control, communications, computers, intelligence, surveillance and reconnaissance with big data [24].

Summary in the sources of the literature review to the content of the manuscript: the one view is bringing similarities in data description and its processing and other view differences between analyzed papers and this manuscript in detail point of view at ISR data. The manuscript is specifying formats of ISR data as NATO standard (STANAG). ISR products are described and explained in relation to the command and control processes, management cycle and to the intelligence loop.

The common focus is in military area with ISR data processing, including even big data. The ISR data has sources from many various activitiers and is of different formats. Procedures of data processing are oriented primarily to data integration (fusion) and then to dissemination for military operations support.

In the manuscript are mentioned experiences and achievements of the Czech Armed Forces (CAF) in the described topic. It is also a comprehensive overview of the use of business intelligence tools.

The manuscript presents the results of research during the elaboration of the dissertation, too.

3. ISR Data Processing

3.1. ISR Data Overview

The Intelligence Cycle, the ISR process and its individual steps were introduced in previous chapters. In these individual steps, many heterogeneous sensor data and information are created, processed and managed. Military orders and tasks are issued to military units, civilian agencies and other cooperation organizations. Data are collected on gathering sensors platforms based on those orders. Operators and analysts process and exploit collected data and based on them they create ISR reports or other needed documents. Finally, ISR results are spread to its customers.

All sensor data, information and documents can be processed, stored and managed in many different ways. They are typically stored in the traditional File System or the Relational Database. This type of storage represents an easy way to store the data but in most cases are proprietary data formats used. There is a lack of interoperability because of proprietary data format utilization so it is difficult or even impossible to share proprietary data among different sensor platforms, units or even allies and nations. As a result, operators, analysts and commanders might not be able to understand gathered data so those data are not used in the decision preparation. NATO to reduce this interoperability lack introduced the concept of Coalition Shared Data (CSD).

CSD was standardized in STANAG 4559 as the NATO Standard ISR Interface (NSILI) [25] and its objective is to ensure interoperability for ISR documents exchange. NSILI defines the interface as a standard to search and access in ISR document libraries (IPL). CSD and NSILI are used as a synonym within NATO ISR community. For simplicity, the term NSILI will be used for the standard itself and the term CSD for its physical implementation in this article.

The CSD aims to ensure interoperable sharing of ISR products but NSILI is not designed and intended for user-friendly data querying, searching, reporting, data mining and analysis. NSILI does not specify client behavior; it is focused only on the server-side functionality and machine-to-machine interfaces, such as other CSDs, sensor platforms and exploitation stations. Moreover, NSILI data model is not universal military data model and therefore it does not cover all relevant military data formats so all ISR data and information cannot be stored in the CSD. NSILI also defines its own dissemination and synchronization mechanism, which affects data accessibility and availability on different nodes.

The operator and analyst needs to search relevant data efficiently regardless of their formats, storage or even CSD implementation details and limits. To achieve this goal, it is proposed to build a data warehouse on top of CSDs, sensor data databases and other military data sources. The data warehouse can help to solve some issues mentioned while other difficulties would remain. Those are connected with strict NSILI data model and streaming nature of some sensors data. The Big Data approach will be utilized to cope with those aspects.

3.2. ISR Data Formats

In the following paragraphs, will be introduced relevant ISR and military data formats, which will be later used as a base for building ISR data warehouse. Many data format standards are applied in a military environment to store ISR specific data. Those data format standards include common military data formats, ISR specific ones as well as common enterprise data formats such as PDF, Office formats, images formats, video formats etc. In the article is focus only on military data formats, both common and ISR specific ones.

3.3. Common Military Data Formats

Common military data formats relevant to ISR data include textual messages, operational pictures, blue force tracks etc. These data are widely used for the exchange and integration of military systems. Relevant data standards are already implemented and adapted in the CAF and its implementation in the CAF has been confirmed on many experiments, trials, etc. Some of those standards cover not only data formats but includes also transport protocol and mechanism of exchange.

Textual messages arte included in the APP-11 [26] message catalogue, which is NATO product, used in exercises and missions to exchange text as a structured information between allied military units. The messages are prepared upon the latest technical standard ADatP-3 that defines the rules of organizations and exchange of the messages. The actual version of the APP-11 message catalogue includes four hundreds messages ISR specific ones. All messages can be exchanged using the XML or slash-separated files.

Operational pictures utilize the NATO Vector Graphic (NVG) for coding and common using overlays. The NVG includes an XML data format for the coding battlespace elements into result messages and report for automated exchange of the overlays [27].

Blue Force Tracking (BFT) is the ability to track flawlessly with 100% accuracy the identification and location of own military forces in NATO missions, exercises and operations in almost real time. NATO C2 are equipped with this system and preparation for working with BFT is carefully required. NATO Friendly Force Information (NFFI) and NFFI in Message Text Format (FFI MTF) are two XML based message forms that are used by Blue Force Tracking capability [28].

3.4. ISR Specific Data Formats

The ISR community developed and adopted a wide set of data format standards for individual intelligence disciplines. The set of those ISR data standards is maintained by STANREC 4777 Ed. 2 - NATO ISR Interoperability Architecture (NIIA) [29]. Although NIAA lists about 25 key ISR standards, only the most significant for the CAF are mentioned in this article:

STANAG 4545 – NATO Secondary Imagery Format (NSIF) provides implementation guidance that is designed for the distribution, storage, and interchange of secondary imagery products.

STANAG 4559 – NATO Standard ISR Library Interfaces (NSILI) promotes interoperability of NATO ISR libraries and provides services for the exchange of shared ISR products.

STANAG 4609 – NATO Digital Motion Imagery Standard (NDMIS) guides the consistent implementation of Motion Imagery Standards to achieve interoperability in both the communication and functional use of Motion Imagery Data.

STANAG 4658 – Cooperative Electronic Support Measure Operations (CESMO) exists to enable the warfighter to rapidly share and receive electromagnetic spectrum threat information.

STANAG 5516 – Tactical Data Exchange - Link 16 provides guidelines on how to ensure interoperable use of Link 16 Tactical Data Links (TDL) to disseminate information.

The most important standard from a data integration perspective is already introduced STANAG 4559 NSILI, which will be evaluated in more details. NSILI defines architecture; use cases (publish, ingest, retrieve, synchronization); two service interfaces, modern web service interface and obsolete Common Object Request Broker Architecture (CORBA) interface; metadata model etc. NSILI covers two main functional areas, ISR Product Library and JISR Workflow.

ISR Product Library covers the first set of data managed by NSILI. It is specified as an AEDP-17 NATO STANDARD ISR LIBRARY INTERFACE. There are about 20 different types of ISR documents that are supported by standard AEDP-17.

They have format as ground moving target indication, tactical data link (TDL) reports, geographic area of interest and geospatial vector, intelligence information requirements, images and video, Microsoft office files, request for information (RFI), operational rules, organization structures (ORBAT), system specifications and assignments; chemical, biological, radiological and nuclear information; electronic order of battle, system deployment status, messages and tactical symbols.

AEDP-17 defines the following services:

- The service to metadata of query entries of ISR documents included within CSD servers.
- The publishing service to publish and modify the content of CSD servers.

The metadata set valid to all ISR products is specified by NSILI. This set contains ISR document status, source library, publisher, date and time of modification, security classification and ability for publication, product type, identifier of mission, used language, comments, related geographical data etc.

NSILI is defining as an additional group of metadata for each ISR document. The additional set of metadata extends the common set of metadata by attributes relevant only to given ISR product type; it is containing logical data model of metadata but it does not prescribe how it should be implemented on the physical level. Query and publish services defines XML based structure of metadata (Catalogue Entry Data Set) which can be exchanged among multiple CSDs.

JISR Workflow is covering the second set of data managed by NSILI. It is specified as an AEDP-19 NATO STANDARD ISR WORKFLOW ARCHITECTURE [30]. AEDP-19 defines Community of Interest (COI) services to manipulate following data entities: organization, requirement, priority, task, request and geographic area of interest (GAOI). COI services utilize underlying Simple Persistence Service (SPS++) to store entities.

SPS++ stores entities as XML artefacts in a content-agnostic way and allows easy synchronization among nodes. On the other hand, SPS++ does not allow efficient search, because the whole entity is stored as one (atomic) XML artefact without any auxiliary attributes.

Both sets of services, ISR product Publish/Query services and JISR Workflow services are intended for operational purposes mainly, not for data analyses. The support for efficient data searching is very limited because

IPL introduces its own specific query language, which enables to search ISR products according to the value of attributes, and its combination but extended data querying is limited. JISR WF services provide even only predefined business services without any extension mechanism available.

3.5. ISR Data from Operational Point of View

In the previous sub-chapter, selected data formats were described from a technology point of view. The following section will focus on its utilization in ISR operations from an operational point of view. The ISR tasks, sensors data, ISR products, and ISR operational pictures will be discussed.

3.6. ISR Tasks

The ISR task is a directive for the appropriate employment of ISR assets. There are software tools utilized to manage ISR tasks and visualize them in different perspectives (time, unit, GAOI) to depict ISR tasks decomposition, dependencies or conflicts. The main goal is to effectively coordinate and manage data gathering activities. [31]

In NATO operations, ISR tasks should be managed within the Information Request Management and Collection Management (IRM&CM) process [32].

IRM&CM process is corresponding with Task and Collect phases of the ISR process and it aims to decompose commanders' tasks in form of Request for Information (RFI) into the hierarchical structure of Priority Information Requirements (PIR), Specific Intelligence Requirements (SIR), Essential Elements of Information (EEI). ISR tasks may be refined into specific (proprietary) orders to allow automated tasking of ISR assets.

The overview lists and plans are created based on this decomposition and it consists o fa list of approved and prioritized ISR collection requirements, ISR tasks and dynamic re-tasking priorities. [33]

Those lists and plans include Intelligence Collection Plan (ICP), Collection Exploitation Plan (CXP) and Collection Tasking List (CTL), which are intended to track and monitor the operation of data collection.

It aims to assist in producing, completing and monitoring intelligence requirements, to indicate how each requirement is to be satisfied, to visualize tasks structure and de-conflict tasks.

The general structure of an ISR task includes identification of the task, name of the task, description of the task, associated geographic area, date and time of task validity, sensor or asset responsible for collecting activity according to organization structure (ORBAT), associated PIRs, SIRs, EEIs and other detailed parameters.

There are a lot of technology standards used to store ISR tasks:

- Structure of requirements, tasks and requests defined in AEDP-19, which should be the preferred option.
- Tasks format defined in AEDP-17.
- APP11 messages stored in the file system or the database.
- Intelligence discipline-specific formats, e.g. CESMO requests.
- Proprietary textual or binary formats stored as files.
- Proprietary database structures.
- Office documents (Word, Excel, PDF), and others.

It is not possible to create an overall overview of all tasks due to diversity of used data formats. To solve this issue data mart could be built on the top of multiple ISR tasks storages. Building this data mart is challenging due to the fact of a variety of data formats used. Tens or even hundred APP11 message types, individual disciplines data formats, office documents, AEDP-17 and AEDP-19 XML structures are in a place for the ETL phase. Due to the number of data sources and used data formats, implementation of ETL phase might include hundreds of transformation rules and its maintenance might be expensive. Also, rigid data structures of data mart might not be flexible enough, more flexible Big Data technologies could be utilized to build a data mart and mine the data.

3.7. Sensor Data

Sensor data are the main output of the Collection phase of the ISR process. Sensor data serves as input for the analyst's work. Although intelligence disciplines cover multiple data sources, this article and experiments focus, only on those sensors data formats which are used in the CAF, especially IMINT and COMMS-ESM ones.

COMMS-ESM data formats

Sensor data formats for COMMS-ESM includes detected activities, direction findings and geolocations. Detected activity (DA) is a logical data entity describing detected radio signals and its parameters. DA is produced by sensor platform based on (multiple) sensors (activity detectors) raw data.

DA consists of physical parameters such as carrier frequency, quality and level of the detected signal, date and time of the detection, transmission time, type of signal, type of the modulation and other technical information depending on the sensor and its capabilities. All parameters are the same within one detected activity, when one of the parameters changes, the new activity is automatically detected and recorded.

Direction finding (DF) is a logical data entity describing direction from an emitter (to the receiver) of the radio signal and its parameters. DF is produced by sensor platform based on (multiple) sensors (direction finders) raw data. DF consists of physical parameters such as direction, focus error, carrier frequency, quality and the level of detected signal, date and time of the detection, transmission time, and other technical information depending on the sensor and its capabilities. DF is associated with one or multiple DAs.

The geolocation is a logical data entity describing a position of the emitter and error ellipse. The geolocation can be produced on the sensor platform or commands post, it can be calculated automatically or set manually, both options are based on the intersection of DFs.

All mentioned entities (DA, DF, geolocation) are standardized in STANAG 4658 CESMO but proprietary formats are widely used on most sensor platforms, as captured physical parameters are always sensor specific. Gathered data are transformed into CESMO format in the case that sensors and assets are running CESMO operation and they are connected into the CESMO network.

CSD does not support those entities (DA, DF, geolocation) as they are in the scope of Electronic Warfare (EW) rather than in the scope of ISR. Due to this fact, EW and SIGINT data are stored in specific (proprietary) databases, which can be utilized for building ISR data mart. Due to the diversity, volume and streaming nature of the data, the Big Data approach can be considered in processing those data.

Imagery data formats

Imagery data formats are intended to store images (photos), video sequences and video streams. Many proprietary and standard-based data formats are used in this area to cover the complexity and diversity of image and video shooting. Imagery data formats contain not only imagery data itself but also image/video metadata such as a sensor position, heading and telemetry, camera technical details (field of view, lens parameters, etc.), camera type (electro-optical, infra-red, thermal, etc.), picture technical parameters (resolution, dimension, bands, encoding, etc.), video technical parameters (resolution, dimension, sampling frequency, encoding, etc.). Imagery analysts can utilize only a few of those metadata; the majority of them is rather technical.

Imagery data and video streams and clips are well supported in NSILI; however, this support is limited only to images compliant to STANAG 4545 NSIF and video clips and stream compliant to STANAG 4609 NDIMS. In the case that other picture or video formats are used, the data should be converted and enriched to

supported formats or it can be stored as a simple document type without taking advantage of imagery and/or video metadata.

STANAG 4545 NSIF is a container for images and its metadata. Imagery data and metadata are combined into one complex binary file. Industry based and proven image formats and codecs such as JPG, JPG2000, not compressed images etc. are utilized for images encoding. Common NSILI metadata set is extended for imagery products with parameters such as imagery category, cloud cover percentage, decompression technique, number of bands, number of rows, number of columns and title.

Common NSILI metadata set is extended for video products with parameters such as average bitrate, category, encoding schema, framerate, number of rows, number of columns, metadata-encoding scheme, MISM level, scanning mode, number of VMTI target reports and VMTI processed. Some of those parameters are very technical ones and they are not suitable for the operator to search and query.

Even in the case that industry data formats are used for imagery data, e.g. JPG captured by the digital camera, they can contain useful metadata stored in EXchangeable Image File Format (EXIF). EXIF is an industry data format for digital photography and enables to enrich raw imagery data by metadata such as geoposition, date, time, camera technical parameters such as pixel dimensions, ISO setting (equivalency of sensitivity of a film to light), aperture, white balance, lens used parameters etc.

Those metadata can be extracted and some of them could be transformed into NSIF or NSILI metadata however, none of those STANAGs support EXIF metadata directly.

STANAG 4609 NDIMS uses industry-based container architecture to store video, audio and metadata components of a video sequence. Industry-based data formats and codecs for video encoding are utilized so STANAG 4609 defines parameters for containers, data formats, codecs and its combination. It also defines a set of metadata using key-length-value (KLV) triples to encode camera position, sensor telemetry, video quality etc. KLV metadata should be extracted when video files or streams are published info CSD.

Not all imagery data are stored in the CSD due to the variety of formats and limited support in NSILI. Imagery data are rather stored directly in the file system, especially in the case that standard JPG pictures or MPEG videos are used. Naming conventions and directory hierarchy are typically used to capture basic metadata as date-time and position.

To build data mart on the top of imagery data, metadata defined in STANAG 4545 NSIF, STANAG 4609 NDIMS and STANAG 4559 NSILI, as well as KLV and EXIF metadata, should be considered in data mart design and they should be supported in the extract phase of ETL.

As was demonstrated in the previous paragraphs, the set of image and video metadata is quite wide and fuzzy which is difficult to maintain in relational databases but Big Data can handle it more appropriately.

Another reason to utilize Big Data technologies is managing a huge volume of multimedia files and geolocation metadata processing.

3.8. ISR Products

ISR products are the output of Processing and Exploitation phases of the ISR process. ISR products include sensor data, ISR reports and other documents created and evaluated by the analyst.

There are about 20 ISR product types with a set of common metadata and ISR product-specific ones defined in the NSILI as was showed in NSILI related chapters. The common structure of ISR report or product consists of the identification of the product, name of the report, associated geographic area, date and time of validity, capturing sensor or asset identification, associated task and other detailed parameters.

There are a lot of technology standards used to store ISR products:

- ISR products format defined in AEDP-17, which should be the preferred option.
- APP11 messages stored in a file system or database.
- Intelligence disciplines specific formats, e.g. COMS-ESM or IMIMT data formats introduced in the previous chapter.
- Proprietary textual or binary formats stored as files.
- Proprietary database structures.
- Office documents (Word, Excel, PDF), and others.

It is not possible to create an overall overview of all ISR products due to diversity of used data formats. To solve this issue data mart could be built on the top of multiple ISR products storages. Building this data mart would be challenging due to the fact of a variety of data formats used, similarly as in the case of ISR tasks data mart. Tens or even hundred APP11 message types, individual disciplines data formats, office documents, AEDP-17 XML structures, NVGs, etc. are in place for the ETL phase. Due to the number of data sources and used data formats, implementation of ETL phase might include hundreds of transformation rules and its maintenance might be expensive. In addition, rigid data structures of a data mart might not be flexible enough, more flexible Big Data technologies could be utilized to build data mart and mine the data.

3.9. ISR Operational Picture

ISR Operational Picture is a specific ISR product, a contribution of the ISR community to Common Operational Picture (COP). COP is [34] a single identical display of relevant operational information (e.g. position of own troops and enemy troops, position and status of important infrastructure such as bridges, roads, etc.) shared by more than one command.

NVG is used for encoding and sharing COPs. NVG standard includes a file format for the encoding of battlespace objects into overlays and protocol for the automated exchange of the overlays. NVG is an XML based format, the elements of NVG represent individual objects and its type (point, line, arrow, etc.), label, position, symbol code and other attributes, see the example in Figure 6.

The symbol code is uniquely defining associated graphic symbol from APP-6 symbology. The code is an alphanumerical sequence (e.g. SFGPUCR---EF***) where each character's

subsequence symbolizes [35] unit identity, dimension, status, function, size of the unit and additional information. There are several encoding standards for military symbology, NATO specific, USA specific and symbology for the dismounted soldier.

```
<?xml version="1.0"?>
<nvg
xmlns="https://tide.act.nato.int/schemas/2012/10/nvg"
version="2.0.0">
        <point label="CZE_ISR_BN_HQ" uri="da8eefcc-
1a3f-403d-9d12-e91f60a19693" modifiers="M:53;T:ISR"
symbol="app6c:SFGPUCR---EF***" y="53.1288105381945"
x="18.00605177595984"/>
        <point label="F_EB_coy_A_ISR" uri="22c7cc78-
e3a1-4bb9-81be-f2d4b3cac037" modifiers="M:ISR;T:A"
symbol="app6c:SFGPU------G***" y="50.20523308421297"
x="12.914633886152924"/>
</nvg>
```

Figure 6: Example of NVG data format

The operational picture represented in NVG format is intended for operational purposes, to visualize a tactical situation in overlays on the map. As its primary usage is visualization, qualitative aspects of the COP such as the size of covered area, count, status and types of units, etc., are not automatically processed or evaluated at all. Those quantitative aspects can be handled by data mart where those values could be extracted during the ETL phase. In fact, it is the task of data mining, because it is about digging hidden information from (source) data. Big Data technologies could be also for managing incoming changes of NVGs due to its streaming nature.

3.10. ISR Data Mart

As was discussed in previous paragraphs, ISR data are stored in multiple CSDs, APP11 messages and files, proprietary databases, proprietary data files, NVG files, office documents, images, videos etc. Although CSD is preferred way to store the ISR data, CSD does not cover all types of ISR products and common military data such as NVGs, NFFIs, APP11s, etc. so there is no central point to collect all ISR related data. Moreover, validation rules applied to the metadata are quite strict in the NSILI, which might conflict with wide and fuzzy nature of a sensor data. Finally yet importantly, as NSILI is focused on the interoperability mainly, it does not cover all requirements of a data analyst, especially requirements related to powerful queries, data analyses and data mining.

Building specific data marts for individual intelligence disciplines was proposed to enable operator and data analyst perform real data analyses on the ISR data. Following steps should be done to design ISR data mart: the identification of relevant data sources, the design of data model for data mart, the definition and implementation of ETL processes, the definition of analytical tasks etc. Data warehouse for all ISR data can be also built on the top of those ISR data marts but its design is out of the scope of this article.

3.11. Data Sources

The possible sources of ISR data presented in previous paragraphs include, but are not limited to:

- CSDs.
- NVG files and servers.
- APP11 messages stored in databases or file system.
- Proprietary databases/file systems for ISR task.
- Proprietary databases/file systems for reports.
- Proprietary ESM-COMMS data databases/file systems.
- Imagery files with NSIF or EXIF metadata.
- Video files with NDISM and/or KLV metadata.
- Office documents.
- CESMO network.
- NFFI network.

3.12. The Data Model for Data Marts

The design of a data model for individual data marts would be a straightforward generalization of logical data models introduced in previous chapters. Common metadata attributes should be added and those could be inspired by NSILI common metadata attributes such as date, time, geolocation, originator, type and security to list the most important ones. APP11 messages would be more challenging as there are about 400 messages types. Common header items could be generalized but individual content of the message should be analyzed with respect of corresponding intelligence discipline. Detail design of individual data marts is out of the scope of the article but some ideas are introduced in the practical part of the article.

There is only a very small subset of common metadata applicable to all individual data marts. The basic idea of its identification was introduced in the previous chapter. Generalization of common metadata applicable to a data warehouse level can be done in the next steps. The data warehouse will then consist of a very small number of common attributes (date, time, position) and a huge number of specific attributes for each database entry, there can be tens or even hundreds of attributes for sensor data. The huge number of attributes can lead to some implementation and performance issues. Columns are used in traditional databases to store attributes so there will be hundreds of mostly empty columns. It may not be effective concerning allocated disk space and memory. Big data technologies, especially column databases are designed and intended for this type of tasks.

3.13. The Extract Task

Extract tasks are to be run periodically; typical business intelligence process runs ETL during the night, which might not be sufficient in the military environment especially on the tactical level. The data extraction interval could be hours or online business intelligence should be employed. Big Data technologies, especially streaming databases are designed and intended for this type of tasks.

Data extraction details will be discussed in the next paragraphs:

CSDs

Data extraction from CSD can be implemented on the top of the AEDP-17 and AEDP-19 services. The query service specified in AEDP-17 uses Catalogue Entry Data Set, which can be used as

an input format for the extraction task. In that case, the extraction task will be universal for each standard based CSD. On the other hand, the web service interface is used so the extraction task does not have to be optimal from a performance perspective. Direct access into the underlying database should be much efficient but it will not be universal as an internal implementation of underlying CSD database differs.

The COI services specified in AEDP-19 to manipulate organization, requirement, priority, task, request and GAOI uses XML structures, which can be used as an input format for the extraction task; there are about 150 entities used. The extraction task will be universal for each standard based implementation as the underlying SPS++ service is content-agnostic. Direct database access is not an option in that case as SPS++ entities are not bound with its type and COI services must be used. On the other hand, AEDP-19 defines its own notification mechanism, which allows spreading out data extraction in the time and thus improving performance.

NSILI defines a synchronization mechanism so the server's topology must be taken into account in the extraction task. It must be specified from which CSDs are data extracted to avoid duplicities and unnecessary extraction of the data. Another approach would be to build a special instance of CSD, which will be synchronized with other CSDs as their copy in read-only mode. The extraction task will be run only on this specific instance so it can utilize direct access into the underlying database and thus provide optimal performance.

NVG

Operational Pictures are stored as NVG files on the file system or in the database and those can be read with the extraction task. The web service interface defined to publish and retrieve NVGs can be also utilized for the extraction task.

The extraction task needs to access NVGs from all the mentioned resources. There could be also built read-only NVG server, which will aggregate NVGs from different NVG sources (servers, files, databases) and notify the extraction task in the case of new or changed NVG.

APP11

An exchange mechanism is not defined for APP11 messages so they can be stored in files, databases or messaging queues. The extraction task needs to access all those storages and to handle the complexity of messages as standard defines more than 400 message types which can be stored in two different formats, XML and slash-separated file.

Proprietary databases and file systems

The standard database connector should be used to extract data from proprietary databases. The definition of the extraction task needs to be done for each individual data source by the source database definition. As the database definition is proprietary there might be legal, technology and other obstacles, which might make the extraction task more complicated or even impossible.

The extraction from the file system is a very similar task, significant difference would be that the parser of the file structure needs to be incorporated into the chain of ETL processes. Parsing of XML or CSV files is mostly trivial or similar to database extraction but there are more complex and/or binary data formats used.

Imagery and video databases and file systems

Imagery and video files are a good example of complex binary file formats where the complex parser needs to be employed. Common ETL tools mostly do not support binary formats such as NSIF or EXIF so user-specific code needs to be implemented. It needs to be considered that parsing of binary structures might be time-consuming.

CESMO network

The concept of CESMO is based on platforms, which periodically contribute to the sensor network. All assets connected into CESMO network are real sensor systems so the connection of read-only data consumers is not supported from the operational perspective although it is technically possible. For that reason, CESMO should not be connected directly to the extraction task; rather some ESM proxy system should be used. This proxy ESM system would collect, filter and aggregate raw ESM data and store them in the ESM database, which would be a source for the extraction task.

3.14. Transform and Load Tasks

Data extracted from different data sources need to be transformed into data mart data model (the transform task) and then loaded into data mart itself (the load task). Data from multiple data marts can be also processed into the overall ISR data warehouse. Transform and load tasks are standard functionality of data marts and data warehouses, so its analyses and design are out of the scope of the article, only a few notes are mentioned.

Many parameters for those tasks, such as time of jobs execution, frequency and period of jobs, error recovery policies, executed actions etc. should be allowed to be configured by a user.

The transform task is typically covered by mapping tool, which declarative prescribes mapping of input data structures to output structures. This mapping can be defined directly in the source code or industry standard based mapping such as eXtensible Stylesheet Language (XSLT) can be used. Datamart implementations utilize a combination of those approaches and powerful GUI enabling the visual mapping of data structures is available to users.

The duration of the load task might take a lot of time as huge volumes of the data are being stored. Some optimization can be implemented on the database level as well as time-intensity could be also covered on the organizational level. For example, spreading activities during the whole day can be considered.

3.15. Visualization Tools

The final step of a building data mart is the design of the presentational layer for end-user. Most BI tools provide this kind of functionality as an integral part of the toolset. Set of dashboards, tables, graphs are typically available to the end-user. The goal is to enable to visualize underlying data in user friendly and easy to understand manner. Some practical observations and experiences with selected tools are presented in the practical part of the article.

3.16. Data Marts and the Tactical Network Limits

Limits of the tactical network must be taken into account in the case that data mart implementation and topology is designed at the military environment. The tactical network cannot be expected as reliable and it could be limited by low bandwidth, high latency, communication windows and distributed topology; mash-up networks need to be also considered.

The data mart should be implemented only on that organizational level where it makes sense. A higher level of command would be an appropriate level, while sensor level would be probably inappropriate. Anyway, carefully evaluation of a data flow is needed for the proper design of data mart topology.

3.17. The Big Data Tools to Process ISR Data

Some issues remain even in the case of using BI tools for ISR data processing; those issues cover:

- Extract phase is typically run once a day, in a given time.
- Streaming nature of sensor data.
- Wide and fuzzy data sources.
- Multimedia data processing.
- The volume of the data.
- Office documents and other unstructured data.

The concept of Big Data can be used to deal with those issues. The nature of the ISR data and identified issues of its processing satisfies already mentioned characteristic of Big Data described by 5Vs. Following technologies can be utilized to mitigate listed issues: Not Only SQL databases (NoSQL), Map-Reduce algorithm, In-memory computing and Streaming databases.

3.18. NoSQL Databases

NoSQL databases are a wide set of technologies and approaches solving or improving selected issues of relational databases. One of the primary impulses for NoSQL movement was the beginning of massive distributed databases and applications on the internet. NoSQL databases include key-value databases, document databases, column databases, and graph databases to name the most significant ones.

NoSQL databases are employed by Big Data techniques or products as an underlying layer and enabler thus they are evaluated in the following chapter of the article.

3.19. Map Reduce

Map-Reduce represents a universal way to parallelize the computation. Map-Reduce is a combination of two functions, map and reduce. Input data are split into chunks based on key-value pairs. Those chunks are distributed to mapper components, mapper component execute given task, which is written in the map function. The output of the map function is collected and grouped by the controller according to keys in the key-value pair. Then reducer functionality is executed and finally, all values are combined into the result. [36]

Development with Map-Reduce principle means split code into map, reduce functionality, and define rules for distribution input data into chunks. Coordination and synchronization of execution map and reduce tasks, data distribution and collection of results is managed by appropriate Map-Reduce framework.

Apache Hadoop is an open-source implementation of Map-Reduce principle and it is intended to process, store and analyze a huge amount of distributed and non-structured data. Although is Hadoop the key technology for the Big Data, Hadoop and the Big Data are not equal. [37]

Hadoop ecosystem consists of more connected projects including Hadoop Distributed File System (HDFS), MapReduce framework, NoSQL storages Cassandra and HBase and other components. HDFS [38] is a distributed file system intended to utilize common (utility) hardware so it was designed to be fault-tolerant to HW and network faults. MapReduce framework is a computational layer enabling parallel execution of computation in each node. Results of the computation are aggregated and stored in the storage layer (HDFS). All those components form an ecosystem enabling vertical scaling; each layer is implemented in a specialized product, which can be easily replaced.

3.20. In-Memory Databases

In-memory databases offers very fast processing of data, over 100.000 operations per second on a single machine. The volume of processed data is limited by the size of RAM, another downside is that data can get lost between subsequent disk flushes (e. g. due to server crashes or by losing power supply). Due to those facts, in-memory databases are recommended to be used on the strategic or operational level, not on the tactical level, thus In-memory databases are out of the scope of the article [39].

3.21. Streaming Databases

Streaming databases are intended to process the data which are streamed continuously without the option of being available for long periods; so the query processor can only respond to queries that have been previously registered with it and can produce results for the data stream by looking at a portion of the stream available within a certain window. Another use case for streaming databases is a situation when the query is executed in parallel with incoming data processing so returned data does not contain data just received. [40]

An alternative to the streaming database could be streaming platform such as Apache Kafka. The streaming platform allows publish and subscribe to streams of records, similar to a message queue or enterprise messaging system, store streams of records in a fault-tolerant durable way and process streams of records as they occur. [41]

Some ISR data have streaming nature, especially COMMS-ESM data, TDL etc. Even so, streaming databases are out of the scope of this article.

4. Building Multiple ISR Data Marts

Selected ISR data and its main characteristic, BI principles and Big Data technologies were introduced in the previous chapters. In this chapter will be proposed how to utilize BI approach to build the ISR data mart. The universal ISR data warehouse will not be built, only specific data marts for each individual area with a focus on EW data, ISR products, ISR operational pictures and operational diary as an overview of ISR tasks.

Data marts were built with the utilisation of RapidMiner Studio, an open-source software platform for analytics teams. The

tool is used for data preparation, machine learning, and predictive model deployment. To achieve that functionality, the tool contains about 400 operators for data manipulation enabling load data from multiple sources, transform and process them and visualize the resulting output.

4.1. The Data Mart for Electronic Warfare Data

The physical nature of radio signal propagation and other factors such as reflections, terrain conductivity, time of day, weather, etc. make the processing of electronic warfare data difficult and challenging. Another challenge is that radio transmission or broadcasting on fixed frequency or band is not widely used as it is easy to monitor, jam and deception by the enemy so sophisticated methods such as frequency hopping, direct sequence, chirp, time-hopping etc. are used instead. Frequency hoping will be considered in the following experiments.

Frequency hopping is a method of transmitting radio signals by rapidly switching a carrier frequency among many frequencies using a pseudorandom sequence known to emitter and receiver only. In practice frequency hopping means separation of one (logical) radio transmission into a huge amount of very short (2-300 milliseconds) physical transmissions on different frequencies. The activity detector does not know parameters of the pseudo-random sequence so it is not possible to determine to which transmission the detected activity belongs, if any, or if it is burst, interference or disturbance. If the emitter, detector or both are moving, the detected frequency can be changed due to the Doppler Effect. [42]

Due to aspects of the real environment, an activity detector or direction finder will produce a lot of data including bursts, disturbances, interferences, reflected signals, fake signals etc. Captured data may look randomly on the first view so further analysis of data is needed. Physical aspects need to be taken into account for these analyses so an experienced operator is needed. Alternative methods based on data mart implementation will be evaluated to reduce the need for operator skills, one example could be employing histogram to analyse DFs [43].

Table 4: The sample of generated data.

Frequency	Quality	Level	Detection Start	Duration
64566928	80	88	2020-06-13 13:34:03.642	10
65449040	90	91	2020-06-13 13:34:03.652	10
66397251	3	92	2020-06-13 13:34:03.662	10
67477144	93	5	2020-06-13 13:34:03.672	10
64161124	83	92	2020-06-13 13:34:03.682	10

Following part of the experiment focus on analyses of detected activities. The first step of the building data mart is to define the data model as different sensors use different sets of attributes. The proposed data model for detected activities consists of (carrier) frequency, quality (of detected signal), level (of detected signal), date and time of the detection and duration of the transmission. A random set of activities was generated in a proposed logical data format, the sample of generated data is shown in Table 4.

The generated set of activities simulates multiple frequency hopping transmissions on different bands including bursts and disturbances. Multiple communication scenarios were tested, in the presented example are two frequency hopping broadcasts within one minute frame; one broadcast is in band 63-68 MHz with duration 10 milliseconds of one physical activity and the second one is in the band 66-73 MHz with duration 200 milliseconds of one physical activity.

Both transmissions partially overlap and random disturbances, noise and bursts are added in the band. The different activities/bursts ratio in the sample data were simulated, the number of activities, bursts and its ratio is shown in Table 5. The volume of generated data was limited to about 5000 to 10 000 thousand rows, which is the limit of a free version of RapidMiner Studio.

Those sets were loaded into RapidMiner Studio to be analyzed and visualized. Basic statistical analysis such as average of individual parameters (frequency, duration, quality) are not useful at all but histogram can be used for a basic overview of data distribution for quick identification of two bands with the most significant traffic as demonstrated in Figure 7.

Table 5: Activities/bursts ratio for a different set of data

Activities	Bursts	Sum	Burst/Activity Ratio
3170	32	3202	1,05%
3185	162	3347	5,1%
3148	320	3468	10,2%
3169	1578	4747	49,8%
3163	3243	6406	102,5%

On the other hand, as the number of bursts increases, activities/bursts ratio also increases and as a result, the histogram became much less clear as demonstrated in Figure 8 for 50% activities/bursts ratio.

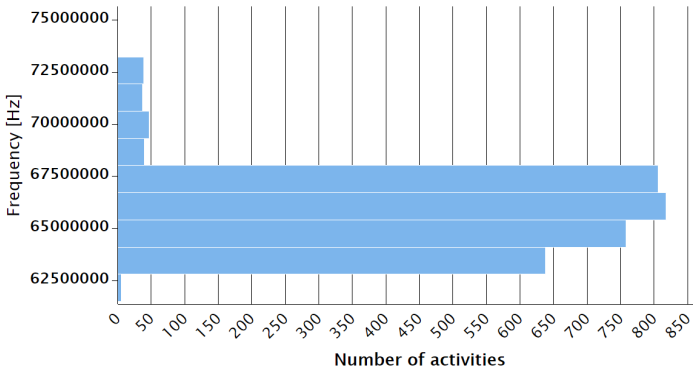


Figure 7: Frequency histogram for sample data with 1% activities/bursts ratio

Distribution of frequencies in the time can be seen on the scatter diagram, examples are shown in Figure 9, 10, and 11 for different activities/bursts ratio. The colour represents the duration of the activity, which can help to analyse data visually.

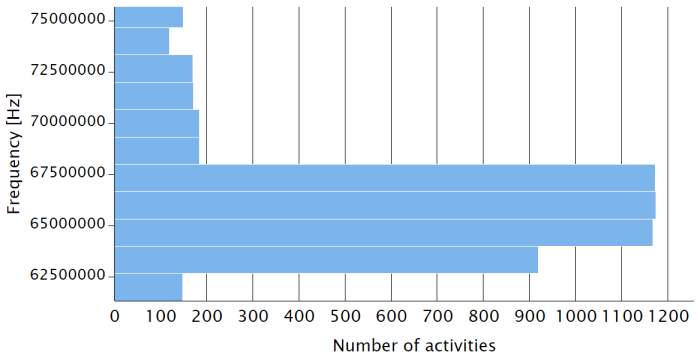


Figure 8 Frequency histogram for sample data with 50% activities/bursts ratio

The data with activities/bursts ratio higher than 10% are quite messy, so it must be filtered in the tool. All activities with quality or level lower than 70 are cut off, the result is in Figure 12 and the resulting picture is much clear to see the relevant activities.

Note: Despite all the data were generated randomly, the output seems to be quite regular and it will be expected that real data are less regular but real measurement data are classified. Similar kind of data analyses and visualization could be done for DF data where azimuth will be used as another dimension for visualization.

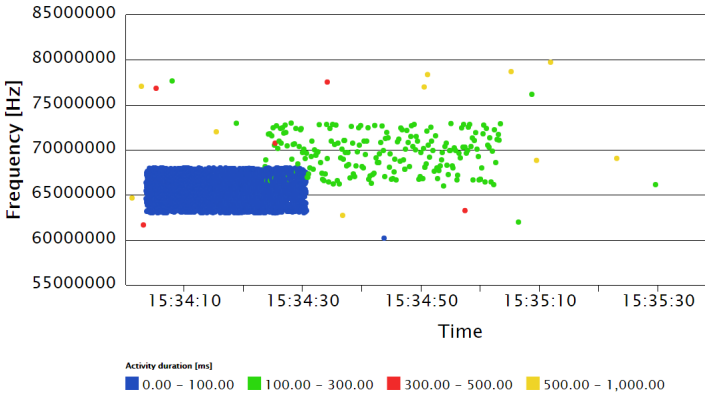


Figure 9: Scatter diagram for the original data with 1% activities/bursts ratio

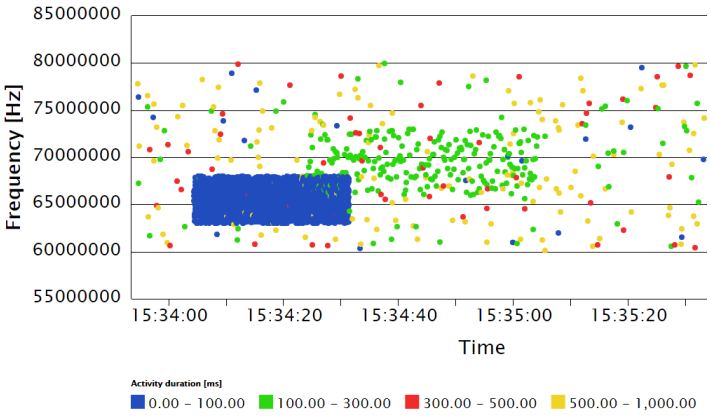


Figure 10: Scatter diagram for the original data with 10% activities/bursts ratio

The main benefit of the presented approach is its easy ability of the graphical presentation and visualization. Although the basic graphical overview of the radio spectrum is an integral part of a sensor controlling SW, its main purpose is to control the sensor, not to provide advanced data analyses.

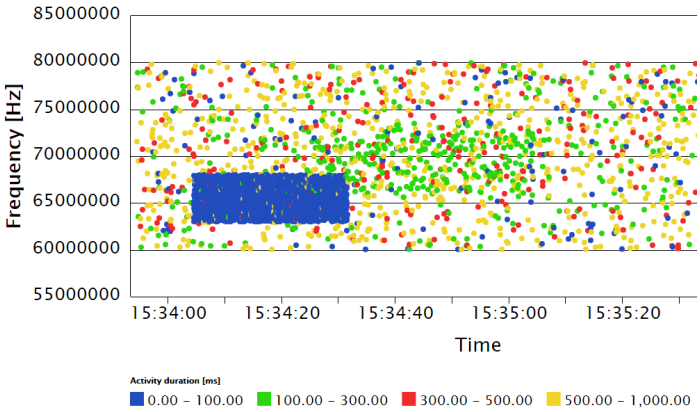


Figure 11: Scatter diagram for the original data with 50% activities/bursts ratio

Moreover, sensor controlling SW is typically tightly coupled with the sensor's data model (and capabilities) so it is hard, or even impossible, to utilize it for processing and visualization of data from different sensors.

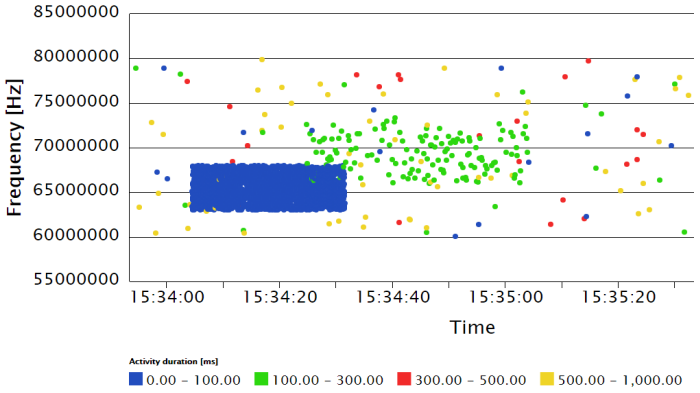


Figure 12: Scatter diagram for the filtered data with 50% activities/bursts ratio

Mentioned disadvantages could be eliminated by the implementation of the data mart and utilisation of its visualisation tools. Those tools enable to load the data from different sensors, to transform them into a unified data model, to clean and filter them and finally to visualize the results. It allows performing much complex data analysis then any sensor controlling SW can with lower effort. Visualisation tools are highly and easily configurable so a different view of the data is easy to achieve. On the other hand, all analyses must be done concerning the physical nature of detected activities and radio waves propagation as well as knowledge of the operational and tactical situation.

4.2. ISR Products Statistics

The common set of metadata for ISR products defined by NSILI includes: ISR product status, publisher, source library, date and time of modification, security (classification, release-ability), ISR product type, mission identifier, language, description, related geographical area (country code or geographic reference box) etc. There was generated a random set of metadata for ISR products in the experiment. Most of the generated ISR products were imagery products, the rest were random products. All mandatory attributes were generated, optional attributes were set optionally. Values of some attributes were set with typos to simulate real data.

The generated set of metadata was load into RapidMiner Studio and processed. In this case, unlike the previous experiment, basic statistics such as the total number of products and the number of different product types could be beneficial, see Figure 13. Note: Number of documents is in logarithmic scale in figure.

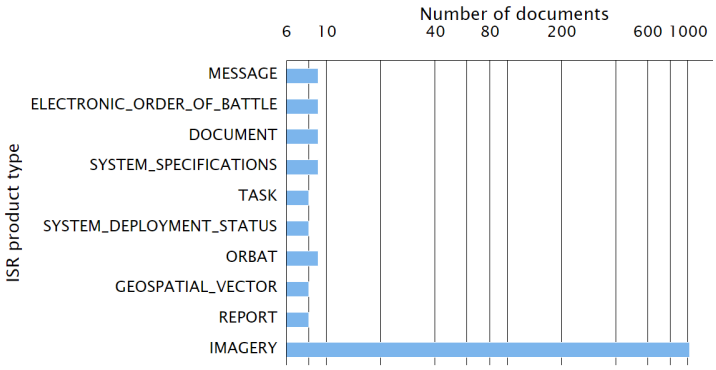


Figure 13: ISR Products types and its number

More advanced statistics were also evaluated, ISR products were grouped by Language and Type, the output is shown in Table 6. It can be easily seen that many imagery products have not set Language attribute, which could be an indicator of the low quality of input data. This observation is an example of how this kind of analyses could easily identify issues in the data and help to improve the process of data collection and processing.

Table 6: The example of grouping ISR products by Language and Type

Language	Type	Number of records
CZE	IMAGERY	511
null	IMAGERY	414
FRA	IMAGERY	37
ITA	IMAGERY	31
ENG	IMAGERY	31
CZE	EOB	7
CZE	DOCUMENT	7
CZE	MESSAGE	6
CZE	REPORT	5
CZE	GEOSPATIAL_VECTOR	5

Queueing three processes generated the report shown in Table 6: data load, aggregation and sort. The chain of data processes is shown in Figure 14, which demonstrate how is data processing managed within RapidMiner Studio.

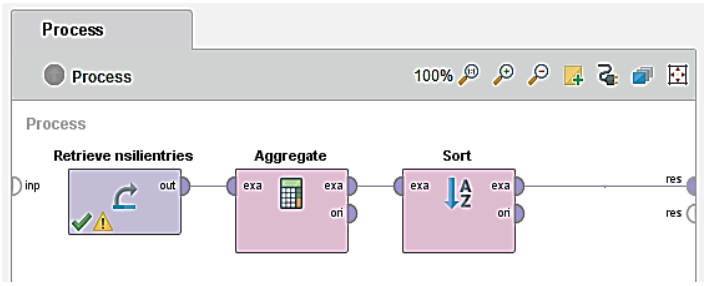


Figure 14: The example of processes design

The data mart built on top of ISR products can be used to generate ISR products statistics. Those statistics and reports to improve collecting tasks and their data quality. The missing data or incorrect values are easy to identify in reports such as in Table 6. The RapidMiner Studio allows drilling down the data to identify the original source of incorrect data. In contrast, NSILI defines the search functionality via its query language only and there is no standard way to see such qualitative and quantitative criteria.

Another benefit would be the improvement of collecting process based on the statistics. The most of ISR products in the shown example were originated from imagery intelligence, based on this simple statistic can commander request data from another intelligence discipline to verify or refute the information by another source.

4.3. The Operational Picture

The operational picture can contain tens or hundreds of battlespace objects such as friendly, enemy and neutral units. In this case, the data model of the data mart would be based on Position, Unit identity, Dimension, Status, Function, Size of the unit, additional modifiers and other attributes. The proposed set of attributes was created as a generalization of the data models of individual military symbology standards. Operational pictures can be load and transform into the data mart, only challenging operation during this step is parsing of military symbol codes. Following analyses, presentation and visualization would be based on the same or similar principles as already presented.

The benefit of BI approach in this scenario is a display of quantitative indicators derived from COP. Those quantitative indicators are not displayed in the COP graphical representation, so their explicit calculation brings new information to the operator. The issue here is that especially the enemy's situation is based on reconnaissance output so it might not correspond to reality.

4.4. The Operational Diary

The operational diary keeps track all activities of the unit and should contain all tasks, requests, and relevant reports. As a side effect, it enables to verify that all RFIs were answered in the given time, the reliability of the sensor can be tracked as well as a response time of the subordinate units, their efficiency etc.

The benefit of BI approach in this scenario is the ability to monitor the performance of units based on automatically tracked tasks and reports. The data model of a data mart would use attributes such as the task and report status, the date and time of creation, the expected time of reporting, the real-time of reporting, collecting platform etc.

The real tactical situation must be taken into account when the operational diary is evaluated. No relevant information might be tracked from tactical reasons or due to the communication infrastructure limitations and issues.

5. Big Data to Process ISR Data

Selected Big Data technologies were evaluated for ISR data processing, especially document databases for metadata storage and Hadoop for ISR data synchronization.

5.1. Document Database as NSILI Metadata Storage

Document databases are an evolution of simple key-value stores, which use the key-value tuples. The key is an identifier of record and the value is content of record that can be both structured

and unstructured. One logical entity can be stored multiple times, each one is optimized for a different type of a query which is very efficient for queries but leads to a rise of the volume of space needed. [44]

Document databases extend this approach with the possibility of structured document storage. Another advantage of a document database is its schema-less nature, which means that structure of stored documents is not validated on the database level so different data structures can be stored within one table.

The experimental CSD implementation with the document database as an underlying layer for AEDP-17 Publish/Query services and AEDP-19 SPS++ services was evaluated and discussed in [45]. The conclusion was that document database usage raised significantly flexibility of the solution as it enables to store any relevant document independently of its content and version. The implementation time and workload were also dramatically lower.

On the other hand, the disadvantage of NoSQL technologies is the fragmentation of its implementation where each particular product is specialized to its specific purpose and/or use case. There is also a lack of standardization so it is very difficult or even impossible to migrate from one NoSQL solution to another.

5.2. Hadoop for ISR Data Synchronization

CSD utilizes two technologies to share and disseminate ISR products and its metadata, CORBA and web service, both of them have some disadvantages.

CORBA as a technology is already obsolete, it is not supported in modern programming languages and it would become obsolete in the future edition of NSILI. Moreover, CORBA protocol is binary so security mechanism based on XML security labelling is not applicable. Real experiences show that CORBA implementation of data synchronization is slow as CORBA is an implementation of Object Request Broker with a lot of overhead functionality from the perspective of CSD.

Web services, in contrast, solve those issues but NSILI does not define all functionality, especially synchronization features are not defined. It is not a problem of web service technology rather of NSILI standard itself.

Although synchronization via Web services is not standardized, it is possible to utilize Web services for synchronization CSDs. Synchronization based on Web services must handle topology of CSDs network, data duplication etc. The use of Hadoop to manage this task is proposed. The basic idea is to export data via Query service to Hadoop File System, then to distribute it to different nodes and finally import the data into CSD via Publish service. Batch export and import need to be implemented which is an implementation detail.

Import or Publish services should handle duplicity. Circle synchronisation is forbidden in NSILI and according to the standard, it is a configuration issue. CSDs synchronisation is a directional point to point and the administrator is responsible for proper configuration and avoiding the circle dependencies. The same rule needs to be satisfied in the case of proposed Hadoop based synchronisation.

Hadoop would be used in the way, that all CSDs has the same data or the same subset of data exported by individual CSDs. One CSD would export for example only IMINT data and do not export EW data, so all CSD in the network will access IMINT data. The data would be separated to those, which should be synchronised with all participants of the CSD network and those data that should not be synchronised. The set of products can be easily configured by the query definition.

The advantage of Hadoop utilisation would be a much easier configuration of synchronisation network and synchronised products. The map-reduce mechanism is not used in that case, only advantages of a distributed file system are utilised. The synchronisation implemented in this way would be resistant to network or hardware failures which is provided by Hadoop itself. This type of the synchronisation over distributed file system would be applicable for other data distribution as well, including EW data etc. but limits of the tactical network need to be taken into account.

5.3. Map-Reduce for ISR Data

Processing of the ISR data can be seen as Map-Reduce task as data collected on sensor level are reduced and distributed to a higher level on command, where are combined with data from other sources, reduced and so on. Although the idea is similar, human interaction is expected, so the process is not fully automated to be understood as Map-Reduce task.

Based on experience, distributed computing is not ideal for the tactical network. The requirement is that systems should work even in the case that connection is lost, some nodes are temporary or permanently not available etc. Distributed computing and especially MapReduce is a specific approach to computation and only selected tasks can benefit from it, especially those which can be easily parallelised. The infrastructure for parallelisation and data distribution is needed which is not a case on the tactical level. Nevertheless, future research of utilisation MapReduce principles is recommended. Big Data principles are applicable for processing the data of individual intelligence disciplines while Map-Reduce principle seems to be not so beneficial on a tactical level due to infrastructure reasons.

6. Conclusion

The ISR process, principles of BI and Big Data technologies were introduced in this paper. Areas in ISR environment where BI and Big Data approach can be beneficial were identified and a way to build ISR data mart was proposed. Experimental data marts were built and loaded by data preparation (data from practice is not possible to get or it is not available due to security reason), but the possibility of data processing, analyses and presentation remain without problems. The utilisation of the document database for ISR products library implementation and Apache Hadoop for its synchronisation was proposed.

Some other promising Big Data technologies were identified and proposed for evaluation in future works. Those technologies include streaming databases and streaming platforms to process sensor data and MapReduce principles.

Proposed utilisation of Big Data technologies is also applicable to other intelligence disciplines, especially OSINT which is based on research on open sources such as newspapers, internet pages, television broadcast etc. would benefit of all aspects of Big Data technologies. Similarly, IMINT analyses would benefit from presented technologies. In any case, the limits of tactical networks must be taken into account when implementing Big Data technologies.

There were demonstrated an advanced BI business analyses that are not used in ISR systems; it is quite original solution. It was also shown that all proposed analyzes have to take into account the real data from military operations, exercises and missions. Regarding this result is awaiting that the domain experts, operators and the data analyst in military environment have sufficient knowledge not only in monitoring operations, but they are able also prepare appropriate results using BI, AI and other relevant technologies.

Acknowledgement

Mr. Novak was participating of development Integrated SoftWare Module for Command, Control, Communications, Computers, Intelligence, Surveillance, Targeting Acquisition and Reconnaissance (ISWM C4ISTAR) project in 2014-2016 focusing on ICT (Information and Communication Technology) support of JISR (Joint ISR) cycle and interoperability of ISR products through STANAG 4559 NSILI compatible implementation.

Mr. Burita presents the results of the research as a part of the project [46] at the Faculty of Military Technologies of the University of Defence, Department of Informatics, and Cyber Operations.

References

[1] A. Novak, L. Burita, "Business Intelligence and ISR Data Processing" in Proceedings of the International Conference on Military Technologies 2019 - (ICMT). Brno, Czech Republic: University of Defence, 1-8, 2019.
[2] OODA Loop [Online]. [Cit. 2020-03-12]. Available at https://www.toolshero.com/decision-making/ooda-loop/
[3] NATO Standardization Office, AJP-2 Allied joint doctrine for intelligence, counter-intelligence and security, 2016.
[4] NATO Standardization Office, AJP-2.7 Allied joint doctrine for intelligence, surveillance and reconnaissance, Edition A, version 1, 2016.
[5] The Global Airborne Intelligence and C2 Community's Annual General Meeting, The Hurlingham Club, London, UK, 2020 [Online]. [Cit. 2020-03-26]. https://www.defenceiq.com/events-airborneisr/
[6] Business Intelligence [Online]. [Cit. 2020-03-15]. Available at https://www.123rf.com/visual/search/9342840
[7] Data Warehousing Solution [Online]. [Cit. 2020-03-16]. Available at http://www.aspiretss.com/services/data-warehousing-solution
[8] OLAP Cubes in SQL Server [Online]. [Cit. 2020-03-17]. Available at https://www.sqlshack.com/olap-cubes-in-sql-server/
[9] CRISP-DM methodology phases [Online]. [Cit. 2020-03-18]. Available at https://www.researchgate.net/figure/CRISP-DM-methodology-phases-Adaptation-of-Chapman-P-3_fig1_318309402
[10] SAS Analytics SW and solutions [Online]. [Cit. 2020-03-20]. Available at https://www.sas.com/en_us/home.html
[11] IBM SPSS software [Online]. [Cit. 2020-03-22]. Available at https://www.ibm.com/analytics/spss-statistics-software
[12] F. L. F. Almeida, "Benefits, Challenges and Tools of Big Data Management", Journal of Systems Integration, **4**(8), 12-20, 2017, DOI: 10.20470/jsi.v8i4.311.
[13] International conference Big Data for Defence, June 23-25, 2020, London, UK.
[14] M.K. Hernandez, C. Howard, R. Livingood, and C. Calongne, „Applications of decision tree analytics on semi-structured north Atlantic tropical cyclone forecasts", International Journal of Sociotechnology and Knowledge Development, **11**(2), 31-531, 2019. DOI: 10.4018/IJSKD.2019040103.
[15] B. Essendorfer, A. Hoffmann, J. Sander, and A. Kuwertz, „Integrating coalition shared data in a system architecture for high level information management", in Proceedings of The International Society for Optical Engineering (SPIE): Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies II, Berlin, Germany, 10802, 2018. DOI: 10.1117/12.2501861.
[16] A. Kuwertz, D. Mühlenberg, J. Sander, and W. Müller, "Applying knowledge-based reasoning for information fusion in intelligence, surveillance, and reconnaissance", Lecture Notes in Electrical Engineering, 501, 119-139, 2018. DOI: 10.1007/978-3-319-90509-9_7.
[17] Z. Kira, A.R. Wagner, C. Kennedy, J. Zutty, and G. Tuell, "STAC: A comprehensive sensor fusion model for scene characterization", in Proceedings of SPIE, Multisensory, Multisource Information Fusion: Architectures, Algorithms, and Applications, 9498, 949804, 2015. DOI: 10.1117/12.2178494.
[18] C. Bowman, G. Haith, A. Steinberg, C. Morefield, and M. Morefield, "Affordable non-traditional source data mining for context assessment to improve distributed fusion system robustness", in Proceedings of SPIE, Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications, 8756, 875603, 2013. DOI: 10.1117/12.2018386.
[19] G.E. Horne, E.A. Bjorkman, and T. Colton, "Putting agent based modeling to work: Results of the 4th international project albert workshop", in Proceedings of SPIE, 4716, 99-107, 2002. DOI: 10.1117/12.474904.
[20] M. Barnell, C. Raymond, C. Capraro, and D. Isereau, "Agile Condor: A scalable high performance embedded computing architecture", in 2015 IEEE High Performance Extreme Computing Conference, Waltham, United States, 2015. DOI: 10.1109/HPEC.2015.7322447.
[21] S. Rogers, J. Culbertson, M. Oxley, H. Clouse, B. Abayowa, J. Patrick, E. Blasch, and J. Trumpfheller, "The QuEST for multi-sensor big data ISR situation understanding", in Proceedings of SPIE: Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VII, Baltimore, United States, 9831, 2016. DOI: 10.1117/12.2229722.
[22] J. Rajamäki, J. Simola, "How to apply privacy by design in OSINT and Big Data analytics", in 18th European Conference on Information Warfare and Security (ECCWS), Coimbra, Portugal, 2019, 364-371.
[23] S. Das, L. Jain, and A. Das, "Deep Learning for Military Image Captioning", in 21st International Conference on Information Fusion, Cambridge, United Kingdom, 8455321, 2165-2171, 2018. DOI: 10.23919/ICIF.2018.8455321.
[24] S.P. Wang, W. Kelly, "IT Professional Conference: Challenges in Information Systems Governance, Gaithersburg, United States, 70293032014, 2014. DOI: 10.1109/ITPRO.2014.7029303.
[25] NATO Standardization Office (NSO), AEDP-17: NATO Standard ISR Library Interface, Edition A Version 1, Brussels, 2016.
[26] APP-11 and ADatP-3. What is APP-11 (NATO Message Catalogue) and ADatP-3 (Allied Data Publication 3)? And what is the difference between the two? [Online]. [Cit. 2019-04-14]. Available from: https://systematic.com/defence/capabilities/c2/interoperability/app-11-and-adatp-3/
[27] ADatP-4733 NATO Vector Graphics Specification, NATO Standardization Office, 2017.
[28] Friendly Force Tracking, US-Army Stand-To [Online]. [Cit. 2019-04-16]. Available from: https://www.army.mil/standto/archive_2016-02-24
[29] AEDP-02 NATO Intelligence, Surveillance, and Reconnaissance (ISR) Interoperability Architecture (NIIA), Edition 1, Volume 1: Architecture Description.
[30] NATO Standardization Office (NSO), AEDP-19: NATO Standard ISR Workflow Architecture, Edition A Version 1, Brussels, 2018.
[31] NATO AJP-2.7 ALLIED JOINT DOCTRINE FOR JOINT INTELLIGENCE, SURVEILLANCE AND RECONNAISSANCE. [Online]. [Cit. 2020-06-01]. Available from: https://standards.globalspec.com/std/10031247/AJP-2.7
[32] Information Requirements Management and Collection Management. [Online]. [Cit. 2020-06-01]. Available from: https://moodle.unob.cz/pluginfile.php/46438/mod_resource/content/1/T3-1%20Information%20Requirements%20Management%20and%20Collection%20Management%20%28lecture%29.pdf
[33] Intelligence Studies: Types of Intelligence Collection. [Online]. [Cit. 2020-06-01]. Available from: https://usnwc.libguides.com/c.php?g=494120&p=3381426´
[34] Common Operational Picture Defined. [Online]. [Cit. 2020-06-01]. Available from: 34https://www.coolfiresolutions.com/blog/common-operational-picture-defined/
[35] NATO joint military symbology APP-6(C), NATO Standardization Office, 2011.
[36] What is Map Reduce? [Online]. [Cit. 2020-05-16]. Available from: https://www.ibm.com/analytics/hadoop/mapreduce
[37] What is Apache Hadoop? [Online]. [Cit. 2020-05-16]. Available from: https://www.ibm.com/analytics/hadoop
[38] HDFS Architecture Guide. Online]. [Cit. 2020-05-16]. Available from: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
[39] In-Memory Database for Time and Mission-Critical Applications. [Online]. [Cit. 2020-05-20]. Available from: https://raima.com/in-memory-dbms-2/?gclid=CjwKCAjwh472BRAGEiwAvHVfGm2FqIjqdfL5ljAyoDD7JL8jeFJu_fIf1KJNjN2H90vWnma3qKwtuxoCgVoQAvD_BwE
[40] What Is a Streaming Database? [Online]. [Cit. 2020-05-20]. Available from: https://hazelcast.com/glossary/streaming-database/
[41] Apache Kafka® is a distributed streaming platform. [Online]. [Cit. 2020-06-01]. Available from: https://kafka.apache.org/intro
[42] Frequency Hopping. [Online]. [Cit. 2020-05-22]. Available from: https://www.sciencedirect.com/topics/physics-and-astronomy/frequency-hopping

[43] „Communication tools PR4G, Retia Company“, Proceedings of the C2 systems in Ministry of Defence Czech Republic, Brno, 2002.
[44] What is a Document Database? [Online]. [Cit. 2020-06-01]. Available from: https://www.mongodb.com/document-databases
[45] A. Novak, “The Comparison of STANAG 4559 NSILI Compliant Tactical Intelligence Database Implementation in Relational Database and NoSQL Database” in 2018 New Trends in Signal Processing Conference, Demanovska Dolina, Slovakia, 2018, 1-7, DOI: 10.23919/NTSP.2018.8524063.
[46] ZRO209, Project of departmental research: Development of systems C4I and cyber security, Brno, University of Defence, 2016-2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 317-323 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

A Hybrid Approach for Intrusion Detection using Integrated K-Means based ANN with PSO Optimization

Jesuretnam Josemila Baby*, James Rose Jeba

Noorul Islam Centre for Higher Education, Department of Computer Applications, Noorul Islam University, Kumaracoil, 629180, India

ARTICLE INFO

Article history:
Received: 04 February, 2020
Accepted: 24 May, 2020
Online: 29 May, 2020

Keywords:
Particle Swarm Optimization
Intrusion Detection System
Artificial Neural Networks

ABSTRACT

Many advances in computer systems and IT infrastructures increases the risks associated with the use of these technologies. Specifically, intrusion into computer systems by unauthorized users is a growing problem and it is very challenging to detect. Intrusion detection technologies are therefore becoming extremely important to improve the overall security of computer systems. In the past decades, most of the intrusion detection systems designed suffer from the problem of high false negative and low efficiency rate. A powerful intrusion detection system (IDS) should be implemented to solve these issues and it is necessary to collect, reduce and analysis the data automatically. The integration of machine learning and artificial intelligence techniques serves this purpose in this paper. A use of particle swarm optimization (PSO) selects the optimal number of clusters and the integration of k-means based artificial neural network (ANN) achieves maximum efficiency when the number of clusters selected optimally. The proposed IDS are t bested with NSL-KD dataset and the experiment result shows the significance of the proposed IDS.

1. Introduction

Due to the advancement of computer and communication technology, the reliance on Internet and worldwide connectivity, damages caused by unexpected intrusions. These crimes related to computer systems have been increased rapidly; a computer system should provide confidentiality, integrity and availability against denial of service; therefore, it is very important that the security mechanisms of systems be designed to prevent unauthorized access to system resources and data [1]. Firewalls are hardware or software systems placed in between two or more computer networks to stop the committed attacks by isolating these networks using the rules and policies determined for them [2]. However, it is very clear that firewalls are not enough to secure a network completely because the attacks committed from outside of the network are stopped whereas inside attacks are not [3]. This is the situation where intrusions detection systems (IDSs) are in charge.

Intrusion detection Systems (IDSs) is a software or device that helps to resist network attacks [4]. The goal of IDS is to have defense wall, which does not allow such types of attacks [5]. It detects unauthorized activities of a computer system or a network. IDS are an active and secure technology there are two categories of intrusion detection system [6].Anomaly detection system creates a database of normal behavior and any deviations from the normal behavior are occurred an alert is triggered regarding the occurrence of intrusions [7]. Misuse Detection system stores the Predefined attack patterns in the database if a similar data and if similar situations occur it is classified as attack. Based on the source of data the intrusion detection system is classified to Host based IDS and Network based IDS [8]. In network-based IDS the individual packet flowing through the network are analyzed [9]. The host-based IDS analyzes the activities on the single computer or host [10]. The main disadvantage of the misuse detection (signature detection) method is that it cannot detect novel attacks and variation of known attacks [11]. To avoid these drawbacks, we proposed anomaly-based detection methods.

Most unsupervised anomaly detection methods are established on two basic assumptions about data [12]. First, the number of normal instances vastly outnumbers that of anomalies. Second, data instances of the same classification (type of attack or normal) should be close to each other in the feature space under some reasonable metrics, and instances of different classifications are far apart [13].Data mining technique is used to find the interesting rules from a large database depending

* Jesuretnam Josemila Baby, Email: josemilakissinger@gmail.com

upon the user defined support and confidence [14]. It will be useful for the decision maker to differentiate between data as useful or irrelevant [15]. Clustering is the unsupervised classification of input items into groups (clusters) without any prior knowledge [16].

Although many kinds of clustering methods, such as Fuzzy C-Means (FCM), K-means, are widely used in intrusion detection, few clustering algorithms guarantee a global optimal solution [17]. Based on this intention we developed a new anomaly intrusion detection mechanism employed with multi-dimensional hierarchical k-means algorithm in this research work in order to overcome all the above issues [18].To integrate K-means and Artificial Neural Network (ANN) technique with optimization to propose an efficient IDS with High True Positive Rate(TPR) and Low False Negative Rate(FNR) [19].

The organization of the paper is as follows: In Section 2, discuss about the related work with emphasis on various methods and frameworks used for intrusion detection. Section 3 covers the Integrated K-Means based ANN with PSO Optimization algorithm. Section 4 presents the experimental results and comparison of the proposed method with other approaches. It is observed that the proposed system. Section 5 gives some conclusions.

2. Related Work

Sharma, Ruby, and Sandeep Chaurasia (2018) [20] proposed an Intrusion Detection System based on the density maximization-based fuzzy c-means clustering (DM-FCC). In that approach, cluster efficiency was improved through a membership matrix generation (MMG) algorithm. Dissimilarity Distance Function (DDF) has been used to compute the distance metric while creating a cluster in proposing IDS. The proposed enhanced fuzzy c-means algorithm has been tested up on ADFA Dataset and the model performs highly appreciable in terms of accuracy, precision, detection rates, and false alarms.

Chung, Yuk Ying, and Noorhaniza Wahid (2012) [21] proposed a new hybrid intrusion detection system by using intelligent dynamics warm based roughest (IDS-RS)for feature selection and simplified swarm optimization for intrusion data classification. IDS-RS was proposed to select the most relevant features that can represent the pattern of the network traffic. In order to improve the performance of SSO classifier, a new weighted local search (WLS) strategy incorporated in SSO was proposed. The purpose of this new local search strategy was to discover the better solution from the neighborhood of the current solution produced by SSO.

Thaseen, IkramSumaiya, and Cherukuri Aswani Kumar (2017) [22] proposed an intrusion detection model using chi-square feature selection and multi class support vector machine (SVM). A parameter tuning technique was adopted for optimization of Radial Basis Function kernel parameter namely gamma represented by 'ϒ' and over fitting constant 'C'. These are the two important parameters required for the SVM model. The main idea behind this model was to construct a multi class SVM which has not been adopted for IDS so far to decrease the training and testing time and increase the individual classification accuracy of the network attacks. The investigational results on NSL-KDD dataset which was an enhanced version of KDDCup 1999 dataset shows that the proposed approach results in a better detection rate and reduced false alarm rate.

Çavuşoğlu, Ünal (2019) [23] developed a hybrid and layered Intrusion Detection System (IDS) that uses a combination of different machine learning and feature selection techniques to provide high performance intrusion detection in different attack types. In the developed system, initially, data preprocessing was performed on the NSL-KDD dataset, then by using different feature selection algorithms, the size of the dataset was reduced. Two new approaches have been proposed for feature selection operation. The layered architecture was created by determining appropriate machine learning algorithms according to attack type. Performance tests such as accuracy, DR, TP Rate, FP Rate, F-Measure, MCC and time of the proposed system are performed on the NSL-KDD dataset.

Aswani, Reema et al., (20-17) [24] introduced a hybrid artificial bee colony approach integrated with k-nearest neighbors to identify and segregate buzz in Twitter. A set of metrics comprising of created discussions, increase in authors, attention level, burstiness level, contribution sparseness, author interaction, author count and average length of discussions are used to model the buzz. The proposed approach considers the buzz discussions as outliers deviating from the normal discussions and identifies the same using the proposed hybrid bio inspired approach. Findings may be useful in domains like e-commerce, digital and influencer marketing to explore the factors that might create buzz along with the difference between the impact of buzz and normal discussions on the consumers.

3. Intrusion Detection Model

Over the last two decades, computer threats and cybercrimes have proliferated at the disadvantage of the general public, and newer threats are introduced each day that compromise the integrity, validity and confidentiality of data. Malicious activities in the internet are also known as intrusion.

Intrusion detection system (IDS) is software and hardware deployed to carry out the process of detecting unauthorized use of, or attack upon, a computer or a telecommunications network which is supposed to bridge the gaps in firewall and anti-viruses. An IDS provides monitoring [25] and analysis of user and system activity, can audit system configuration and vulnerabilities, assess the integrity of critical system and data files, provide statistical analysis of activity patterns based on the matching with known attacks, analyze abnormal activity, and operate system audit.

One advantage of the IDS is its ability to document the intrusion or threat to an organization, thereby providing bases for informing the public regarding the latest attack patterns through system logs. The proposed IDS model is the integration of K means-based ANN and PSO. Initially the features of the intruded networks are extracted from a benchmark dataset and it was trained to the proposed classifier. But it is difficult to find the number of clusters and the detection accuracy will be maximum if the number of clusters equal to the number of data types in the

dataset. This problem is formulated as a multi-objective function in PSO which optimally finds the cluster numbers that maximizes the detection accuracy and lower the false negatives.

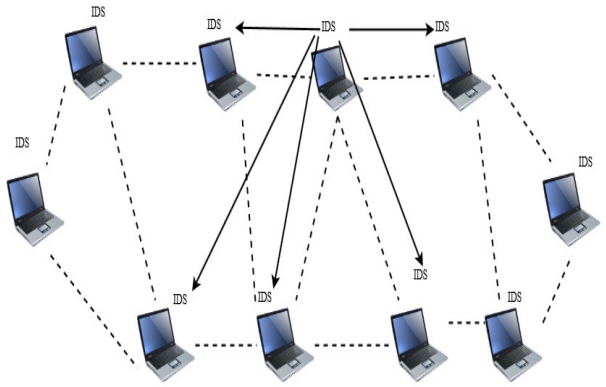


Figure 1: The IDS Architecture for Wireless Ad-Hoc Network

4. Proposed IDS Model

The study integrates artificial intelligence and machine-learning techniques with k-means data mining algorithm to develop an IDS [26] model with higher efficiency and lower false negatives. A common problem shared by current IDS is the high false positives and low detection rate. In the proposed work, we integrate the advantage of Artificial intelligence and machine learning techniques to overcome this issue. The proposed IDS model uses K-means algorithm based Artificial Neural Network (ANN) [27] integrated with Particle swarm optimization (PSO) algorithm [28] to increase the efficiency rate.

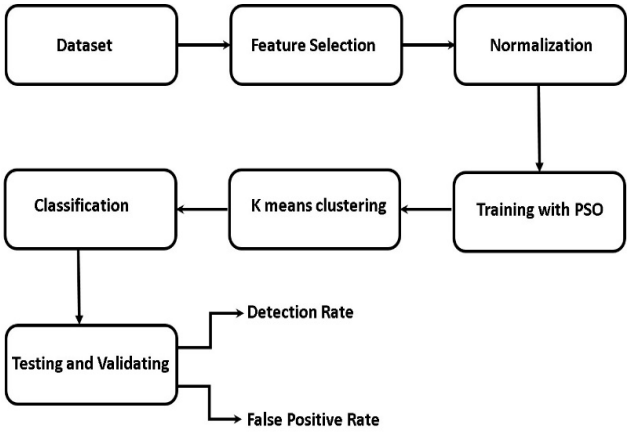


Figure 2: Proposed Block Diagram

4.1. Pre-processing

Pre-processing involves cleaning the data of inconsistencies and/or noise and combining or removing redundant entries. Pre-processing also involves converting the attributes of the dataset into numeric data and saving in a format readable because k-means works only on numerical data.

4.2. Feature Extraction [29]

After pre-processing the initial data, the useful information's or features from each data are extracted. We employed the simple features (attributes) that are extracted from the header's area of the selected network packets. These intrinsic features are available in many networks, for example, the duration (length of the connection), source host, destination host, source interface, and destination interface. We also used three features in each 2 seconds time interval:

- Total number of packets sent from and to the given interface in the considered time interval,
- Total number of bytes sent from and to the given interface in the considered time interval,
- Number of different source- destination pairs matching the given hostname-interface that are observed in the considered time interval.
- The number of packets and bytes allows to detect anomalies in traffic volume, and the third features allows detecting the network and the interface scans as well as the distributed attacks, which both result in an increased number of source-destination pairs.
- An efficient classifier further utilizes the extracted features in order to detect Intrusion present in the network.

4.3. K-Means Clustering Algorithm

Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other. Until now, the clustering algorithms can be categorized into four main groups partitioning algorithm, hierarchical algorithm, density-based algorithm and grid-based algorithm. Partitioning algorithms construct a partition of a database of N objects into a set of K clusters. Usually they start with an initial partition and then use an iterative control strategy to optimize an objective function. K-means represents a type of useful clustering techniques by competitive learning, which is also proved promising techniques in intrusion detection. K-Means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The objective is to classify a given data set into a certain number of clusters (assume initial clusters) fixed a priori.

Algorithm 1 The pseudo code for the adapted K-Means clustering

1. Choose random k data points as the initial Cluster Centroids.
2. Repeat
3. For each data point x from D
4. Then compute the distance of x from each cluster mean (centroid)
5. Assign x to the nearest cluster.
6. End for loop.
7. Again compute the mean for current cluster collections.
8. Until reaching stable cluster
9. Use these centroids for normal and anomaly traffic.
10. Calculate the distance of centroid from normal and anomaly centroid points.
11. If distance$(d, Dm) >= 5$
12. Then anomaly found, exit 2.
13. Else 3.
14. d is a normal and it is not an Intrusion;

K-means clustering module can be summarized as follows:

$$K = \sum_{m=1}^{x} \sum_{n=1}^{y} \left\| d_n^{(m)} - c_m \right\| \quad (1)$$

$\left\| d_n^{(m)} - c_m \right\|$is a chosen distance measure between a data point and the cluster centroid, is an indicator of the distance of the n data points from their respective cluster Centroids. In order to apply the K-means algorithm to intrusion detection system, we design and realize the K-means algorithm analyze module, the graph shows the process flow:

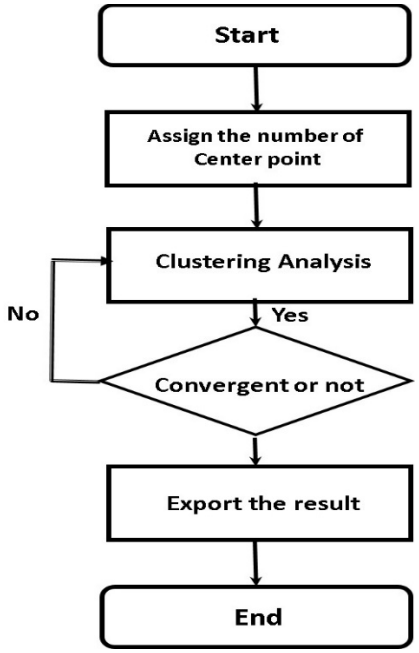


Figure 3: Working flow for K- means Algorithm

4.4. Review of Artificial Neural Network [30]

Feed forward neural network training is usually carried out using the called back propagation algorithm. Training the network with back propagation algorithm results in a non-linear mapping between the input and output variables. Thus, given the input/output pairs, the network can have its weights adjusted by the back propagation algorithm to capture the non-linear relationship. After training, the networks with fixed weights can provide the output for the training the network is based on the minimization of an energy function representing the instantaneous error.

ANN analysis was carried out using software Easy NN version 8.01. The network software uses back propagation algorithm and logistic function as activation function. The ANN used has three layers: an input layer that consists of five nodes (variables), one hidden layer consisting of five hidden nodes and an output layer that has one output node. To train an ANN model, a set of data containing input nodes and output nodes are fed. Once the training is over, ANN is capable of predicting the output when any input similar to the pattern that it has learned is fed. The ANN is tested for the remaining set of experimental data. The learning rate and momentum value of the network is set to optimize with a targeted error value of 0.05.

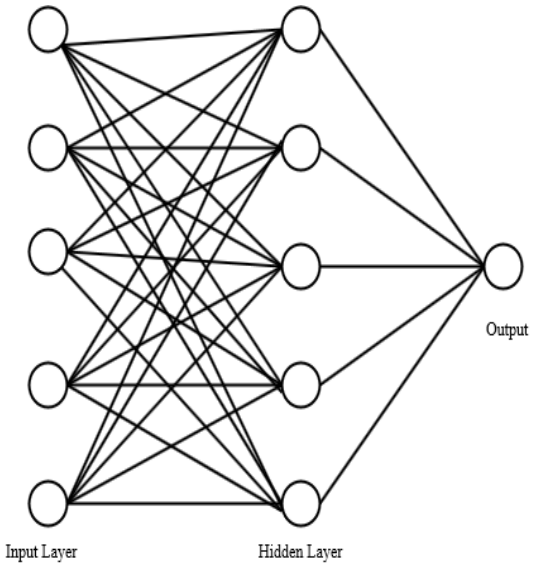


Figure 4: Schematic diagram of artificial neural network

4.5. K-Means based ANN-PSO

For implementation of ANN algorithm on the dataset, training data is divided into several subsets using k-means clustering technique. Subsequently, it trains different ANN using different subsets. Then, it determines membership grades of all these subsets and combines them using a new ANN to get final results. The whole framework of K-Means ANN is illustrated in Fig. 1. As typical machine learning framework, K-Means ANN incorporates both the training phase and testing phase.

Algorithm 2 The pseudo code for the adapted K-Means ANN

1. Start
2. Get the input features and Initialize the neural network parameters.
3. Define the relationships between input and output.
4. Divide the input for testing and training.
5. Train ANN using the training set.
6. Execute the Neural Network for the testing set.
7. Set j=1;
8. While(j<x)
9. Cluster the ANN outputs into K-clusters.
10. Calculate the weight of each cluster.
11. Calculate the error between the target output and ANN estimated output.
12. Increase j by 1.
13. Calculate false negatives.
14. End

The results of K-means ANN is optimized using PSO. In PSO, each particle is a point of N-dimensional solution space and has a speed$(N - dimensional\ vector)$. Different particle has individual fitness associated with objective function. Each particle adjusts their flight path according to its flying experience and flying experience of group and move closer to optimal point. The position of i-particle is denoted as,$X_i = (x_{i1}, x_{i2} \ldots . x_{iN})$ Flight speed is denoted as: $V_i = (v_{i1}, v_{i2} \ldots . v_{iN})$the best position which $i$particle passed is denoted as$P_{ibest} = (p_{i1}, p_{i2} \ldots . p_{iN})$the

groups' best position which it can get is denoted as $G_{best} = (g_1, g_2 \ldots . g_N)$. Particle Swarm has two primary operators: Velocity update and Position update. During each generation each particle is accelerated toward the particles previous best position and the global best position. At each iteration a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position, and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle in the search space. This process is then iterated a set number of times or until a minimum error is achieved. In each step, according to PSO algorithm formula, which is proposed by Kenddy, particles update their velocity and position according to the following formula:

$$V_i(t) = wV_i(t-1) + C_1 r_1\big(P_i - X_i(t-1)\big) + C_2 r_2\big(G - X_i(t-1)\big) \quad (2)$$

$$X_i(t) = X_i(t-1) + V_i(t-1) \quad (3)$$

C_1 And C_2 denote accelerating factor. According to the experience of PSO algorithm, they are usually set$C_1 - C_2 = 2$. r_1 and r_2 are two random number between zero and one, w is called inertia weight. Researchers often use a constant V max to limit the speed of particles and improve search results. W plays a role which balance global search ability and local search ability. It is essential for the success of the algorithm. Shi and Eberhart, study on the effect of the W for optimize performance. They found that the larger the W is, the more easily escape from local minima, and the smaller the W is, the more favorably algorithm converges. Then they present a method, which makes inertia weight decrease linearly according to number of iterations.

In the beginning algorithm uses large inertia weight, it has a strong overall search capability. The later smaller inertia weight is used, and local search ability is improved. W is calculated as follows:

$$w = (w_1 - w_2) * \frac{Max_i - 1}{Max_i} + w_2 \quad (4)$$

w_1 and w_2 are the initial value and final value of inertia weight. Max_i and i are the maximum number of iterations and the current number of iterations for the algorithm, W reduces from 0.9 to 0.4 with the conduct of iteration. The objective function for PSO is set as

$$f(x) = \max\left(\frac{Detection\ Accuracy}{False\ Negatives}\right) + \max\,(1/Time) \quad (5)$$

The PSO algorithm proceeds as follows:

Algorithm 3 The pseudo code for the adapted PSO

1. Begin
2. Initialize the total number, velocity and position of the particles.
3. Calculate pbest and gbest.
4. Calculate the objective function.
5. Update the velocity and position of the particle.
6. Update pbest and gbest .
7. Repeat the steps until the termination condition reached.
8. End

5. Performance Analysis

NSL-KDD dataset is an improved version of the popular KDD Cup'99 dataset. It solves some inherent problems of the KDD'99 dataset (Tavallaee et al. 2009; McHugh 2000).Due to lack of public datasets for network-based IDS, the current version of NSL-KDD may be applied as an effective benchmark dataset for this work. Furthermore, major improvements are carried out on KDD'99 to obtain NSL-KDD and this is more advantageous over KDD'99. The number of instances in the NSL-KDD train and test sets is reasonable, which makes experimentation bias less by running experiments on whole dataset instead of running on randomly selected short portion of KDD'99 dataset. It is free of redundant records in train and test dataset, so the classifiers will not be biased towards repeated instances in the dataset.

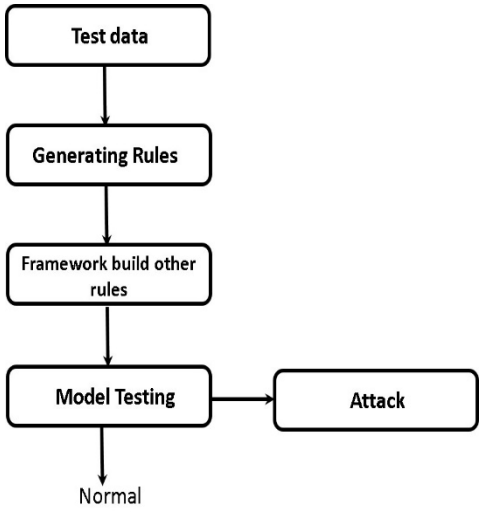


Figure 5: Implementation Setup Model

The NSL-KDD dataset contains two different files for training and testing, and hence, there is no overhead of dividing the dataset into training and testing which also makes a slight contribution towards performance evaluation of the learning techniques.

Table 1: Datasets with Attributes

Dataset	No. of records	No of attributes
KDDTrain+.txt	125,973	42
KDDTest+.txt	22,544	42

Table1shows the properties of the NSL-KDD train and test datasets (both are obtained as .txt files). It should be noted that there are no missing values in any attribute, and number of attributes in the table includes a class attribute. One approach during the training phase refers to processing of all the patterns or instances present in the KDDTrain+.txt dataset. However, size of testing set is 15% of size of whole dataset. Since these datasets are large and computation in ordinary machines might take too much time or sometimes might not support the memory requirements, randomly chosen 10% of the training dataset was used during training. The random selection of 10% of the training dataset was repeated for different executions so that there would be comparatively less chance of repeating the training data in different simulations. There are different types of attacks in the

dataset. However, in this work, it is considered as a two-class problem where patterns may belong to either 'normal' or 'anomaly' class.

5.1 Normalization of dataset

In the NSL-KDD datasets, the values for each attribute are often not distributed uniformly. It is wise to maintain a uniform distribution of each input attributes in the dataset before processing in the neural network. Hence, to ensure that the input values were compatible despite significant differences in their values, the dataset is normalized with respect to each input value where$d_i$the original value is;$d_{max}$and$d_{min}$are the maximum and minimum value, respectively, in the input attribute from whichdiis obtained. Then, normalized value of$d_i$is denoted as$d_i^{'}$. However, it was seen that the normalized dataset some-times contained majority of zeros and in such cases, it was preferred to use the in-built data normalization function of MATLAB called *mapminmax, which* normalizes data in the range$[-1,1]$.

$$d_i^{'} = \frac{d_i - d_{min}}{d_{max} - d_{min}} \quad (6)$$

5.1.1 Testing and Validation

For our experiments, we are using NSL-KDD dataset. NSL-KDD contains 42 fields as an attribute. In our algorithm, we have taken selected features. The performances of each method are measured according to,

- Accuracy
- False Positive Rate

A false positive occurs when the system classifies an action as anomalous (a possible intrusion) when it is a legitimate action. Although this type of error may not be completely eliminated, a good system should minimize its occurrence to provide useful information to the users. A false-negative occurs when an actual intrusive action has occurred but the system allows it to pass as non-intrusive behavior. While the true-positives (TP) and true-negatives (TN) are correct classifications. Recall Rate measures the proportion of actual positives that are correctly identified.

a. Accuracy

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Where FN is False Negative, TN is True Negative, TP is True Positive, and FP is False Positive

b. False Alarm Rate

The false positive rate is the number of normal connections that are misclassified as attacks divided by the number of normal connections in the data set.

$$False\ alarm = \frac{FP}{FP+TN} \quad (8)$$

5.1.2 Performance Comparison

The performance of the proposed IDS model is compared with the performance of the K-means algorithm and K-means based ANN algorithm in this section in terms of accuracy and false alarm rate.

Figure 6 shows the accuracy comparison between proposed IDS model with K-means algorithm and K-means based ANN algorithm.

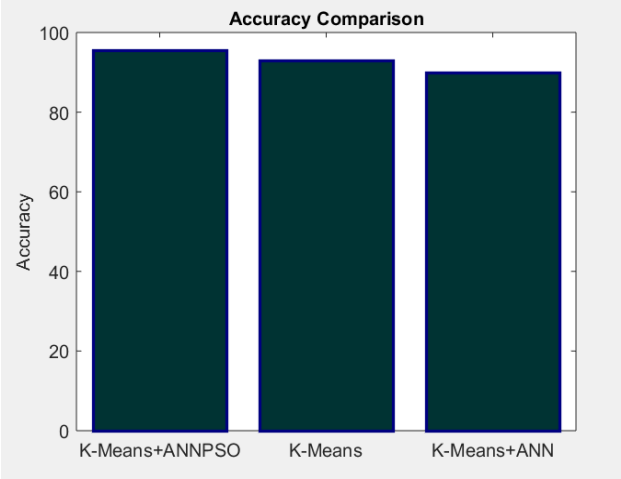


Figure 6: Accuracy comparison between existing techniques

Figure 6 shows the accuracy of proposed IDS model is 88.2321% and the K-means algorithm attains 85% accuracy and K-means based ANN algorithm attains 83% accuracy with NSL-KDD dataset. This shows the significance of the proposed IDS model than the existing IDS models.

Figure 7 shows the false alarm rate comparison between proposed IDS model with K-means algorithm and K-means based ANN algorithm.

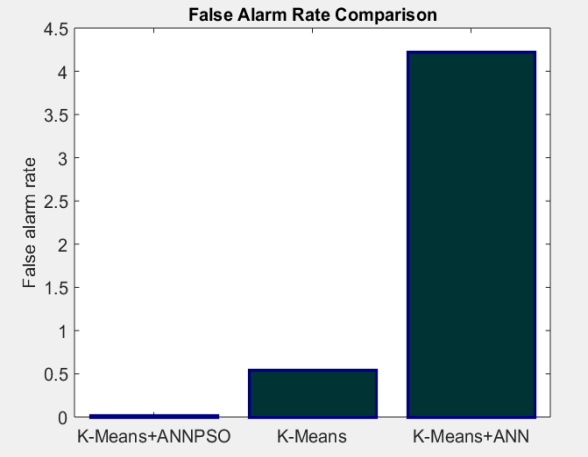


Figure 7: False Alarm Rate Comparison between existing techniques

Figure 7 shows the false alarm rate of proposed IDS model is ranging between 0-0.5 and the K-means algorithm attains false alarm rate of 0.5 and K-means based ANN algorithm attains maximum false alarm rate above 4 with NSL-KDD dataset. The significant reduce in false alarm rate shows the proposed IDS model is efficient.

6. Conclusion

Intrusion Detection is a process of detecting Intrusion in a computer system in order to increase the security. Intrusion detection is an area in which more and more sensitive data are stored and processed in networked system. After reading several research works, we come with several advantage and disadvantage. In this paper, a novel method for Intrusion detection was proposed. The proposed Intrusion detection system combines the advantages of machine learning and artificial intelligence to overcome the general issues present in the intrusion detection systems. Using the PSO algorithm, a multi-objective problem is formulated to find the optimal clusters and the issue is solved. The proposed system is tested in NSL-KD dataset and it achieved maximum detection accuracy of 88%. In addition to this, it is observed that the false negative rate of the proposed IDS is significantly reduced when compared with the results obtained with K-means algorithm and K-means based ANN algorithm.

References

[1] Sobh, Tarek S. "Wired and wireless intrusion detection system: Classifications, good characteristics and state-of-the-art." Computer Standards & Interfaces 28, no. 6 (2006): 670-694.

[2] Aydın, M. Ali, A. Halim Zaim, and K. GökhanCeylan. "A hybrid intrusion detection system design for computer network security." Computers & Electrical Engineering 35, no. 3 (2009): 517-526.

[3] Raymond, Jean-François. "Traffic analysis: Protocols, attacks, design issues, and open problems." In Designing Privacy Enhancing Technologies, pp. 10-29. Springer, Berlin, Heidelberg, 2001.

[4] Tsai, Chih-Fong, Yu-Feng Hsu, Chia-Ying Lin, and Wei-Yang Lin. "Intrusion detection by machine learning: A review." expert systems with applications 36, no. 10 (2009): 11994-12000.

[5] Debar, Hervé, Marc Dacier, and Andreas Wespi. "A revised taxonomy for intrusion-detection systems." In Annales des télécommunications, vol. 55, no. 7-8, pp. 361-378. Springer-Verlag, 2000.

[6] Liao, Hung-Jen, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. "Intrusion detection system: A comprehensive review." Journal of Network and Computer Applications 36, no. 1 (2013): 16-24.

[7] Patcha, Animesh, and Jung-Min Park. "An overview of anomaly detection techniques: Existing solutions and latest technological trends." Computer networks 51, no. 12 (2007): 3448-3470.

[8] Depren, Ozgur, Murat Topallar, EminAnarim, and M. Kemal Ciliz. "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks." Expert systems with Applications 29, no. 4 (2005): 713-722.

[9] Labib, Khaled, and V. Rao Vemuri. "An application of principal component analysis to the detection and visualization of computer network attacks." In Annales des télécommunications, vol. 61, no. 1-2, pp. 218-234. Springer-Verlag, 2006.

[10] Verwoerd, Theuns, and Ray Hunt. "Intrusion detection techniques and approaches." Computer communications 25, no. 15 (2002): 1356-1365.

[11] Syarif, Iwan, Adam Prugel-Bennett, and Gary Wills. "Unsupervised clustering approach for network anomaly detection." In International conference on networked digital technologies, pp. 135-145. Springer, Berlin, Heidelberg, 2012.

[12] Jiang, ShengYi, Xiaoyu Song, Hui Wang, Jian-Jun Han, and Qing-Hua Li. "A clustering-based method for unsupervised intrusion detections." Pattern Recognition Letters 27, no. 7 (2006): 802-810.

[13] Fan, Cheng, Fu Xiao, Yang Zhao, and Jiayuan Wang. "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data." Applied energy 211 (2018): 1123-1135.

[14] Aggarwal, Charu C., and S. Yu Philip. "Data mining techniques for associations, clustering and classification." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 13-23. Springer, Berlin, Heidelberg, 1999.

[15] Stewart, Theo J. "A critical survey on the status of multiple criteria decision making theory and practice." Omega 20, no. 5-6 (1992): 569-586.

[16] Abraham, Ajith, Swagatam Das, and Sandip Roy. "Swarm intelligence algorithms for data clustering." In Soft computing for knowledge discovery and data mining, pp. 279-313. Springer, Boston, MA, 2008.

[17] Zhang, Zhongxing, and Baoping Gu. "Intrusion detection network based on fuzzy c-means and particle swarm optimization." In Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation, pp. 111-119. Atlantis Press, Paris, 2016.

[18] Patel, Ahmed, Mona Taghavi, KavehBakhtiyari, and Joaquim CelestinoJúNior. "An intrusion detection and prevention system in cloud computing: A systematic review." Journal of network and computer applications 36, no. 1 (2013): 25-41.

[19] Folino, Gianluigi, and Pietro Sabatino. "Ensemble based collaborative and distributed intrusion detection systems: A survey." Journal of Network and Computer Applications 66 (2016): 1-16.

[20] Sharma, Ruby, and Sandeep Chaurasia. "An enhanced approach to fuzzy C-means clustering for anomaly detection." In Proceedings of First International Conference on Smart System, Innovations and Computing, pp. 623-636. Springer, Singapore, 2018.

[21] Chung, Yuk Ying, and Noorhaniza Wahid. "A hybrid network intrusion detection system using simplified swarm optimization (SSO)." Applied Soft Computing 12, no. 9 (2012): 3014-3022.

[22] Thaseen, IkramSumaiya, and CherukuriAswani Kumar. "Intrusion detection model using fusion of chi-square feature selection and multi class SVM." Journal of King Saud University-Computer and Information Sciences 29, no. 4 (2017): 462-472.

[23] Çavuşoğlu, Ünal, Shirin Panahi, AkifAkgül, SajadJafari, and SezginKaçar. "A new chaotic system with hidden attractor and its engineering applications: analog circuit realization and image encryption." Analog Integrated Circuits and Signal Processing 98, no. 1 (2019): 85-99.

[24] Aswani, Reema, S. P. Ghrera, Arpan Kumar Kar, and Satish Chandra. "Identifying buzz in social media: a hybrid approach using artificial bee colony and k-nearest neighbors for outlier detection." Social Network Analysis and Mining 7, no. 1 (2017): 38.

[25] Duque, Solane, and Mohd Nizam bin Omar. "Using data mining algorithms for developing a model for intrusion detection system (IDS)." Procedia Computer Science 61 (2015): 46-51.

[26] Benmouiza, Khalil, and Ali Cheknane. "Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models." Energy Conversion and Management 75 (2013): 561-569.

[27] Abd-El-Wahed, W. F., A. A. Mousa, and M. A. El-Shorbagy. "Integrating particle swarm optimization with genetic algorithms for solving nonlinear optimization problems." Journal of Computational and Applied Mathematics 235, no. 5 (2011): 1446-1453.

[28] Hart, Christopher G., and Nickolas Vlahopoulos. "An integrated multidisciplinary particle swarm optimization approach to conceptual ship design." Structural and Multidisciplinary Optimization 41, no. 3 (2010): 481-494.

[29] Lunt, Teresa F. "A survey of intrusion detection techniques." Computers & Security 12, no. 4 (1993): 405-418.

[30] Manohar, Balaraman, and SoundarDivakar. "An artificial neural network analysis of porcine pancreas lipase catalysed esterification of anthranilic acid with methanol." Process Biochemistry 40, no. 10 (2005): 3372-3376.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 323-330 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Logistics Solutions in the Preparation Phase for the Appearance of Disasters

Erika Barojas-Payán*,[1], Diana Sánchez-Partida[2], Miguel-Josué Heredia-Roldan[3], Victorino Juárez-Rivera[1], Jesús Medina-Cervantes[1]

[1]*Faculty of Engineering, Veracruz University, Xalapa 94466, Mexico*

[2]*Department of Logistics and Supply Chain Management, Universidad Popular Autónoma del Estado de Puebla, Puebla 72410, Mexico*

[3]*Gustavo A. Madero Technological Institute, Technological National of Mexico, Mexico 07470, Mexico*

ARTICLE INFO

Article history:
Received: 14 April, 2020
Accepted: 03 June, 2020
Online: 28 July, 2020

Keywords:
Warehouse location
Inventory levels
Routing Problem

ABSTRACT

Natural disasters have caused not only economic but also human losses. These events bring with them, among other things, deficiencies in the supply of food, clothing, health and cleaning products, to name a few. This situation makes it imperative to locate facilities that can supply the needs, as mentioned earlier, to the victims in the shortest possible time. This document presents the evaluation of a logistic model of the literature whose foundations are: a) the classic p-median problem for the location of a pre-positioning warehouse; b) an extension of the (q-R) model for calculating inventories of different products, according to different types of demand, and c) the problem of multiple vehicles routing with the capacity to establish delivery routes from warehouses to the affected municipalities. This model is evaluated with 90 municipalities belonging to Veracruz, Mex. The results show that the location of the warehouse falls in the municipality of Fortín de las Flores, the inventory levels for five demands and four product kits, and product delivery routes, which are a total of 12, with which a favorable cost minimization is obtained.

1. Introduction

This document is an extension of an original work presented at the "*2019 IEEE International Conference on Engineering Veracruz (ICEV)*" [1], which derives from a global problem, such as the impact of *natural phenomena of hydrometeorological* type on the world. The Office of the United Nations Organization (UN) for Humanitarian Affairs (2020), mentions that from the year 2000 to date, 152 million Latin Americans and Caribbean people have been affected by 1205 disasters, including floods, hurricanes and storms, earthquakes, avalanches, droughts, fires, extreme temperatures and volcanic events [2]. It should be mentioned that natural phenomena do not affect society in the same way; the so-called hydrometeorological phenomena such as heavy rains and floods have a higher frequency compared to other types of natural phenomena and tend to affect the same region in relatively frequent periods. In the last 20 years, these were responsible for causing 90% of economic losses worldwide.

In Mexico, the effects of hydrometeorological phenomena appear depending on the distribution of the population throughout its territory. In a broad sense, northern Mexico is characterized by low levels of precipitation, frequent droughts, and recurring heat waves. In contrast, the south of the country has the highest records of recurrent precipitation and flooding, in addition to being the region most affected by the effects of hurricanes and other tropical storms. During 2016 these were responsible for causing 86% of the damage in Mexico [3, 4].

The National Civil Protection System (SINAPROC, for its acronym in Spanish), was created to respond to this and other types of natural phenomena. The said system, together with various federal and state agencies, organizations, and entities, creates the necessary instruments and mechanisms in order to carry out coordinated actions in the area of civil protection [5].

Within SINAPROC, there is the National Fund for Natural Disasters (FONDEN, for its acronym in Spanish), which is an inter-institutional mechanism that aims to execute actions, authorize and apply resources to mitigate the effects of a disturbing natural phenomenon, in regarding of the SINAPROC. For access to that fund, the Federal Government's Ministry of the Interior may issue: *a)* an *emergency declaration,* i.e., an act by which the

*Corresponding Author: Erika Barojas-Payán, Ixtaczoquitlán, Ver., Mex., Email: eribarojas@gmail.com

https://dx.doi.org/10.25046/aj050438

Ministry recognizes that one or more territorial demarcations or municipalities are in imminence, high probability of presence of an abnormal situation generated by a disturbing agent. Therefore, immediate assistance is required to the population whose safety and integrity are at risk, or *b)* a *disaster declaration,* i.e., the act by which the Secretary recognizes the presence of a natural agent. Severe disturbance in particular municipalities, whose damage exceeds the local financial and operational capacity for their care, in order to access resources from the financial instrument for natural disaster care [6, 7, 8].

Based on the previous information, it should be noted that the southern states and the Gulf of Mexico coasts such as Veracruz, Guerrero, Tabasco, Chiapas, and Oaxaca concentrate the expenses to support the reconstruction of damaged or destroyed infrastructure by FONDEN due to the effects of hydrometeorological events. While, in the northern region of Mexico, the states of Nuevo León and Baja California Sur were the ones that required the largest amounts [9].

Likewise, Table 1 contains a synthesis of the declarations issued during the period covered from 2000 to January 2020, by FONDEN, for the states of the southern zone and the coast of the Gulf of Mexico. It can be seen: *a)* the number of municipalities that the state houses; *b)* the number of disaster declarations; and *c)* the number of declarations of emergency. In that Table, it can be identified that the most significant number of declarations are issued on the state of *Veracruz.*

Table 1: Declarations from 2000 - 2020

State	# of municipalities	disaster declaration	emergency declaration
Chiapas	118	59	75
Guerrero	81	34	36
Oaxaca	570	61	95
Tabasco	17	14	31
Veracruz	212	114	205

Own elaboration with information from [9, 10].

Based on that information, the present document evaluates the model presented in [1], whose bases fall on the problems of *a)* location of facilities, *b)* inventory levels, and *c)* multiple vehicle routing with capacity. All this, to generate a feasible solution that allows the establishment of a pre-positioning warehouse, with product inventory levels that allow supplying people affected by a hydrometeorological phenomenon, as well as their delivery routes. In order to provide a decision-making tool to institutions in the face of disaster.

Unlike the conference document, this research expands the geographical space, carrying out the zoning of the state of Veracruz, and the application of the model to one of them. This central area is made up of 90 municipalities, of which 66 of them were affected by some hydrometeorological phenomenon during 2016, as well as the expansion of products to be supplied, introducing the health one.

The document is made up of the following sections from this: *a)* Literary Review, contains studies carried out by researchers immersed in the area of emergency logistics; *b)* Methodology, shows step by step how the research was carried out until the results were obtained; *c)* Conclusions, to which the authors arrived, after an analysis of results and d) Bibliographic sources, for support and foundation of the research.

2. Literate Review

Relief operations often focus on response rather than preparedness due to the unpredictability of natural disasters. It leads to reactive rather than proactive systems. The effectiveness of the response may be determined by the structure of the supply chain [11]. These supplies can be said to be preparatory elements if some (or many) of them are available when needed, within strategic establishments [12]. Since the main cause and magnitude of the event of most disaster scenarios in Mexico is unknown, it is very complex to make some kind of forecast, but it is possible to prepare for it.

In recent years, many researchers from various countries have addressed the problem of the location of facilities in various fields and areas, an example of this is the health area, in which optimal location of hospitals, clinics, warehouses of medicines, and natural disaster centers that allow them to provide their services to the highest number of people in the shortest time. Decision making regarding the optimal or feasible location of a facility involves both significant investment and operating costs from the start of the project, which is why this is a crucial issue to be studied [13].

Research has been carried out over time regarding the issue of humanitarian logistics, and these documents reflect the application of different methodologies, methods, and tools in any of the stages of the disaster: (1) prevention; (2) mitigation; (3) preparation; (4) alert; (5) response; (6) rehabilitation, or (7) reconstruction [14].

These inquiries take as a basis for study one or several elements of logistics, such as the location of facilities, which have been captured in [15], where researchers examine the problem through a survey focused on four different models: 1. Deterministic; 2. Stochastic; 3. Dynamic, and 4. Robust.

In the same way in [16], they present two models, one deterministic, which considers the dependence of the damages about the distance between the disaster response facilities and the location of the affected population, and another stochastic, which extends the first concerning the intensity of the damage. A cost advantage of the second model over the first is demonstrated. Likewise, the researchers of [17], raise the impact of the mean time between disasters (MTBD) in the strategy of pre-positioning of medical supply inventory. The authors carry out a compensation model implemented on a platform based on a spreadsheet. In this area, we can mention the authors in [18], who propose a relocation of sites for the delivery of support in Mexico, through the problem of the p-median modeled in the LINGO software, thereby achieving the minimization of the current path at 25%. Meanwhile, in [19], the authors develop a stochastic programming model for shelter location in disaster preparedness, introducing a Genetic Algorithm (GA) for larger instances.

In addition to the above, some authors carry out studies focused on the location of facilities, adding *the capacity term*, in order to supply the victims assigned to it, such as the case of [20] whose study focuses on the location of warehouses for the distribution of humanitarian aid and its capacity, through a mixed-integer linear programming model.

Another element of logistics through which different investigations have been carried out under different optimization models, heuristics and hybrids, is the problem of vehicle routing, among which, it is worth mentioning [21], where multiple vehicle routing is presented with stochastic demand, in which researchers develop a nonlinear mathematical model with a probability constraint approach. This model is linearized with separable programming methods. While researchers in [22] develop a greedy iterative heuristic for vehicle routing with stochastic post-disaster times. Likewise, the researchers in [23] propose a model that determines the policy of resupply of products in disaster containers and generates the optimal route for vehicles to carry out the activity, maximizing the total benefits from sales revenue and costs of transport in a planning horizon. It is worth mentioning the investigation of [24], It is developed in a municipality in the state of Veracruz, Mexico, within which the heuristics of the closest neighbor (NN) generate product delivery trajectories to localities in the emergency phase of the disaster.

Traditional approaches had analyzed production, inventory, and transport operations separately rather than considering the entire operation as a whole [25, 26]. For example, when deciding inventory levels and product shipments in a given logistics network, inventory levels were generally determined first, and then product shipments were determined. However, as industries have become global and more competitive, simultaneous analysis of complex interactions between production and distribution operations has become essential to achieve efficiency [27].

Other integrated network models include [28], which studies an integrated production and transportation planning model with multiple retailers, products, and time frames. He used a two-phase heuristic model to solve the resulting mixed-integer program. Heuristics first gets an initial solution for distribution and production decisions.

However, the literature is scarce in terms of integrative approaches that consider a comprehensive plan that covers both aspects: plant location, inventory levels, and product supply. Only recent works take into account these two dimensions of the problem [29, 30].

In this scale, three logistical elements are introduced, such as location, inventories, and routing, such as the case [31], in which researchers address a plan to transfer of people affected by a flood to safe areas, they formulate a linear and mixed programming model, which allows defining the location and number of shelters and distribution centers with a pre-positioned inventory of aid packages, in addition to the evacuation routes for the victims.

3. Methodology

The research considers two of the regions immersed in the subject state of Veracruz that is used as a case study, called the *Capital Region* and *the High Mountains Region*, which are located in the central area of the entity. This region is composed of 90 municipalities. Moreover, 66 of its municipalities have been emergency declarations by the National Fund for Natural Disasters during 2016. Figure 1, shows the methodology to be followed, begins with the collection, selection, and analysis of the information, continues with the programming of the logistics model that allows the location of a host municipality for the installation of the pre-positioned warehouse, the calculation of the levels of each product for said warehouse and the set of routes for the delivery of products, to end with the analysis of the results obtained.

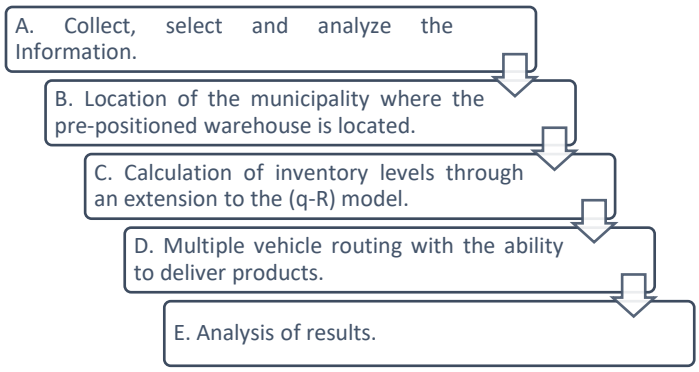


Figure 1: Applied methodology

3.1. Collect, select, and analyze the information.

An analysis of the information necessary to feed the model from [1] and its evaluation is carried out, which is described in the following paragraphs.

- *Matrix of distances between municipalities with the emergency declaration and municipalities without the emergency declaration,* using the Google Earth Pro® application, the distances are obtained that will allow the creation of the base matrix to locate a municipality without an emergency declaration in 2016 for the installation of the pre-positioned warehouse and the routes to the municipalities affected by the natural phenomenon in the same year, for the delivery of products.
- *Positional weight,* is the value made up of the degree of satisfaction that a municipality meets according to its attributes of *a)* highway infrastructure, highway type; *b)* services, water, network, telephone, drainage, electricity, *c)* not being a reason for issuing an emergency declaration, that is, not having been impacted by the hydrometeorological phenomenon, and *d)* constructed buildings. This variable, together with the distance, will allow finding the municipality where the pre-positioning warehouse is located. The data was extracted from the pages of [32, 33, 34].
- *Demand*, five types of demand: *a)* early childhood; *b)* childhood; *c)* adolescence; *d)* adulthood, and *e)* old age; taking as a reference to the World Health Organization, ruling out the peculiar stages such as pregnancy. The calculation of victims is based on the emergency declarations issued and the number of people by type of demand that lives in the area [35].
- *Products,* the items to be supplied, will be calculated in the form of kits: *a)* food; *b)* equipment (clothing, hygiene, and protection); *c)* health, and *d)* water. The first two, calculated according to the type of demand and the last two for total demand.
- *Total Inventory cost includes a)* ordering cost, *b)* purchasing cost, *c)* cost per shortage, and *d)* holding cost, [1, 36, 37, 38].

3.2. Location of the municipality where the pre-positioned warehouse is located

For the location of the host municipality of the pre-positioned warehouse, the p-median problem is taken as a base, adding the

positional weight variable, with which not only minimum routes are sought, but also adequate conditions for the management of the warehouse. Below is the mathematical formulation:

$$\boldsymbol{min} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} w_i d_{ij}\, x_{ij} \tag{1}$$

subject to

$$\sum_{j=1}^{m} x_{ij} = 1, \quad \forall_i\,, \tag{2}$$

$$x_{ij} \;\leq\; y_j, \quad \forall_{ij}, \tag{3}$$

$$\sum_{j=1}^{m} y_j \;=\; p, \tag{4}$$

$$x_{ij}, y_j \;\in\; \{0,1\} \tag{5}$$

Where: *equation (1)* objective function, seeks to minimize the sum of the distances d_{ij} traveled for each location of the demand, w_i being the variable that qualifies the attributes of each location with the possibility of hosting the facility; *equation (2)* confirms that each client is assigned to a single facility *n*; *equation (3)* confirms that each customer is supplied by a single facility; *equation (4)* assign *p* facilities, and *equation (5*) indicates that the variables are binary [18].

3.3. Inventory level calculation through an extension to the (q-R) model

The following mathematical formulation shows how the calculation of inventory levels for the different types of demand is carried out, which is an extension of the (q-R) model.

$$\textbf{min} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{p=1}^{P}\left[\frac{Co_i^p D_j^p}{Q_{ij}^p} + Cs_i\left(\frac{Q_{ij}^p}{2} + z_{cls} * S'^p_j\right) + C_i^p D_j^p + \frac{D_j^p}{Q_{ij}^p}\left(Cf_i^p + S'^p_{D_j} * E^p_{z_i}\right)\right] \tag{6}$$

subject to

$$S'^p_{D_j} = S'^p_{D_j}\ \sqrt{L_i} \quad i = 1,2,\dots,I \quad j = 1,2,\dots,J \tag{7}$$

$$R^p_{D_j} = D_j^p * L_i + z_{CSL} * S'^p_{D_j} \quad i = 1,2,\dots,I \;\; j = 1,2,\dots,J \;\; p = 1,2,\dots,P \tag{8}$$

$$Q_{ij}^p = \sqrt{2Co_i \frac{\sum_{j=1}^{J} Y_{ij}^p D_j^p}{Cs_i^p}} \; i = 1,2,\dots,I \quad j = 1,2,\dots,J \;\; p = 1,2,\dots,P \tag{9}$$

$$E^p_{z_i} = z\,[F_s^p(z) - 1] + f_s^p(z) \quad i = 1,2,\dots,I \quad p = 1,2,\dots P \tag{10}$$

$$FR_i^p = 1 - \frac{S'^p_j\, E^p_{z_i}}{Q_{ij}^p} \quad i = 1,2,\dots,I \quad p = 1,2,\dots,P \tag{11}$$

$$Q_{ij}^p \;\in\; Z^+ \tag{12}$$

The equation *(6),* seeks to minimize costs and is made up of*: a)* ordering cost, $\frac{Co_i^p D_j^p}{Q_{ij}^p}$; *b)* holding cost $Cs_i^p\left(\frac{Q_{ij}^p}{2} + z_{cls} * S'^p_{D_j}\right)$; *c)* purchasing cost, $C_i^p D_j^p$, and *d)* Stockout cost $\frac{D_j^p}{Q_{ij}^p}\left(Cf_i^p + S'^p_{D_j} * E^p_{z_i}\right)$; the *equation* (7), performs the calculation of the adjusted standard deviation for each demand ($S'^p_{D_j}$). It is mentioned that the demand has a normal distribution of its data; the *equation* (8), performs the calculation to obtain the point to which the inventory of products *(p)* of each of the demands of the affected municipalities *(D_j)* must reach, to issue a purchase order for these; *equation* (9), determines the number of products *(p)* to order for the pre-positioning warehouse *(i)* that will allow supplying the different demands of the affected municipalities *(j)*; *equation* (10), evaluates the probability of shortages to victims of products *(p)*; *equation* (11) measures the level of product delivery service *(p)* of the pre-positioned warehouse located in the municipality (*i)* for municipalities *(j)*, and finally, *equation* (12), indicates that (Q_{ij}^p) is the integer variable.

3.4. Multiple vehicle routing with the ability to deliver products

Through the problem of vehicle routing with capacity, the necessary routes are obtained that will allow supplying products to people affected by the natural phenomenon of hydrometeorological type, and the mathematical formulation starts from:

The two-index vehicle flow formulation type linear programming model uses $O(n^2)$ binary variables of x_{ij} to indicate with a value of 1 if the arc $(i,j) \in A$ belongs to the optimal solution and takes a value of 0 otherwise [39, 40].

$$\boldsymbol{min} \quad \sum_{i \in V}\sum_{j \in V} c_{ij} x_{ij} \tag{13}$$

subject to

$$\sum_{i \in V} x_{ij} = 1 \qquad \forall\, j \in V \setminus \{0\} \tag{14}$$

$$\sum_{j \in V} x_{ij} = 1 \qquad \forall\, i \in V \setminus \{0\} \tag{15}$$

$$\sum_{i \in V} x_{i0} = K, \tag{16}$$

$$\sum_{j \in V} x_{0j} = K, \tag{17}$$

$$\sum_{i \notin S}\sum_{j \in S} x_{ij} \geq r(S) \quad \forall\, S \subset V \setminus \{0\}, S \neq \emptyset \tag{18}$$

$$x_{ij} \in \{0,1\} \qquad \forall\, i,j \in V \tag{19}$$

Where: $\mathcal{V} = \{v_0, v_1, \dots v_n\}$ is the set of vertices of the graph, where v_0 corresponds to the warehouse; *C* is the distance or cost matrix c_{ij} between customers v_i and v_j; *K* is the number of capacity *C* vehicles needed to load all the demand; *S* is a set of vertices of the graph where $S \subset V \setminus \{0\}$, and $r(S)$ is the minimum number of vehicles necessary to serve all customers in *S*.

In this way, *equations* (14-15), ensure that a single arch enters and a single arch leaves each vertex, clients are visited only once; while *equations* (16-17) ensure that the number of vehicles leaving the warehouse (vertex) return to it; while *equation* (18) satisfies connectivity and non-violation of vehicle capacity.

3.5. Analysis of the results

- Impacted and non-impacted municipalities.

Table 2 shows the zoning of the state of Veracruz that is established to carry out this document, as can be seen, part of its regionalization. Table 2 shows the number of municipalities that each area houses and the number of municipalities impacted by a natural phenomenon of hydrometeorological type during 2016, the percentage of emergency declarations issued by FONDEN of a total is shown in the last column of 147 [41]. The central zone is taken as a reference for evaluating the model, derived from the fact that it is the one with the most significant number of municipalities and more than 50% of declarations issued.

- Vehicle capacity.

The same vehicular capacity is set as in [1], 4913 kits.

Table 2: Zoning of the state of Veracruz

Zone	Regions	# municipalities	Impacted municipalities	% declarations
North	Huasteca Alta, Huasteca Baja, Nautla, Totonaca	59	26	21.77%
Center	Capital, Altas Montañas	90	66	51.70%
South	Los Tuxtlas, Olmeca, Papaloapan, Sotavento	63	31	26.53%
Total		212	123	100%

- Location of the host municipality of the pre-positioned warehouse.

For the location of the host municipality, four different scenarios are proposed for the location of host municipalities of the pre-positioned warehouse. The first scenario designates the municipality of Fortín de las Flores as the headquarters of the aforementioned warehouse, which is obtained from the 24 municipalities without issuing an emergency declaration and the 66 municipalities impacted in 2016. The second scenario, part of rule out the municipality obtained in the first scenario (Fortín de las Flores) to give rise to the municipality of Emiliano Zapata as headquarters, the discard is from the assumption that the municipality first obtained, is impacted at a particular moment by a natural phenomenon of the type hydrometeorological. The third scenario is carried out in the event that the two previous municipalities cannot house the pre-positioned warehouse, Paso del Macho is obtained as the host municipality. A fourth scenario is captured, suppressing the variable of positional weight, that is, without taking into account the attributes of each municipality except that of affectation by a natural hydrometeorological phenomenon, thus the municipality of Atoyac, Miguel Alemán becomes the headquarters of the warehouse, which is located at a distance of 29.6 km from the first seat municipality, that is, Fortín de las Flores.

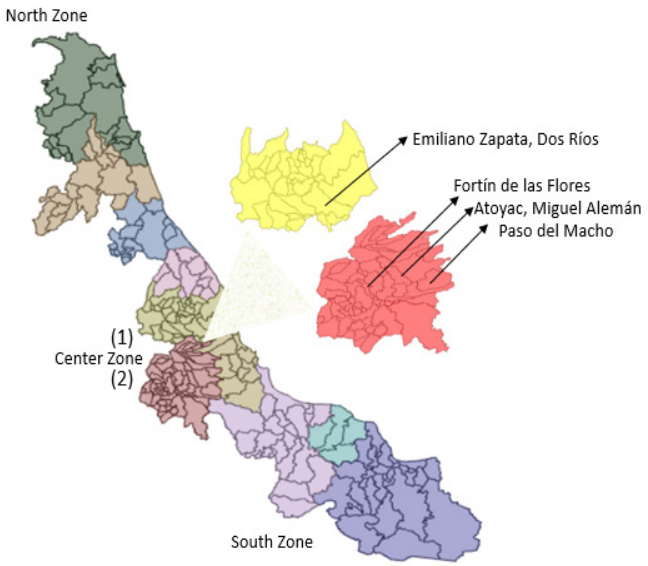


Figure 2: Municipalities headquarters of the pre-positioned warehouse [42]

While Figure 2 shows the regionalization of the state in the study, it should be remembered that in previous paragraphs, the state was zoned, based on this regionalization. The regions that make up the central area subject to study are the Capital region (1) and the High Mountains Region (2).

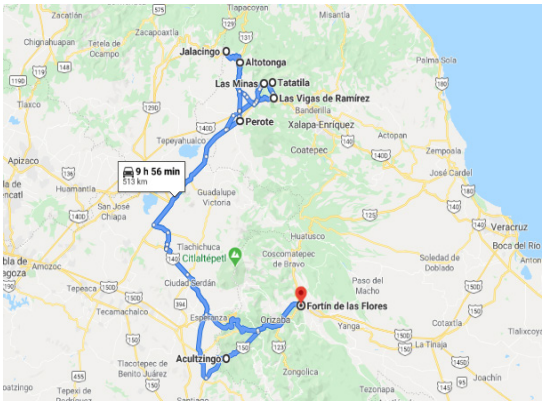
The location of each of the selected municipalities is shown in Figure 2: *1.* Emiliano Zapata; *2.* Fortín de las Flores; *3.* Atoyac, Miguel Alemán, and *4.* Paso del Macho.

- Inventory levels.

The levels of inventory of food, equipment, water, and medicines grouped in kits are shown in Table 3. It shows the economic order quantities, the reorder point, the safety inventory, and the number of people not supplied.

Table 3: Inventory levels

TYPE OF DEMAND	Kits	Economic order quantity (Q_{ij}^p)	Re-order point (R_j^p)	Safety stock	None-supplied
Demand_1 D_j^p (Early childhood)	Food	15732	6545	653	1
	Equipment	15521	6545	653	1
Demand_2 D_j^p (childhood)	Food	32793	13999	1443	3
	Equipment	32793	13999	1443	3
Demand_3 D_j^p (adolescence)	Food	48396	20353	1781	3
	Equipment	48396	20353	1781	3
Demand_4 D_j^p (adulthood)	Food	37000	15460	1275	2
	Equipment	37000	15460	1275	2
Demand_5 D_j^p (old age)	Food	9078	3755	327	1
	Equipment	9050	3755	327	1
Demand_6 D_j^p (General)	Water	227931	60014	5379	8
Demand_6 D_j^p (General)	Medicine	60014	28324	5379	8
Inventory Cost		640,501,050 MXN			



- Delivery routes

Table 4 shows 4 of the 12 delivery routes obtained with a pre-positioned warehouse within the municipality of Fortín de la Flores. The overall results in the said municipality showed a total of 1,964.68 km traveled for delivery, an arithmetic mean of 163.72 km traveled per route, and arithmetic means of supplying victims of 6,291.45. The delivery of products can be carried out in a minimum time, under the assumption of not affecting the road infrastructure.

Figures 3, 4, 5 y 6, They show the delivery routes in Table 4, in order to be able to observe the routes to travel. For this, the application of Google Maps®.

Table 4: Delivery routes

#	ROUTE								
1	Acultzingo	Perote	Las Vigas Ramírez	Tatatila	Las Minas	Villa Aldama	Altotonga	Jalancingo	
5	Coatepec	Xalapa	Tlacolulan	Acajete	Teocelo	Xico	Cosautlan de Carvajal	Ixhuacan de los Reyes	Ayahualulco
7	Huatusco	Zentla	Sochiapa	Tlacotepec de Mejia	Tenampa	Totutla	Tlaltetela		
12	Tlilapan	San Andrés Tenejapan	Atlahuilco	Tlaquilpa	Xoxocotla	Astacinga	Tehuipango		

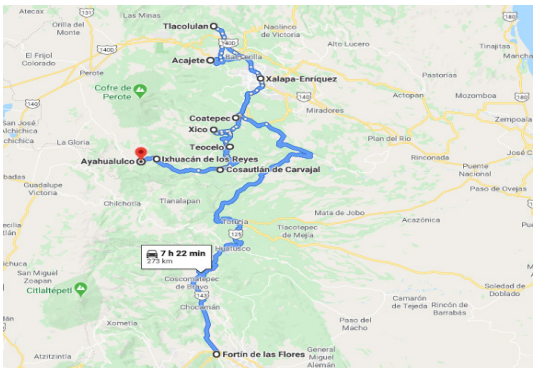


Figure 4: Route 5, 385 km traveled, nine municipalities supplied, extracted Google Maps ®

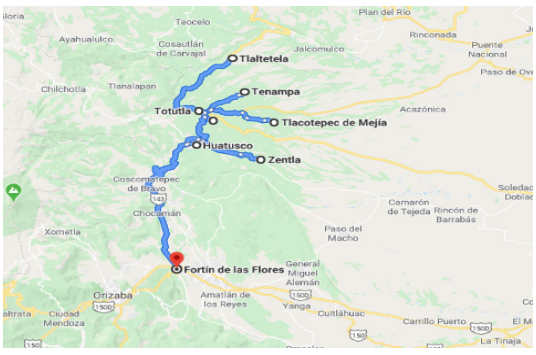


Figure 5: Route 7, 237 traveled, seven municipalities supplied, extracted from Google Maps ®

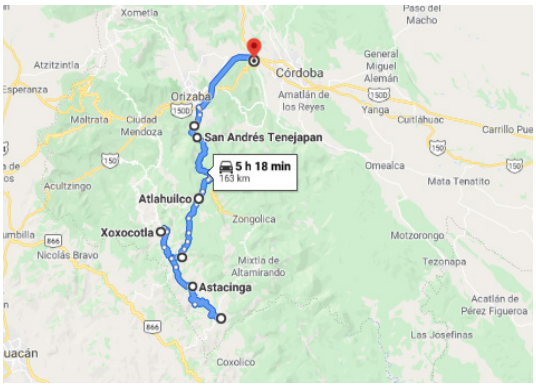


Figure 6: Route 7, 163 km traveled, seven municipalities supplied, extracted from Google Maps ®

4. Conclusions

Natural disasters have caused not only economic but also human losses. Mexico is not far from the effects of these events, much of its territory is exposed to the impact of natural phenomena of a hydrometeorological and geological nature, specifically the state of Veracruz, which has been considerably affected by hydrometeorological phenomena. This situation makes it imperative to locate facilities that can supply the needs, as mentioned above, to the victims in the shortest time.

The present document carries out the extension of a conference paper, evaluating the presented logistics model that has as its basis: *a)* the classic problem of the p-median for the location of a pre-positioning warehouse; *b)* an extension of the (q-R) model for calculating inventories of different products, according to different types of demand, and *c)* the problem of multiple vehicle routing with the capacity to establish delivery routes from warehouses to the affected municipalities. The evaluation is carried out in the same federal entity, with a more significant number of instances, in total 90, which are in the central area of the state, the results show that the location of the warehouse is assigned to the municipality of Fortín de las Flores, the delivery routes to the affected municipalities are presented, obtaining a total of 12 with a total of 1964 km traveled and an average travel distance of 163.72 km.

Similarly, the discarding of the municipality obtained is carried out, under the assumption that it has been affected by the natural event, obtaining another feasible location for the pre-positioned warehouse, this being the municipality of Emiliano Zapata, under the same. Of course, the latter municipality is discarded, obtaining Paso del Macho as the headquarters municipality of the warehouse. A fourth run of the program is carried out, in which the variable of positional weight and the previous municipalities are ruled out for having been affected, obtaining Atoya, Miguel Aleman, as the host municipality. In the same way, the increase in the number of products to be supplied was carried out. Therefore, the inventory levels obtained present a total cost of 640,501,050 Mexican pesos for the five demands: *a)* early childhood; *b)* childhood; *c)* adolescence; *d)* adulthood, and *e)* old age; with four product kits: *1.* Food; *2.* Equipment; *3.* Health, and *4.* Water.

It was carried out using the (q, R) model because the kits are supplied from 1, 2 or up to 3 times in specific time, according to the duration of the support bulletin´s opening or, in accordance to

the need of the people. We did not consider the newsboy model because it considers a single purchase. Even so, it is suggested to look for some other model that can answer this problem.

People's lives are priceless, natural phenomena are not controllable, but actions can be carried out to reduce the damage that they may cause to people immersed in impacted areas at a particular time, and maintain a quality of life for them. Until his return to his healthy life. The state of Veracruz, year after year is hit by different phenomena, but derived from its location, those of the hydrometeorological type occurs with a specific frequency, even more so in the period covered between the months of June-November, which is why the research carried out, looks for feasible locations that allow the location of a pre-positioned warehouse in municipalities that meet conditions for its efficient management, and that this warehouse, in turn, contains food products, clothing, health, and clean water to ingest, at adequate inventory levels, with the aim of a resupply until the return to normality, with a delivery routing with minimum distances, so that attention is fulfilled in the shortest possible time in terms of food and health, the latter derived that diseases such as typhoid, tetanus, and lung can occur to name a few. In the literature, especially in studies carried out in this state, minimal research was visualized in the field of humanitarian logistics.

As future work, a extend model is proposed that can to locate multiple locations with their delivery routes and travel distance constraints.

References

[1] E. Barojas-Payán, D. Sánchez-Partida, V. Juárez-Rivera, R. Villafuerte-Díaz, J. Medina-Cervantes, "Plant location, inventory levels and supply of products to areas affected by a natural phenomenon" in 2019 IEEE International Conference on Engineering Veracruz (ICEV), Boca del Rio, Ver., Mex., 2019. https://doi.org/10.1109 / ICEV.2019.8920691

[2] United Nations. UN News. 2010 [Online]. Available at https://news.un.org/es/story/2020/01/1467501 Last accessed February 1, 2020.

[3] J.M. Rodríguez, C.M. Welsh, M.L. Romo, A.C. Travieso, Riesgo de desastres en México: eventos hidrometeorológicos y climáticos. Instituto Mexicano de Tecnología del Agua, 2018.

[4] Cred-Usaid, Cred crunch, Brussels, Centre for Research on the Epidemiology of Disasters. United States Agency International Development. 2016. https://www.emdat.be/publications?page=1

[5] Secretaria de Gobernación, Manual de organización y operación del sistema nacional de protección civil, 2018.

[6] Secretaria de Hacienda y Crédito Público, Acuerdo por el que se emiten las reglas generales del fondo de desastres naturales, Capítulo I, Disposiciones generales, Articulo 1º, 2010.

[7] Cámara de Diputados del H. Congreso de la Unión, Ley General de Protección Civil, Nueva Ley publicada en el Diario Oficial de la Federación el 6 de junio de 2012, 12(59), 1-6, 2018.

[8] Cámara de Diputados del H. Congreso de la Unión, Ley General de Protección Civil, Nueva Ley publicada en el Diario Oficial de la Federación el 6 de junio de 2012, Capítulo XII, 2018.

[9] Secretaría de Gobernación - Centro Nacional de Prevención de Desastres, Impacto socioeconómico de los principales desastres ocurridos en la República Mexicana en 2016, Dirección de Análisis y Gestión de Riesgos Subdirección de Estudios Económicos y Sociales México, 2018.

[10] Instituto Nacional de Estadística y Geografía. Veracruz Ignacio de la Llave. 2019. https://www.inegi.org.mx/app/areasgeograficas/?ag=30

[11] S. Saeyeon, S. Pettit, I. Harris, A. Beresford, "The pre-positioning of warehouses at regional and local levels for humanitarian relief organization" International Journal of Production Economics, 170(1) Part B, 616–628, 2015.

[12] C.G. Rawls, M.A. Turnquist, "Pre-positioning of emergency supplies for disaster response" Transportation Research Part B, 44(4), 521–534, 2009.

[13] M. S. Daskin, Network and Discrete Location: Models, Algorithms, and Applications, John Wiley and Sons, Inc., 2013.

[14] Veracruz, Gobierno del Estado. Ciclo de los desastres. 2020. http://www.veracruz.gob.mx/proteccioncivil/ciclo-de-los-desastres/

[15] C. Boonmee, M. Arimura, T. Asada, "Facility location optimization model for emergency humanitarian logistics" International Journal of disaster risk Reduction, 24(2017), 485-498, 2017. https://doi.org/10.1016/j.ijdrr.2017.01.017

[16] A. Vermaa, G, Gaukler, "Pre-positioning disaster response facilities at safe locations: An evaluation of deterministic and stochastic modeling approaches" Computers & Operations Research, 62(2015), 197-209, 2015. https://doi.org/10.1016/j.cor.2014.10.006

[17] T. Saputra, O. Pots, K.D. de Smidt-Destombes, "The impact of mean time between disasters on inventory pre-positioning strategy" Disaster Prevention and Management, 24(1), 115-131, 2015. https://doi.org/10.1108/DPM-11-2013-0197.

[18] I. Landa-Cruz, D. Sánchez-Partida, E. Barojas-Payán, J.L. Martínez-Flores, "Optimization model for relocation of distribution sites of economic support in areas of high marginalization" DYNA Management, 4(1), [15], 2016. http://dx.doi.org/10.6036/MN8104

[19] M. Mostajabdaveh, W. Gutjahr, F. Salman, "Inequity-averse shelter location for disaster preparedness." IISE Transactions, 51(8), 809-829, 2019. https://doi.org/10.1080/24725854.2018.1496372

[20] C. Cornejo-Sánchez, J. Vargas-Florez, L. Aragón-Casas, V. Serpa-Oshiro, "Innovation in engineering, technology and education for competitiveness and prosperity" in Eleventh LACCEI Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2013), Cancún, Q. Roo, Mex, 2013.

[21] A. USLU, C. ÇETİNKAYA, SK. İŞLEYEN, "Vehicle routing problem in post-disaster humanitarian relief logistics: a case study in Ankara" Sigma J Eng & Nat Sci. 35(3), 481-499, 2017.

[22] M.E. Bruni, S. Khodaparasti, "A fast heuristic for routing in post-disaster humanitarian relief logistics" Transportation Research Procedia, 30, 304-313, 2018. https://doi.org/10.1016/j.trpro.2018.09.033

[23] M.A. Gökçe, E. Ercan, "Multi-period vehicle routing & replenishment problem of neighbourhood disaster stations for pre-disaster humanitarian relief logistics" IFAC-PapersOnLine, 52(13), 2614-2619, 2019. https://doi.org/10.1016/j.ifacol.2019.11.601

[24] E. Barojas-Payán, D. Sánchez-Partida, D.E. Riaño, J.L. Martínez-Flores, W. Velázquez, S.O. Caballero, "Heuristic of the nearest neighbor in the delivery of supports within the state of Veracruz" International Journal of Combinatorial Optimization Problems and Informatics, 10(2), 26-38, 2018.

[25] L. Bertazzi, M. Speranza, "Models and algorithms for the minimization of inventory and transportation costs: a survey" New Trends in Distribution Logistics, 480, 137-157, 1999.

[26] C.H. Glock, "The joint economic lot size problem: A review" International Journal of Production Economics, 135(2), 671-686, 2012.

[27] A. Sarmiento, R. Nagi, "A review of integrated analysis of production-distribution systems" IIE Transactions, 31(11), 1061-1074, 1999.

[28] Y.B. Park, "An integrated approach for production and distribution planning in supply chain management" International Journal of Production Research, 43(6), 1205-1224, 2005.

[29] G. Ghiani, A. Manni, E. Manni, M. Toraldo, "The impact of an efficient collection sites location on the zoning phase in municipal solid waste management" Waste Management, 34, 1949-1956, 2014.

[30] V.C. Hemmelmayr, K.F. Doerner, R.F. Hartl, D. VIGO, "Models and algorithms for the integrated planning of bin allocation and vehicle routing in solid waste management" Transportation Science, 48, 103-120, 2014.

[31] J. Gaytán-Iniestra, P.E. Arroyo-López, R. Enríquez-Colón, "A bi-criteria model to locate shelters as part of an evacuation plan in the case of floods" Revista Ingeniería Industrial, 2, 35-56, 2012.

[32] Instituto Nacional para el Federalismo y Desarrollo Municipal, Secretaria de Gobernación. Sistema Nacional de Información Municipal: Municipios en cifras. 2017. http://www.snim.rami.gob.mx/

[33] Secretaría de Comunicaciones y Transportes. Traza tu ruta. 2017. http://app.sct.gob.mx/sibuac_internet/ControllerUI?action=cmdEscogeRuta

[34] Secretaria de Desarrollo Social. Programa de abasto rural. 2017. http://catalogo.datos.gob.mx/dataset/programa-de-abasto-rural-tiendas-2015

[35] Organización Mundial de la Salud. Biblioteca electrónica de documentación científica sobre medidas nutricionales (eLENA), "Ciclo de vida". 2017. https://www.who.int/elena/life_course/es/

[36] Procuraduría Federal del consumidor. Quién es quién en los precios. 2018. https://www.profeco.gob.mx/precios/canasta/home.aspx?th=1

[37] Servicio de Administración Tributaria. Salarios Mínimos. 2017. www.sat.gob.mx/informacion_fiscal/tabla_indicadores/Paginas/salarios_minimos.aspx

[38] Secretaria del trabajo y previsión social. Salarios Mínimos. 2017. www.gob.mx/cms/uploads/attachment/file/175865/Tabla_de_salarios_minimos_vigentes_a_partir_de_01_enero_2017.pdf

[39] P. Toth, D. Vigo, An Overview of Vehicle Routing Problems, Monographs on Discrete Mathematic and Applications, SIAM, 2000.
[40] J.P. Orrego, D. Ospina, E.M. Toro, "Solución al problema de ruteo de vehículos con capacidad limitada (CVRP) usando una técnica metaheurística" Scientia Et Technica, 21(3), 225-233, 2016. https://doi.org/10.22517/23447214.9013
[41] Sistema Nacional de Protección Civil, México DDGR (FONDEN). Insumos autorizados por declaratoria de emergencia. 2016. https://www.proteccioncivil.gob.mx/es/ProteccionCivil
[42] National Institute for Federalism and Municipal Development, SEGOB, Secretary of the Interior. National Municipal Information System: Municipalities in numbers. 2010. http://www.snim.rami.gob.mx/

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 323-330 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Multi-Model Security and Social Media Analytics of the Digital Twin

Jim Scheibmeir[1], Yashwant Malaiya[2]

[1]*Colorado State University, Systems Engineering, 80523, United States*

[2]*Colorado State University, Computer Science, 80523, United States*

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 28 October, 2020
Online: 10 November, 2020

Keywords:
Internet of Things
Digital Twin
Cybersecurity
Contact Tracing
Social Media Analytics

ABSTRACT

Digital twins act through application programming interfaces to their physical counterparts to monitor, model, and control them. Beyond these traditional functions of digital twins, they must also act to secure their physical counterparts. A multi-model scheme is presented to help digital twins towards the task of securing the physical system. Additionally, this work includes an analysis of more than four hundred thousand tweets each relating to digital twin technology and cybersecurity which were collected during June and July in 2020. Of the first corpus of tweets collected by searching for #digitaltwin during the research period, only a small population of 10% reference security concepts. In the second and larger corpus of collected tweets, the top mentioned industries were health, education, and public. A naïve Bayes model reached a 70.3% accuracy at differentiating tweets that were either related to cybersecurity or the internet of things. The study also indicates that cybersecurity tweets are consistently more negative in many areas of sentiment when compared to tweets about the internet of things. The sentiment findings of cybersecurity tweets will reinforce the need to address culture in cybersecurity posture while the security multi-model schema contributes to the state of the art.

1. Introduction

An API development model for digital twins has been proposed recently [1]. This paper augments that work to address the cybersecurity concerns of the digital twin within the context of the various implementations and industries where the twin may be utilized. Furthermore, the adoption of current security models, such as state machine, lattice, non-interference, and information flow models are proposed for digital twins.

In the original paper, a development model for the APIs of the digital twin was created. The model was proposed to establish the context of the environment, the system, relationships between the two, and the functional and non-functional requirements.

This research includes the mining of Twitter tweets using appropriate search criteria for digital twins, the internet of things (IoT), and cybersecurity. The analysis includes mention analysis of industries, exploration of the sentiment, natural language processing (NLP) implementation towards the classification of tweets, and modeling to predict the twitter user behavior.

*Corresponding Author: Jim Scheibmeir, Email: jimscheibmeir@gmail.com

Three sets of tweets were collected for this study with the latter two sets being joined into a combined corpus. Each set was collected through the execution of scripts written in R. The first collection is smaller and has 3,102 tweets about digital twins. The second and third collections are larger. In total, 422,963 industrial internet of things (IIoT) tweets and 497,174 cybersecurity tweets were collected. These larger sets were combined into one corpus for analysis. The search criteria for the tweet collections are listed in Table 1.

Table 1: Tweet collection search criteria

Script Topic	Criteria used in actual R code
Digital Twin	"#digitaltwin"
IIoT	"#industry40 OR #IIOT OR #IOT"
Cybersecurity	"#cybersecurity OR #infosec OR #hacking"

This study is organized as follows. Section 2 reviews the literature and the background information on the subject. Section 3 describes our research method, including the collection of tweets as well as the processing. Sections 4 and 5 present the results from

social media analysis. Section 6 proposes a multi-model security scheme for digital twins. Finally, section 7 presents the conclusion of this study.

2. Background and related work

Digital twins are a logical conduit between physical entities and humans, geared towards monitoring and controlling a system [2]. The design of the interfaces themselves may utilize structured standards such as OpenAPI to list the various objects, paths, and operations [3].

Digital twins can be used during system design for vulnerability detection or during physical system operations to assist with intrusion detection [4]. Digital twins can utilize and manage the Internet of Things (IoT) devices and IoT systems of systems [5]. Wireless Sensor Networks (WSN) and attached IoT sensor components can be energy and cost-effective monitoring approaches [6]. WSNs are also frequently constrained to limited computing and power resources and thus security in WSNs can be challenged [7]. To maintain the integrity of the sensor data and overall system state awareness, secure links with proper authentication must be established [8].

Vulnerable digital twins offer hackers a blueprint of the physical counterpart as well as other backend systems that may be a part of system integration [9]. The digital twin can help secure the physical systems or be yet another vulnerability as IoT growth has created a large cyber-attack surface [10]. IoT has revealed security vulnerability ranging from trust management, authentication, privacy, to access control of embedded systems [11]. While digital twins may present such risk, they are still getting popular. The smart factory market, an economy of integrated automation solutions adopted to streamline manufacturing, is predicted to be worth approximately $205 billion-dollar by 2022 [9]. Hearn and Rix also cite another 2019 study whereof 220 security leaders in industrial and manufacturing who were a part of the research survey, 79% indicated they had experienced an IoT cyberattack within that past year. IoT and digital twins-based systems can be attractive targets for malicious actors.

Hearn and Rix identify advantages, such as the prevention of downtime and monitoring attacks against cyber-physical systems, as potential benefits of digital twins [9]. Risks of digital twins in cyber-physical systems (CPS) include the intellectual property incorporated in the digital twins as well as critical information about the CPS itself. For such reasons, the digital twin itself must go through rigorous software hardening routines, such as fuzz and penetration testing [9]. Security of the digital twin must start from the ground-up, inception to retirement. Security concerns must be incorporated with the organization's culture [9].

IoT systems and CPS are being deployed in the energy industry, and there have been instances of cyber-attacks on them. Some known attacks include the US electrical grid in 2009, Ukraine power outage in 2015, and Stuxnet attacks in 2005 and 2010 [10]. Common problems can facilitate such attacks, such as low power IoT devices have minimal resources and thus they are not resilient to denial of service attacks. A digital twin may create yet another vulnerability of an industrial or energy system.

Atalay & Angin in their 2020 study proposes that the digital twin can help to secure a cyber-physical system by establishing a security framework. The framework should contain an extensible digital twin that represents the smart grid, a cyber-threat database for applicable attack vectors, attack simulation tools, and an analysis and reporting module [10].

The framework suggested by Atalay & Angin starts with the creation of the digital twin through specification and reuse of other models. The framework separates the modeling of the CPSs into two phases, identification of the elements and their relationships, and building the network model using simulation. By creating the digital twin and using models, it is possible to achieve standardization through model reuse, utilization of a central threat intelligence database of smart grid attacks, and to achieve continuity of security evaluation [10].

Digital twins may secure physical systems through rule-based controls as well as intrusion detection as proposed by Eckhart and Ekelhart [4]. They identify a wide range of definitions for digital twins. Some definitions omit simulation and focus on visualization, such as implementing augmented reality (AR). AR applications have their security and quality concerns. Eckhart and Ekelhart also explain how a digital twin be used to implement intrusion detection for an industrial control system (ICS) or CPS. They identify concerns beyond intrusion detection, such as:

- the origin of digital twins was in the simulation of systems, although some authors omit simulation as a benefit of digital twins
- digital twins may have fidelity issues which may limit their use
- digital twins have a lifecycle that follows the lifecycle of their physical counterpart
- proper retirement of a digital twin system is imperative as it may contain private and confidential data
- digital twins may be used for security personnel training or in a cyber range as the attack target
- cost of digital twin creation can still be a prohibitive factor

Eckhart and Ekelhart refer to their 2018 publication, where, it was suggested that previously established specifications of the physical system could be modeled with AutoML, virtualized within Mininet and other tools, which can reduce the cost of creating digital twins [12]. They further proposed that digital twins can assist in intrusion detection by offering a baseline of normal behavior that can be analyzed and compared to the behavior of the physical system to detect outlying and potentially malicious activity [12].

3. Methodology

3.1. Collection

This paper focuses on the analysis of collections of Twitter tweets, as well as academic papers. Scripts have been written in R and executed throughout the research period. More R code has been utilized to clean the collected tweets, transform new classification data fields, and analyze them for n-grams, mention,

and sentiment analysis. The scripts use various libraries, such as the rtweet library as an API to the Twitter social media platform. Figure 1 below omits various keys and passwords yet illustrates the basics of library import and the implementation of a search and data frame store of tweets.

```
library(rtweet)
library(ggplot2)
library(dplyr)
library(tidyr)

c_key <- ''
c_secret<- ''
access_token <- ''
access_secret <- ''
appdt <- ''

twitter_token <- create_token(
  app = appdt,
  consumer_key = c_key,
  consumer_secret = c_secret,
  access_token = access_token,
  access_secret = access_secret)

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

digitaltwin_tweets <- search_tweets(q = "#digitaltwin",since = '2020-06-08', n = 50000)
```

Figure 1: Screenshot of example usage of rtweet library within R Studio

Three scripts were written to collect tweets for this research. The first used a single search criterion, "#digitaltwin", which has collected 3,102 tweets within 21 days of June 2020. The second and third scripts ran for the entire month of July 2020. 422,963 industry internet of things (IIoT) tweets and 497,174 cybersecurity tweets were collected. The tweets are stored with metadata made available via the rtweet API, but not all data has been retained. Additional data has been generated and stored such as the industry a tweet may mention and the sentiment score (positive or negative) of the text within the tweet. Table 2 lists the acquired and generated data.

Table 2: Tweet data fields

Field Name	Short Description
created_at	The date the tweet was originally posted to the Twitter platform
text	Character contents of the tweet
favorite_count	An integer representing the number of times a tweet has been marked as a favorite by Twitter users
retweet_count	An integer representing the number of times a tweet has been retweeted by Twitter users
country	Frequently NA within the dataset and sometimes providing the country of origin of the Twitter user posting the tweet
retweet_location	Frequently NA within the dataset and sometimes providing the country of origin of the Twitter user retweeting a tweet
Industry and count	A list of industries and the count of mentions from the gathered tweets
Sentiment	The individual sentiment score of a tweet's text
Retweet	Binomial value indicating if retweet_count > 0
Favorite	Binomial value indicating if favorite_count is > 0

3.2. Data Preparation and Processing

Data preparation activities may account for 80 percent of the time invested in data science effort [13]. In this study, the collection took many days, but the Twitter data was usable practically from the beginning. This is due to the limited search criteria used and the nature of the required data (such as we did not need the country or retweet_location fields for this research, both having many empty or not applicable entries). The longest time in preparation was determining industry mentions and labeling the tweets. Industry and identification keywords for cross-reference and classification efforts were collected from the International Labor Organization's web site [14]. Preprocessing was done cross-referencing the ILO keywords to the tweet's text to determine the tweet's topics. This enabled mention analysis by the labor industry.

In the second corpus of tweets, the labeling of tweets allowed the naïve Bayes models to be trained and then tested. Tweets were labeled based upon the search criteria that was used to collect the tweets. This label would consist of either IoT or cybersecurity. Eighty percent of the data was used to train the models and twenty percent was utilized to test the model accuracy. Various models and parameters were used to predict the correct tweet label. Other preprocessing for the model input parameters included sentiment analysis and word lemmatization. The package textstem was utilized for word lemmatization and the packages tidytext and sentimentr were utilized for the sentiment analysis.

4. Results of First Corpus Analysis

The word cloud illustrated in Figure 2 highlights common bigrams found within the first corpus of collected tweets, those 3,102 found with search criterion #digitaltwin, after removing common stop words. While concerning the word 'security' is not paired with another common word frequently enough to make this bi-gram chart, we do see that conversations on Twitter around defense, risk, and the industrial use of IoT are occurring, each having relevance towards this research.



Figure 2: Word Cloud of Bigrams within the Tweet Collection Text

The word cloud used in figure 2 omits counts of the actual instances of the bigrams. Table 3 outlines a few common security terms and the number of instances they occur within our first

corpus, those 3,102 tweets collected by searching #digitaltwin. Such terms may occur in one tweet multiple times. Overall, terms such as "encrypt", or "vulnerability" have almost no mention in such tweets and do not appear as a part of the social media conversation which is a concerning observation.

Table 3: Common security terms and their occurrence

Terms	Count of occurrences
"security"|"secure"	88
"risk"	212
"defense"|"defensive"	179
"iiot"|"industrial iot"	750
"encrypt"|"decrypt"|"cipher"	1
"vulnerability"|"threat"	1

Looking at this first corpus of tweets, the percentage of tweets containing common security terms make up 10% of the total volume of tweets collected using the #digitaltwin search criteria within the specified time range, as illustrated in Figure 3. We observe a dearth of conversation within the social media analysis towards securing digital twins, which could be used to implement the security models needed.

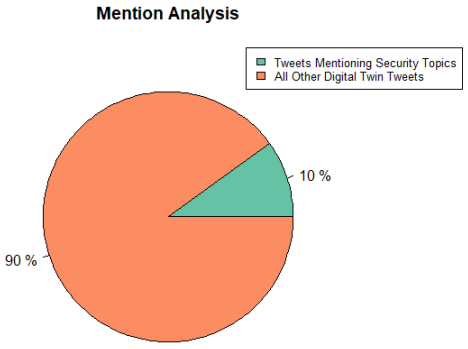


Figure 3: Most IoT tweets do not mention security topics

4.1. Most Favored and Most Retweeted Tweets about Digital Twins

Tweets can be favored as well as re-tweeted. The number of favorites indicates the count of unique user accounts that like or agree with the content of the tweet [15]. In the first corpus, the most favored tweet is also the most retweeted. It will be discussed later in this section. To provide more breadth of the findings, the second most favored tweet is illustrated in Figure 4 [16]. This tweet refers to Intel technology and links to a 2015 white paper where Intel published best practices to help with IoT sensor technology implementation against challenges such as cost of ownership, maintaining security, and designing for scalability [17].

Within the scope of this first corpus, the most retweeted tweet is about the use of digital twin used in contact tracing to help prevent the spread of Covid-19. The digital twin was specifically built and used to help contain the spread of the virus in Dharavi, a location within Mumbai that houses nearly one million residents. Many of these residents rely on daily wages. Living and working conditions may reduce or prevent social distancing. During the time of the tweet creation, epidemiologists had credited the Dharavi community for their efforts towards containment [18]. Figure 5 is a screen capture of this most favorited tweet within the first corpus [19]. While we cannot confirm the accuracy of the tweet, we can recognize from the tweet contents the broad application of digital twin usage to monitor, model, and control a pandemic. The nature of such a system and the associated data also provide support for why such the system requires a mature security posture, to protect individuals and their privacy rights.

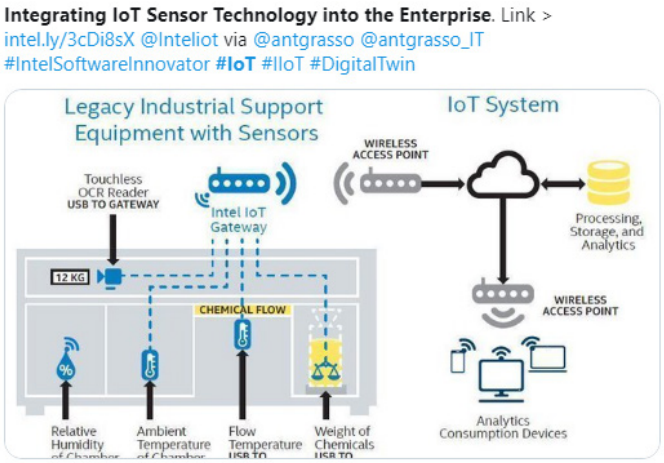


Figure 4: Second Most Favored Tweet within the Scope of this Research



Figure 5: Most retweeted tweet within the research scope

4.2. Mention and Sentiment Analysis

Studies show digital twins being utilized across industries; mention analysis from social media provides an additional perspective. To analyze industry mention among the first tweet collection, we crossed referenced the tweets' text with a list of known industry categories [14]. The simple algorithm reads the tweet texts and counts industry mentions, including when multiple

industries are mentioned within a single tweet's text. Figure 6 illustrates mention analysis results with health, education, and public as the top three topics with most mentions within the tweets' text.

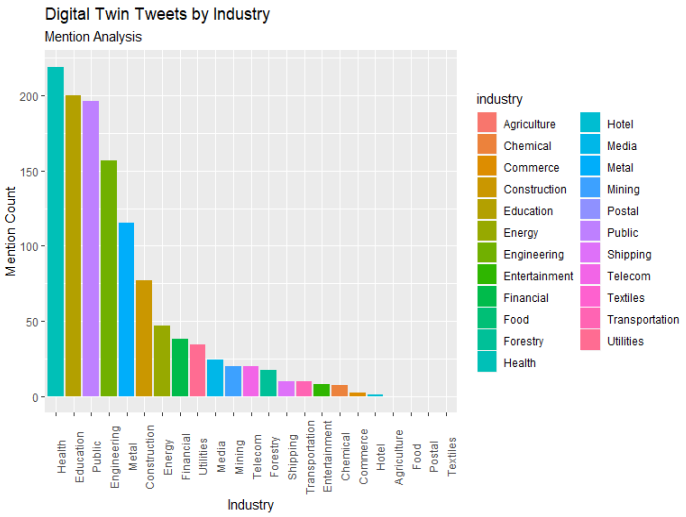


Figure 6: Industry mention analysis: health, education, and public are the top three mentioned industries

Utilizing the sentimentr package, each industry group has the individual tweet text sentiment scores averaged and placed on the chart found in Figure 7. The sentiment scores from the sentimentr package may range from a negative 2.0 to a positive 2.0. While the health and public industries were high in the mention-analysis, the sentiment scores for those industry groups are negative. The three industry groups having the highest average sentiment scores are engineering, commerce, and energy.

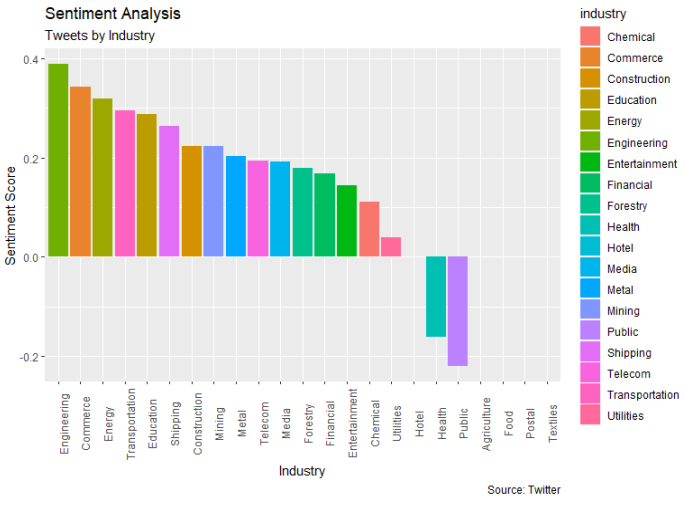


Figure 7: Aggregated sentiment scores grouped by industry

The tweet in the engineering group having the highest sentiment score is given in Figure 8 [20]. This tweet indicates the usage of digital twins to assist in the simulation of electric vehicles for optimal energy management.

Today's electric vehicles market is high stakes. The current pandemic threatens the automotive supply chain [21], a Tesla truck recently marketed towing a competitor's vehicle (F-150) up a hill [22], and Tesla competitor Nikola Motors saw a one-day stock jump of 104% [23]. Security posture is required around the intellectual property in this dynamic and competitive market, and therefore must be applied to such a digital twin.



Figure 8: Tweet having highest sentiment score within the engineering group

5. Results of Social Media Analytics of IIoT and Cybersecurity Tweets

Multiple prediction models were created to further inspect the tweets. This included attempts to predict whether a given tweet was towards the topic of cybersecurity or IIoT. Methods such as word stems, lemmatization, term frequency-inverse document frequency (TFIDF) were utilized in naïve Bayes and general linear models.

A naïve Bayes model that attempted to predict the label, cybersecurity or IIoT, of a given tweet using word lemmatization and sentiment analysis could predict the label correctly with a 70.3% accuracy. While this prediction accuracy is not high, perhaps the sentiment coefficient correlations are more telling. As previously stated, security must start with the organization's culture [9]. An organization's culture consists of the beliefs and values of the employees. Communication incorporating the sentiment of the topic may be more influential towards influencing behavior or belief. The sentiment level of the cybersecurity tweets, when compared to the IIoT tweets, was higher in the areas of anger, disgust, fear, negative, and sadness, whereas the IIoT tweets sentiment level, when compared to cybersecurity tweets, was higher in anticipation, joy, positive, surprise, and trust. Table 4 lists the average sentiment scores for each classification of sentiment by the cybersecurity and IIoT labels.

Given a tweet is about Cybersecurity, there is a 14.5% probability that the tweet is negative, an 11.6% probability that the tweet provokes anticipation, and an 11.3% probability that the tweet has the sentiment of fear. This information may be key for communicating cybersecurity strategies and plans within an organization. We should consider the sentiment of our security posture messaging, and how it fits into the culture, behaviors, and values of the communication's target audience.

The Term Frequency-Inverse Document Frequency (TFIDF) score utilization within a naïve Bayes model had less accuracy,

56.3%, compared with the model using both word lemmatization and sentiment analysis. TFIDF does allow us to see important words in a corpus, we observe that "bugbountytip" had the highest TFIDF score within the cybersecurity corpus at 0.00127.

Table 4: Sentiment coefficient probabilities on tweet labels

Conditional Probabilities of IIoT and Cybersecurity Tweet Sentiment		
Sentiment	Cybersecurity	IIoT
Anger	0.07145	0.03765
Anticipation	0.11663	0.14300
Disgust	0.01868	0.01400
Fear	0.11321	0.06652
Joy	0.05369	0.07048
Negative	0.14541	0.07757
Positive	0.25503	0.35884
Sadness	0.04404	0.02833
Surprise	0.03586	0.03683
Trust	0.14597	0.16677

6. Application of a Digital Twin as a Multi-Model Security Architecture

Digital twins are commonly used to help monitor, model, and control cyber-physical systems. Engineering a digital twin should be done using a systematic approach by performing a needs analysis from the physical device and context of the operating environment [1]. Needs analysis and environmental context must include non-functional requirements, including cybersecurity. Securing information confidentiality of these cyber-physical systems is a growing concern [24]. Beyond the common monitor, model, and control mechanisms, digital twins may also be utilized for incorporating multiple security models.

Several security control models exist. The information flow model is a common mechanism to maintain information confidentiality [25]. We outline in Table 5 how the information flow model may also work to maintain the integrity of a digital twin and its physical counterpart. Multiple models will need to be implemented in the system engineering effort to establish a secure posture. Each model has a strength, such as Bell-LaPadula working towards the confidentiality of information [26]. Furthermore, in Table 5, we indicate how the lattice model, incorporating Biba, can work to ensure information integrity.

The digital twin may also incorporate concepts and techniques such as automated policy exchange (APEX) to prevent data leakage. APEX technology works to keep honest people working within the guidance of security best practices and policies, as large systems and enterprises may have complex policies that not everyone is aware of [27]. Using APEX, the digital twin would monitor labeled files throughout their lifecycle and use control mechanisms, such as deployed agents within the cyber-physical system, to prevent or warn on file activities by users.

Another possible avenue would be for the monitoring portion of the digital twin to utilize a System Call Intercept (SCI) framework. SCI works to reduce data leakage by reviewing policies of subjects and objects, trapping some specific activities before an operating system completing the execution of the activity and perhaps suspending the activity completely [28]. Using the SCI framework, the monitor and modeling functions of the digital twin would use policies to review specific subjects or objects (perhaps all users and system components within a zero-trust approach) and trap some actions for review before the OS or another system component completing the action. An inventory of objects, subjects, and system calls that could be trapped would be required to implement such a framework, which will improve data protection. This inventory could be complemented with behavior analysis by machine learning models to identify malicious activity. Machine learning can be used for countering cybersecurity threats and vulnerability mitigation using regression, prediction, and classification techniques.

Table 5 provides high-level examples of how a digital twin implementation may satisfy or utilize common security models such as State Machine, Lattice, Interference, and Information Flow models.

Table 5: Application examples of a digital twin as a multi-model security schema

Security Model	**Example Digital Twin Application**
State Machine	Digital twin control input is first executed within the model. Validation occurs within the model for an appropriate output state. The input may be rejected before entering the digital twin control mechanism for actual processing on the cyber-physical system.
Non-Interference Model	Monitoring and control mechanisms must obfuscate users reading (monitoring) and writing (controlling) across domains. This reduces the probability of users being influenced by the actions of subjects with greater/lesser clearance. This approach may utilize common discretionary access control (DAC) across file systems, mandatory access control (MAC) through rule-based systems, and role-based control. DAC and MAC limit the availability of information objects to certain subjects to retain the confidentiality of information [29].
Lattice Model	Biba protects integrity by preventing a subject from reading lower integrity or lower-level access data [29]. Using the Biba Model approach, the digital twin modeling process accesses only known secured sensors (having a level of integrity through controls). While we may monitor many sensors, decision dashboards and predictive analytics should utilize only trusted and secured sensors. A matrix of such processes and secure sensors would instruct inputs for such high integrity algorithms.
Information Flow Model	Sensor information flow is one way. The Digital twin may read sensor data for monitoring or modeling purposes. No data is transmitted to a sensor to be later sent back, as sensor-originated data. This prevents levels of spoofing sensor data, by not allowing writes to the sensor memory. Any PUT type of action to sensors APIs would be rejected and logged as malicious intent.

Figure 8 illustrates a situation where a population, Bluetooth sensors from mobile devices, network paths, and system logic are used for the creation of a population's digital twin. In this case, the digital twin of the population would be used by public health officials and the application subscribers for contact tracing. The diagram's purpose is to provide examples illustrating the concepts given in Table 5 applied to a hypothetical contact tracing system.

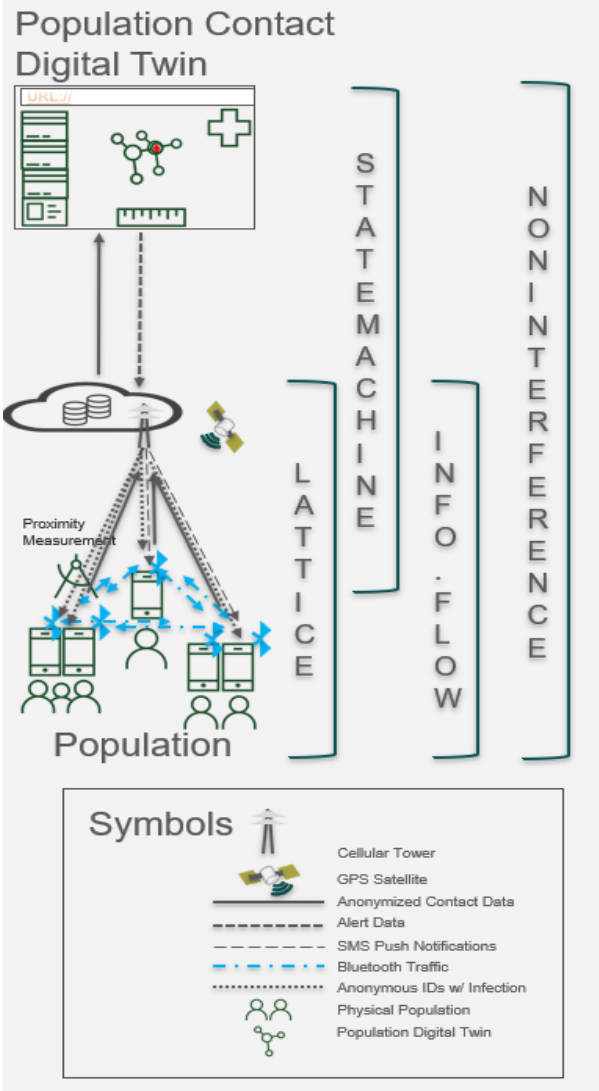


Figure 8: Applying the multi-model in digital twin security

If contact tracing systems were injected with malicious data, it could alter the behaviors of a population. Applying the lattice model helps assure that a system having integrity does not read from a system or sensor with low integrity. The lattice model can prevent a malicious actor's device farm from spoofing the contact system. To validate the integrity of a sensor or mobile device in this case, the system may validate a history of GPS or use human interaction to assure the device is not a simulator or otherwise driven by a malicious bot.

Public health officials may instruct the system to deploy warning notifications to participants who have opted in to use the feature. However, before deploying the notification, the digital twin should utilize the state machine model. Again, a notification could be a malicious message being sent to alarm a population. The logic should first model that notification within the digital twin. The modeling should evaluate the message accuracy and may even predict how the population may behave to help public health officials craft response plans. If the notification messages were not related by a contact tracing scheme, the message may be malicious and should be stopped.

The information flow model is utilized to protect the population and their privacy. The mobile device's application logic allows Bluetooth connections to write to ledgers indicating contacts. This data is anonymized and does not indicate a person's identity. Writing to this ledger is only allowed by the local mobile device which is in the possession of the user. The centralized system does not allow writing to this contact ledger. The mobile device application will pull anonymized alert data from the central system. This data is written to a separate ledger. The only mechanism for the centralized system to send data to the host devices is through the SMS push notification mechanism, which has levels of security such as users being required to opt-in and the state machine model.

Users of the Population Contact Digital Twin dashboard will utilize roles, authentication, and authorization to access the digital twin of the population. Low entitlement users would not have the ability to interfere with the population by spying on them using inference techniques. Entitled users, such as medical doctors, may be granted access to information having more fidelity to help protect the population's health. This protects a population using the non-interference model with access controls.

7. Conclusions and Future Work

7.1. Conclusions

Digital twins operate from and create new cyber-physical system information. A proper design approach for a digital twin should model its cyber-physical system, including its future states. In addition to monitoring, modeling, and controlling the cyber-physical system, a digital twin must also provide security. This fourth element as applied to the prior three digital twin requirements allows for a layered approach to securing cyber-physical systems by implementing multiple security models. While past researchers have focused on intrusion detection, this article proposes that common security control models such as state machine, non-interference, lattice, and information flow models can be utilized in digital twin implementation to secure itself the associated cyber-physical system.

In the first corpus of tweets collected in this study by searching for #digitaltwin, those referencing security concepts represented a small population of 10%. In the second and larger corpus of the collected tweets, the top mentioned industries or business sectors were health, education, and public in that order. The top three industries having the highest tweet sentiment were engineering, commerce, and energy. While naïve Bayes models reached only a 70.3% accuracy at differentiating a tweet that was about cybersecurity versus a tweet about digital twins, the sentiment findings of the messages leaned towards negative messaging from the cybersecurity tweets.

7.2. Future Work

Most of our research utilized social media, which can be time-consuming to prepare such data and has limits in what it can

contribute. Future research should include detailed case studies of direct implementation which would strengthen the approaches proposed in this work. The sentiment findings of cybersecurity tweets deserve further research as other researchers have stated that security must start with an organization's culture. The implication of negative messaging on an organization's cybersecurity posture and culture needs to be studied.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J. Scheibmeir, Y. Malaiya, "An API Development Model for Digital Twins," in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, 518-519, 2019, doi: 10.1109/QRS-C.2019.00103.

[2] E. A. Saddik, "Digital Twins: The Convergence of Multimedia Technologies," IEEE Multimedia Mag., 25, 87-92, Apr. 2018.

[3] B. Martino. M. Di, M. Rak, A. Ficco, S. Esposito, S. Nacchia, "Internet of Things Reference Architectures, Security and Interoperability: A Survey." Internet of Things, 1-2, 99-112, 2018, doi:10.1016/j.iot.2018.08.008.

[4] M. Eckhart, A. Ekelhart, "Digital Twins for Cyber-Physical Systems Security: State of the Art and Outlook." Security and Quality in Cyber-Physical Systems Engineering. Springer, 2019, doi:10.1007/978-3-030-25312-7_14.

[5] A. Canedo. "Industrial IoT Lifecycle via Digital Twins." Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/software Codesign and System Synthesis. ACM, 1-1, 2016, doi: 10.1145/2968456.2974007

[6] D. Yadav, S. Jayanthu, S. Das, S. Chinara, P. Mishra, "Critical Review on Slope Monitoring Systems with a Vision of Unifying WSN and IoT," IET Wireless Sensor Systems, 2019, doi:10.1049/iet-wss.2018.5197.

[7] Y. Wang, A. Garhan, R. Byrav. "A Survey of Security Issues in Wireless Sensor Networks." IEEE Communications Surveys & Tutorials, **8**(2), 2-23, 2006, doi:10.1109/comst.2006.315852.

[8] P. Porambage, C. Schmitt, P. Kumar, A. Gurtov, M. Ylianttila, "Two-phase Authentication Protocol for Wireless Sensor Networks in Distributed IoT Applications," IEEE Wireless Communications and Networking Conference, 2014, doi:10.1109/WCNC.2014.6952860.

[9] M. Hearn, S. Rix, "Cybersecurity Considerations for Digital Twin Implementations report". IIC Journal of Innovation, 2019

[10] M. Atalay, P. Angin, "A Digital Twins Approach to Smart Grid Security Testing and Standardization," 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT" 435-440, 2020, doi: 10.1109/MetroInd4.0IoT48571.2020.9138264.

[11] A. Girma, "Analysis of Security Vulnerability and Analytics of Internet of Things (IoT) Platform," in Information Technology - New Generations. Advances in Intelligent Systems and Computing, 738, 2018, doi:10.1007/978-3-319-77028-4_16

[12] M. Eckhart, A. Ekelhart, "Towards Security-Aware Virtual Environments for Digital Twins," in Proceedings of the 4th ACM Workshop on Cyber-Physical System Security, ACM, 61–72, 2018, doi:https://doi.org/10.1145/3198458.3198464

[13] A. Cirillo, RStudio for R Statistical Computing Cookbook. Birmingham Packt Publishing, 2016.

[14] International Labor Organization. "Industries and Sectors." Industries and sectors, https://www.ilo.org/global/industries-and-sectors/lang--en/index.htm.

[15] R.S. Perdana, A. Pinandito, " Combining Likes-Retweet Analysis and Naive Bayes Classifier within Twitter for Sentiment Analysis", Journal of Telecommunication, Electronic, and Computer Engineering, **10**(1-8), 41-46, 2018.

[16] A. Grasso [@antgrasso]. (2020, June 8). Integrating IoT Sensor Technology into the Enterprise. [Twitter moment]. Retrieved from: https://twitter.com/antgrasso/status/1269962528530071552

[17] K. H. Gaiser, V. P. Harrington, G. T. Loughrin, S. J. Meyer, J. M. Sartini. "Integrating IoT Sensor Technology into the Enterprise." December 2015. https://www.intel.com/content/www/us/en/it-management/intel-it-best-practices/integrating-iot-sensor-technology-into-the-enterprise-paper.html.

[18] P. M.N., S. Bengali. "How an Indian Slum Became an Unlikely COVID-19 Success Story." Los Angeles Times. June 24, 2020. https://www.latimes.com/world-nation/story/2020-06-24/dharavi-slum-in-mumbai-india-contained-covid-19.

[19] Zuppa [@ZuppaGeoNavTech]. (2020, June 17). "Offence Is The Best Defense" [Twitter moment]. Retrieved from: https://twitter.com/ZuppaGeoNavTech/status/1273449554797408257

[20] Simcenter [@Simcenter_] (2020, June 12). Learn more about #simulation for optimal vehicle energy management in our new on-demand webinar series. [Twitter moment] Retrieved from: https://twitter.com/Simcenter_/status/1271442013586624512

[21] L. Harbour, "The Coronavirus' Impact On The Global Automotive Supply Chain." Forbes. March 14, 2020. https://www.forbes.com/sites/laurieharbour1/2020/03/13/the-coronavirus-impact-on-the-global-automotive-supply-chain/#17fd6b9f444e.

[22] M. Wayland, L. Kolodny, "Tesla Cybertruck Tug-of-war with a Ford F-150 Is High Stakes and Highly Unlikely." CNBC. November 27, 2019. https://www.cnbc.com/2019/11/26/tesla-cybertruck-tug-of-war-with-a-ford-f-150-is-high-stakes-unlikely.html.

[23] M. Fox, "Tesla Competitor Nikola Motors Skyrockets 104% after Setting Reservation Date for Its Electric Truck (NKLA, TSLA) | Markets Insider." Business Insider. June 8, 2020. https://markets.businessinsider.com/news/stocks/nikola-motors-stock-price-soars-setting-truck-reservation-date-competitor-2020-6-1029290291.

[24] A. Humayed, J. Lin, F. Li, B. Luo, "Cyber-physical systems security—A survey", IEEE Internet Things J., **4**, 1802-1831, 2017, doi: 10.1109/JIOT.2017.2703172.

[25] A. Sabelfeld, A. C. Myers, "Language-based information-flow security", IEEE J. Sel. Areas Commun., **21**(1), 5-19, 2003, doi: 10.1109/JSAC.2002.806121.

[26] X. Ma, Y. Huang, D. Li, "A Security Model Based on Lattice," International Conference on Electrical and Control Engineering, 4958-4961, 2010, doi: 10.1109/iCECE.2010.1199.

[27] S.J., Simske, and H. Balinsky. "Apex." Proceedings of the 10th ACM Symposium on Document Engineering, 10, 2010, doi:10.1145/1860559.1860587.

[28] H. Balinsky, D. S. Perez, and S. J. Simske. "System Call Interception Framework for Data Leak Prevention." 2011 IEEE 15th International Enterprise Distributed Object Computing Conference, 2011. doi:10.1109/edoc.2011.19.

[29] C. Schou, & S. Hernandez, Information assurance handbook:effective computer security and risk management strategies. New York: McGraw-Hill Education, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 331-343 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Distributed Microphone Arrays, Emerging Speech and Audio Signal Processing Platforms: A Review

Shahab Pasha[*, 1], Jan Lundgren[1], Christian Ritz[2], Yuexian Zou[3]

[1]*STC Research Centre, Mid-Sweden University, Sundsvall 85230, Sweden*

[2]*School of Electrical, Computer and Telecommunication Engineering, University of Wollongong, NSW 2500, Australia*

[3]*ADSPLAB/Intelligent Lab, School of ECE, Peking University, Shenzhen 518055, China*

ARTICLE INFO

Article history:
Received: 14 April, 2020
Accepted: 03 June, 2020
Online: 28 July, 2020

Keywords:
Distributed signal processing
Machine learning
Spatial selectivity
Speech enhancement

ABSTRACT

Given ubiquitous digital devices with recording capability, distributed microphone arrays are emerging recording tools for hands-free communications and spontaneous tele-conferencings. However, the analysis of signals recorded with diverse sampling rates, time delays, and qualities by distributed microphone arrays is not straightforward and entails important considerations. The crucial challenges include the unknown/changeable geometry of distributed arrays, asynchronous recording, sampling rate mismatch, and gain inconsistency. Researchers have recently proposed solutions to these problems for applications such as source localization and dereverberation, though there is less literature on real-time practical issues. This article reviews recent research on distributed signal processing techniques and applications. New applications benefitting from the wide coverage of distributed microphones are reviewed and their limitations are discussed. This survey does not cover partially or fully connected wireless acoustic sensor networks.

1. Introduction

With new portable devices, such as smartphones and tablets, conventional microphone arrays are no longer the main signal and speech acquisition platform; rather, distributed microphone arrays (also called ad hoc microphone arrays) formed by the joint analysis of randomly located independent recording devices such as laptops and cell phones are emerging recording platforms for various applications [1]. Conventional compact microphone arrays and recording devices are now processed within ad hoc arrays and are jointly analyzed with other such devices.

For this reason, distributed microphones are popular and have been used in a wide range of applications, such as speaker tracking and speech recognition systems [2]. Currently, there is substantial potential for applications that use digital recording devices collaboratively as a virtual array [3]. By positioning such devices at random locations within the acoustic scene, the array geometry is no longer limited to standard structures. Distributed microphone nodes provide wide spatial coverage and are ideal for capturing multiple speakers located meters away from one another. This advantage in the means of signal (i.e., sound and other cues derived from it) acquisition allows different audio scenes to be captured to give a more accurate picture of the user environment. To enable such ubiquitous and flexible teleconferencing and multimedia applications within the distributed signal processing context, several important technical and theoretical problems should be addressed. Some of the main challenges are the unknown/changeable array structure, inconsistent sampling frequencies, varying gains (due to varying source-to-microphone distances), and unsynchronized recordings. Although speech and signal processing applications are well studied and straightforward [4] in the context of known-geometry compact microphone arrays, existing methods are not directly applicable to distributed scenarios.

This survey article reviews the strengths and limitations of recently proposed distributed signal processing methods. The

[*]Corresponding Author: Shahab Pasha, Email: shahab.pasha@miun.se

focus is on scenarios in which the microphones are independent and do not communicate within the array.

2. Literature Selection Methodology

The authors have reviewed the most recent theoretical and practical literature on distributed signal processing and related, overlapping fields of study. More than 250 research papers, theses, patents, and online tools and resources were studied and compared over a five-year period. One hundred peer-reviewed published works were chosen for final comparison and investigation according to the novelty of the work and the credibility of the journal/conference. The current set of references covers the literature and online resources as of 2020, though most of the main methods were proposed between 2013 and 2017.

3. Distributed Signal Processing

Ad hoc microphone arrays [1, 3] consist of a set of recording devices (referred to as nodes [5]) randomly distributed in an acoustic environment to record an unknown acoustic scene with wide spatial coverage (Figure 1). The nodes can be identical [5] or different [6] in terms of their structure and number of elements [7, 8] (Table 1). Ad hoc arrays eliminate the restrictions on microphone and source placement at fixed locations and facilitate dynamic and flexible recording experiences.

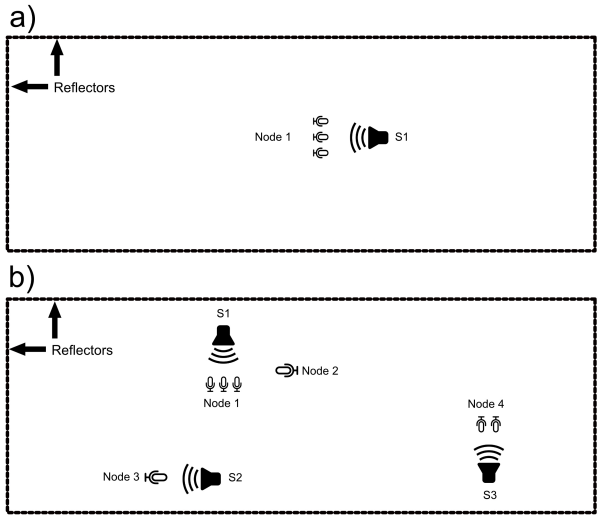


Figure 1: a) A three-element linear microphone array recording a single source; b) a distributed acoustic scene comprising three sound sources, two single-channel nodes, and two multi-channel linear arrays.

Table 1: Ad hoc microphone arrays compared with compact arrays

Distributed	***Compact***
Unknown structure [9]	Known structure
Changeable microphone locations [10]	Fixed topology
Unknown inter-channel time delays [11]	Known inter-channel time delays
Inconsistent gain within the array [12]	Consistent gain
Uncertain direction of arrival (DOA) definition [13]	Straightforward DOA definition
Large phase differences (i.e., spatial aliasing)	Negligible spatial differences
Inconsistent signal quality [5]	Consistent signal quality

3.1. Definition

M recording devices (both single- and multi-channel devices) located randomly to capture N sound sources form an ad hoc microphone array. Node m contains M_m channels and the signal picked up by the ith channel forming the mth node is modelled as

$$x_{m,i}(n) = \sum_j \sum_i s_j(n) * h_{m,i,j}(n) + v_{m,i}(n) + w_{m,i}(n), \quad (1)$$

where $s_j(n)$ is the jth sound source, $h_{m,i,j}(n)$ represents the room acoustic response between the ith channel forming the mth node and source j, and $v_{m,i}(n)$ is the additive noise; $w_{m,i}(n)$ is the non-coherent component and represents reverberation and diffuse noise. Each channel location is modelled as

$$\mathbf{r}_{m,i} = [x_{m,i}, y_{m,i}, z_{m,i}], \quad (2)$$

and the jth source location is

$$\mathbf{r}_{s,j} = [x_j, y_j, z_j]. \quad (3)$$

in the Cartesian coordinate system. $h_{m,i}(n)$, $\mathbf{r}_{m,i}$, and $\mathbf{r}_{s,j}$ are assumed unknowns for all the values of m, i, j, and n. The total number of channels in the array is

$$M_{ch} = \sum_{m=1}^{M} M_m. \quad (4)$$

For ad hoc scenarios where all the nodes are single channel

$$M_{ch} = M. \quad (5)$$

The truncated RIRs ($\mathbf{h}_{m,i,j}$) of length L are modelled as follows:

$$h_m(n) = \sum_{k=0}^{L} a_{m,k}\delta(n - \tau_{m,k}), \quad (6)$$

with time delays, $\tau_{m,k}$, and amplitudes, $a_{m,k}$. L is chosen based on the application and the reverberation time [14]. $\tau_{m,0}$ represents the time of arrival (TOA) at node m.

Assuming that the distances between the channels at each node are relatively small and that each node forms a compact microphone array [5, 7], each node can deliver a single-channel output, so there will be only M distributed recordings [15]. Having only one active source during the short frames simplifies (1) to

$$x_m(n) = s(n) * h_m(n) + v_m(n) + w_m(n). \quad (7)$$

3.2. Significance and applications

Tablets, smartphones, sound recorders, and other portable and wearable digital devices are becoming prevalent in workplaces, homes, and lecture halls, redefining how we communicate and record our communications. Consequently, these devices are becoming key tools for daily activities, including teleconferencing and hands-free speech communication [16]; ad hoc signal processing is therefore inevitable.

Advances in ad hoc signal processing technologies can also relax the highly demanding constraints on network layer design [17]. Ad hoc arrays also improve the quality of speech communication, acoustic scene analysis, and speech recognition (Figure 2) due to the multiple observations they make within a wider area [18].

3.2.1. Improved Distributed Meetings

Meetings are an important part of everyday life for many types of educational and professional workgroups. The main application of ad hoc microphone arrays is for distributed meetings (DMs) [19], as such meetings are recorded by microphones at unknown and varying environments with different signal qualities. DM systems enable the high-quality broadcasting and recording of meetings by utilizing the flexible spatial coverage of ad hoc arrays. Distributed meetings using signal processing techniques are not necessarily online meetings, but could be group meetings [20]. Ad hoc arrays are more suitable recording tools for distributed meetings than are compact microphone arrays, as the recording devices can be spread out in the meeting room. Joint analysis of the distributed microphone signals yields more accurate results for signal processing applications such as active source detection (i.e., localization) [21].

3.2.2. Hearing Aids

Reduction of interference and noise is important in hearing aids (HAs) to provide intelligible speech signals in noisy environments [22]. Using an array of microphones, it is possible to exploit the spatial characteristics of the acoustic scenario to obtain more information about the target scene. Although the scenario investigated by Bertrand and Moonen [22] was a fully connected binaural network, the unknown geometry of the array and the random recording setup (i.e., the source locations) form an ad hoc acoustic scene.

3.2.3. Hands-Free Communication

Wearable recording nodes are essential to hands-free communication systems, and the node structure can also vary. As the node and source locations in such systems are unknown and changeable, classic array-processing methods cannot be applied. Ad hoc signal processing offers a solution for the joint analysis of the nodes in hands-free voice communication systems for improved speech enhancement and source localization applications. The general scenario of hands-free audio recording and voice recognition with distributed nodes has been described and investigated by Jia et al. [23].

3.2.4. Ambient Intelligence and Smart Homes

Ambient intelligence advances are based on advances in sensors, pervasive computing, and artificial intelligence [24]. Very little can be done by an ambient intelligent system without detecting the user's presence and movements. The ambient intelligence system must be aware of the users' locations in each time period. One way to do this is by tracking the sound sources and acoustic cues. Spatially distributed microphone arrays are effective tools for monitoring user movements [25]. The insights gained provide important clues as to the type of activities the user is engaged in, and make the system responsive to the user's

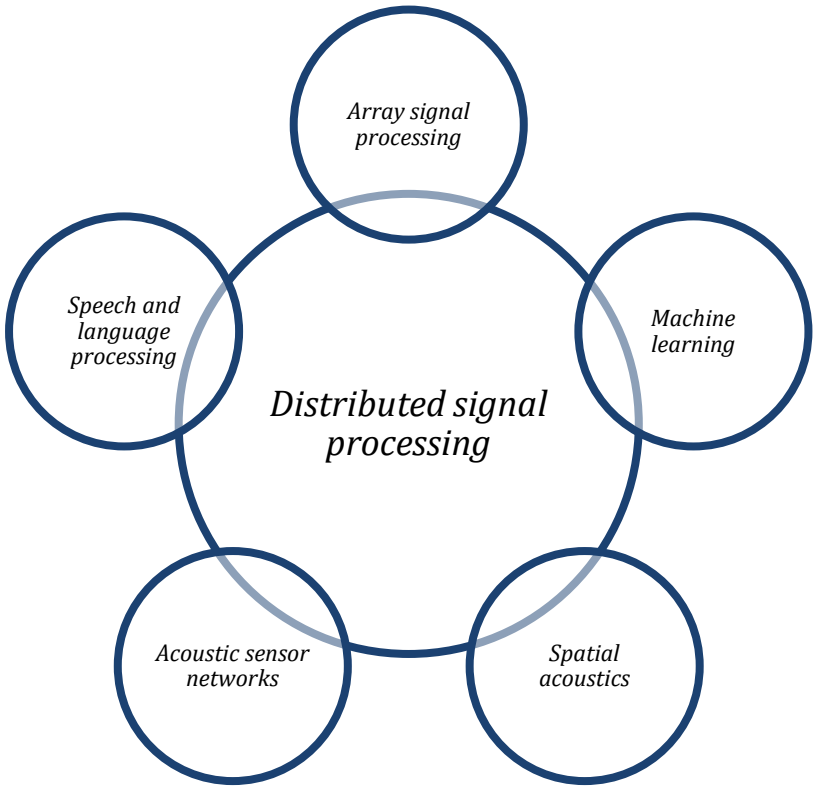


Figure 2: Distributed signal processing and overlapping fields of study

location [26]. Speaker verification and identification by ad hoc microphone arrays and speech recognition for automated human–computer interaction [18] are important fields of research that can help customize ambient intelligence applications.

3.2.5. Monitoring

Some recent research focuses on analyzing the environment by means of acoustic sensing due to its unobtrusive nature. Acoustic monitoring covers a wide range of applications, from security tasks (e.g., intrusion detection, malicious activity early detection, and traffic monitoring [27]) to whale migration tracking [28] and bird behavior studies [29].

3.2.6. Medical Signal Processing

Unconventional distributed microphone arrays have been used in medical applications for accurate and high-quality chest sound pick-up [30] and vital sound separation [31]. Compared with conventional compact arrays, ad hoc microphone arrays provide more accurate recordings of closely located organs such as the lungs and heart. The design and development of ad hoc acoustic sensors and microphones are essential for medical and E-health research.

3.3. Room Acoustics and Distributed recording

Unlike conventional compact microphone arrays, in which the noise and reverberation levels are consistent, in distributed signal processing each microphone has its own unique reverberation level and room acoustic response. Researchers have recently shown that the unique RIR and echo pattern at each ad hoc microphone location contain location and distance information even if the recording setup is unknown [32, 33]. This idea is applicable to ad hoc microphone arrays for microphone clustering [11], room geometry reconstruction [33], and microphone localization [34].

The framework for recording using distributed microphone nodes was formulated by Tavakoli et al. [16] (1), covering both reverberation and noise. The signals recorded by all M_{ch} channels (4) within an ad hoc array are modeled as

$$\mathbf{x}(n) = [x_1(n), \dots, x_M(n)]^T. \quad (8)$$

As each ad hoc microphone receives its own unique distorted version of the source signal, the RIR and reverberation at each microphone location contain important information [11, 35]. Each RIR ($h_m(n)$) (6) can be represented by sets of reverberation times and reverberation amplitudes [11], as follows:

$$[a_{m,0}, \dots, a_{m,L}], \quad (9)$$

where a_0 is the direct path impulse amplitude, and

$$[\tau_{m,0}, \dots, \tau_{m,L}]. \quad (10)$$

Researchers have divided the RIR (1) into two segments: early and late echoes [36]. This segmentation is specifically important in the context of ad hoc microphone arrays and is the basis for defining discriminative features. The clarity feature is used for sound source localization by ad hoc microphones [37]. Table 2 summarizes the acoustic features applied to signal processing applications.

Table 2: Room acoustics features applied to distributed signal processing

Method/features	*Application*	*Reference(s)*	*Year*
Energy	Localization	Liu et al. [12]	2007
Noise coherence	Clustered beamforming	Himawan et al. [49]	2011
Voice activity detection	Speech enhancement	Sakanashi et al. [50]	2013
Sparsity analysis	Localization	Asaei et al. [21]	2014
Time delay and sound level	Speech separation	Souden et al. [7]	2014
Signal power	Traffic monitoring	Toyoda et al. [27]	2014
Euclidian distance matrix (EDM)	Microphone localization	Dokmanic et al. [51]	2015
Kurtosis of the linear prediction residuals	Clustering	Pasha et al. [8]	2015
Generalized cross-correlation	Speaker tracking	Tian et al. [52]	2015
Reverberation and echoes	Clustering	Pasha et al. [11]	2015
Non-negative matrix factorization (NMF)	Calibration	Asaei et al. [53]	2015
Pseudo coherence vector	Node selection for speech enhancement	Tavakoli et al. [5]	2015
Reverberation	DOA estimation	Pasha et al. [54]	2015
Cepstral features	Interference suppression	Gregen et al. [9]	2016
C_{50} (short-time reverberation)	Multi-talk detection	Pasha et al. [37]	2016
Magnitude-squared coherence (MSC)	Crosstalk and multi-talk detection	Pasha et al. [37]	2016
Echoes	Room geometry reconstruction	Dokmanic et al. [33, 55]	2016
Coherent-to-diffuse ratio	Multi-talk and crosstalk detection	Pasha et al. [56]	2017
Power spectral density (PSD)	Spotforming	Habet et al. [57]	2017
Signal to interference and noise ratio	Noise cancellation	Tavakoli et al. [20]	2017
Time difference	Microphone localization	Woźniak et al. [58]	2019
Distributed unscented Kalman particle filter (DUKPF)	Source localization	Zhang et al. [59]	2020

3.4. Challenges and Limitations

Large scale distributed arrays are inherently asynchronous [38]. Inconsistent sampling rates [39], gain differences [12], and different signal-to-noise ratios (SNRs) at different locations [5, 12] are challenges with distributed signal processing. The main differences between ad hoc array and compact array signal processing are summarized in Table 1.

4. Distributed Signal Processing

The most recent existing distributed signal processing techniques proposed in the literature are reviewed here in terms of their target scenarios and requirements. Distributed signal processing overlaps with array processing [4, 40], wireless sensor networks [41] (not reviewed here), feature extraction and machine learning [42], hands-free speech communication [43], and guided/informed signal and speech processing [44, 45].

4.1. Microphone Calibration

Microphone calibration is especially important in the context of large arrays [46, 47] and distributed microphones, as the area covered by the microphones can be large [48]. These large distances and attendant time delays should be considered when time aligning the signals for beamforming and speech enhancement applications. Representing the gain of the array by $\boldsymbol{g} = \{g_1, \dots, g_M\}$, (1) can be rewritten as

$$x_m(n) = s(n) * g_m \bar{h}_m(n) + v_m(n), \tag{11}$$

for one active source, where g_m and $\bar{h}_{m,i}(n)$ are the gain and normalized RIR of microphone m, respectively. Calibration often consists of estimating the distances between the pairs of microphones and reconstructing the array geometry given all the pairwise distances [53, 60]. Microphone array calibration in general suffers from reverberation, noise, and complicated mathematical computations (making calibration infeasible for real-time applications); specifically, sampling frequency mismatch, inconsistent microphone gain, and non-stationary array geometry make calibration an ongoing process throughout the recording session. Due to calibration difficulties, some methods prefer to avoid microphone calibration altogether, if possible applying methods inherently robust to microphone placement and steering error [61].

For microphone array calibration, some advanced mathematical methods use joint source and microphone localization methods [62] and incorporate matrix completion constrained by Euclidean space properties [63, 64]. Such methods require partial knowledge of pairwise microphone distances [65]. It has been shown that using sound emissions for self-calibration can result in a calibration method more robust to sampling frequency mismatch [66]; however, the method is only applicable to devices that have both recording and sound emitting capabilities.

4.2. Signal Synchronization

Signal synchronization is an essential task for more advanced applications such as dereverberation and speech enhancement. Improperly addressing the synchronization issue leads to poor dereverberation performance. The different types of delays, including microphone internal, TOA, and onset time delays (Figure 3), have been investigated and explained [67, 68]. The lack of a reference channel and inconsistent sampling frequencies [69] in ad hoc arrays lead to critical phase differences [70] between the channels, fundamentally challenging most speech and audio processing applications. Existing solutions to clock drift [71] require limiting assumptions, such as unmoving sources and stationary amplitudes.

If the array topology is available, the time difference of arrival (TDOA) (13) between any two microphones in the array can be calculated as

$$TOA_{s,m} = \frac{|\mathbf{r}_s - \mathbf{r}_m|}{c} + \delta_m \tag{12}$$

$$TDOA_{m,\acute{m}} = \frac{|\mathbf{r}_s - \mathbf{r}_m|}{c} - \frac{|\mathbf{r}_s - \mathbf{r}_{\acute{m}}|}{c} + (\delta_m - \delta_{\acute{m}}) + (T_{om} - T_{o\acute{m}}), \tag{13}$$

where δ_m and T_{om} represent the internal delay and onset time of microphone m, respectively (Figure 3) [67, 68], and $\mathbf{r}_s =$

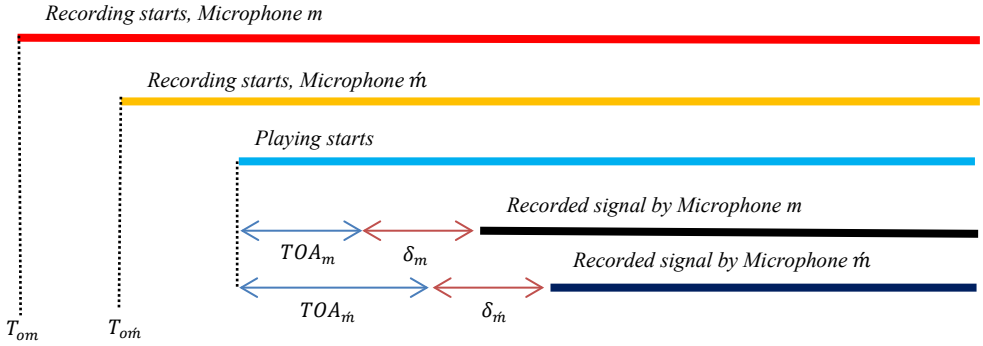


Figure 3: Time delays

$[x_s, y_s\ z_s]^T$, $\mathbf{r}_m = [x_m, y_m\ z_m]^T$, and $\mathbf{r}_{\acute{m}} = [x_{\acute{m}}, y_{\acute{m}}\ z_{\acute{m}}]^T$ are the source, microphone m, and microphone $\acute{m}$ locations in Cartesian coordinate space, respectively. The objective of synchronization is estimating the overall time delay between every pair of microphones. The time delays are given as

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \cdots & \tau_{1M} \\ \vdots & \ddots & \vdots \\ \tau_{M1} & \cdots & \tau_{MM} \end{bmatrix}, \tag{14}$$

where $\tau_{mm} = 0$ for $m = 1$ to M and $\tau_{m\acute{m}} = \tau_{\acute{m}m}$ for all m and $\acute{m}$ values.

Researchers have used the time alignment of ad hoc channels for source localization by means of generalized cross correlation (GCC) [17] and for defining the parametric squared errors of time differences [72].

Some advanced techniques use the least squares for the temporal offset estimation [73] and for audio fingerprinting [38]. These methods are developed based on the clustering and synchronizing methods applied to unorganized multi-camera videos [74] which are based on matching the time-frequency landmarks between two channels.

The effect of synchronization on blind source separation (BSS) by a wireless acoustic sensor networks (WSAN) was investigated by Lienhart et al. [75], who concluded that full synchronization increases the BSS cost function by an average of 4 dB.

Most of the proposed synchronization methods can time-align the signals accurately and calculate the TDOA with an error of 1–10 milliseconds. The other important factor is the computational cost. The watermark-based algorithms [38] are shown to be more efficient than GCC methods [74].

4.3. Spatial Multi-Channel Linear Prediction (LP)

Multi-channel LP was developed for compact arrays and has applications, such as dereverberation [76] and compression [77, 78].

Using x_m from (1), the autocorrelation $r_m(c)$ is obtained for channel m from

$$r_m(c) = E\big(x_m(n)x_m(n+c)\big), \quad c = 0,1,2,\ldots \tag{15}$$

where E is the mathematical expectation.

$$\bar{r}(c) = \frac{1}{M} \times \sum_{m=1}^{M} r_m(c). \tag{16}$$

The baseline autocorrelation function (16), can be formulated in the more general form of a weighted average autocorrelation ($\bar{r}_w(c)$) (17). Assuming that the applied weights are $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$, the weighted average autocorrelation function is calculated as

$$\bar{r}_w(c) = \frac{1}{\sum_{m=1}^{M} \beta_m} \times \sum_{m=1}^{M} \beta_m r_m(c), \tag{17}$$

where β_m is the weights given to $r_m(c)$.

$$\mathbf{w}_s = \begin{bmatrix} \bar{r}_w(0) & \cdots & \bar{r}_w(P_{short}-1) \\ \vdots & \ddots & \vdots \\ \bar{r}_w(P_{short}-1) & \cdots & \bar{r}_w(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{r}_w(1) \\ \vdots \\ \bar{r}_w(P_{short}) \end{bmatrix}, \tag{18}$$

and the pre-whitened signal is

$$\tilde{e}_m(n) = x_m(n) - \sum_{k=1}^{P_{short}} w_{s,k} x_m(n-k), \tag{19}$$

where $w_s = \{w_{s,1}, \ldots, w_{s,P_{short}}\}$ [13]. Having the source-to-microphone distances $\{q_{1,s}, \ldots, q_{M,s}\}$, the ideal distance weights are $\mathbf{q} = \{\frac{1}{q_{1,s}}, \ldots, \frac{1}{q_{M,s}}\}$. It is observed that using $\mathbf{q}$ as the weights significantly improves the autocorrelation function estimation in (18).

4.4. Beamforming

Beamforming, i.e., the process of focusing on a specific signal (based on the DOA or other characteristics), is widely used as part of multi-channel speech enhancement methods [6, 79]. Three beamforming techniques (listed below) have been applied to ad hoc microphone arrays. Generally, delay and sum beamforming (DSB) [8, 49] is more flexible and does not require limiting requirements, whereas more advanced beamforming methods assume some prior knowledge, which might not be the case for general scenarios.

4.4.1. Delay and Sum Beamforming (DSB)

This beamforming technique has been successfully applied to ad hoc microphone arrays [49]. Using $\boldsymbol{y}(n)$ (8), the DSB output is calculated as

$$\bar{x}_{DSB} = \sum_{m=1}^{M} x_m\left(n - \tau_{m,ref}\right), \tag{20}$$

where $\tau_{i,ref}$ (14) is the time delay between channel i and the reference channel. The beamformer filter coefficients are obtained by:

$$\widehat{W} = \arg\min_{w} \mathbf{w}^H \varphi_x \mathbf{w}, \tag{21}$$

where $\varphi_x = E\{xx^H\}$ is the covariance. The solution is

$$\widehat{\mathbf{W}} = \frac{\varphi_x{}^{-1}\mathbf{h}}{\mathbf{h}^H \varphi_x{}^{-1}\mathbf{h}}. \tag{22}$$

4.4.2. Minimum Variance Distortionless Response beamforming (MVDR)

An optimization method for the MVDR beamformer using the pseudo-coherence model of the array (24), based on the coherence function (20), is proposed and successfully tested by Tavakoli et al. [5]:

$$P_{X_m X_{ref}} = \frac{E\left[\mathrm{X}_m X_{ref}{}^*\right]}{E\left[\left|X_{ref}\right|^2\right]}, \quad (23)$$

$$X_m = \sum_{p=1}^{P} P_{x_m^p X_{ref}} X_{ref}^p, \quad (24)$$

where X_m is the frequency domain signal of $\mathbf{x_m}$ (1) and * is the complex conjugate. The MVDR beamformer requires knowledge of the source DOA and of the steering vector of the array when applied to compact microphone arrays; however, under certain assumptions, it is possible to modify the MVDR beamformer and apply it to the distributed scenarios. The assumptions include connection between the channels [80] and an ad hoc array with nodes of known geometry [5].

4.4.3. Linearly Constrained Minimum Variance (LCMV)

The LCMV was applied to distributed scenarios by Wood et al. [81] in experimental setups covering a wide range of random meeting scenarios. However, the LCMV beamformer requires knowledge of the RIR at each microphone location for each source in the meeting room. Himawan [82] proposed and successfully tested a clustered approach to blind beamforming for ad hoc microphone arrays, the applied features being the coherence between the diffuse noise and TDOA. The fact that the noise coherence between two microphones depends on the inter-microphone distance is exploited to estimate how close two microphones are. It is also well known that microphones located near each other have lower TDOAs, whereas microphones located farther away (i.e., metres) from each other have larger TDOAs.

4.5. Speech Enhancement

Speech enhancement can cover applications, such as noise cancellation [22, 83], beamforming [5], and echo cancellation [8]. These applications can be used separately or jointly as a combined speech enhancement method. The state-of-the-art speech enhancement techniques proposed for conventional arrays of known geometry are inapplicable to ad hoc microphone arrays, and existing approaches are confined to basic beamforming techniques [5, 49]. Some basic techniques apply centralized multi-channel Wiener filters and the so-called distributed adaptive node-specific signal estimation (DANSE) algorithm [22] to remove noise in distributed hearing aid systems. These methods assume that the channels can communicate and transmit time stamps. Other speech enhancement methods proposed for ad hoc arrays require limiting supervision requirements, such as user identification of the target speech [50].

A general scenario with *M* ad hoc microphones is rewritten as

$$\mathbf{x}(n) = \mathbf{h}(n) * s(n) + \mathbf{v}(n), \quad (25)$$

$$\mathbf{h}(n) = \begin{bmatrix} h_1(n) \\ \vdots \\ h_M(n) \end{bmatrix}, \quad (26)$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ (8) contains the multi-channel recording of all M microphones in the array, $\mathbf{h}(n)$ is the RIR matrix at each microphone's location for source j, and $\mathbf{v}(n)$ is the diffuse noise. The goal is to retrieve $s(n)$ from $\mathbf{x}(n)$(25). A clustered speech enhancement approach based on beamforming and auto-regressive (AR) modelling of the speech signal was proposed and tested by Pasha and Ritz [8]. They showed that removing the microphone located far from the active source and exclusively applying the multi-channel dereverberation method for the microphones located nearer the source improved the speech enhancement performance.

$$e_{x_m}(n) = x_m(n) - \sum_{k=1}^{p} b_k x_m(n-k). \quad (27)$$

where $x_m(n)$ (7) is the single-channel recording of an ad hoc microphone array and b_k is the LPC coefficient of order P. The kurtosis of the LP residual signals is then calculated as

$$k_{x_m}(n) = \frac{E\{e_{x_m}^4(n)\}}{E^2\{e_{x_m}^2(n)\}} - 3 \quad (28)$$

where $E\{\}$ denotes the mathematical expectation and $k_{x_m}(n)$ is the kurtosis of the LP residual signals.

4.6. Source Localization and DOA Estimation

DOA estimation refers to one-dimensional source localization (i.e., angle of arrival) [84], whereas source localization can have a more general meaning, such as pinpointing the source location in a room (i.e., two- or three-dimensional localization) [85]. In this section, DOA estimation and source localization are reviewed together as they both require location feature estimation [21].

Source localization methods proposed for microphone arrays are based on extracting location features from the recorded signals and analyzing them to localize the active source [7]. This approach has limiting assumptions when applied to ad hoc arrays; for instance, if the relative distance between the microphones is unknown, extracted features such as amplitude attenuation and time delays cannot be accurately translated to location features. Researchers have tried to address this issue for the gain feature by proposing a relative attenuation feature [12, 62] that can localize collocated sources and microphones when the microphones have different gains. It has also been shown that if some sources (i.e., three out of seven) are not collocated with any microphones, the method can still localize all the sources and microphones accurately.

It has been shown that arbitrarily arranged sensors (forming a network of acoustic sensors) can be effectively applied for source DOA estimation [13]. The results indicate that the proposed method detects the source angle of arrival in degrees with less than 2% error.

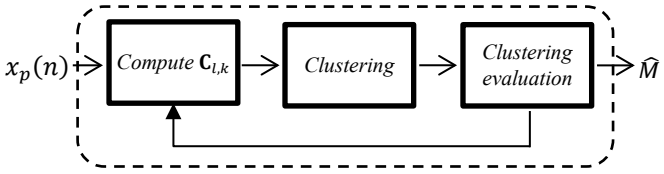


Figure 4: The proposed source counting system

Features derived from RIRs, such attenuation [54], and the clarity feature (C_{50}) [37], are also applied for the two-dimensional localization of sources.

4.7. Source Counting and crosstalk detection

Speech processing methods use voice activity detection (VAD) to detect the periods with an active speaker. In scenarios with more than one microphone, VAD can be applied in source counting (Figure 4) and multi-talk detection applications as well [9].

Inspired by VAD algorithms, researchers have proposed multi-talk detectors in which the source and microphone locations are not available (Figure 4). Moonen and Bertrand [86] suggested a multi-speaker voice activity detection method that tracks the power of multiple simultaneous speakers. Coherent-to-diffuse ratio (CDR) values (32) calculated or estimated at dual microphone node locations are also applied for source counting [56].

For two-element nodes where $\tilde{x}_{n,p}(t)$ (29) represents the signals recorded by the two channels, $p \in \{1,2\}$, at node $n \in \{1,\dots,N\}$, $\tilde{s}(t)$ is the source signal and $\tilde{h}_{n,p}(t)$ is the RIR at the nth node as in (6). The CDR features are calculated using the following:

$$\tilde{x}_{n,p}(t) = \tilde{s}(t) * \tilde{h}_{n,p}(t) + u_{n,p}(t), \tag{29}$$

$$C_{v_n}(f) = \frac{\left|\varphi_{v_{n,1}|v_{n,2}}(f)\right|^2}{\varphi_{v_{n,1}|v_{n,1}}(f)\,\varphi_{v_{n,2}|v_{n,2}}(f)}, \tag{30}$$

$$C_{x_n}(l,f) = \frac{\left|\varphi_{x_{n,1}|x_{n,2}}(l,f)\right|^2}{\varphi_{x_{n,1}|x_{n,1}}(l,f)\,\varphi_{x_{n,2}|x_{n,2}}(l,f)}, \tag{31}$$

where $C_{v_n}(f)$ and $C_{x_n}(f)$ are the noise and the signal magnitude squared coherence (MSC) values respectively and l represents the frame index. The CDR is

$$CDR_n(l,f) = \frac{C_{u_n}(f) - C_{x_n}(l,f)}{C_{x_n}(l,f) - C_s(l,f)}, \tag{32}$$

from which the use of the average CDR over the entire frequency band and L frames is given by

$$\overline{CDR}_n = \frac{1}{L(f_B - f_0)} \int_{f=f_0}^{f_B} \sum_{l=1}^{L} CDR_n(l,f)\ \mathrm{d}f. \tag{33}$$

The main limitation of the method proposed by Pasha et al. [56] is that all the nodes must be of the same structure, which limits the method's applicability. The MSC is found using the cross-power spectral density (CPSD) as presented by Pasha et al. [87] (Figure 4):

$$c(l,f) = \frac{\left|\varphi_{x_1|x_2}(l,f)\right|^2}{\varphi_{x_1|x_1}(l,f)\,\varphi_{x_2|x_2}(l,f)}, \tag{34}$$

where $f \in \{1,\dots,F\}$ is the frequency index of F total frequencies. The CPSD function used in (34) is defined as

$$\varphi_{x_1|x_2}(l,f) \triangleq \frac{1}{AK} \sum_{a,b} (x_1 \star x_2)(Al + a, b) e^{\frac{-j2\pi kb}{K}}, \tag{35}$$

where $a \in \{1,\dots,A\}$ is a frame index and $j = \sqrt{-1}$ represents the imaginary unit. The cross-correlation is calculated by:

$$(x_1 \star x_2)(\cdot, b) \triangleq \sum_n x_1(\cdot, n) x_2(\cdot, n+b), \tag{36}$$

where b is the displacement and $x_p(n)$ framed is $x_p(Al + a, n)$.

4.8. Source Separation

Crosstalk and speaker overlap decrease the signal quality and intelligibility in scenarios such as teleconferencing and meetings [88, 89].

The problem is mathematically formulated for M microphones and N sources (1) as

$$y_m(n) = \sum_j s_j(n) * h_{m,i,j}(n) \tag{37}$$

where $y_m(n)$ is the speech mixture recorded by the mth microphone from the array, $s_i(n)$ is the jth-source speech signal, and $h_{m,i,j}(n)$ is the RIR for the mth microphone and the jth source (1). The goal is to obtain $\mathbf{s}(n) = \{s_1(n),..,s_N(n)\}$ from $\mathbf{y}(n) = \{y_1(n),..,y_M(n)\}$.

Independent component analysis (ICA) was applied to blind speech separation in the online teleconferencing applications by Dmochowsky et al. [90]. Although the proposed ICA method is formulated for a general scenario, the experimental setup does not cover various scenarios and is limited to a two-element array. The noise and reverberation levels are low in the experimental setup and challenging scenarios (e.g. reverberation times higher than 800ms and very low SNRs) are not investigated.

More advanced sound source separation methods take into account the spatial coverage of the distributed arrays [91]. The signals obtained by the sub-arrays are then filtered by a geometric filter to achieve the highest output SIR. It is concluded that the proposed method can suppress (reject) interference by up to 40 dB in a reverberant environment. The novelty of this method is its use of passing and rejecting masks in the time-frequency domain to partition the microphones based on their power spectral density (PSD). The experimental setup is confined to one scenario with three sources and three microphones located near each other in pairs.

4.9. Speech Recognition

Speech recognition as a main aspect of human–machine interaction has attracted significant attention in recent years. Distributed microphone arrays have significant advantages over compact microphone arrays as they provide unobtrusive and spatially flexible interaction between humans and personal devices spread out within a room. A series of ad hoc signal processing techniques for speech recognition applications, such as spatial directivity, beamforming, and speech feature extraction, was discussed by Himawan [18], the main focus being on beamforming for speech enhancement. Generalized side lobe cancelling techniques [92] and linear prediction (LP)-based speech enhancement [79] have proven to be successful speech recognition methods in ad hoc scenarios.

5. Machine Learning Applied to Distributed Scenarios

Machine learning techniques have been widely used in different areas of speech and audio signal processing, such as emotion recognition and source localization [93]. Machine learning and data mining techniques have been shown to be effective for learning and predicting nonlinear patterns. They have been widely used for beamforming via support vector machines (SVMs), source localization via neural networks, and other applications. As machine learning techniques are highly sensitive to the training set and parameters, using them in the flexible and uncertain distributed scenarios is very challenging. However, researchers have managed to define informative discriminative features for clustering and classifying microphones and signals, features that are independent of any specific setup [8, 91] and can discriminate among the microphones within an ad hoc array regardless of array topology. These methods flexibly exclude a subset of the microphones (nodes) from the multi-channel process (e.g., multi-channel speech enhancement [8]) and are based on certain predefined selection criteria [5].

5.1. Microphone Clustering

Clustering is an unsupervised machine learning technique the goal of which is to assign objects (e.g., microphones) to groups with small intra-group differences and large inter-group differences [94]. The problem of clustering microphones based on their spatial locations was investigated by Gregen et al. [95]. It is important to cluster the microphones based on their spatial distances to select an optimal subset of microphones. Clustered speech processing approaches are applied to take advantage of the spatial selectivity of beamforming [49], dereverberation [8], and interference suppression [9]. The mel-frequency cepstral coefficients (MFCCs) [95], the coherence feature [3], the Legendre polynomial-based cepstral modulation ratio regression (LP-CMRARE) [95] (38)–(39), and the kurtosis of the LP residual [8] are used as features in various clustering methods reported in the literature. Along with the discriminative clustering feature, the clustering technique is important as it determines the limitations of the overall clustering method. For instance, the fuzzy clustering method presented by Gregen et al. [95] requires prior knowledge of the number of sources, whereas a flexible codebook-based clustering method based on RIRs was applied by Pasha et al. [11]:

$$\hat{X}_C = |\mathrm{F}(X_c)| \tag{38}$$

where X_c is the recorded signal in the cepstral domain and F denotes the Fourier transform. The average magnitude over the window length (C_T) of the modulation spectrum is calculated using

$$\bar{X}_c = \frac{1}{C_T}\sum_{c=0}^{C_T-1}|\hat{X}_C|, \tag{39}$$

as the clustering feature. A practical clustering method in which the numbers of microphones and clusters are not known was proposed by Pasha and Ritz [8], who applied clustering to exclude highly distorted and reverberant microphones from the dereverberation process. It is concluded that the clustered approach improves the direct-to-reverberant ratio (DRR) by 5 dB. Although it is impossible to evaluate the clustering methods decisively [95] when there is no ground truth, researchers have proposed clustering performance measures such as the success rate (SR) [11] and purity [49].

5.2. Signal Classification

Supervised machine learning methods such as classification are highly sensitive to the training set and setup, and are not always applicable to uncertain changeable scenarios such as ad hoc microphone arrays and meetings [96]. However, it has been shown that under certain assumptions (e.g., the availability of a clean training set) [95], it is possible to classify the recorded signals based on certain predefined classes (e.g., speech, noise, and music). It is concluded that cepstral features such as MFCC and LP-CMRARE are reliable features with which to discriminate speech, music, and noise signals [95].The microphone clustering method uses MFCC and LP_CMRARE as the features and divides microphones into two (i.e., the number of sources, which is assumed to be known) clusters. The signals recorded by the clustered microphones are then classified based on the predefined classes. It is also assumed that clean training data for each class are available [97].

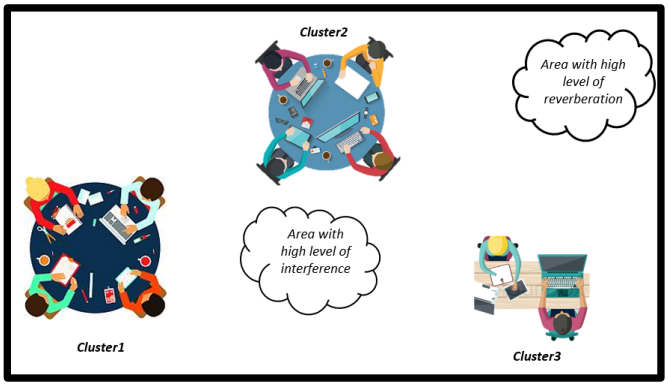


Figure 5: Microphone clusters [3]

6. Distributed Signal Processing Resources

An extensive database recorded using ad hoc microphones was reported and applied by Wood et al. [81] for real-world beamforming experiments. Twenty-four microphones were positioned in various locations on a central table in a reverberant room, and their outputs were recorded while four target talkers seated at the table read some text or had natural conversations. The recorded speech signals have been made publicly available [98]. Distributed beamforming tools and tutorials covering distributed speech enhancement and random microphone deployment have been made available [99]. Materials and resources associated with the distributed speech recognition (DSR) research conducted by IBM have been archived and sourced [100]. A coherence-based (31) source counting method for distributed scenarios has been made available by Donley [101]; the method counts the number of active speakers (up to eight) in a spontaneous meeting in reverberant environments.

University of Illinois have provided a dataset which facilitates distributed source separation and augmented listening research [102]. The dataset is recorded using 10 speech sources and 160 microphones in a large, reverberant conference room. The applied microphone array includes wearable sensors and microphones connected to tablets.

7. Conclusion

This review paper discussed recent advances in the context of distributed microphone arrays and signal processing. Standard dereverberation, speech separation, and source counting methods have been successfully adapted to the context of distributed signal processing using novel features and machine learning. Most existing ad hoc beamforming methods suffer from limiting assumptions that make them niche applications. Issues such as real-time source localization and DOA estimation are still challenging.

References

[1] S. Pasha, C. Ritz and J. Lundgren, "A Survey on Ad Hoc Signal Processing: Applications, Challenges and State-of-the-Art Techniques," in *International Symposium on Signal Processing and Information Technology (ISSPIT)*, Ajman, United Arab Emirates, 2019, DOI:10.1109/ISSPIT47144.2019.9001860.

[2] M. J. Taghizadeh, Enabling speech applications using ad hoc microphone arrays: PhD dissertation, Lausanne: EPFL, 2015,https://publications.idiap.ch/downloads/papers/2015/Taghizadeh_THESIS_2015.pdf.

[3] S. Pasha, Analysis and Enhancement of Spatial Sound Scenes Recorded using Ad-Hoc Microphone Arrays, Doctor of Philosophy thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2017, https://ro.uow.edu.au/theses1/450.

[4] J. Benesty, J. Chen and J. Huang, Microphone Array Signal Processing, Verlag Berlin Heidelberg: Springer, 2008, DOI: 10.1007/978-3-540-78612-2.

[5] V. M. Tavakoli, J. R. Jensen, M. Christenseny and J. Benesty, "Pseudo-coherence-based MVDR beamformer for speech enhancement with ad hoc microphone arrays," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, 2015, DOI: 10.1109/ICASSP.2015.7178453.

[6] N. D. Gaubitch, J. Martinez, W. B. Kleijn and R. Heusdens, "On near-field beamforming with smartphone-based ad-hoc microphone arrays," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014, DOI: 10.1109/IWAENC.2014.6953345.

[7] M. Souden, K. Kinoshita, M. Delcroix and T. Nakatani, "Location Feature Integration for Clustering-Based Speech Separation in Distributed Microphone Arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 22(2), 354-367, 2014, DOI: 10.1109/TASLP.2013.2292308.

[8] S. Pasha and C. Ritz, "Clustered multi-channel dereverberation for ad-hoc microphone arrays," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Hong Kong, 2015, DOI: 10.1109/APSIPA.2015.7415519.

[9] S. Gregen and R. Martin, "Estimating Source Dominated Microphone Clusters in Ad-Hoc Microphone Arrays by Fuzzy Clustering in the Feature Space," in *Speech Communication; 12. ITG Symposium*, Paderborn, Germany, 2016, Print ISBN: 978-3-8007-4275-2.

[10] C. Evers, Y. Dorfan, S. Gannot and P. Naylor, "Source tracking using moving microphone arrays for robot audition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, DOI: 10.1109/ICASSP.2017.7953337.

[11] S. Pasha, Y. Zou and C. Ritz, "Forming ad-hoc microphone arrays through clustering of acoustic room impulse responses," in *China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, Chengdu, 2015, DOI: 10.1109/ChinaSIP.2015.7230367.

[12] Z. Liu, Z. Zhang, L. W. He and P. Chou, "Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays," in *International Conference on Acoustics, Speech and Signal Processing*, Honolulu, 2007, DOI: 10.1109/ICASSP.2007.366347.

[13] S. Araki, H. Swada, R. Mukai and S. Makino, "DOA Estimation for Multiple Sparse Sources with Arbitrarily Arranged Multiple Sensors," *Journal of Signal Processing Systems,* 63(3), 265–275, 2011, https://doi.org/10.1007/s11265-009-0413-9.

[14] L. Krishnan, P. Teal and T. Betlehem, "A robust sparse approach to acoustic impulse response shaping," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, DOI: 10.1109/ICASSP.2015.7178067.

[15] T. v. Waterschoot, "Distributed estimation of cross-correlation functions in ad-hoc microphone arrays," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, 2015, DOI: 10.1109/EUSIPCO.2015.7362385.

[16] V. M. Tavakoli, J. R. Jensen, M. G. Christensen and J. Benesty, "A Framework for Speech Enhancement With Ad Hoc Microphone Arrays," *ACM Transactions on Audio, Speech, and Language Processing,* 24(6), 1038-1051, June 2016, DOI: 10.1109/TASLP.2016.2537202.

[17] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *18th Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Ghent, 2011, DOI: 10.1109/SCVT.2011.6101302.

[18] I. Himawan, Speech recognition using ad-hoc microphone arrays, PhD dissertation, Queensland university of technology, 2010, https://pdfs.semanticscholar.org/e95b/ad880c47be6e146dd75b82bc72e51e97db2f.pdf.

[19] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He and A. Colburn, "Distributed meetings: A meeting capture and broadcasting system," in *ACM Multimedia*, Juan-les-Pins, France, 2002, DOI: 10.1109/ICASSP.2003.1202753.

[20] V. Tavakoli, J. Jensen, R. Heusdens, J. Benesti and M. Christensen, "Distributed max-SINR speech enhancement with ad hoc microphone arrays," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, DOI: 10.1109/ICASSP.2017.7952136.

[21] A. Asaei, H. Bourlard, M. Taghizadeh and V. Cehver, "Model-based sparse component analysis for reverberant speech localization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, DOI: 10.1109/ICASSP.2014.6853835.

[22] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on*

*Advances in Signal Processing(*14, 2009, https://doi.org/10.1155/2009/530435.

[23] Y. Jia, L. Yu and I. Kozintsev, "Distributed Microphone Arrays for Digital Home and Office," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, 2006, DOI: 10.1109/ICASSP.2006.1661463.

[24] Diane J. Cooka, Juan C. Augustob and Vikramadit, "Ambient intelligence: Technologies, applications, and opportunities," *Pervasive and Mobile Computing,* 5, 277–298, August 2009, https://doi.org/10.1016/j.pmcj.2009.04.001.

[25] A. Brutti, M. Omologo and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *NTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, Corpus ID: 2927411.

[26] J. Lee, C. H. Lee, D. W. Kim and B. Y. Kang, "Smartphone-Assisted Pronunciation Learning Technique for Ambient Intelligence," *IEEE Access,* 2016, DOI: 10.1109/ACCESS.2016.2641474.

[27] T. Toyoda, N. Ono, S. Miyabe, T. Yamada and S. Makino, "Traffic monitoring with ad-hoc microphone array," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014, DOI: 10.1109/IWAENC.2014.6954310.

[28] K. Tsuji, "Migration monitoring of fin whales in the southern Chukchi Sea with acoustic methods during 2012–2015," in *Techno-Ocean (Techno-Ocean)*, Kobe, 2016, DOI: 10.1109/Techno-Ocean.2016.7890746.

[29] D. Stowell, E. Benetos and L. Gill, "On-Bird Sound Recordings: Automatic Acoustic Recognition of Activities and Contexts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 25(6), 1193-1206, June 2017, DOI: 10.1109/TASLP.2017.2690565.

[30] A. M. McKee and R. A. Goubran, "Chest sound pick-up using a multisensor array," in *SENSORS*, Irvine, CA, 2005, DOI: 10.1109/ICSENS.2005.1597816.

[31] J. Chien, M. Huang, Y. Lin and F. Chong, "A Study of Heart Sound and Lung Sound Separation by Independent Component Analysis Technique," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY, 2006, DOI: 10.1109/IEMBS.2006.260223.

[32] M. krekovic, I. Dokmanic and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, DOI: 10.1109/ICASSP.2016.7471627.

[33] I. Dokmanic, L. Daudet and M. Vetterli, "From acoustic room reconstruction to slam," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, DOI: 10.1109/ICASSP.2016.7472898.

[34] I. Dokmanic, L. Daudet and M. Vetterli, "How to localize ten microphones in one finger snap," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, 2014, Electronic ISBN: 978-0-9928-6261-9.

[35] F. Antonacci, "Inference of Room Geometry From Acoustic Impulse Responses," *IEEE Transactions on Audio, Speech, and Language Processing,* 20(10), 2683-2695, 2012, DOI: 10.1109/TASL.2012.2210877.

[36] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda and P. Naylor, "A Single-Channel Non-Intrusive C50 Estimator Correlated With Speech Recognition Performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 24(4), 719-732, April 2016, DOI: 10.1109/TASLP.2016.2521486.

[37] S. Pasha, C. Ritz and Y. X. Zou, "Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, DOI: 10.1109/APSIPA.2016.7820873.

[38] T. K. Hon, L. Wang, J. D. Reiss and A. Cavallaro, "Fine landmark-based synchronization of ad-hoc microphone arrays," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, 2015, DOI: 10.1109/EUSIPCO.2015.7362600.

[39] S. Araki, N. Ono, K. Kinoshita and M. Delcroix, "Estimation of Sampling Frequency Mismatch between Distributed Asynchronous Microphones under Existence of Source Movements with Stationary Time Periods Detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019), 785-789, DOI: 10.1109/ICASSP.2019.8683192.

[40] S. Pazors, M. Hurtado and C. Muravchik, "On Sparse Methods for Array Signal Processing in the Presence of Interference," *Antennas and Wireless Propagation Letters,* 14, 1165-1168, 2015, DOI: 10.1109/LAWP.2015.2394233.

[41] M. Taseska, S. Markovich-golan, E. Habets and S. Gannot, "Near-field source extraction using speech presence probabilities for ad hoc microphone arrays," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014, DOI: 10.1109/IWAENC.2014.6954000.

[42] A. D. Sarwate and K. Chaudhuri, "Signal Processing and Machine Learning with Differential Privacy: Algorithms and Challenges for Continuous Data," *IEEE Signal Processing Magazine,* 30(5), 86-94, Sept. 2013, DOI: 10.1109/MSP.2013.2259911.

[43] R. Wehrmann, "Concepts of improving hands-free speech communication," in *IEEE International Symposium on Circuits and Systems*, San Diego, CA, 1992, DOI: 10.1109/ISCAS.1992.230435.

[44] P. Chevalier and A. Maurice, "Blind and informed cyclic array processing for cyclostationary signals," in *9th European Signal Processing Conference (EUSIPCO 1998)*, Rhodes, 1998, Print ISBN: 978-960-7620-06-4.

[45] E. Vincent, N. Bertin, R. Griboval and F. Bimbot, "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *EEE Signal Processing Magazine,* 31(3), 107-11, May 2014, DOI: 10.1109/MSP.2013.2297440.

[46] C. Vanwynsberhea, P. Challande, J. Marchal, R. marchiano and F. Ollivier, "A robust and passive method for geometric calibration of large arrays," *The Journal of the Acoustical Society of America,* 139, p. 1252, 2016, https://doi.org/10.1121/1.4944566.

[47] I. McCowan, M. Lincoln and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *Transactions on Audio,Speech and Language Processing,* 16(3), 666-670, 2008, DOI: 10.1109/TASL.2007.911428.

[48] D. McCarthy and F. Boland, "A method for source-microphone range estimation, using arrays of unknown geometry, in reverberant room environments," in *15th European Signal Processing Conference*, Poznan, 2007, Print ISBN: 978-839-2134-04-6.

[49] I. Himawan, I. McCowan and S. Sridhan, "Clustered Blind Beamforming From Ad-Hoc Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing,* 19(4), 661-676, May 2011, DOI: 10.1109/TASL.2010.2055560.

[50] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Kaohsiung, 2013, DOI: 10.1109/APSIPA.2013.6694323.

[51] I. Dokmanic, J. Raineri and M. Vetterli, "Relax and unfold: Microphone localization with Euclidean distance matrices," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, 2015, DOI: 10.1109/EUSIPCO.2015.7362386.

[52] Y. Tian, Z. Chen and F. Yin, "Distributed Kalman filter-based speaker tracking in microphone array networks," *Applied Acoustics,* 89), 71-77, 2015, https://doi.org/10.1016/j.apacoust.2014.09.004.

[53] A. Asaei, N. Mohamadiha, M. J. Taghizadeh, S. Doclo and H. Bourlard, "On application of non-negative matrix factorization for ad hoc microphone array calibration from incomplete noisy distances," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, 2015, DOI: 10.1109/ICASSP.2015.7178460.

[54] S. Pasha and C. Ritz, "Informed source location and DOA estimation using acoustic room impulse response parameters," in *International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, 2015, DOI: 10.1109/ISSPIT.2015.7394316.

[55] I. Dokmanic, Y. Lu and M. Vetterli, "Can one hear the shape of a room: The 2-D polygonal case," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, DOI: 10.1109/ICASSP.2011.5946405.

[56] S. Pasha, J. Donley, C. Ritz and X. Y. Zou, "Towards real-time source counting by estimation of coherent-to-diffuse ratios from ad-hoc microphone array recordings," in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, 2017, DOI: 10.1109/HSCMA.2017.7895582.

[57] M. Taseska and A. P. Habets, "Spotforming using distributed microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2013, DOI: 10.1109/WASPAA.2013.6701876.

[58] S. Woźniak and K. Kowalczyk, "Passive Joint Localization and Synchronization of Distributed Microphone Arrays," *Signal Processing Letters,* 26(2), 292-296, Feb. 2019, DOI: 10.1109/LSP.2018.2889438.

[59] Q. Zhang, W. Zhang, J. Feng and R. Tang, "Distributed Acoustic Source Tracking in Noisy and Reverberant Environments With Distributed Microphone Networks," *IEEE Access,* 8), 9913-9927, 2020, DOI: 10.1109/ACCESS.2020.2965210.

[60] S. Vesa, "Binaural Sound Source Distance Learning in Rooms," *Transactions on Audio, Speech, and Language Processing,* 17(8), 1498-1507, 2009, DOI: 10.1109/TASL.2009.2022001.

[61] I. Himawan, S. Sridharan and I. McCowan, "Dealing with uncertainty in microphone placement in a microphone array speech recognition system," in *International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 2008, DOI: 10.1109/ICASSP.2008.4517922.

[62] M. Chen, Z. Liu, W. He, P. Chou and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, 2007, DOI: 10.1109/ASPAA.2007.4393035.

[63] M. J. Taghizadeh, R. Parhizkar, P. N. Garner and H. Bourlard, "Euclidean distance matrix completion for ad-hoc microphone array calibration," in *18th International Conference on Digital Signal Processing (DSP)*, Fira, 2013, DOI: 10.1109/HSCMA.2014.6843239.

[64] M. J. Taghizadeh, N. Carner, H. Bourlard and A. Asaei, "Ad Hoc Microphone Array Calibration: Euclidean Distance Matrix Completion Algorithm and Theoretical Guarantees," *Signal Processing,* p. https://doi.org/10.1016/j.sigpro.2014.07.016, 2014.

[65] M. Taghizadeh, A. Asaei, P. Garner and h. Bourlard, "Ad-hoc microphone array calibration from partial distance measurements," in *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Villers-les-Nancy, 2014, DOI: 10.1109/HSCMA.2014.6843239.

[66] N. Ono, K. Shibata and H. Kameoka, "Self-localization and channel synchronization of smartphone arrays using sound emissions," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, DOI: 10.1109/APSIPA.2016.7820778.

[67] L. Wang, T. K. Hon and A. Cavallaro, "Self-Localization of Ad-Hoc Arrays Using Time Difference of Arrivals," *IEEE Transactions on Signal Processing,* 64(4), 1018-1033, Feb.15, 2016, DOI: 10.1109/TSP.2015.2498130.

[68] N. D. Gaubitch, W. B. Kleijn and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, DOI: 10.1109/ICASSP.2013.6637618.

[69] E. Robledo, T. S. Wada and B. Juang, "On Dealing with Sampling Rate Mismatches in Blind Source Separation and Acoustic Echo Cancellation," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2007, DOI: 10.1109/ASPAA.2007.4393044.

[70] H. Chiba, N. Ono, S. Miyabe, Y. Takashaahi, T. Yamada and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan les pins, France, 2014, DOI: 10.1109/IWAENC.2014.6954007.

[71] S. Miyabe, N. Ono and S. makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing,* p. 185–196, 2015, https://doi.org/10.1016/j.sigpro.2014.09.015.

[72] N. Ono, H. Kohno, N. Ito and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2009, DOI: 10.1109/ASPAA.2009.5346505.

[73] P. Pertila, M. S. Hamalainen and M. Mieskolainen, "Passive Temporal Offset Estimation of Multichannel Recordings of an Ad-Hoc Microphone Array," *IEEE Transactions on Audio, Speech, and Language Processing,* 21(11), 2393-2402, 2013, DOI: 10.1109/TASLP.2013.2286921.

[74] N. J. Bryan, P. Smargadis and G. J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012, DOI: 10.1109/ICASSP.2012.6288396.

[75] R. Lienhart, I. Kozintsev, S. Wher and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 2003, DOI: 10.1109/ASPAA.2003.1285842.

[76] T. Nakatani, T. Yoshioka, K. Kinoshita and M. Miyoshi, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, 2008, DOI: 10.1109/ICASSP.2008.4517552.

[77] S. Pasha, C. Ritz and Y. X. Zou, "Spatial multi-channel linear prediction for dereverberation of ad-hoc microphones," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, 2017, DOI: 10.1109/APSIPA.2017.8282306.

[78] S. Pasha and J. Lundgren, "Multi-Channel Compression and Coding of Reverberant Ad-Hoc Recordings Through Spatial Autoregressive Modelling," in *30th Irish Signals and Systems Conference (ISSC)*, Maynooth, Ireland, 2019, DOI: 10.1109/ISSC.2019.8904918.

[79] M. Delcroix, T. Yosghioka, A. Ogawa and Y. Kubo, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," *Proceedings of REVERB Workshop ,* 2014.

[80] V. M. Tavakoli, J. R. Jensen, R. Heusdens, J. Benesty and M. G. Christensen, "Ad hoc microphone array beamforming using the primal-dual method of multipliers," in *24th European Signal Processing Conference (EUSIPCO)*, Budapest, 2016, DOI: 10.1109/EUSIPCO.2016.7760416.

[81] W. S. Wood, E. Hadad, I. Merks, B. Xu and S. Gannot, "A real-world recording database for ad hoc microphone arrays," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2015, DOI: 10.1109/WASPAA.2015.7336915.

[82] I. Himawan, I. McCowan and S. Sridharan, "Clustering of ad-hoc microphone arrays for robust blind beamforming," in *International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, DOI: 10.1109/ICASSP.2010.5496201.

[83] T. C. Lawin, S. Stenzel, J. Freudenberger and S. Doclo, "Generalized Multichannel Wiener Filter for Spatially Distributed Microphones," in *Speech Communication; 11. ITG Symposium*, Erlangen, Germany, 2014, Print ISBN: 978-3-8007-3640-9.

[84] T. Ballal and C. J. Bleakley, "DOA estimation of multiple sparse sources using three widely-spaced sensors," in *17th European Signal Processing Conference*, Glasgow, 2009, Print ISBN: 978-161-7388-76-7.

[85] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Edinburgh, 2011, DOI: 10.1109/HSCMA.2011.5942378.

[86] A. Moonen and M. Bertrand, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, DOI: 10.1109/ICASSP.2010.5496183.

[87] S. Pasha, J. Donley and C. Ritz, "Blind speaker counting in highly reverberant environments by clustering coherence features," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, 2017, DOI: 10.1109/APSIPA.2017.8282303.

[88] Z. Liu, "Sound source separation with distributed microphone phone arrays in the presence of clock synchronization errors," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, 2008.

[89] S. Araki, N. Ono, K. Kinoshita and M. Delcroix, "Projection Back onto Filtered Observations for Speech Separation with Distributed Microphone Array," in *8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Le gosier, Guadelope, 2019, DOI: 10.1109/CAMSAP45676.2019.9022666.

[90] J. Dmochowsky, Z. Liu and P. Chou, "Blind source separation in a distributed microphone meeting environment for improved teleconferencing," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Las Vegas, NV, USA, 2008, DOI: 10.1109/ICASSP.2008.4517553.

[91] V. M. Tavakoli, J. R. Jensen, J. Benesty and M. G. Christensen, "A partitioned approach to signal separation with microphone ad hoc arrays," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, DOI: 10.1109/ICASSP.2016.7472272.

[92] S. Golan, S. Gannot and I. Cohen, "Distributed GSC beamforming using the relative transfer function," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, 2012, Print ISBN: 978-1-4673-1068-0.

[93] D. Salvati, C. Drioli and L. Foresti, "On the use of machine learning in microphone array beamforming for far-field sound source localization," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Vietri sul Mare, 2016, DOI: 10.1109/MLSP.2016.7738899.

[94] B. Clarke, E. Fokoue and H. Zhang, Principles and Theory for Data Mining and Machine Learning, Springer-Verlag New York Inc, 1st August 2009, https://doi.org/10.1007/978-0-387-98135-2.

[95] S. Gregen, A. Nagathil and R. Martin, "Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Vancouver, BC, 2013, DOI: 10.1109/ICASSP.2013.6638347.

[96] "IEEE DCASE challenge," [Online]. Available: http://dcase.community/. [Accessed 30 3 2020].

[97] S. Gergen and R. Martin, "Linear Combining of Audio Features for Signal Classification in Ad-hoc Microphone Arrays," in *Speech Communication; 11. ITG Symposium*, Erlangen, Germany, 2014.

[98] "Ad-hoc recording signals," http://www.eng.biu.ac.il/gannot/speech-enhancement

[99] "Prof. Gannot speech enhancement tutorial," http://www.eng.biu.ac.il/gannot/tutorials-and-keynote-addresses/introduction-to-distributed-speech-enhancement-algorithms-for-ad-hoc-microphone-arrays-wireless-acoustic-sensor-networks/.

[100] IBM, "Distributed Speech Recognition," https://www.ibm.com/blogs/research/2019/10/asr-deep-learning/.

[101] J. Donley. https://github.com/JacobD10/CoherenceBasedSourceCounter.

[102] R. M. Corey, M. D. Skarha and A. C. Singer, Composers, *Massive Distributed Microphone Array Dataset.* [Sound Recording]. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-6216881_V1 . 2019.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 331-340 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Contribution of Wind Energy Capacity on Generation Systems Adequacy Reliability using Differential Evolution Optimization Algorithm

Athraa Ali Kadhem[*,1], Noor Izzri Abdul Wahab[1], Ahmed Abdalla[2]

[1]Department of Electrical and Electronic Engineering, University Putra Malaysia, Selangor 43400, Malaysia

[2]Faculty of Electronics Information Engineering, Huaiyin Institute of Technology, Huai'an 223003, China

ARTICLE INFO

Article history:
Received: 07 May, 2020
Accepted: 30 October, 2020
Online: 20 November, 2020

Keywords:
Optimization Algorithms
Intelligent Search
Adequacy Systems

ABSTRACT

Currently, increasing penetration of the Wind Energy Conversion System (WECs) in Power generation systems has influenced the supply of electrical power reliability for the consumer in comparison with other traditional sources. In this paper, the performance and efficiency of a new optimization approach referred to as the "Differential Evolution Optimization Algorithm" (DEOA) to measure the reliability of power adequacy systems (RPAs). The proposed intelligent algorithm which relies on the Population-based Intelligent Search (PBIs) technique is viewed as a feasible alternative for the Monte Carlo simulation (MCS) method in the assessment of the non-chronological system. The benefit of utilizing this algorithm is apparent in the manner it expedites the calculation and achieves greater precision with less calculation effort. Additionally, there is a deeper understanding of the effect of the increasing levels of wind energy on generation adequacy from the WECs sources to satisfy future power electricity demands. In addition, the effectiveness of the suggested algorithm in assessing the RPAs was compared to the analytical and MCS method.

1. Introduction

The inconsistent nature of wind energy presents challenges to the planners and operation personnel of power plants. Therefore, it is now crucially important to assess the extent that power impacts the dependability and sufficiency of the power being generated by existing plants. In assessing the penetration of the wind power from the WECs based on the nature of the wind can be complicated as it involves a number of critical issues that have to be taken into account when designing the WECs, such as the speed and power of the wind. To design a method for estimating wind production necessitates a massive amount of historical data of wind velocity so that the stochastic nature of the wind at a particular location can be accurately simulated. Another possible option could be to use a dependable stochastic simulation technique for the generation of the time series artificial of the wind speed.

Analyzing how adequate the PS is possible through the identification and evaluation of failure states, which are situations when customer demand for power cannot be met by the power system, hence, making it necessary to practice load shedding to ensure the maintenance of the system integrity. Due to the fact that there are numerous potential failure situations [1], [2], it may not be practicable to directly count and estimate it all even in cases of medium-size power systems.

Currently, RE sources usually combined with the conventional electrical grid for generating less polluting energy; as for example WTGs [3]. As a result, the estimation of the reliability index for unconcentrated resources such as wind energy is complicated [4]. As such, new tools need to be used to assess PS with the wide presence of wind energy sources [5].

Evaluating the dependability of PGS is typically performed by simulation or analysis approaches. The majority of authors in [6] focused on assessing the dependability of PGS in wind farms' basis of analytic approaches. The MCS method offers an accurate estimate of the dependability indices. Therefore, the MCS would be beneficial for such a purpose, but it requires considerable computational work which is time-consuming and should be avoided if convergence efficiency is critical [7]. This research

[*]Corresponding Author: Athraa Ali Kadhem, University Putra Malaysia Email: athraaonoz2007@yahoo.com

offers a replacement for the MCS and PBIs method [8],[9]. This optional approach method is used to identify a num ber of possible failure situations to facilitate the adequacy indices of the system.

Numerous PBIs approaches are possible for intelligent research, epitome for these algorithms like the BPSO, ISSPs, OGA, EC, BPSO, and BCA [10]–[13]. Although some research has been done on the dependability of the generating systems, there is still a need for approaches that are more appropriate and which offer greater computational flexibility and are more realistic in representing the desired PGS adequacy [14]–[17].

Models are needed for the simulation of the stochastic characteristic of PGS behavior, which is not a new issue for population-based intelligent research approaches. However, it has emerged as an important issue when wind sources and the PS are combined and taken into account. This paper suggested a novel algorithm known as "DEOA" to lower the time amount cost of the computation to simulation pattern in the non-SMCS and other Optimization Algorithms by enhancing the checking efficiency in a given area state-space employing a PBIs approach. DEOA is based on the meta- heuristics searching employed in simulating the generating system's operation and taking into account the not expecting failure of current systems and the instability of "WECs" energy. DEOA exhibited benefits in the form of higher computational speed, utilizing a relatively uncomplicated structure, providing ease-of-use, and was shown to be highly robust. Besides, the numerically simulated solution was compared with the MCS-related Work. It also explains the impacts of wind energy capacity from the WECs on power systems to attain at the level of the incremental peak load capability.

Nomenclature

WECs	Wind Energy Conversion System
RPAs	Reliability of Power Adequacy Systems
PBIs	Population-Based Intelligent Search
DEOA	Differential Evolution Optimization Algorithm
MCS	Monte Carlo simulation
SMCS	Sequential Monte Carlo Simulation
PS	Power System
RE	Renewable Energy
WTGs	Wind Turbine Generators
PGS	Power Generation Systems
ISSP	Intelligent State Space Pruning
OGA	Optimization Genetic Algorithm
EC	Evolutionary Computation
PSO	Particle Swarm Optimization
BCA	Bee Colony Algorithm
LOLE	Loss of Load Expectation
LOEE	Loss of Energy Expectation
LOLF	Loss of Load Frequency
LDC	Load Duration Curve
RI	Reliability Indices
LH	load levels
PNS	Power Not Supplied
LOLP	Loss of Load Probability
WDM	Weibull distribution model
WPG	Wind Power Generation
BPSO	Binary Particle Swarm Optimizations
MSGA	Modified Simple Genetic Algorithm
RTS	Reliability Test System

2. Assistant Theoretical Materials

2.1. Assessment of Conventional Generation Systems Adequacy

Analyzing the reliability of electrical energy systems is important to determine PGS adequacy and for possible future increase in the system capacity. This is to ensure adequate availability of power to satisfy consumer requirements for particular installed system capacity. As a result, the could be utilized as an element of the quantitative conditions in a power generation system in planning for the system's risk paradigm. In this study, the essential evaluated were utilized in facilitating the approximation of the PS dependability level of their "Loss of Load Frequency (LOLF)", "Loss of Energy Expectation (LOEE)", and "Loss of Load Expectation (LOLE)".

Load designs system are normally non-sequential or sequential paradigms that can be used with various Optimization Algorithms approaches. The sequential Load Duration Curve scheme (" LDC") can, for epitome, can create values of load demand hourly states, the implication being that around "8736 Hours" separate states can be logged per annum [18]. The annual LDC approach for IEEE-RTS-79 - 96 is presented in Figure 1.

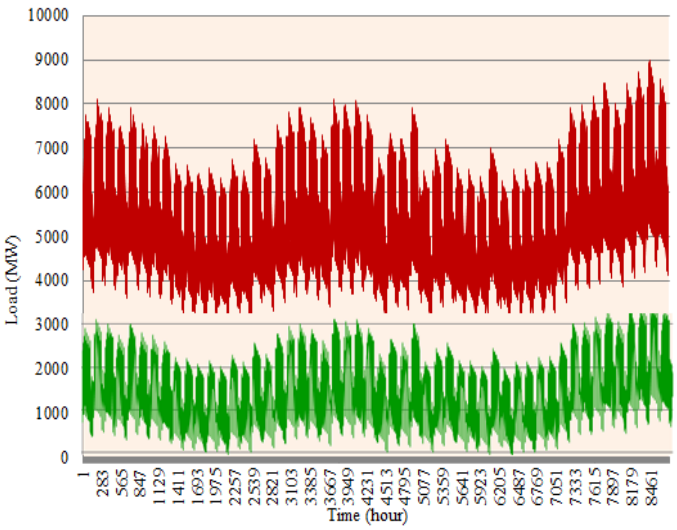


Figure 1: Annual LDC model for the RTS

To attain elevated levels of RPAs and the PGS valuation, the best design of RI has an important role in designing and developing a successful generation system [19]. Estimates of the LOLE, LOEE, and LOLF are made based on the attained capacity states' collections and load characteristics. In this study, denoted the discrete states for the load levels (LH) at a time (t). The "Loss of Load Probability" (LOLP) is estimated as equation (1):

$$LOLP(LH_i) = \sum_{j=1}^{sa} S_j \cdot P_j \cdot Copy_j \quad (1)$$

where:

sa = total number of state arrays, and

Sj = status of the system state.

In a failure state, the status value will=1; while in a successful state, the status will=0, this refers to the status of the load demand satisfaction [20]. equation (2) displayed the values of the LOLE annually:

$$LOLE = \sum_{j=1}^{8736} LOLP(LH_j) \quad (2)$$

Estimated the megawatts of the PNS by used equation (3):

$$PNS(LH_i) = \sum_{j=1}^{sa} S_j \cdot P_j \cdot Copy_j \cdot (LH_i - Cap_j) \quad (3)$$

$$LOEE = \sum_{j=1}^{8736} PNS(LH_j) \quad (4)$$

Equations employed to calculate the components of the LOLF index [19] as:

$$LOLF(LH_i) = \sum_{j=1}^{sa} S_j \cdot FG_j \cdot Copy_j \quad (5)$$

$$FG = \sum_{j=1}^{8736} LOLF(LH_j) \quad (6)$$

$$FL = \sum_{j=1}^{8736} V_j \cdot \left[LOLP(LH_j) - LOLP(LH_{j-1})\right] \quad (7)$$

$$LOLE = [FG_{(LH)}] + [FL_{(LH)}] \quad (8)$$

2.2. Wind Speed Paradigm

Simulating of wind speed/velocity to study WECs performance and dependability was done and reported by [21]. The sequential WECs simulation model involves the generation of hourly wind speed data over a given perio and specified site. The main challenge in maintaining the reliability of PGS is the fluctuation nature of wind power. It is therefore necessary to develop an appropriate wind velocity model before proceeding to analyze the system reliability.

The majority of research recently done dependent on the Weibull distribution model (WDM) to signify the hourly fluctuations in the annual mean wind velocity. The level of wind turbine generated power at a given wind site depends on the mean and standard variation of the wind velocity. Based on the assumption that the WDM is able to change particular variables such as the shape (b) and the scale (a), it is normally used for simulating wind speed fluctuations. Thus, the inverse transform function for the WDM in [22] in equation (9) can be employed to simulate the wind speed s.

$$s = a\left[-Ln(H)^{1/b}\right] \quad (9)$$

where; a and b = two parameters for the WDM;

H = random variable;

s = wind speed.

By applying the wind speed that generated by using the equation (9) to the wind turbine units output model, the turbine units power output is as shown in equation (10).

$$P_{WTG} = \begin{cases} 0 & ws < V_{ci} \\ P_r \times (A + B_x + Cx^2) & V_{ci} \le ws < V_r \\ P_r & V_r \le ws < V_{co} \\ 0 & ws > V_{co} \end{cases} \quad (10)$$

where

ws = wind speed (m/s),

Vci = WTG cut-in speed (m/s),

Vco = WTG cut-out speed (m/s),

Vr = WTG rated speed (m/s),

Pr = WTG rated power output (MW).

The constants A, Bx, and Cx have previously been calculated by [22].

3. Methodology

Two programs based on the simulation method have been developed using MATLAB. The first simulation program was developed to simulate the variations of wind speed using WDM, which was utilized for an estimation of the parameters “a” and “b”, that were deuced form wind speed data. Consequently, during the simulation process, the two parameters (a, b) will generate wind speed variations at any time. Additionally, the program can estimate a wind power result from the wind turbine. Figure 2 shows models for wind speed generation and output power simulation. The second simulation program is based on the DEOA technique to evaluate the dependability of the PGS and, wind power penetration consideration is incorporated in the generating capacity adequacy evaluation. According to the steps taken as stated in reference [22], the RI was calculated utilizing DEOA algorithm as shown in Figure 3.

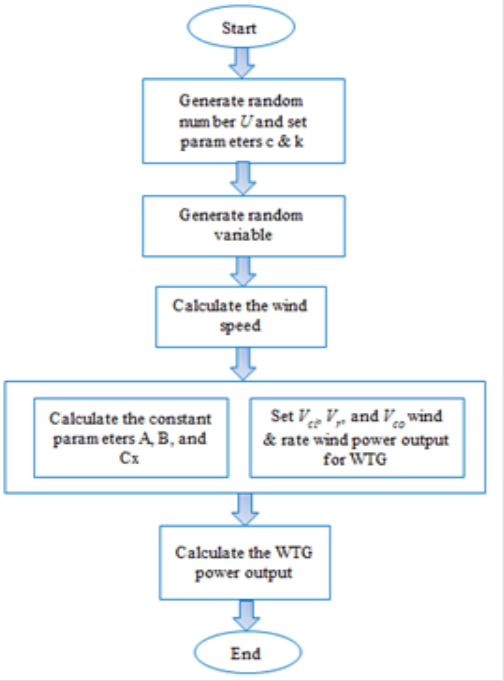


Figure 2: Diagram of speed and wind power for the turbine unit

4. Outcomes Analysis

4.1. Simulation Outcomes of Speed & Power Wind

This study explains the possibility to generate wind energy at different levels which parallel the probability of wind speed fluctuation, therefore, the wind power model can generate wind energy with different abilities states as instant at hourly. One of the features of the wind speed is that the turbine output power is determined subsequently when the hourly wind speed is obtained. The simulation of the profiles of the speed of the wind was employed to simulate ability states from the simulated generation of wind output power.

There is a need to examine the power output from the wind farm, to obtain a better picture of the selected site so that the WDM parameter, a, is set as the mean wind speed and b=2, is the usual wind characteristic [23]. The particular WTG characteristics employed here are the 4, 25, and 19 m/s (cut-in-speed, cut-out-speed, and rate speed) and at a rated power of 2 MW [22]. Figure 4 shows the forecast for hourly wind speed over a year, and also indicates the simulation wind power output for WTG (with 85 WTG) with a power rate of 2 MW for a similar period.

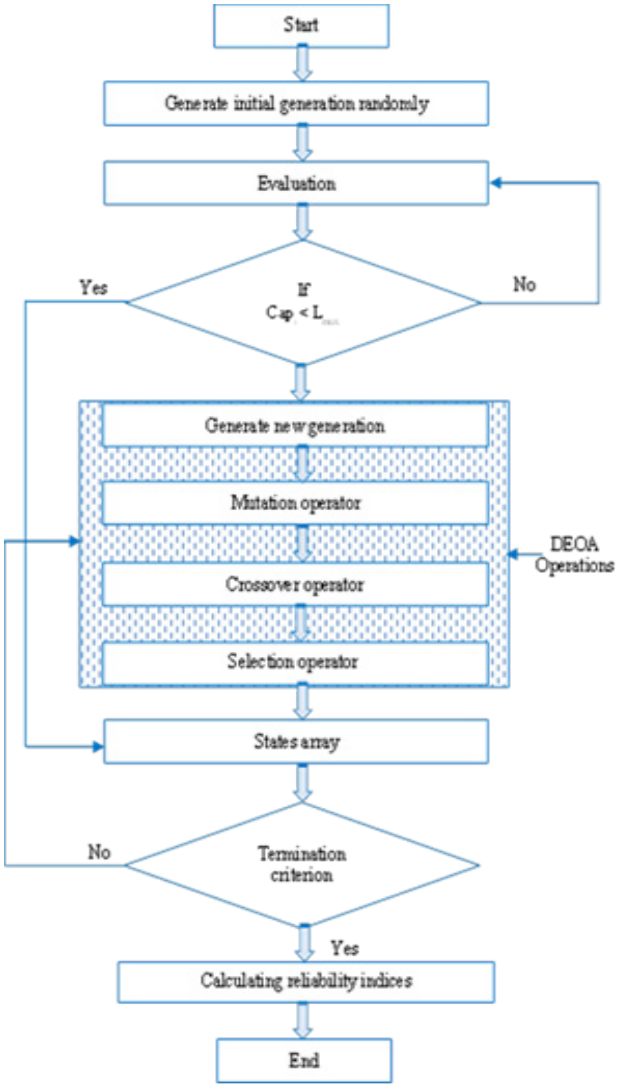


Figure 3: Diagram of reliability indices assessment using DEGA

4.2. Cases Running the Algorithm

This study develops and tests a method for evaluating the reliability of PGS using WECs with two dissimilar test systems ("IEEE-RTS-79" & "IEEE-RTS-96"), to validate how the DEOA performs to assess the dependability and sufficiency of the PGS. Meanwhile, it tests the efficiency of the proposed algorithm to evaluate multiple penetration-levels of wind energy.

The test systems "IEEE-RTS-79 and RTS -96" have 32 and 96 generating units respectively, with the capability to generate power ranging approximately from (12- 400 MW), with a total power supply of 3405 and 10215 MW respectively, at approximately (2850 MW) and (9000 MW) peak load [24], [25]. Figure 5 showing the diagram of the single area of the system RTS-79, and Figure 6 showing the outline of the RTS-96 of three-areas interconnected through merging three single areas from the RTS-79 system.

Run I: *Valuation of Algorithm Performance for the "RTS-79" System*

The designs of the variable control values for the proposed DEOA algorithm was documented as displayed in Table 1 [26], during applied at the "Test System- 79 & 96"

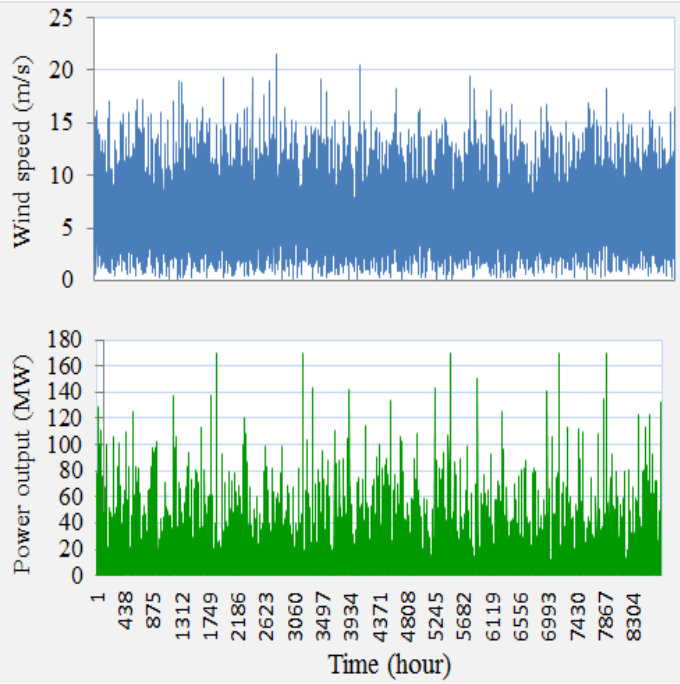


Figure 4: Wind forecasting and simulation of the wind MW generation for one year

The DEOA algorithm was applied for 100 iterations run consecutively in matching conditions as previously stated, to verify the robustness and confidence of this algorithm. The results of the various implementations are presented in Table 2. The results achieved from the DEOA were compared with those of [27]. Figures 7 and 8 show that following runs of 100 iterations, the values were wavering in the region of the real value. Repetitions of the runs of the experiment applying the algorithm displayed that this algorithm was highly accurate when determining the RI.

Table 1: The control values for the proposed algorithm

Parameter	DEOA (RTS-79)	DEOA (RTS-96)
Population size	60	100
Generation limit	100	150
Crossover probability	0.6	0.6
System Capacity (MW)	3405	10215
Mutation probability High	0. 1	0.6
Mutation probability Low	0.6	0.05
Load max	2850 MW	9000 MW
Threshold probability	1e-15	1e-20
Stopping Repetitions	100	150

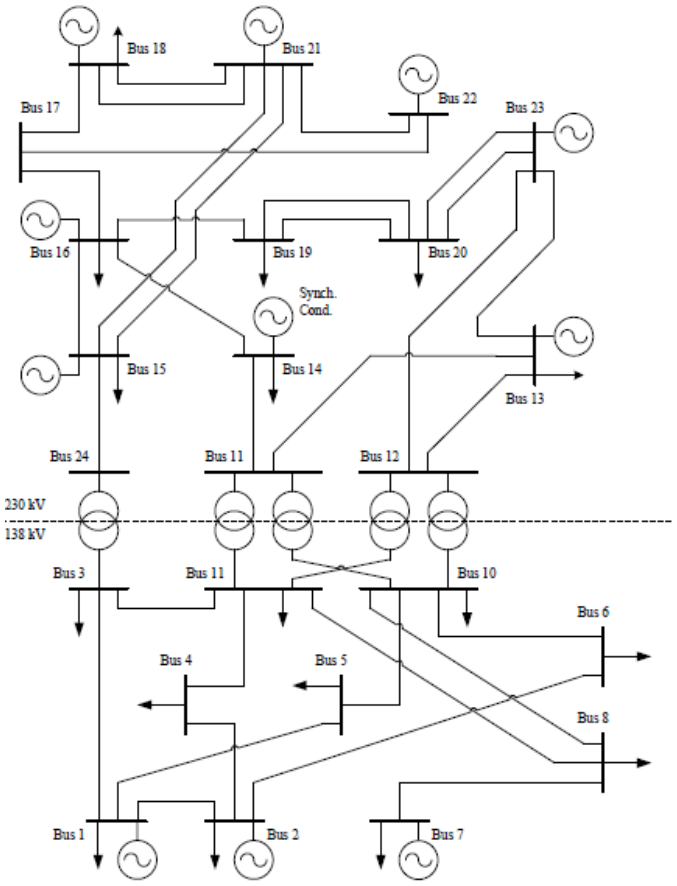


Figure 5: Diagram of single area for the system RTS-79

Table 2: The DEOA algorithm compare with analytical method through runs of the algorithm (100) times

Method	Reliability Indices		Error (%)
	LOLE (hour/year)	LOEE (MWh/year)	
(Accepted value) Analytical method [30]	9.39418	9.3670	0.28
(Experimental value) DEOA method	1176.00	1120.00	4.76

The consequences outcomes of the DEOA algorithm were compared with those of other optimization genetic algorithms earlier mentioned in [28], [29], as shown in Table 3. Comparisons of these outcomes were made. The absolute values utilized for these comparisons were from a single turn one mploying "MSGA", & "DEOA".

The computation effort of the proposed method was also assessed and a comparison was made with those of "MSGA", "MCS", and "DEOA" as shown in Table 4. On the other hand, comparison of the simulated proposed method with MCS was halted when the coefficient of variation level was approximately 5% [29]. This comparison was performed with absolute values obtained from a single turn on of the DEGA and MSGA. Evaluating the computation time of the DEOA, as presented in Table 4 was shorter compared to other methods. Hence, the suggested DEOA exhibits benefits such as enhanced computation speed, by employing a simple structure.

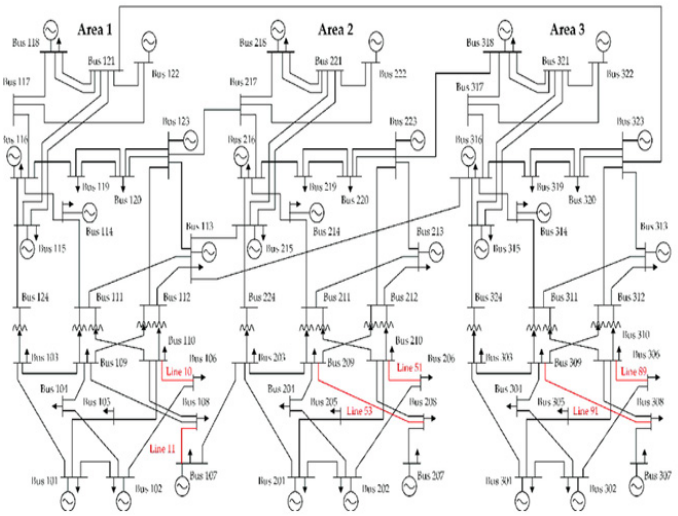


Figure 6: Diagram of three-areas interconnected for the system RTS-96

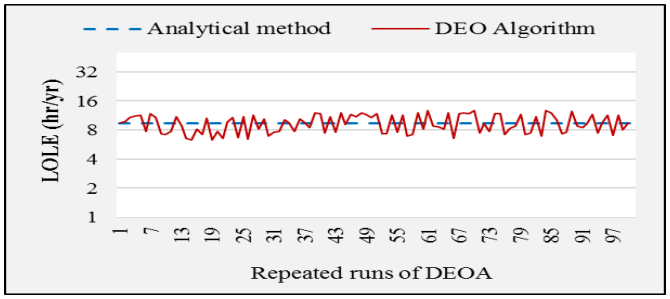


Figure 7: Calculation of estimated LOLE with 100 repeated runs for the RTS-79 System

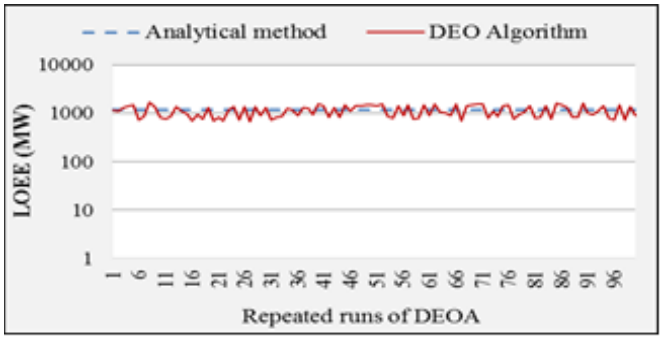


Figure 8: Calculation of estimated LOEE with 100 repeated runs for the system 'RTS-79'

Table 3: Compare the DEOA algorithm RI results for RTS-79 with accepted value

Reliability Indices	Experimental value		
	MCS [26]	MSGA [28]	DEOA
LOLE (hour/year)	9.371	9.324	9.370
Error (%)	0.244	0.745	0.255
LOEE (MWh/year)	1197	1163	1178
Error (%)	-1.785	1.105	-0.170
LOLF (occ/year)	1.9190	2.0030	2.300
Error (%)	0.00	4.37	19.85

Table 4: Comparison of the runs time between MSGA, DEOA, and MCS methods

Techniques	Experimental value		
	MCS	MSGA	DEOA
	Runs= 200	Runs= 750	Runs= 100
LOLE (hour/year)	9.541	9.324	9.37
Time (sec)	372	177	4

Run II: *Valuation of Algorithm Performance for the "RTS-96"*

In this run, the control parameters values had been setting beforehand as shown in Table 1. The tuning for the failures data for Generating stations was in line with previous data provided by [30]. The result of the DEOA was compared to those of Optimization algorithms declared by [31], as shown in Table 5.

In this paper, the percentage of error was considered, where the percentage error for individual values of the proposed DEOA algorithm was compared with individual values of the other techniques, as displayed in Tables 2, 3, and 5. On the other hand, the error percentage shows the effectiveness of the proposed DEOA algorithm in presenting separate values, large positive values of the error indicate a considerable deviation in the predicted value from the real one.

Table 5: Results of the RI for the system RTS-96 utilizing the DEOA method

Techniques	Reliability Indices			
	LOLE (hour/year)	Error (%)	LOEE (MWh/year)	Error (%)
Analytical method	0.1380	-	24.26	-
MCS Algorithm	0.1400	-1.44	23.97	1.195
DEOA Algorithm	0.140	-1.44	26.21	- 8.037

Run III: *Valuation of Performance SMCS with WPG for "RTS-79" System*

In this study, the SMCS method was employed to recompute the RI (for RTS-79 System) for comparison with the DEOA algorithm. From the simulation, Figure 9 shows what the system is capable of based on the conventional and the unconventional units. Figure 10 however presents a whole PGS and what it is capacity of, based on the simulation of the process overlaid together with the sequential load model and Figure 11 shows the LOLE, LOEE, and LOLF for 200 selected SMCS sampling years.

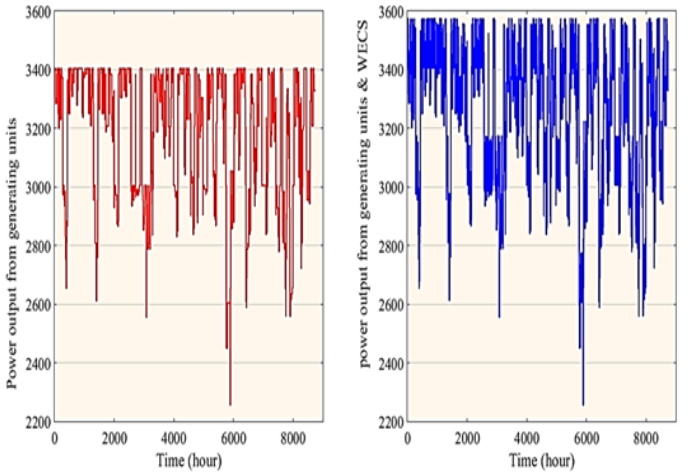


Figure 9: Available capacities obtained from conventional and unconventional units

Some intersections observable in Figure 10, which means that the available electrical power generation from the PGS insufficient for meeting the required load.

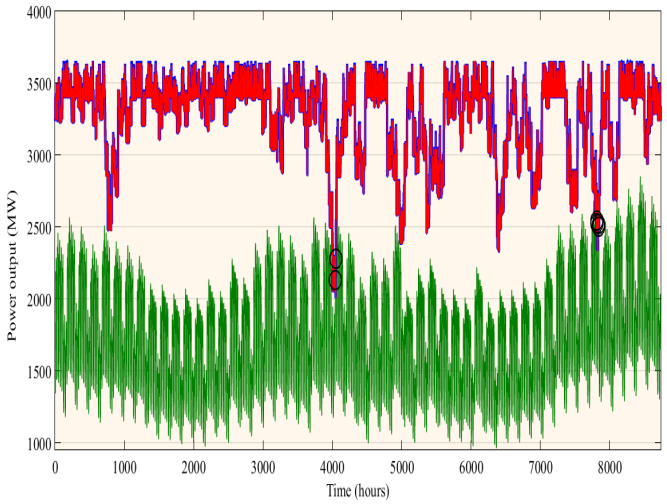


Figure 10: Simulation process shows the available capacities of the system, superimposition with the load paradigm

Run IV: *Valuation of Reliability Indices for DEOA with WPG for "RTS-79" System*

For peak demand load (2850MW) and 170 MW combined with wind energy, the RI the for "RTS-79" derived was from the Optimization Algorithm and the new proposed DEOA approach

as shown in Table 6. The MCS might be utilized for choosing failure states for the systems. However, due to its reliance on proportionate sampling, its effectiveness in the detection of failure states could be diminished. According to the results indicated in Table 6, the SMCS method could lead to extremely poor convergence time. Additionally, the DEOA provides the most efficient computation as well as easy usage in comparison with other algorithms.

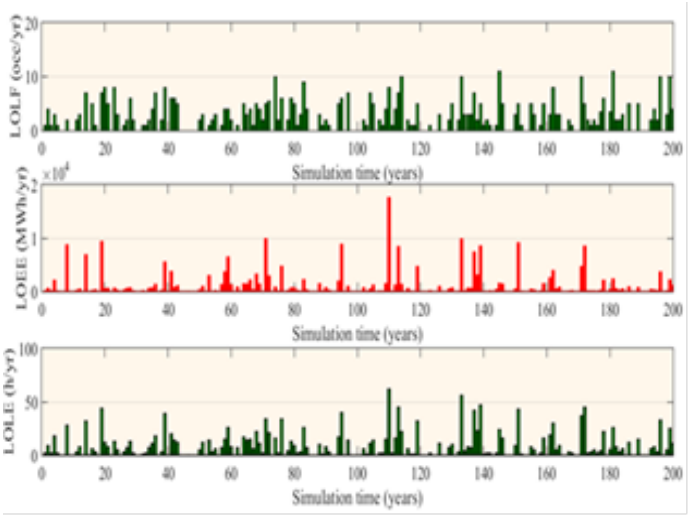


Figure 11: The reliability indices LOLE, LOEE, and LOLF for 200 samplings

Table 7 presents the outcomes of five models previously discussed [32] in comparison with those of DEOA. These techniques, computed the RI employing the MCS approach with variable wind Speed models (the actual wind speed recording data, Markov method, ARMA distribution model, Weibull distribution model, and Normal distribution). It was obvious that the outcomes derived from the use of the DEOA algorithm had values which were about the same as those obtained from using real wind speed data in the MCS approach.

Table 6: Comparison of the runs time between DEOA, BPSO, and SMCS Method

Proposed Methods	LOLE (hour/year)	LOEE (MWh/year)	Time (Sec)
SMCS Algorithm	7.550	941.070	64.641
BPSO Algorithm	7.430	823.780	8.979
DEOA Algorithm	7.450	1318.340	4.00

Table 7: Comparison between suggested methods in literature and proposed DEOA method

Techniques	LOLE (hour/year)	LOEE (MWh/year)
MCS technique with Actual wind data	7.450	908.70
MCS technique with Markov model	7.470	918.10
MCS technique with ARMA method	7.120	884.90
MCS technique with Weibull model	7.780	976.70
MCS technique with Normal model	6.950	858.50
Proposed DEOA with Actual wind data	7.450	1318.34

In real PGSs the LOLE index is used more often for the generation of adequacy analysis to determine sufficiency. On the other hand, the LOEE indicator is used for complex generating/transmitting of the adequacy analysis [33]. This means that outcomes presented in the tables are for assessing the PGS's dependability and sufficiency, which took into account the level of precision of LOLE from the suggested DEOA approach with shorter computation time in comparison with other approaches.

The proposal of an intelligent algorithm to assess the PGS is shown as a feasible replacement for the Non-SMCS in the evaluation of the non-sequential systems. Thus, this algorithm lowered the computation time needed to calculate the Non-MCS approach, which is efficient but does not take into account time-based phenomena, including the frequent computation interruptions of the LOLF and the duration indicatots (LOLD). This means that outcomes in tables that assess the LOLF indicate poor precision from the suggested PBIs approaches with shorter computation time in comparison with other approaches.

Run V: *Valuation of Reliability Indices with Wind Power Penetration for the "RTS-79" System Using DEOA*

In this section, the DEO algorithm is applied to examine the effects of wind power capabilities generated from WTGs on power systems to attain dependability indicators at a similar level to additional peak load capacity. The Weibull distribution approach was employed to simulate hourly repeating wind speed. Figure 12 shows the the simulation results of the forecast forecast for hourly wind speed and wind power output for several installed WTG units with a power rate of 2 MW for a year. The number of additional wind turbines that have been attached to a PGS is commensurate with the proportions of the wind power incorporation. Where the rate of penetration of wind energy ranges from 5% - 30% generates the output of wind energy ranges from 170 - 1020 MW from number WTG units (from 85 – 510).

This paper also examines the choice an appropriate penetration level of wind energy to attain a level similar to that of traditional PGSs. The DEO algorithm was used to analyze the RTS-79 system cases with changes in the number of WTG units, and changes in the penetration level of wind energy.

The effect of the wind energy penetration rate as a result of the number of available WTG units was tested on the values of reliability indicators for generation systems. Table 8 shows the different wind energy penetration levels (%), the number of WTG units and the output power of the WTG units, as well as the values of reliability indicators LOLE, LOEE.

Table 8: The wind power penetration beside on reliability indexes and percentage of installed capacity

Wind Pen. Rate %	No. of WTG	Product Power (MW)	LOLE	LOEE
5%	85	170	7.45	1318
10%	170	340	6.38	1364
15%	255	510	5.44	1214
20%	340	680	4.78	1256
25%	425	850	3.93	1079
30%	510	1020	3.60	1035

It is seen that the values of LOLE and LOEE are decreased for the RTS-79 system with additional WTG units. Figures 13 and 14 show LOLE and LOEE indicayors and the role of some of the WTG units added to RTS-79, where each WTG unit has a rated capacity of 2 MW.

Besides, it is noted that the LOLE value is lowered considerably in the event the wind energy penetration rate ranges from 5% to 20%; after which the rate of decline is relatively insignificant, implying that increasing the wind energy penetration to higher rates will have less effect in reducing the values of reliability indicators.

5. Conclusions

Both MCS and PBIs approaches have been used to assess the reliability of generating system combined with wind energy.

This paper presented a comprehensive performance clarification of the suggested technique for assessing the electric power systems dependability, such as; DEOA. The suggested intelligent algorithm which depends on the PBIs approach is viewed as a feasible alternative for the MCS in the assessment of non- sequential systems.

Furthermore, this study introduced the novel DEOA algorithm with WDM to assess the dependability of the PGSs combined with WTGs. It was shown that with the application of the proposed approach, it was likely that there would be simulation of the wind speed at a given hour for assessing the dependability of the power generation system. The main improvements achieved in this study are as presented below:

- Using the DEOA as an alternative to MCS for guided intelligent searches.
- The repeated runs of the experiment with the algorithm displayed the algorithm's elevated precision level when determining the RI.
- Incorporating WTGs into the power generating system, to revise the dependability evaluation process to adapt to this change.

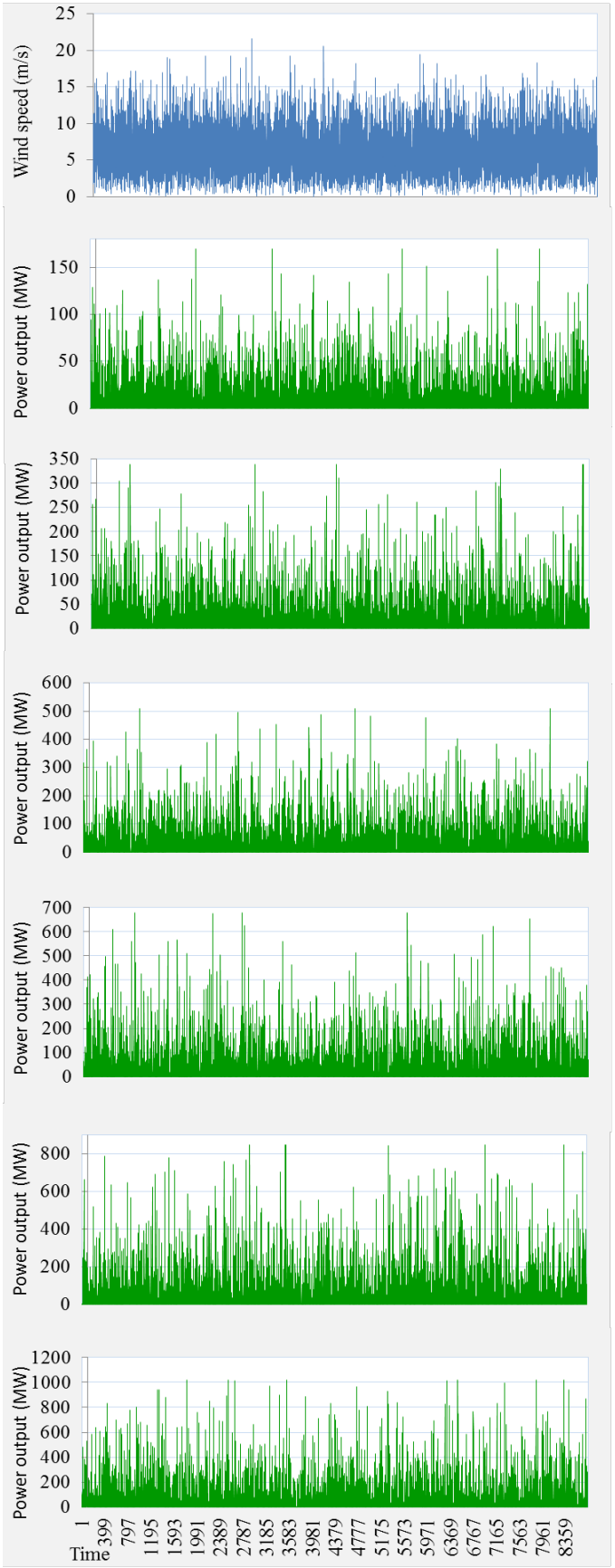


Figure 12: Wind forecasting and simulation of the generated wind energy for a year

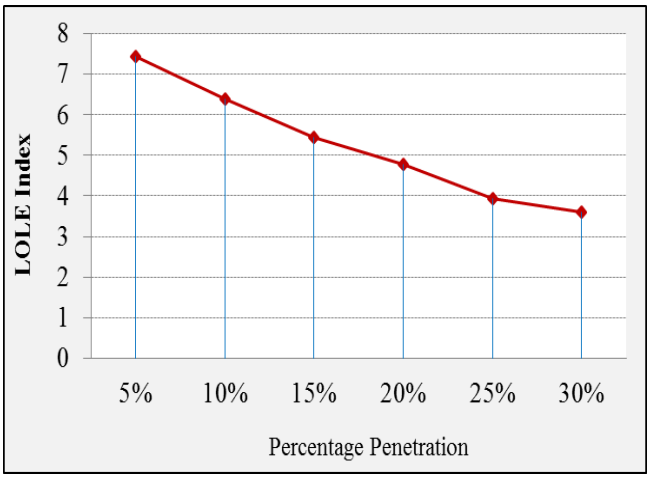


Figure 13: Shows the number of WTG units added to RTS-79 as a function for an estimate of the LOLE indices

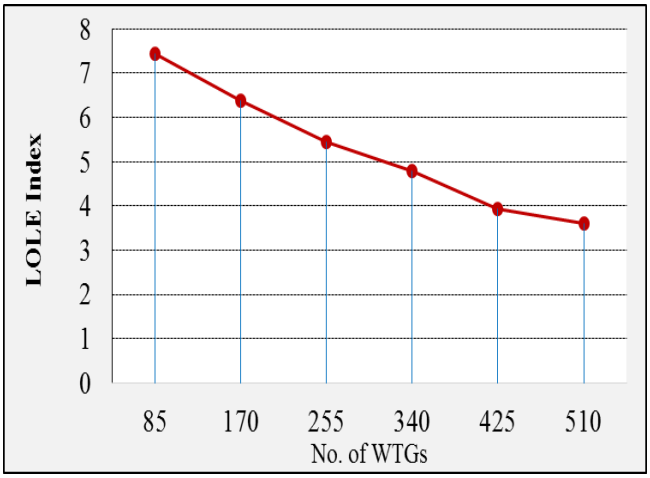


Figure 14: Shows the number of WTG units added to RTS-79 as a function for an estimate of the LOEE indices

On the other hand, RTS test system was used to make a comparison of the achievement of the DEOA with the conventional MCS to indicate the effectiveness of the proposed algorithm.

Besides, selecting an appropriate penetration level of wind power and determining its effect on the PGS and the essential element of investigating the potential of the wind energy source as a replacement for sustainable energy generation. Also, increasing the penetration rates of wind energy to higher rates will have less effect on reducing the values of RI.

References

[1] R. A. González-Fernández, A. M. Leite Da Silva, L. C. Resende, and M. T. Schilling, "Composite systems reliability evaluation based on Monte Carlo simulation and cross-entropy methods," IEEE Transactions on Power Systems,**28**(4), 4598–4606, 2013, doi: 10.1109/TPWRS.2013.2267154.

[2] A. M. L. Da Silva, R. A. Gonzalez-Fernandez, S. A. Flavio, and L. A. F. Manso, "Composite reliability evaluation with renewable sources based on quasi-sequential Monte Carlo and cross entropy methods," International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 1-6. 2014, doi: 10.1109/PMAPS.2014.6960618.

[3] M. H. Moradi and M. Abedini, "A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems," Electrical Power & Energy Systems, **34**(1), 66-74, 2012. doi: 10.1016/j.ijepes.2011.08.023.

[4] Y. Zhang, A. A. Chowdhury, and D. O. Koval, "Probabilistic wind energy modeling in electric generation system reliability assessment," IREP-Bulk Power System Dynamics and Control-VII Conference, 1–8, 2010, doi: 10.1109/TIA.2011.2127435.

[5] R. Karki, D. Dhungana, and R. Billinton, "An Appropriate Wind Model for Wind Integrated Power Systems Reliability Evaluation Considering Wind Speed Correlations," Appled Scinces, **3**, 107–121, 2013. doi: 10.3390/app3010107.

[6] S. H. Karaki, B. A. Salim, and R. B. Chedid, "Probabilistic model of a two-site wind energy conversion system," IEEE Transactions on Energy Conversion, **17**(4), 530–536, 2002. doi: 10.1109/TEC.2002.805215.

[7] H.J.Jayatheertha, "Evaluation of Composite Electric System Performance Indices Using Sequential Monte Carlo Simulation," International Journal of Advanced Engineering Research and Studies, **21**(2), 1–4, 2012.

[8] L. Wang, C. Singh, and K. C. Tan, "Reliability evaluation of power-generating systems including time-dependent sources based on binary particle swarm optimization," in 2007 IEEE Congress on Evolutionary Computation (CEC), 3346–3352, 2007. doi: 10.1109/CEC.2007.4424904.

[9] L. Wang and C. Singh, "Population-based intelligent search in reliability evaluation of generation systems with wind power penetration," IEEE Transaction on Power System, **23**(3), 1336–1345, 2008. doi: 10.1109/TPWRS.2008.922642.

[10] M. Benidris and J. Mitra, "Use of intelligent search methods in performing sensitivity analysis of power system reliability indices," The PES General Meeting, Conference & Exposition, National Harbor, MD, 2014, doi: 10.1109/PESGM.2014.6939915.

[11] R.C. Green II, L. Wang, and M. Alam, "Intelligent State Space Pruning with Local Search for Power System Reliability Evaluatio," in 2012 IEEE PES Innovative Smart Grid Technologies Europe (ISGT), 1-8, 2012.

[12] S. J. Huang and X. Z. Liu, "Application of artificial bee colony-based optimization for fault section estimation in power systems," Electrical Power & Energy Systems, **44**(1), 210–218, 2012. doi: 10.1016/j.ijepes.2012.07.012.

[13] R. C. Green, Z. Wang, L. Wang, M. Alam, and C. Singh, "Evaluation of loss of load probability for power systems using intelligent search based state space pruning," in 2010 IEEE the Probabilistic Method Applied to Power Systems (PMAPS), 319–324, 2010, doi: 10.1109/PMAPS.2010.5528892.

[14] A. Ali Kadhem, N. I. Abdul Wahab, I. Aris, J. Jasni, and A. N. Abdalla, "Computational techniques for assessing the reliability and sustainability of electrical power systems: A review," Renewable and Sustainable Energy Reviews, **80**, 1175–1186, 2017. doi: 10.1016/j.rser.2017.05.276.

[15] A. Almutairi, M. H. Ahmed, and M. M. A. Salama, "Probabilistic generating capacity adequacy evaluation: Research roadmap," Electric Power Systems Research, **129**, 83–93, 2015. doi: 10.1016/j.epsr.2015.07.013.

[16] J. Lin, L. Cheng, Y. Chang, K. Zhang, B. Shu, and G. Liu, "Reliability based power systems planning and operation with wind power integration: A review to models, algorithms and applications," Renewable and Sustainable Energy Reviews, **31**, 921–934, 2014. doi: 10.1016/j.rser.2013.12.034.

[17] D. Saxena, S. N. Singh, and K. S. Verma, "Application of computational intelligence in emerging power systems," International Journal Engineering Science Technology, **2**(3), 1–7, 2010. doi: 10.4314/ijest.v2i3.59166.

[18] A. A. Kadhem, N. I. Abdul Wahab, and A. N. Abdalla, "Wind energy generation assessment at specific sites in a Peninsula in Malaysia based on reliability indices," Processes, **7**(7), 1-23, 2019. doi: 10.3390/pr7070399.

[19] R. A. Bakkiyaraj and N. Kumarappan, "Optimal reliability planning for a composite electric power system based on Monte Carlo simulation using particle swarm optimization," Electrical Power & Energy Systems, **47**(1), 109–116, 2013. doi: 10.1016/j.ijepes.2012.10.055.

[20] Ž. B. Rejc and M. Čepin, "Estimating the additional operating reserve in power systems with installed renewable energy sources," Electrical Power & Energy Systems, **62**, 654–664, 2014. doi: 10.1016/j.ijepes.2014.05.019.

[21] M. Mosadeghy, T. K. Saha, R. Yan, and S. Bartlett, "Reliability evaluation of wind farms considering generation and transmission systems," the IEEE PES General Meeting Conference & Exposition, 1-5, 2014, doi: 10.1109/PESGM.2014.6939156.

[22] A. A. Kadhem, N. I. Abdul Wahab, I. Aris, J. Jasni, A. N. Abdalla, and Y. Matsukawa, "Reliability assessment of generating systems with wind power penetration via BPSO," International Journal on Advanced Science, Engineering and Information Technology, **7**(4), 1248–1254, 2017. doi: 10.18517/ijaseit.7.4.2311.

[23] A. K. Azad, M. G. Rasul, M. M. Alam, S. M. Ameer Uddin, and S. K. Mondal, "Analysis of wind energy conversion system using Weibull

distribution," International Conference on Mechanical Engineering (ICME), **90**, 725–732, 2014, doi: 10.1016/j.proeng.2014.11.803.

[24] A. A. Kadhem, N. I. Abdul Wahab, I. Aris, J. Jasni, and A. N. Abdalla, "Reliability assessment of power generation systems using intelligent search based on disparity theory," Energies, **10**(3), 1-13, 2017. doi: 10.3390/en10030343.

[25] C. Grigg and P. Wong, "The IEEE reliability test system -1996 a report prepared by the reliability test system task force of the application of probability methods subcommittee," IEEE Transactiona on Power Systems, **14**(3), 1010–1020, 1999. doi: 10.1109/59.780914.

[26] A. A. Kadhem, N. Izzri Abdul Wahab, and A. N. AbdAlla, "Differential evolution optimization algorithm based on generation systems reliability assessment integrated with wind energy," 5th International Conference on Power Generation Systems and Renewable Enegy Technologies (PGSRET), 1-6, 2019, doi: 10.1109/PGSRET.2019.8882705.

[27] C. Singh and V. Gubbala, "An alternative approach to rounding off generation models in power system reliability evaluation," Electric Power Systems Research, **7796**(95), 37–44. 1996.

[28] N. Samaan and C. Singh, "Adequacy assessment of power system generation using a modified simple genetic algorithm," IEEE Transactions on Power Systems, **17**(4), 974–981, 2002. doi: 10.1109/TPWRS.2002.804994.

[29] R. Billinton and W. Li, Reliability Assessment of Electric Power Systems Using Monte Carlo Methods. Plenum Press, New York, 1994.

[30] R. N. Allan, R. Billinton, and N.M.K. Abdel-Gawad "The IEEE Reliability Test System—Extensions to and Evaluation of the Generating System," IEEE Transactions on Power Systems, **PWRS(4)**, 1–7. 1986.

[31] V. Miranda, L. de Magalhaes Carvalho, M. A. da Rosa, A. M. L. da Silva, and C. Singh, "Improving power system reliability calculation efficiency with EPSO variants," IEEE Transactions on Power Systems, **24**(4), 1772–1779, 2009. doi: 10.1109/TPWRS.2009.2030397.

[32] S. Soleymani, M. E. Mosayebian, S. Mohammadi, and D. G. Gomes, "A combination method for modeling wind power plants in power systems reliability evaluation," Computer and Electrical Engineering, **41**, 28–39, 2015. doi: 10.1016/j.compeleceng.2014.12.005.

[33] P. Henneaux, F. X. Bouchez, and L. Rese, "Confidence intervals for adequacy assessment using Monte Carlo sequential simulation," in 2016 IEEE International Energy Conference, 1-6, 2016, doi: 10.1109/ENERGYCON.2016.7514142.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 344-350 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Computational Intelligence and Statistical Learning Performances on Predicting Dengue Incidence using Remote Sensing Data

Nittaya Kerdprasop[*,1], Kittisak Kerdprasop[1], Paradee Chuaybamroong[2]

[1]*School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand*

[2]*Department of Environmental Science, Thammasat University, Rangsit Campus, Pathum Thani, 12120, Thailand*

ARTICLE INFO

Article history:
Received: 02 June, 2020
Accepted: 13 July, 2020
Online: 28 July, 2020

Keywords:
Dengue incidence
Data modeling
Computational intelligence
Statistical learning
Satellite-based index
Climate variability

ABSTRACT

Dengue is a viral infection disease transmitted to people through the bite of specific mosquito species living in a tropical zone. According to the World Health Organization, dengue has been listed among the top-ten diseases for 2019 as it makes 3.9 billion people in 128 countries be at risk of infection. One major cause of substantial dengue widespread is the globally warm climate that accelerates rapid growth of mosquito vectors. In this research, we aim to build data-driven models to predict dengue cases using satellite index data to represent temperature, humidity, and greenness over the surface area of Bangkok, which is our target area of dengue prediction because of its high infection cases. Oceanic Niño Index is also used as a predictor variable to represent climate variability. The modeling methods employ seven algorithms from two broad schemes of the machine learning field. Artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) are algorithms from the subfield of computational intelligence, whereas multiple linear regression (MLR), generalized linear model (GLM), support vector regression (SVR), classification and regression tree (CART), and chi-squared automatic interaction detection (CHAID) are from the statistical learning subfield. Performances of these algorithms are evaluated on the same set of out-of-sample data. The results are that ANFIS is the best model for predicting dengue outbreak in the capital city of Thailand.

1. Introduction

Dengue is a communicable disease that can indirectly transmit from an infected person to another through the bite by a female mosquito in the two major *Aedes* species, *Ae. aegypti* and *Ae. albopictus*. These small mosquitoes have white stripes on their back and legs. They bite during day-time and live in and around human habitats. These mosquitoes can transmit not only dengue virus, but also chikungunya, Zika, and yellow fever [1]. Illness from dengue infection is caused by one of the four types of viruses [2]: DENV1, DENV2, DENV3, and DENV4. Patients recovered from infection by a specific type of dengue virus are immune to that virus type for life. Unfortunately, they are still vulnerable to the other three types of dengue viruses. The many types of viruses is just one reason for a widespread of dengue fever. The other causes for a rapid spread of dengue are global warming phenomenon and poorly or unplanned urbanization. The rise in temperature globally makes mosquitoes spreading from a tropical region into subtropical zone such as China, Japan, U.S.A., and many countries in Europe [3]. A fast expansion of big cities is also another factor accelerating dengue spread as the *Ae. aegypti* mosquitoes can breed and lay eggs in small water holders such as vessels and abandoned containers.

In 1950, there were only 908 dengue cases reported [4] to the World Health Organization (WHO). The endemic areas are mostly countries in the Southeast Asia including Cambodia, Philippines, Thailand, and Vietnam. In 2016, the number of dengue infection cases reported to WHO has been substantially jumped to over 3.34 million cases. The latest reports [5] during September to November 2019 of the European Centre for Disease Prevention and Control (ECDC) list countries with most dengue cases in descending order as Brazil, Philippines, Mexico, Nicaragua, Thailand, Malaysia, and Colombia. The current number of dengue infections in Brazil is over two million cases which is almost ten times higher than the same period in the previous year. In Thailand as of November 2019, more than 120,000 cases have been reported and 126 cases are fatal [5]. The highest number of dengue

[*]Corresponding Author Nittaya Kerdprasop, nittaya@sut.ac.th

infections in Thailand is in the Bangkok metropolitan area, which is the capital city with massive urbanization and warm atmosphere.

Based on the dengue case classification criteria released by WHO in 1997 [6], clinical infections can be broadly categorized into two distinct classes: dengue fever (DF) and dengue haemorrhagic fever (DHF)/dengue shock syndrome (DSS). DF is a classic dengue fever that infected person may either show no signs of illness, or having mild illness with symptoms such as abdominal pain and mucosal bleeding [7]. DHF is a type of severe dengue with manifestation of vascular permeability resulting in plasma leakage, a decrease in platelets in the blood, and an increase of red blood cells in relative to the plasma that may cause the fatal shock syndrome, DSS [7]. The classification of dengue infection as DF/DHF/DSS is based on historical cases in Thailand that infections normally occur in children younger than 15 years old [7]. At the present decade that dengue infections have extended to older age groups and spatially invade to different zones, clinical symptoms of infected patients show some variations from the WHO's descriptions [8]. Therefore, in 2009 WHO developed new criteria and guideline for dengue case diagnosis and treatment. The new criteria [8] classify dengue infection into three main groups based on the ascending levels of severity as dengue without warning signs, dengue with warning signs such as abdominal pain and mucosal bleeding, and severe dengue that includes severe plasma leakage leading to shock, severe bleeding, and severe organ failure.

Fatal cases of severe dengue in Thailand are increasing from 36 deaths in 2014 to 127 deceases in 2019 [9]. The rise in dengue mortality in Thailand is more or less coincident with the occurrence of medium to strong level of El Niño events. This relation can be noticed from the high number of dengue fatality in El Niño years of 2015 and 2019. The coincidence of El Niño and dengue endemic, especially around the area of capital city, is of particular interest to our research team. We thus plan our modeling method to employ climate variability through the Oceanic Niño Index (ONI) as one independent factor to predict the number of dengue cases in Bangkok. Other independent variables are amount of rainfall and remote sensing indices to represent temperature, humidity, brightness, and greenness of vegetation. The modeling algorithms are those from the computational intelligence (artificial neural network -- ANN, adaptive neuro-fuzzy inference system -- ANFIS) and statistical learning schemes (multiple linear regression -- MLR, support vector regression -- SVR, generalized linear model -- GLM, classification and regression tree -- CART, chi-squared automatic interaction detection -- CHAID).

A review of these and other learning techniques applied to predict dengue incidence is presented in Section 2. Details of dataset and our modeling method are explained in Section 3. Experimental results of model performance comparison are illustrated in Section 4. The conclusion is in Section 5.

2. Related Work

Historically, dengue has been identified as an endemic locally found in the Southeast Asian countries. After World War II, severe type of dengue (DHF) firstly appeared in Philippines in 1954 and later emerged in Thailand in 1958 and became an outbreak in Malaysia, Cambodia, Singapore, and Vietnam in the 1960s [9]. Currently, dengue expands its boundary from a local communicable disease to become a global epidemic. Disease prevention and efficient mechanisms to monitor and control infection are urgent actions needed to be implemented instantly by governments and health organizations worldwide. Scientists and researchers attempt to estimate in advance the outbreak of dengue infection with the purpose of precise warning leading to proper preparation and management.

Research teams from Singapore estimate dengue incidence with nonlinear regression model [10], time-series Poisson regression model [11], and neural network model [12]. Scientists in Malaysia apply satellite imagery to study the pattern of dengue incidence distribution with correlation analysis against land-use patterns in the geographic information system (GIS) [13]. To predict dengue cases in Malaysia, variations of network modeling methods such as ANFIS [14], self-organizing map [15], and a hybrid of genetic algorithm and ANN [16] are applied. Various research groups in Philippines study dengue incidence patterns through regression analysis [17], ANN [18], fuzzy association rule mining [19], and many descriptive studies [20]. The sophisticated methods such as a combination of flower pollination algorithm and least squares support vector machines (FPA-LSSVM) [21] and ANFIS with various parameter tuning methods [22] are applied to forecast dengue incidence in Indonesia. Modeling dengue incidence in Thailand includes a wide range of techniques such as exploratory analysis through the remote sensing GIS [23], Poisson autoregressive [24], cellular automata with the inclusion of genetic algorithm for chromosome weighting and ANN for determining probability in state change [25], SVR [26], Bayesian network [27], and generalized additive model (GAM) [28].

Besides the Southeast Asia, dengue has become a public health threat in more than hundred countries worldwide. Such global pandemic urgently requires collaboration among researchers and practitioners at both regional and global scales. The awareness regarding an attempt to monitor and control dengue in Taiwan can be noticed through many proposals adopting various state-of-the-art machine intelligence techniques including fuzzy inference system [29], multivariate logistic regression analysis [30], and decision tree induction [31]. To forecast dengue outbreak in southern China, researchers apply GAM [32], SVR as well as variations of regression modeling [33], and the factor-detector modeling module of the ArgGIS software [34]. Researchers in Sri Lanka adopt fuzzy system with trapezoidal membership function to predict the developing from classic dengue fever to the severe level using clinical evidences as input data [35].

In Latin America that dengue has spread massively during this decade, researchers apply numerous machine learning techniques to build an efficient predictive model. The applied techniques include SVR and k-nearest neighbor regression to predict cases in the northwest city of Argentina [36]. SVR and ANN are adopted to predict dengue infection in Paraguay and ANN shows the best result [37]. In Brazil which confronts the most serious dengue outbreak in 2019, GAM [38] and Pearson correlation analysis [39] are the main methods applied to predict dengue cases in Sao Paulo, whereas nearest-neighbor regression analysis [40] is the technique that researchers use for predicting cases in Rio de Janeiro. To project the spread of dengue virus globally towards the year 2020, 2050 and 2080, an international team of researchers [41] has applied the ensemble of boosted regression tree to show the trend

in dengue transmission that around 60% of the global population living in Asia, Africa and Americas, in descending order, are at risk of infection.

It can be concluded from the literature review that dengue modeling techniques based on the mature machine learning methodology include regression analysis, ANN, ANFIS, SVR, GAM, and regression tree. We thus apply these main methods to derive prediction models to estimate number of dengue infection in the capital and also the most populous city of Thailand. The main focus is to observe performances of these popular modeling methods, and then select the most accurate model for predicting dengue outbreak in Bangkok.

3. Data and Modeling Method

3.1. Data Sources

In our modeling process, the target for prediction is number of patients in Bangkok metropolitan area that had been diagnosed as infected with dengue virus. These historical data are monthly dengue cases in Bangkok during the years 2003-2017. Data between 2003-2015 are used for training the models, whilst data of 2016-2017 are held out for testing performance of the models.

Independent variables to be used as predictors are time of the year represented by number of month (1 to 12) and satellite-based data obtained from the National Oceanic and Atmospheric Administration (NOAA), U.S.A. Rainfall data are from the Centers for Environmental Information [42]. Satellite indices to reflect vegetation health are obtained from the NOAA's global vegetation health products [43]. The indices to be used as predictors in this research include smoothed and normalized difference vegetation index (SMN, or greenness), smoothed brightness temperature index (SMT, a thermal condition), vegetation condition index (VCI, a proxy for greenness anomaly), temperature condition index (TCI), and vegetation health index (VHI, a combined moisture and thermal conditions).

Another satellite-based index to reflect climate variability is also used in our modeling process. This climate index is Oceanic Niño Index (ONI) that can be accessed from the NOAA's Weather Service [44]. The ONI is computed from the anomaly of sea surface temperature (SST) at the Niño3.4 region in the Pacific Ocean at the equator line near Indonesia and north Australia. The SST is monthly collected to compute ONI by averaging from the 3-month periods: the current one, the previous, and the following month. This 3-month average value is then compared to a thirty-year average to check temperature deviation. If the ONI values in five consecutive months are +0.5 °C or higher than the normal value, then the El Niño (warm event) phenomenon is announced. On the contrary, if the ONI values in five consecutive periods are –0.5 °C or lower than the normal value, the opposite phenomenon called La Niña (the cold event) is announced.

3.2. A Framework of Modeling Method

Main steps in our data modeling process are shown in Figure 1. In data preparation and integration step, remote sensing indices have to be transformed from weekly to be monthly by averaging weekly values of each month. We then select only the attributes contributing the most to accurate prediction in step 2. This is done by evaluating predictor importance value. Step 3 is the train-test separation. Training data are used in the model building step. Test data are used in the model evaluation step.

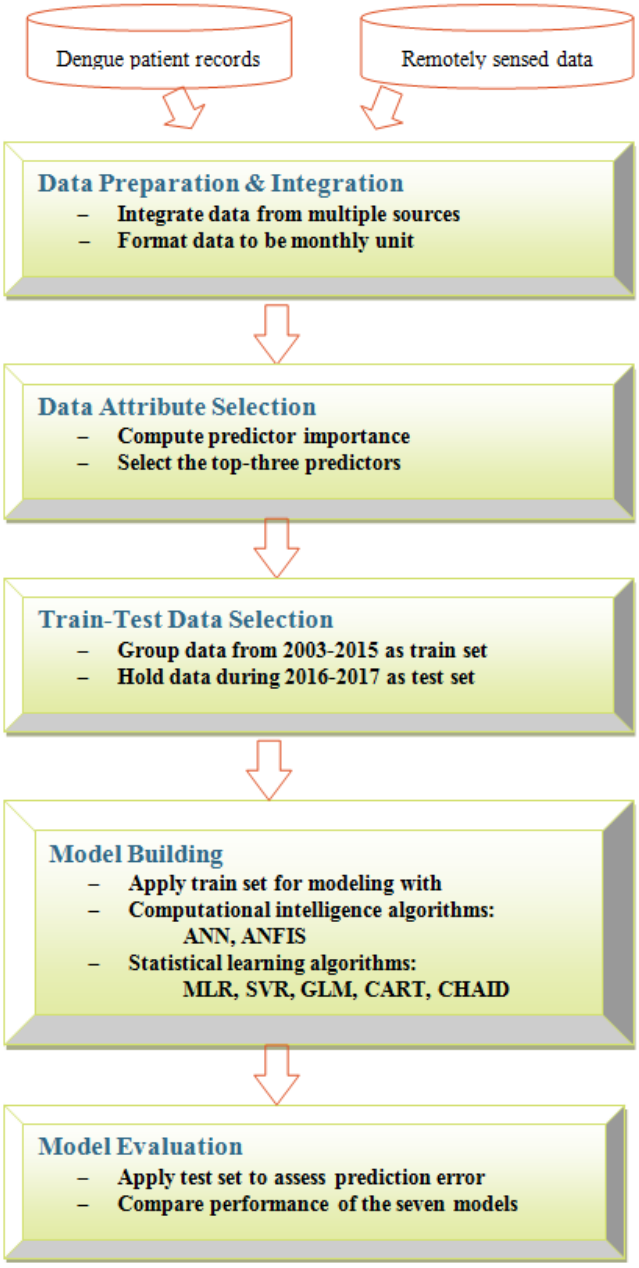


Figure 1: Steps in Data Modeling for Dengue Incidence Prediction

4. Performance Comparison Results

We perform experiments following the steps shown in Figure 1. Software tool used for running ANFIS algorithm is MATLAB, whereas other learning algorithms are run with IBM SPSS Modeler. All algorithms are tested with the same set of testing data that are unseen by the algorithms during the training phase to build model.

Data attributes used as model predictors to estimate number of dengue cases in Bangkok include the satellite-based indices (ONI, SMN, SMT, VCI, TCI, VHI), amount of rainfall, and number of month (1 to 12, in which 1 is January and 12 is December) in each year. Representation of each satellite-based index is as follows:

ONI -- climate variability, SMN -- surface greenness, SMT -- thermal condition, VCI -- greenness anomaly, TCI -- temperature condition, VHI -- moisture and thermal conditions. Number of month can represent season in that months 5 to 8 are rainy season in Thailand. Correlations among the eight predictors are presented in Table 1. Strong correlations are highlighted with bold font.

Prior to building the model with training data set, we analyze importance of each predictor to select only the top-three ones contributing most toward the final predictive model. To compute the importance value [45], each independent variable is assessed through sensitivity analysis to observe the reduction in variance of the target field, which is the dengue incidence in this specific domain. The computed sensitivity values are then ranked to evaluate importance of each predictor; the higher the value, the more important the variable. The ranking result of all seven predictors is shown in Figure 2. The attribute SMT is missing from the figure because the coefficient of variation in this attribute is below the threshold. The top-three variables, which are ONI, month of the year, and SMN as a proxy of greenness, are then selected as important features to be used further in the modeling phase.

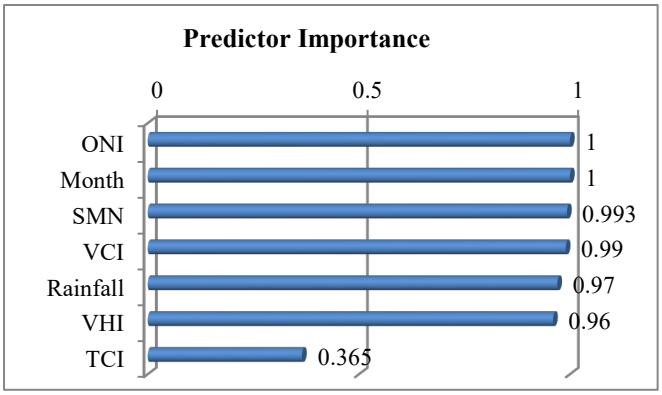


Figure 2: Importance Value of Independent Predictors

Table 1: Pearson Correlation among Predictor Attributes

	ONI	SMN	SMT	VCI	TCI	VHI	Month
SMN	0.078						
SMT	0.118	**0.735**					
VCI	0.135	**0.656**	0.152				
TCI	**-0.362**	-0.107	-0.143	**-0.331**			
VHI	**-0.282**	**0.304**	-0.046	**0.289**	**0.807**		
Month	-0.046	**-0.251**	**-0.460**	0.020	**-0.287**	**-0.281**	
Rainfall	-0.115	0.120	**0.398**	-0.131	**0.236**	0.161	**-0.439**

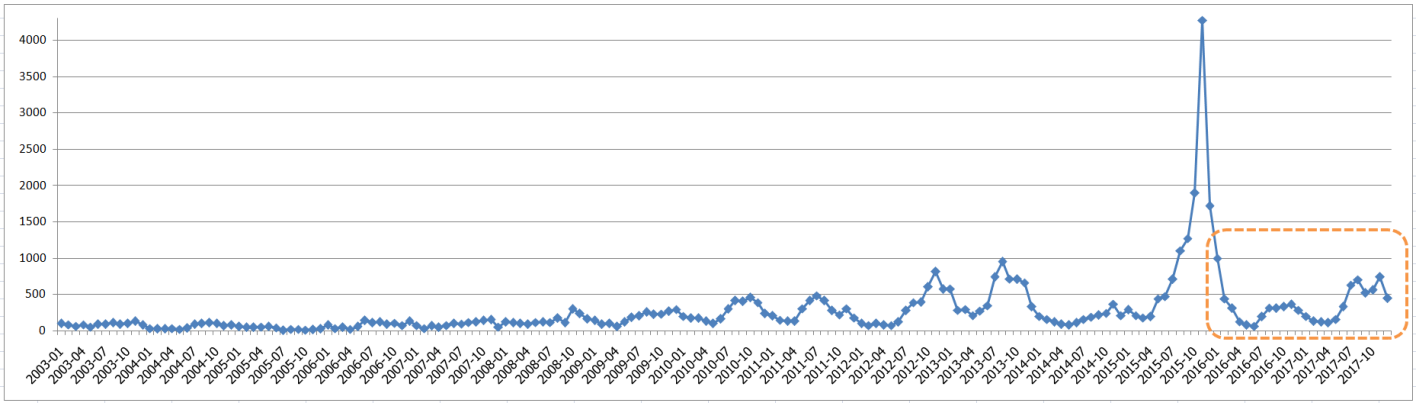


Figure 3: Distribution of Dengue Infection in Bangkok During 2003 to 2017 in which 2016-2017 are Test Data

Table 2: Performance of Learning Algorithms

Algorithm	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Model Correlation	Model Building Time (seconds)
Computation Intelligence				
ANFIS	151.511	216.536	0.835	4
ANN	211.296	308.277	0.827	8
Statistical Learning				
MLR	200.681	236.613	0.501	4
GLM (log function)	217.706	280.076	0.501	4
CHAID	217.018	292.826	0.515	3
SVR	236.732	325.958	0.501	4
CART	278.939	381.517	0.815	3

After data attribute selection step, we prepare two subsets of data: training data and test data. Training data are the records during the years 2003-2015, containing 156 records. Test data are cases from 2016-2017, comprising of 24 records. Dengue case distribution in Bangkok between 2003-2017 is displayed in Figure 3. Vertical axis is the number of dengue infected patients. There exists an extreme outbreak in November 2015, which is the month that ONI value is as high as 2.5. ONI normally ranges between +2.5 to -2.5. The ONI value of -2.5 represent the strong El Nino event that makes the weather in Bangkok hotter and drier than normal years. The dashed area in the figure marks distribution of dengue cases in the test data.

The training data are used as input for each of the learning algorithms to build predictive model. The models are then tested for their performance using the out-of-sample test data of the years 2016-2017. Performance evaluation results are illustrated in Table 2. The measurements for performance evaluation of each model include the mean absolute error (MAE), root mean square error (RMSE), correlation between predictors and the target of the model, and time to build model. Consider from predictive errors, ANFIS shows the best performance with the minimum values in both error metrics. MLR shows the second best performance with errors a little bit higher that ANFIS. On the metric of model correlation, ANFIS is also the best one, while ANN comes as the second best model and CART is in the third place. CART and CHAID, which are the tree-based modeling methods, take less time to build model than the other methods. Based on the time aspect, ANN is the worst one.

Figure 4 shows comparative performance of the seven models based on four measurements in which the scales are normalized to be between 0 to 1. The three measurements (MAE, RMSE, time) seek for the lowest one to be the best, while correlation requires the highest one to outperform the others. Therefore, we transform the normalized correlation metric to be the reverse one by applying the formula (1- normalized_correlation) to make this metric in accordance with the others. Thus, model with all measurements closest to zero is the best one. From this criteria, we can rank the models from the best to the worst ones as follows: (1) ANFIS, (2) MLR, (3) CHAID, (4) ANN, (5) CART, (6) GLM, (7) SVR.

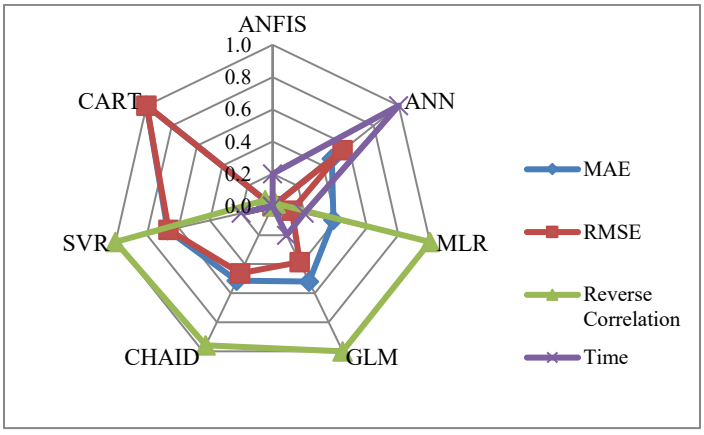


Figure 4: Comparison of Model Performance using Normalized Measurements

The objective of our modeling process is to predict the number of dengue cases in a specific region. It is thus a macro analysis aiming at dengue outbreak monitoring rather than a micro scope of classifying patients as infected with dengue virus or not. Moreover, our target variable is continuous data. Therefore, traditional measurements such as sensitivity, specificity, F-measure, recall, precision, and so on, cannot be applied. The most suitable metric is the error made by the model to predict number of dengue incidence. The plot in Figure 5 shows the difference between actual number of dengue cases (circle) against the predicted number of cases forecasted by ANFIS (star). It can be noticed that at extreme level of dengue infected cases higher than 700, ANFIS forecasts quite inaccurate. For the average level of infection around 100 to 300 cases, ANFIS can make a good prediction.

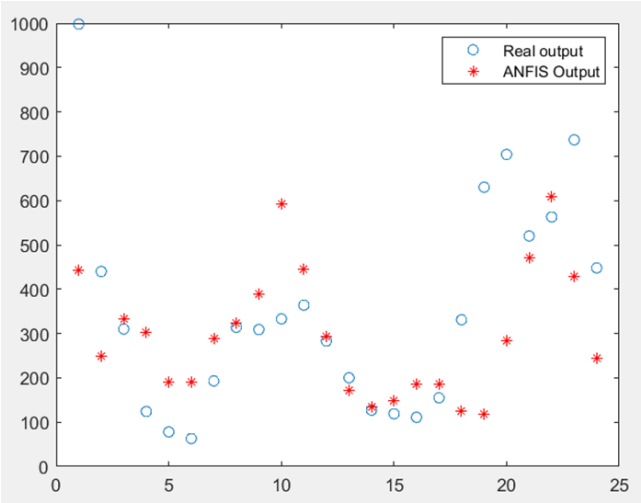


Figure 5: Predicting Performance of ANFIS on Test Data

5. Conclusion

Dengue is a fatal communicable viral disease threatening life of population in Thailand as well as many countries in the tropical zone with warm and humid weather providing a good condition for the spread of *Ae. aegypti* and *Ae. albopictus* mosquitoes that are the main carriers of dengue virus. The outbreak of dengue is seasonal in that infection is abnormally high in every two or three years. Many researchers observed that this spasmodic infection is coincident to the rise of sea surface temperature in the equatorial Pacific region. We thus include Oceanic Niño Index (ONI) as one main factor to derive a model to predict dengue cases in Bangkok. Other predictors used in our modeling phase are smoothed and normalized difference vegetation index (SMN), which is a proxy for vegetation greenness, and month of the year (1 to 12) that is used to capture season as dengue fever in Thailand normally widespread during months 6 to 11. The satellite features (SMN and ONI) and month attribute are selected from the original seven variables through the computation of predictor importance to select only the top-three features showing the best variance reduction against the target attribute, which is the number of dengue infected patients.

The selected features (ONI, SMN, month) in addition to the target field (dengue cases) are then used for the next step of data-driven model building. Modeling algorithms applied in this work are of two categories: computational intelligence and statistical learning. Computational intelligence algorithms are ANN and ANFIS. The other five algorithms (multiple linear regression, generalized linear model, support vector regression, regression

trees -- CART and CHAID) are in a group of statistical learning. We evaluate performance of all algorithms using the same set of out-of-sample data for a fair comparison. The experimental results reveal that ANFIS is the best predictor model. Multiple linear regression is the second best one. In the future, we plan to further our empirical study to cover every province of Thailand.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been supported by grant from Suranaree University of Technology through the funding of Data and Knowledge Engineering Research Unit in which the first and second authors are principal researchers.

References

[1] WHO, World Health Organization, "Dengue and severe dengue" Fact Sheet, updated 4 November 2019. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[2] CDC, Centers for Disease Control and Prevention, "Dengue" Access 10 May 2020. https://cdc.gov/dengue/about/index.html

[3] N.E.A. Murray, M.B. Quam, A. Wilder-Smith, "Epidemiology of dengue: past, present and future prospects" Clinical Epidemiology, 5, 299-309, 2013.

[4] N. Gyawali, R.S. Bradbury, A.W. Taylor-Robinson, "The epidemiology of dengue infection: harnessing past experience and current knowledge to support implementation of future control strategies" Journal of Vector Borne Diseases, 53, 293-304, 2016.

[5] ECDC, European Centre for Disease Prevention and Control, "Dengue worldwide overview" Access 21 April 2020. https://ecde.europa/en/dengue-monthly

[6] S. Bandyopadhyay, L. Lum, A. Kroeger, "Classifying dengue: a review of the difficulties in using the WHO case classification for dengue haemorrhagic fever" Tropical Medicine & International Health, 11(8), 1238-1255, 2006.

[7] S.R.S. Hadinegoro, "The revised WHO dengue classification: does the system need to be modified?" Paediatrics and International Child Health, 32(S1), 33-38, 2012.

[8] B.A. Ajlan, M.M. Alafif, M.M. Alawi, N.A. Akbar, E.K. Aldigs, T.A. Madani, "Assessment of the new World Health Organization's dengue classification for predicting severity of illness and level of healthcare required" PLoS Neglected Tropical Diseases, 13(8), e0007144, 2019. https://doi.org/10.1371/journal.pntd.0007144

[9] Vector Borne Disease Division, Ministry of Public Health, "Dengue infection (in Thai)" Access 10 May 2020. https://ddc.moph.go.th/index.php

[10] B. Loh, R.J. Song, "Modeling dengue cluster size as a function of Aedes aegypti population and climate in Singapore" Dengue Bulletin, 25, 74-78, 2001.

[11] Y. Hii, J. Rocklov, N. Ng, C. Tang, F. Pang, R. Sauerborn, "Climate variability and increase in intensity and magnitude of dengue incidence in Singapore" Global Health Action, 2, 124-132, 2009.

[12] H.M. Aburas, B.G. Cetiner, M. Sari, "Dengue confirmed-cases prediction: a neural network model" Expert Systems with Applications, 37, 4256-4260, 2010.

[13] N.C. Dom, A.H. Ahmad, Z.A. Latif, R. Ismail, B. Pradhan, "Coupling of remote sensing data and environmental-related parameters for dengue transmission risk assessment in Subang Jaya, Malaysia" Geocarto International, 28(3), 258-272, 2013.

[14] T. Faisal, M.N. Taib, F. Ibrahim, "Adaptive neuro-fuzzy inference system for diagnosis risk in dengue patients" Expert Systems with Applications, 39, 4483-4495, 2012.

[15] F. Ibrahim, T.H.G. Thio, To Faisal, M. Neuman, "The application of biomedical engineering techniques to the diagnosis and management of tropical diseases: a review" Sensors, 15, 6947-6995, 2015.

[16] N.A. Husin, N. Mustapha, M.N. Sulaiman, R. Yaacob, H. Hamdan, M. Hussin, "Performance of hybrid GANN in comparison with other standalone models on dengue outbreak prediction" Journal of Computer Sciences, 12(6), 300-306, 2016.

[17] G.L. Su, "Correlation of climatic factors and dengue incidence in metro Manila, Philippines" AMBIO: A Journal of the Human Environment, 37(4), 292-294, 2008.

[18] H.I. Datoc, R. Caparas, J. Caro, "Forecasting and data visualization of dengue spread in the Philippines Visayas island group" in 7th Int. Conf. on Information, Intelligence, System & Applications (IISA), Chalkidiki, Greece, 2016. https://doi.org/10.1109/ IISA.2016.7785420

[19] A.L. Buczak, B. Baugher, S.M. Babin, L.C. Ramac-Thomas, E. Guven, Y. Elbert, P.T. Koshute, J.M.S. Velasco, V.G. Roque Jr, E.A. Tayag et al., "Prediction of high incidence of dengue in the Philippines" PLoS Neglected Tropical Diseases, 8(4), e2771, 2014. https://doi.org/10.1371/journal.pntd.0002771

[20] K.A. Agrupis, M. Ylade, J. Aldaba, A.L. Lopez, J. Deen, "Trends in dengue research in the Philippines: a systematic review" PLoS Neglected Tropical Diseases, 13(4), e0007280, 2019. https://doi.org/10.1371/journal.pntd.0007280

[21] Z. Mustaffa, M.H. Sulaiman, F. Ernawan, Y. Yusof, M.F.M. Mohsin, "Dengue outbreak prediction: hybrid meta-heuristic model" in 19th IEEE/ACIS Int. Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Busan, South Korea, pp. 271-274, 2018.

[22] B. Rachmat, O.D. Nurhayati, Suryono, "Prediction the number of patients at dengue hemorrhagic fever cases using adaptive neural fuzzy inference system (ANFIS)" International Journal of Innovative Research in Advanced Engineering, 3(4), 23-28, 2017.

[23] A. Chaikoolvatana, P. Singhasivanon, P. Haddawy, "Utilization of a geographical information system for surveillance of Aedes aegypti and dengue haemorrhagic fever in north-eastern Thailand" Dengue Bulletin, 31, 75-82, 2007.

[24] M. Tipayamongkholgul, C. Fang, S. Klinchan, C. Liu, C. King, "Effects of the El Nino-Southern Oscillation on dengue epidemics in Thailand, 1996-2005" BMC Public Health, 9, article 422, 2009. https://doi.org/10.1186/1471-2458-9-422

[25] T. Soemsap, S. Wongthanavasu, W. Satimai, "Forecasting number of dengue patients using cellular automata model" in 2014 Int. Electrical Engineering Congress (iEECON), Chonburi, Thailand, 2014. https://doi.org/10.1109/iEECON.2014.6925876

[26] N. Sumanasinghe, A.R. Mikler, J. Muthukudage, C. Tiwari, R. Quiroz, "Data driven prediction of dengue incidence in Thailand" in Int. Conf. on Computing and Information Technology (IC2IT 2017), Bangkok, Thailand, pp. 95-107, 2017.

[27] C. Ruangudomsakul, A. Duangsin, K. Kerdprasop, N. Kerdprasop, "Application of remote sensing data for dengue outbreak estimation using Bayesian network" International Journal of Machine Learning and Computing, 8(4), 394-398, 2018.

[28] R. Jain, S. Sontisirikit, S. Iamsirithaworn, H. Prendinger, "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data" BMC Infectious Diseases, 19, article 272, 2019. https://doi.org/10.1186/s12879-019-3874-x

[29] D. Saikia, J.D. Dutta, "Early diagnosis of dengue disease using fuzzy inference system" in 2016 Int. Conf. on Microelectronics, Computing and Communications (MicroCom), Durgapur, India, 2016. https://doi.org/10.1109/MicroCom.2016.7522513

[30] C. Huang, C. Hsu, H. Guo, S. Su, H. Lin, "Dengue fever mortality score: a novel decision rule to predict death from dengue fever" Journal of Infection, 75, 532-540, 2017.

[31] C. Wu, S. Kao, C. Shih, M. Kan, "Open data mining for Taiwan's dengue epidemic" Acta Tropica, 183, 1-7, 2018.

[32] X. Qi, Y. Wang, Y. Li, Y. Meng, Q. Chen, J. Ma, G.F. Gao, "The effects of socioeconomic and environmental factors on the incidence of dengue fever in the Pearl River Delta, China, 2013" PLoS Neglected Tropical Diseases, 9(10), e0004159, 2015. https://doi.org/ 10.1371/journal.pntd.0004159

[33] P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang, W. Ma, "Developing a dengue forecast model using machine learning: a case study in China" PLoS Neglected Tropical Diseases, 11(10), e0005973, 2017. https://doi.org/10.1371/journal.pntd. 0005973

[34] Y. Chen, Z. Zhao, Z. Li, W. Li, Z. Li, R. Guo, Z. Yuan, "Spatiotemporal transmission patterns and determinants of dengue fever: a case study of Guangzhou, China" International Journal of Environmental Research and Public Health, 16(14), article 2486, 2019. https://doi.org/10.3390/ijerph16142486

[35] S.D.P. Jayasundara, S.S.N. Perera, G.N. Malavige, S. Jayasinghe, "Mathematical modelling and a systems science approach to describe the role of cytokines in the evolution of severe dengue" BMC Systems Biology, 11, article 34, 2017. https://doi.org/ 10.1186/s12918-017-0415-3

[36] J.M. Scavuzzo, F. Trucco, M. Espinosa, C.B. Tauro, M. Abril, C.M. Scavuzzo, A.C. Frery, "Modeling dengue vector population using remotely sensed data and machine learning" Acta Tropica, 185, 167-175, 2018.

[37] J.D. Mello-Roman, J.C. Mello-Roman, S. Gomez-Guerrero, M. Garcia-Torres, "Predictive models for the medical diagnosis of dengue: a case study in Paraguay" Computational and Mathematical Methods in Medicine, 2019, article 7307803, 2019. https://doi.org/10.1155/ 2019/7307803

[38] E.C. Farinelli, O.S. Baquero, C. Stephan, F. Chiaravalloti-Neto, "Low socioeconomic condition and the risk of dengue fever: a direct relationship" Acta Tropica, 180, 47-57, 2018.
[39] I. Ogashawara, L. Li, M.J. Moreno-Madrinan, "Spatial-temporal assessment of environmental factors related to dengue outbreaks in Sao Paulo, Brazil" GeoHealth, 3(8), 202-217, 2019.
[40] A.P.R. Dalvi, J.U. Braga, "Spatial diffusion of the 2015-2016 Zika, dengue and chikungunya epidemics in Rio de Janeiro Municipality, Brazil" Epidemiology & Infection, 147, article e237, 2019. https://doi.org/10.1017/s0950268819001250
[41] J.P. Messina, O.J. Brady, N. Golding, M.U.G. Kraemer, G.R. William Wint, S.E. Ray, D.M. Pigott, F.M. Shearer, K. Johnson, L. Earl et al., "The current and future global distribution and population at risk of dengue" Nature Microbiology, 4(9), 1508-1515, 2019.
[42] NOAA, National Centers for Environmental Information, "Climate Data Online" Access 11 January 2020. https://www.ncdc.noaa.gov/cdo-web/
[43] NOAA STAR, Center for Satellite Applications and Research, "Global Vegetation Health Products" Access 11 January 2020. https://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/vh_browseByCountry_province.php
[44] NOAA National Weather Service, Climate Prediction Center, "Cold & Warm Episodes by Season" Access 11 January 2020. http://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php
[45] IBM, IBM SPSS Modeler 17 Algorithms Guide, IBM Corp, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 347-352 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

DC-DC Buck Converter Driver with Variable Off-Time Peak Current Mode Control

Osvaldo Gasparri[*,1], Paolo Del Croce[2], Andrea Baschirotto[1]

[1]*Department of Physics "G.Occhialini", University of Milano, Bicocca, 20126, Italy*

[2]*IFAT DC ATV BP PD ADL, Infineon Technologies, Villach, 9500, Austria*

ARTICLE INFO

Article history:
Received: 29 August, 2020
Accepted: 01 November, 2020
Online: 20 November, 2020

Keywords:
DC-DC Converter
Peak Current Mode Control
Simulink
dSpace

ABSTRACT

Power converters in automotive industries are needed to source electronic devices. DC-DC Buck converter allows to drive loads ensuring their safety operation taking the car battery voltage as an input to generate a sub-voltage or current. By means of a control loop, the system makes sure that the output variable has the desired value in each operation condition. The Peak Current Mode Control (PCMC) is used to control the output current. Whenever the current reaches the reference maximum value (by load specs), the system acts to lower it. Intrinsically over-current protection is guaranteed. The paper presents an improvement of the basic PCMC, where, instead of just controlling the peak value, the loop controls also the average, more delicate feature, resulting in a more reliable driver system. The circuit is able to source 3A average-4.5V from a 13.5V nominal battery voltage, with a peak current of 3.3A and a 0.6A maximum ripple (3±10%A). The concept has been designed and simulated on Simulink and then tested on hardware using dSpace, a Rapid Control Prototyping set-up.

1 Introduction

Electronic devices in Automotive industries need for power converters to be correctly sourced. Indeed, battery voltage could range between 4.5 to > 27V, making it unsuitable for driving any load without provoking damages or malfunctions. DC-DC Converters allow to take as an input the car battery voltage and return as an output a controlled voltage or current required by the load. What is more, the DC-DC protects the load against over-current and over-voltage, ensuring its correct operation. Meanwhile, integrated circuits are required to occupy small space and reduce external components to minimize the associated cost. The core of a Buck DC-DC converter is a switch, or Power-DMOS, which is repeatedly turned on and off, allowing the voltage/current to increase/decrease quickly in the neighborhood of a desired value. The circuit that turns on and off the switch is the control circuit [1]. The control could be either made on the output current or voltage. This paper presents a current control loop able to control all the current features (peak, average and ripple) at once. This paper is an extension of work originally presented in the 26th IEEE International Conference on Electronics Circuits and Systems (ICECS) [2].

The solution arises from the well-known Peak-Current-Mode Control (PCMC), [3] operating in Continuous Conduction Mode (CCM). Instead of having a constant switching period, as it happens in the typical PCMC, the presented control loop has a variable period. In particular, the time in which the power-DMOS stays off, changes accordingly to the calculated current average value, to make it closer to a target desired value. The loop will be now on referred to it as Variable T_{off}. Moreover, the typical PCMC intrinsically suffers of instability issues directly attributable to the fixed switching frequency [4]. Therefore, by making it variable, such issues are consequently solved. The use of stabilization techniques such as Slope Compensation [4] can be avoided, which increase the complexity of the system without improving the control precision (indeed current features will change accordingly to the input voltage). Also protection against over-current is already guaranteed: indeed the maximum current, namely the peak, is fixed. The proposed system operates with 500kHz maximum switching frequency, delivering an average output current of 3 ± 10%A with a load-fixed 4.5V output voltage. The current peak will be limited to 3.3A. The application case could be, for instance, a LED driver, where the LED itself, as soon as the correct current is provided, fixes the output voltage. The circuit has been simulated and then tested in laboratory by using an emulator called dSpace. Instead of a LED, in first place has been used as a load a linear component: a 1.5Ω resistor.

Finally, the current sensing is integrated, with no need for

[*]Corresponding Author: Santa Maria Hoè (LC) Italy, o.gasparri@campus.unimib.it

an external pad to sense the current, resulting in a lighter circuit. Compared to the most common solutions [5], the here used fully-integrated Current Sensing technique requires less external components, with consequent lower cost in terms of BoM and board fabrication.

The paper is organized as follows. The algorithm principle is presented in Section II, together with the formula useful for its implementation, described in Section III. The simulation results are shown in Section IV and the conclusions are drawn in Section V.

2 Variable-T_{off} PCMC Algorithm

A Buck DC-DC converter is a device which in principle allows to convert a given input voltage to a lower output voltage by turning on and off a Power-DMOS, able to withstand to high power. In the application at stake from [4.5, 27]V are obtained 4.5V in output by controlling the current, I_L, flowing into the load. A simple scheme of a Buck converter is given in Figure 1.

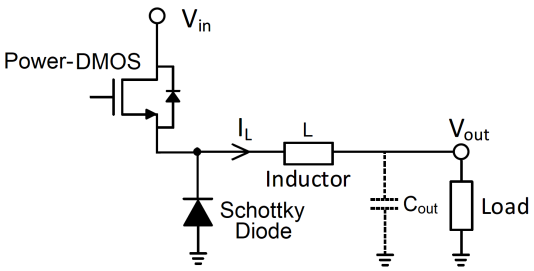


Figure 1: Buck converter scheme

Rather than other diodes, a Schottky diode has been used for its lower threshold (0.5V) and higher speed. The presence of the output capacitor, which minimizes the voltage ripple, can be neglected when controlling the current (and the output voltage is fixed by the load) leading to a lighter circuit. The switching period is divided into two phases: the ON-phase when the Power-DMOS conducts (T_{on}) and the OFF-phase when the diode conducts (T_{off}). Referring to the Duty-Cycle (D):

$$\begin{cases} T_{on} = D * T \\ T_{off} = (1 - D) * T \quad with \quad T = T_{on} + T_{off} \end{cases} \tag{1}$$

The Buck converter should sink an optimum average current I_{avg} to make the load work properly and, to avoid damages, the current must fluctuate limitedly. The current ripple should be indeed limited to the average value±10%, namely 0.3A. Consequently, the maximum current, the peak value, should be 3.15A. The rule of the control loop is to ensure that these specifications are met. The relation between the current features can be expressed as follows:

$$\Delta I_L = 2 * (I_{max} - I_{avg}) \tag{2}$$

In standard PCMC, the period T is fixed. Conversely, in Variable T_{off}, the period is variable. In particular, T_{on} changes automatically according to the input voltage, while T_{off} is forced to vary every cycle by the control loop, based on the vicinity between the calculated output average value of that cycle and the desired value.

The load current during the ON-phase increases with positive slope given by:

$$\left(\frac{dI_L}{dt}\right)_{ON} = \frac{\Delta I_L}{T_{on}} = \frac{V_{in} - V_{out}}{L} \tag{3}$$

Once I_L reaches I_{max}, the power-DMOS turns OFF and the output current decreases with a negative slope:

$$\left(\frac{dI_L}{dt}\right)_{OFF} = \frac{\Delta I_L}{T_{off}} = -\frac{V_{out}}{L} \tag{4}$$

After a time T_{off}, the DMOS is turned ON again and so on. The output current waveform (equal to the inductor current) has the typical triangular CCM shape as in Figure 2.

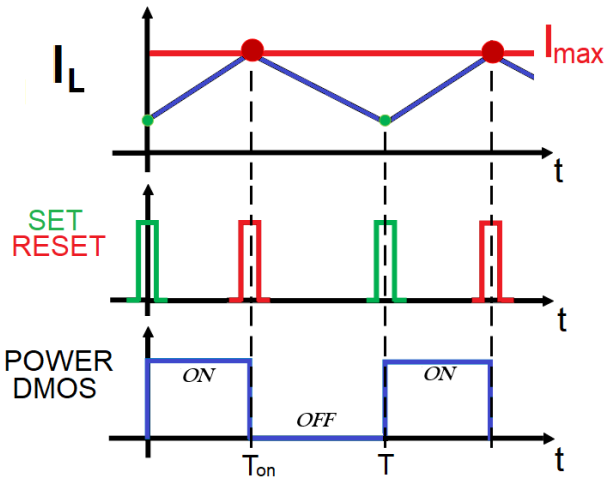


Figure 2: Output current

While I_{max} is fixed, the I_{min} is free to change. The loop, by means of a sensing technique, is able to read I_{min}. Indeed, the current sensing happens during the ON phase to insure the current doesn't grow beyond I_{max}. Therefore, to get I_{min}, it just needed a register to store the first current value when the ON phase begins. Once I_{min} is known, the loop calculates:

$$I_{avg} = \frac{I_{min} + I_{max}}{2} \tag{5}$$

Then, the difference between I_{avg} and $I_{mean,ref}$ = 3A gives an error with sign that, once properly amplified by a factor K, is summed to the previous T_{off} value. The following T_{off} value will be:

$$T_{off,next} = T_{off,prev} + k \cdot (I_{avg} - 3A) \tag{6}$$

In this way, if e.g. I_{avg} is higher than 3A, the T_{off} value will be increased (causing a stretching of the switching period) and allowing the current to decrease more during the OFF phase, reaching a lower I_{min} with the aim of reducing the average value. If otherwise I_{avg} is lower than 3A, the error will be negative, meaning that the T_{off} of the next switching cycle must decrease to make the average value higher and matching with the target. The loop operation is resumed in the flow-chart in Figure 3. The total number of regressions in the algorithm depends on the current rising slope, on the sensing technique sampling frequency and on the counter frequency. Every time a switching cycle ends, the operation restarts from the beginning giving rise to the next switching cycle.

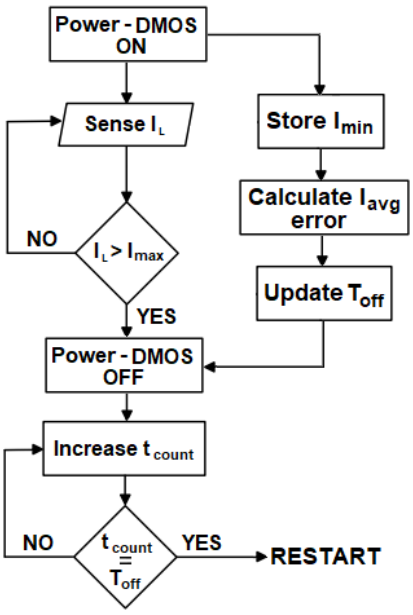


Figure 3: Variable-T_{off} PCMC algorithm flowchart

While T_{off} is decided by the loop, the T_{on}, from (3), changes not only with the the ΔI_L -loop dependent- but also with the battery voltage:

$$T_{on} \propto \frac{1}{(V_{in} - V_{out}) \cdot \Delta I_L} \tag{7}$$

It should be noticed that a first pre-memorized T_{off} value guess must exist (to be later overwritten by the cycle) to avoid malfunction at the cranking of the system during the very first switching cycle. This first guess is based on the application: if for instance the desired average is 3A, the peak is 3.15A, the inductance 30μH and $V_{out} = 4.5V$, it follows that the T_{off} value, from (2) and (4), could be grossly chosen as:

$$T_{off} = \frac{L \cdot \Delta I_L}{V_{out}} = 4\mu \tag{8}$$

Then, given some V_{out} variations due to the load or some uncertainties related to the inductance value, the references etc., the loop will act as described above, varying the T_{off} value to better adapt the system to the load specifications.

Moreover, the T_{off} value has an inferior limit, due to the maximum power-DMOS frequency (f_{max}) allowed. Ideally, the f_{max} can be evaluated for T_{on} tending to zero, namely:

$$f_{max} \simeq \frac{1}{T_{off}} \tag{9}$$

In realty, the minimum T_{on} value won't be zero, but, from (3), for a $V_{in} = 27$V it will be about 0.8μ s.

3 Instability digression

The typical PCMC loop is just able to control the maximum load current, no matter what ripple and average current are. Therefore, such a solution is not suitable in applications where loads require a precise average current with small shifts around its value. However, the biggest issue of the typical PCMC comes from its instability when the duty cycle grows over 50% [6]. Every time $T_{on} > T_{off}$ or, equivalently, $V_{out} > V_{in}/2$, the system would be unstable. In the application case, where $V_{in} = [4.5, 27]$ V and $V_{out} = 4.5$ V, problems would arise when $V_{in} < 9$V. Figure 4 shows what happen when crossing the limit case $D = 0.5$: if the duty cycle is increased the PCMC is no longer helping in controlling the current. The reason behind this issue is that, in typical PCMC, after a fixed period T, the power-DMOS is forced to be turned ON. Instead, in Variable-T_{off}, the power-DMOS turns ON T_{off} seconds after the current has peaked. Contrarily to the typical PCMC, if the input voltage changes, i.e. T_{on} changes, the T_{off} doesn't change (4), the ripple doesn't change and therefore the current features are not impacted. The system stays stable. Techniques as Slope compensation [4] are no longer needed.

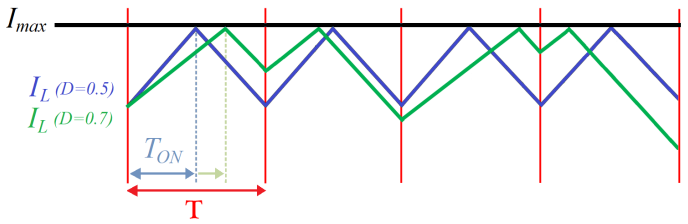


Figure 4: Output current in typical PCMC

4 Constant-T_{off} Algorithm Implementation

The circuit implementation is shown in Figure 5. The only external components are the inductor (L) and the Schottky Diode. As internal components, enclosed in blocks, there are: the Current Sensing, the Control Logic blocks, the Protection, the Internal Reference (including a bandgap voltage reference to generate I_{max} and I_{mean}), the Gate Driver to turn ON and OFF the Power-DMOS [7] and the Power Supply to generate a low voltage domain for the low-side circuit.

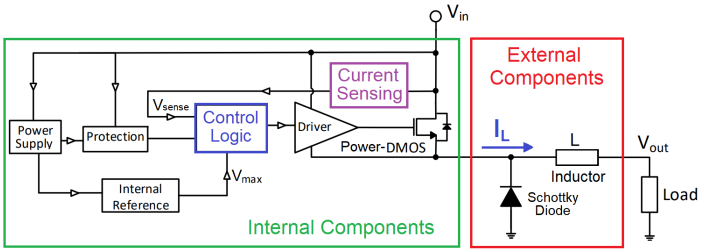


Figure 5: Concept top level

The Current Sensing block is used to read the current flowing I_L. The Power-DMOS current is here mirrored with a factor $N = 1/1000$ and flows through a sensing resistance. Based on the read voltage drop, the system is able to determine the magnitude of the current. Since the proposed technique only operates during the ON-phase (when the Power-DMOS conducts) the power consumption is lower than in the case on which the current is sensed on

the output node, always operating. Furthermore, an internal sensing block also allows to minimize the area occupation on silicon.

4.1 Control Logic block

Figure 6 shows the concept Control Logic block implementation. The result of the comparison between I_{sense} and I_{max} gives the Reset of the SR Latch. When $I_{sense} \geq I_{max}$, the Reset activates and the Power-DMOS turns OFF through the Gate Driver. The Set input is instead connected to a counter consisting of an oscillator and a chain of four D Flip-Flops. As soon as the Power-DMOS turns OFF, the counter starts counting up to T_{off} duration. Afterwards the Set is activated and the Power-DMOS turns ON. The current value at this moment, I_{min}, is stored into a register. The loop then calculates the average error on respect to the desired I_{mean} value. The error, summed to the previous $T_{off,prev}$ value returns the $T_{off,next}$ as in (6).

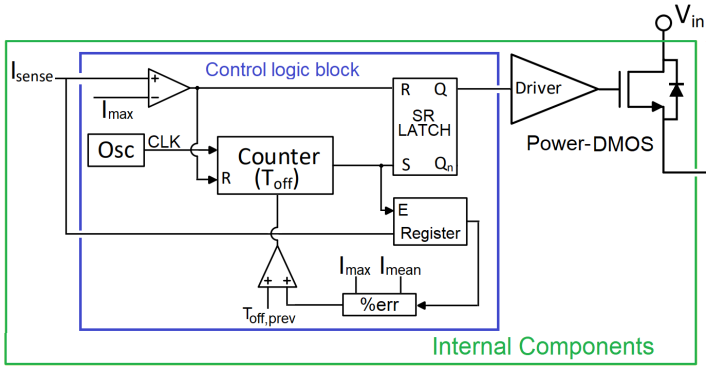


Figure 6: Variable T_{off} Current Control Loop implementation

5 Design example and Simulations results

Table 1 resumes the application specs.

Table 1: Summary of specifications

Technology	0.35 μm
V_{in}	4.5-27 V
V_{out}	4.5 V
f_{max}	500 kHz
Maximum Current Level	3.3 A
Mean Current Level	3 A
Output Current Ripple	0.60 A

By applying the formulas given in Section I, the design parameters are obtained. (9) gives the lower T_{off} limit:

$$T_{off} \simeq \frac{1}{500kHz} = 2\mu s \tag{10}$$

In fact, $1/f_{max} = T_{on,min} + T_{off}$ and (3) becomes:

$$V_{out} \cdot T_{off} = (V_{in,max} - V_{out}) \cdot (1/f_{sw,max} - T_{off}) \tag{11}$$

Therefore a more precise estimation of the minimum T_{off} value is given by:

$$T_{off} = \frac{27V - 4.5V}{27V} \cdot \frac{1}{500kHz} = 1.7\mu s \tag{12}$$

Once the ripple (Table 1) and T_{off} (12) are given, both the peak and the average current are fixed. Furthermore, by using these parameters in (4), the inductance value can be chosen as:

$$L > \frac{V_{out} * T_{off}}{\Delta I_L} = 13\mu H \rightarrow 30\mu H \tag{13}$$

(to stay large, considering also 20% of inductance value uncertainty on the nominal value and considering that actually the typical operation is centered around V_{in} = 16V).

5.1 Simulation results

The concept of Figure 5 has been firstly implemented on Simulink (MATLAB) to ideally simulate the behaviour of the high-level concept design. Simulink works with a block approach system, shown in Figure 7, which follows exactly from Figure 5. Other software are actually exploited in the design phase (such as Virtuoso-cadence). However, there, the circuit must be designed in a deeper transistor-level. Since the design may take large amounts of time, prototyping must be done previously ensuring that the chosen topology is the most suitable one for the application in question and that its worthwhile to proceed with transistor-level design and future fabrication.

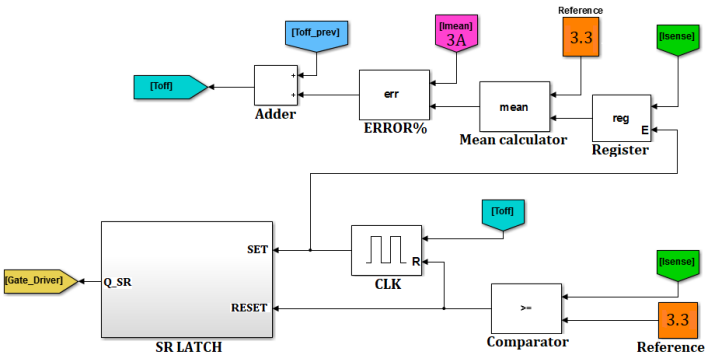


Figure 7: Simulink implementation

The simulations include also worst case scenarios to ensure the correct operation in each condition. To simulate the presence of a load which fixes the 4.5V output voltage with a 3A average current, a 1.5Ω resistor has been used.

Figure 8 (left) shows how the system works with a 7V input voltage and a peak reference reduced from 3.3A to 3.2A, i.e. with a Δ_{I_L} = 0.4A. As expected, the operation is not affected by instability issues which, in a typical PCMC, would surely arise whenever V_{in} < 9V. Later, the input voltage has been abruptly increased to 18V to also prove that a V_{in} change (as could happen during the cranking of the car) has no relevant effect on the operation. The T_{off} remains unchanged (no dependency on the V_{in}) and equal to 2.7μs as expected from (7) (with Δ_{I_L} = 0.4A).

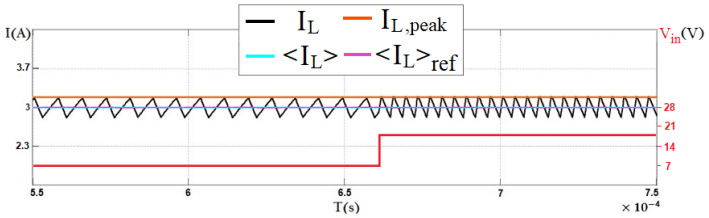


Figure 8: System response changing the input voltage

Figure 9 shows instead the transient simulation plot for V_{in} = 18V with the peak reference at 3.3A as from specification. As expected from (7), the T_{off} = 4μs. To further prove the system robustness, an average reference positive step (from 3A to 3.15A) has been added. In order to achieve the new average current, the loop gradually reduces the T_{off} to make the calculated average equal to the target. Since the peak reference stays equal, according to (8) and (9), the ripple reduces and the switching frequency increases up to 260kHz.

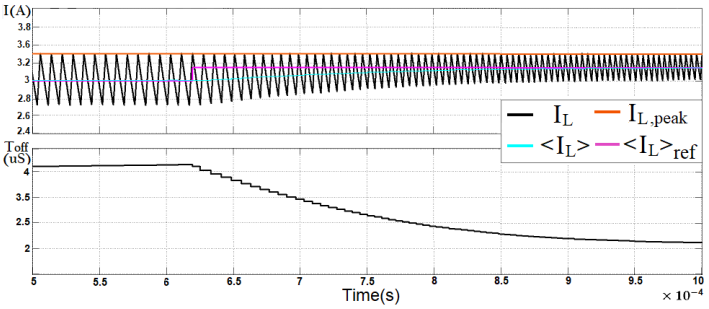


Figure 9: System response changing the average reference

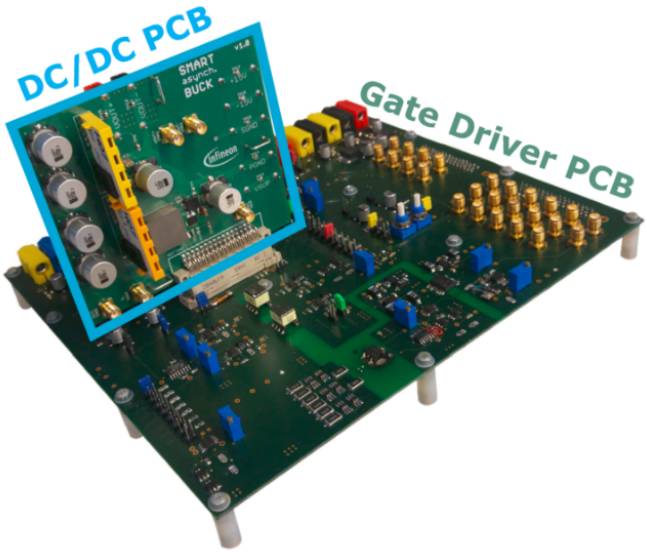


Figure 10: Buck Converter and Gate Driver PCBs

Once proven the concept works as expected, the system can be tested in laboratory by means of a rapid control prototyping called dSpace. Here, the power-DMOS and external components are no longer digital, but located on a DC-DC converter Printed Circuit Board (PCB), which is connected to a Gate driver board (Figure 10), connected in turn to dSpace.

dSpace, through its FPGA, will embody the Simulink bloch approach control logic circuit (uploaded by PC), also providing the current sensing, the internal reference and all the devices to communicate with the hardware (ADCs, DACs). dSpace allows to test whatever control circuit including all the un-idealities of the system, verifying moreover the correct operation of the power-DMOS and that one of a real load. In this way it can be eventully proven that, not only the circuit works as expected, but that is suitable for the application at stake. Through a Graphic User Interface, in Figure 11, the user can monitor the system operation using the PC. Not only to insert the protection to avoid OC and OV, but also to vary the references and other parameters and observe the system response and waveforms in each condition.

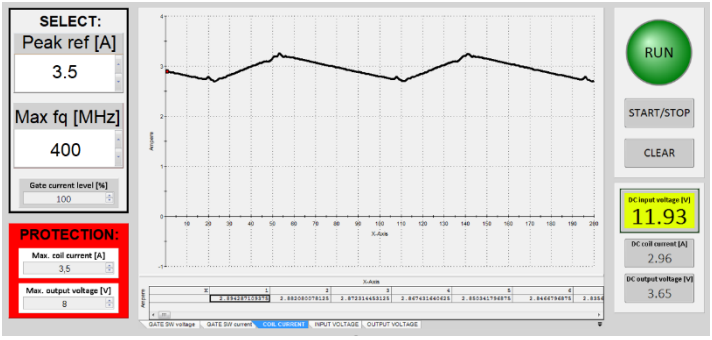


Figure 11: Graphic User Interface

Of course, for having a higher sampling precision, the same waveform can be actually taken from oscilloscope, shown in Figure 12. As expected, the circuit is able to drive the load maintain the desired current features over time and with the same theoretical behaviour of Figure 2.



Figure 12: Output current (violet) and gate driver (green) Waveforms from oscilloscope

6 Conclusions

The presented DC-DC Buck Converter Variable-T_{off} loop allows to precisely and smartly control the current features, returning a reliable car battery sourced driver. Other than the current peak, the

loop is able to control the average current. Moreover, given (2), also the ripple is indirectly controlled. With respect to other solutions, the switching period is variable, intrinsically solving the instability issues. Indeed, the T_{on} changes with the input voltage while T_{off} increases/decreases at each switching cycle accordingly to the comparison between I_{avg} and the target value, making (2) smaller/bigger (i.e. increasing/decreasing I_{avg}).

The system exploits an integrated low-side sensing resistor circuit to minimize the Printed Circuit Board usage space, avoiding extra pads to sense the current. To make the chip lighter and cheaper, the Buck Converter C_{out} has been neglected, as superfluous for the application.

The concept has been simulated with Simulink and then a more rigorous test has been conducted in laboratory with dSpace where worst condition simulations shown an excellent dynamic behaviour. The circuit appears to be robust, proving that sudden perturbations on V_{in}, I_{max} and T_{off} do not upset the operation of the circuit. Compared to [8], the presented solution is lighter and works with an input voltage range of $[4.5, 27]$V, able to drive a 3A LED with 500kHz maximum switching frequency, allowing extremely good average regulation precision and speed. A more detailed comparison of all Buck performances, including efficiency, with the state of the art is provided in [7].

Acknowledgment A part of the work has been performed in the project iDev40. The iDev40 project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 783163. The JU receives support from the European Union's Horizon 2020 research and innovation programme. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain and Romania.The information and results set out in this publication are those of the authors and do not necessarily reflect the opinion of the ECSEL Joint Undertaking.

References

[1] L. Dixon, "Average Current Mode Control of Switching Power Supplies," 1999.

[2] O. Gasparri, R. Di Lorenzo, P. Del Croce, A. Pidutti, A. Baschirotto, "Variable off-Time Peak Current Mode Control (VoT-PCMC) as method for average current regulation in Buck Converter Drivers," 2019.

[3] I. J. Prasuna, M. Kavya, K. Suryanarayana, B. Shrinivasa Rao, "Digital Peak Current Mode Control of Boost Converter," 2014.

[4] Z. Jiuming, L. Shulin, "Design of slope compensation circuit in peak-current controlled mode converters," 1310–1313, 2011.

[5] J. Ejury, Buck Converter Design, Infineon Technologies North America (IFNA), 2013.

[6] Fang, Chung-Chieh, Redl, "Subharmonic Instability Limits for the Peak-Current-Controlled Buck converter With Closed Voltage Feedback Loop," Power Electronics, 1085–1092, 2015.

[7] R. Di Lorenzo, O. Gasparri, A. Pidutti, P. Del Croce, A. Baschirotto, "On-Chip Power Stage and Gate Driver for Fast Switching Applications," 2019.

[8] V. Anghel, C. Bartholomeusz, A. Vasilica, G. Pristavu, G. Brezeanu, "Variable Off-Time Control Loop for Current-Mode Floating Buck Converters in LED Driving Applications," 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 351-362 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design and Optimization of a Three Stage Electromechanical Power Unit using Numerical Methods

Yashwant Kolluru[1,*], Rolando Doelling[1], Lars Hedrich[2]

[1]*Robert Bosch GmbH, eBikes Department, Kusterdingen, 72770, Germany*

[2]*Johann Wolfgang Goethe-University, Institute of Informatics, Frankfurt am Main, 60054, Germany*

ARTICLE INFO

Article history:
Received: 15 May, 2020
Accepted: 17 July, 2020
Online: 28 July, 2020

Keywords:
Vibro-acoustics
Numerical Methods
Compliance
Stress

ABSTRACT

The advent of electric vehicles has changed the face of the automobile industry. The drive system properties of vehicles such as eBikes or electric cars differ fundamentally from those of a diesel engine. The lack of a conventional internal combustion engine has made the vehicles considerably silent. Nevertheless, previously hidden sources of vibration and noise have become more dominant. In addition to these emissions, other structural properties such as compliance and deformation also appear as relevant factors for the original equipment manufacturer. Usually, deterioration of these variables affects the efficiency of the power unit. In this paper, a simulation template is created to understand and analyze these properties of the drive unit. Furthermore, new enhancements to improve the key indicators, such as strain energy, natural frequencies, etc., are shown, thereby creating a potential method flow to develop better performing drive units. Numerical optimization tools are used to simulate structures with complex shapes that exactly meet the mechanical constraints and use as little material as possible. In this work, two optimized variants of electromechanical drives are presented. The first scenario illustrates the optimized model with an objective of minimizing the strain energy of the structures, whereas the second task aids in the development of a variant with superior dynamic properties than the current drive units. Ultimately, several numerical calculations are validated using experiments.

1. Introduction

This paper is an extension of work that was originally presented in ISSE 2019 conference [1]. Originally, paper [1] focused on creation of a multi-body model for analyzing the vibro-acoustic characters inside the eBike. The current work illustrates the extension of the previous setups to the optimization domain using the finite element techniques.

In the current pedelec industry, the core problems faced by manufacturers are the structural behaviors, such as vibration and deformation, due to loading. Lately, researchers have mainly focused on the development of the electric motors, batteries and power electronics, which has laid a solid foundation for building advanced electric powertrains [2]-[4]. Improving the energy efficiency of the entire electromechanical drives was the essential parameter, while areas related to vehicle dynamics and structural properties, such as vibration, noise, stiffness, compliance, etc., are not extensively examined. However, several findings show that the vibrations of electric drives can lead to several problems in regard to stability and driving comfort of the vehicle [5, 6]. In addition, the presence of stresses in structures beyond the yield limits leads to larger deflections and thereby reduce the efficiency of the components [7]. Understanding the static and dynamic behavior of the electromechanical drives therefore remains an essential task for the Original Equipment Manufacturers (OEMs). This paper attempts to develop a method that can be used to capture structural behavior early in the product to avoid unnecessary costs in the later stages of the product lifecycle. The method was developed largely with the help of the Finite Element Modeling techniques (FEM).

The numerical discretization of known structures is a popular topic in computational mechanics. The end user uses the software iteratively to optimize a structure. It is up to him/her to improve the design in order to identify a good geometry [8]. In contrast, topology optimization software outputs the optimal geometry using structured mesh, directly after the optimization task has been defined. Structural topology optimization helps the designer in modeling the type of structure which best fits to satisfy the operating conditions for the problem [9]. It changes the general

*Corresponding Author: Yashwant Kolluru, Gerhard-Kindler-Str. 3, Kusterdingen 72770, Germany. +49-17656830708, yashwant.kolluru@de.bosch.com

component layout to find designs with goals such as minimization compliance or maximization of natural frequencies. Stress and compliance optimization are key indicators in the structural design for a range of engineering applications. Optimization helps in the design of mechanical components with lower stress concentrations [10]. Once the results are extracted from the optimization module, they are enhanced using the smoothing functions provided by the optimization tool. It can be seen as a procedure to optimize the rational arrangement of the available elements in the design volume and to eliminate the unnecessary elements.

In a similar way, the structure borne emission from housings must also be optimized. Vibrations are undesirable movements in the powertrain and are responsible for the problematic noise and energy losses in the system. In general, they are generated from excitations at structural components and would later affect the overall performance of the Drive Unit (DU). In order to maintain the efficiency of the drive, vibration and noise emission should be reduced below the threshold values. One way of achieving this is through the dynamic optimization technique [11]. As part of this approach, the simulation template is created to analyze the surface acceleration and acoustic quantities. Later, a dynamic optimization module is developed to strengthen the structure and reduce the vibrations in the drivetrain.

2. State of the Art

2.1. Stress Analysis in Drivetrain Components

Structural analysis is generally performed to check the durability of certain parts for a particular load and support condition. For the component to be structurally safe, stresses in the body should not be more than the yield strength of the material. Considering the possibility of fatigue failure, the component must be optimized such that the Von-mises stress generated does not exceed the endurance strength of the material. There are several articles that generally describe the stress and deformation of various components such as gears, motors, etc. For instance, paper [12] written by Mao discusses the importance of transmission error in gears and its influence in powertrain structures. Similarly, Chung published his results in paper [13] that focused on stress analysis of helical gears and their deformation profile for different load cases in normal drive trains. Electromechanical drives have distinct structural characteristics. Most of these papers principally illustrate the gear forces in static load cases or influences in the performance of normal drivetrains due to errors at the gears. However, there are no proper references available for analyzing electromechanical drives as entire unit, for example, considering the influences of multiple domains such as axial and electromagnetic forces, static and linear dynamic cases together, or the impact of excitations, forces at the bearings and the structural behavior of the housing combined. Moreover, adding optimization tasks makes the existing scenario more complex. This paper tries to solve this problem by examining and simulating each juncture responsible for deformations, critical stresses etc. In this way, a holistic simulation template is provided with which the entire drivetrain can be analyzed and a future variant with optimized properties can be developed.

2.2. Modal Analysis of Complex Structures

Vibrations are nothing but mechanical oscillations of a system/component at an equilibrium point. A statically balanced object oscillates at a certain frequency depending on its properties such as mass and stiffness. This frequency is referred to as the natural frequency of that structure [14]. The calculation of these frequencies and their respective mode shapes is essential for researchers to understand the vibro-acoustic properties of the components. There are usually two sets of models used to analyze the vibrations of structures, continuous and discrete. The continuous models are mathematical models applied to continuous data and are represented by the Partial Differential Equations (PDE). The coefficients of these PDE's depend on the geometrical and material properties. Solutions to these equations are obtained using numerical methods such as Dunkerley's method, Rayleigh-Ritz method, Newton–Raphson method, Galerkin approach, etc., [15 - 17]. The discrete models are those models with finite degrees of freedom and are generally described by ordinary differential equations. Two main techniques that deal with the discrete models are finite element and rigid element methods.

The structure of the DU is complex and consists of more than 60 components and 800000 Degrees Of Freedom (DOF), which makes it difficult to analyze either with the continuous models or the rigid element method. FEM is used as a simulation technique considering the factors such as complexity (large number of DOF's), modeling (more flexibility to include the effects like contact, hybrid modeling, etc.), adaptability and accuracy (better precision) for calculating the vibrations. FEM generates the physical response of the system at any given position, including some that may have been neglected in an analytical approach. In addition, FEM has an advantage of modeling noise using acoustic finite, infinite elements and the impedance boundaries [18]. The Eigen forms are calculated using the solution technique described by Grimes, Lewis, and Simon in paper [19].

2.3. Relevance of Optimization

Researchers have developed numerous solutions with regard to the structural topology optimization problems. Below is a list of few approaches developed within the topology optimization domain [8]:

- Ground structure approach discussed in the Shiehin's paper [20].
- Homogenization method described by the Bendsoe and Suzukiin in papers [11, 21], respectively.
- Bubble method shown by the Eschenauer's paper [9].
- Fully stressed design technique depicted in Xiein's paper [10].

The first three approaches listed above have some characteristics in common. These are optimization methods with an objective function and a set of design variables and design constraints. They solve the optimization problem, either by a sequential, quadratic programming algorithm (approach 1) or by an optimality criterion concept (approaches 2 and 3). In contrast, the fully stressed design technique, although not an optimization algorithm in the conventional sense, proceeds by removing inefficient material and thereby optimizes the use of the remaining material in the structure, in a step-by-step process.

Within this paper, the methods majorly from approach 1 are used as guidelines for the design of drivetrain structures to improve the stiffness, dynamic characteristics and to reduce the weight of the components. In addition, the optimization design can be

regarded as the topology optimization problem of continuum structures. Structural optimization attempts to find the best path for transferring all types of loads.

This paper deals with the optimization of drivetrain structures for both static and dynamic load cases. Using a topology optimization approach to design new drivetrains should take less time than the traditional trial-and-error experiments.

2.4. Definition of the Optimization Task

Optimization can be used to shorten the development phase by upgrading insight and intuition of the designers through an automated process. To optimize the powertrain model, the variable that needs to be optimized must first be determined. Optimizations to minimize stress or to maximize the Eigen frequency are simply not sufficient. The task must be more specific, for example minimizing the maximum node stresses that occur during a static load case with a limitation of the existing weight, or similarly maximizing the sum of the first natural frequencies with limit on volume of the components. The goal of the optimization task is referred to as the objective function. In addition, certain conditions can be enforced during the optimization, for instance, the displacement of a given node can be curtailed to a certain value. This enforced condition is often called a constraint.

In addition, the structural domain called the reference domain can be divided into the design and non-design areas. The non-design domain includes key regions such as supports, boundaries and areas where loads are applied. Therefore, these regions cannot be modified throughout the entire optimization process. With the constraint on weight, objective functions, such as optimizing stress, stiffness, etc., are used for the optimization of the static load case. In linear dynamic analysis, the visco-elastic material definitions are used to simulate the drivetrain. Here, the goal of optimization is to curtail the surface vibrations and acoustic noise below the limits.

Overall an optimization task is represented using three characteristics. Below is a list of the characteristics, including concerning questions:

- Design variables - How can the structure be modified? Values, such as frequencies, weight, volume and stiffness, can be altered to obtain better vibrations.
- Design constraints – Which restrictions must be implemented in the model? Variables, such as mass and volume, can be limited to certain ranges.
- Design objective functions - What should be minimized, maximized, optimized, etc.? Function of compliance with a goal in increasing the stiffness and reducing the stress. Function of natural frequencies (variable) to push the critical spectrum domains, thereby avoiding vibrations.

2.5. Mathematical formulations

2.5.1. Gearmesh Excitations

The parametric formulation for gears characterized as lumped mass is shown in (1) (1 – gear1, 2 –gear2). It represents the dynamic Equation of Motion (EOM) of one of the gear with to gearmesh excitations. Here the term $e(\psi(t))$ includes the effects of parametric excitation. It is developed as sum of mean excitation plus sine curves (considering the harmonics of the tooth cycle) with corresponding amplitudes. $\psi_i(t)$ (i = 1, i = 2 – gear2) and its time derivatives represent the rotation angle, angular velocity and angular acceleration respectively. M, J, r, z, ψ, e, k_z and d_z represent torque, inertia, radius of pitch circle, number of tooth, rotation angle, excitation, stiffness and damping respectively [1].

$$J_1\ddot{\psi}_1 = M_1(t) - r_1 k_z(\psi_1(t))[r_1\psi_1(t) + r_2\psi_2(t) + e(\psi_1(t))] - r_1 d_z(\dot{\psi}_1(t))[r_1\dot{\psi}_1(t) + r_2\dot{\psi}_2(t) + \dot{e}(\psi_1(t))] \quad (1)$$

Furthermore, the gear forces are described in (2), (3) and (4). Here, $\overline{F}$ consists of three components, namely the tangential, radial and axial force of the gear. $\overline{F_n}$,$T_ß$, $T_{\alpha n}$, c and s describe the normal force, rotation matrix for the helix angle, rotation matrix for the normal pressure angle, cosine angle and sine angle, respectively [1]. These gear forces cause the shafts and bearings to vibrate, which in turn sets the structure borne emission in rest structures.

$$T_{\alpha n} = \begin{bmatrix} c(\alpha n) & s(\alpha n) & 0 \\ -s(\alpha n) & c(\alpha n) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$T_ß = \begin{bmatrix} c(ß) & 0 & -s(ß) \\ 0 & 1 & 0 \\ s(ß) & 0 & c(ß) \end{bmatrix} \quad (3)$$

$$\overline{F} = T_ß T_{\alpha n} \overline{F_n} \quad (4)$$

2.5.2. Fluid Structure Coupling

The housing-air coupling occurs at the interface at which the acoustic medium interacts with the structure. This is called the fluid structure interface. This coupling between the housing structure and air medium is modeled by conservation of linear momentum.

Equation (5) shows the general formulation of the housing and surrounding fluid at its interface. The terms u, M, M_f, M_{fs} K, K_f, K_{fs}, F_s and F_a indicate the displacement of housing, mass of the housing, mass of the fluid/ air, mass of the fluid structure interface, stiffness of the structure, stiffness matrix of fluid, stiffness matrix of fluid structure interface, force applied by the structure and force exerted by the air respectively. Detailed information about the coupling finite element models can be found in papers [22, 23].

$$\begin{bmatrix} M & 0 \\ M_{fs} & M_f \end{bmatrix} \begin{bmatrix} \ddot{u} \\ \ddot{p} \end{bmatrix} + \begin{bmatrix} K & K_{fs} \\ 0 & K_f \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} F_s \\ F_a \end{bmatrix} \quad (5)$$

2.5.3. Optimization formulations

Furthermore, (6) represents the mathematical form of the optimization task. The objectives are defined as minimizing the compliance (C) and the surface accelerations (A). Volume is considered as a design constraint (in this example) and is limited to V_{max}. ρ is a density vector containing the element densities ρ_e. When ρ_e is 1 we consider an element to be filled whereas an element with ρ_e as 0 is considered to be a void element. Ω is considered as the design volume of the optimization task. The volume constraint prevents the optimized structure from ending with the full design volume when searching for its maximum structural stiffness.

$$\begin{aligned} &Minimize\ C(\rho)\ or\ A(\rho) \\ &Subject\ to\ 0 \le \rho \le 1\ and \\ &\int_\Omega \rho\ d\Omega - V_{max} \le 0 \end{aligned} \tag{6}$$

One way to maximize a structure's global stiffness is to minimize its compliance. The compliance is therefore defined as the equivalent strain energy of the FE solution, which yields higher stiffness when minimized. The compliance is defined in (7) and u is obtained by solving the equilibrium equation (8).

$$C(\rho) = f^T u \tag{7}$$

$$K(\rho)u = f, \text{ with } K(\rho) = \sum_{e=1}^{nel} \rho_e K_e^0 \tag{8}$$

K_e^0 is the initial elemental stiffness matrix. Using the gradient based approach, derivatives with respect to C(ρ) are evaluated. Finally, the case of multiple loads as (9) can be included in the structural optimization task by using weights and objectives that are subjected to a particular load case with the index k. M is the total amount of load cases. Equations (7) and (8) can also be further modified for dynamic optimization of surface accelerations [24].

$$f = \sum_{k=1}^{M} f_k w_k \tag{9}$$

In previous models, optimality criterion, a controller based algorithm was utilized for the optimization. This is faster than the sensitivity based algorithm since no gradient information is required [11, 24]. However, the controller based algorithm is based on stresses and is limited to optimizing stiffness under a volume fraction constraint. Sensitivities are not evaluated since the controller uses strain energy and stresses as an input. In this paper, the optimization tasks are executed using sensitivity based Method of Moving Asymptotes (MMA) approach. It solves non-linear optimization problems in the sense that it uses sequences with sub-problems that are approximations of the original problem [25]. For MMA, these sub-problems are constructed by gradient information and it is also assumed that these approximations are convex. Different design responses can also be combined, for example if the objective is to minimize compliance, both the volume and effective stress can be included as constraints.

The primary focus of the paper lies in the development and optimization of a three stage transmission. The modeling and simulation section describes the architectural layout, intricacies involved in the development of the static simulation template, the forces in the linear dynamic step, coupling between the vibration and acoustic domains, etc., are described. In the optimization section, housing weights are used as design variables. The vehicle performance and indicators, such as compliance, acoustic pressure, etc., are considered as design objectives.

The design procedure is defined in three steps. First, the FE model is used to analyze the original structure. The stiffness and Von-mises stress obtained from the analysis are then treated as the design variables in the topology optimization. The constraint of the optimization task is the reduction in total mass while improving the performance parameter such as vibration in comparison to the original. Eventually, two different optimized variants are developed: one for the static scenario and the other for the linear dynamic load case.

3. Development of the Numerical Model

Before the optimization task is illustrated, it is essential to create a simulation template that takes into account all the characteristics, such as excitation mechanisms, loads at the bearings, deformations of the critical components, etc. Considering the necessary complexities within the modeling technique enables the user to build a model that is close to reality. Detailed information regarding the simulation templates with analyses at multiple domains, such as the Lumped Parameter Model (LPM), Multi-Body Dynamics (MBD), etc., can be found in papers [1, 23].

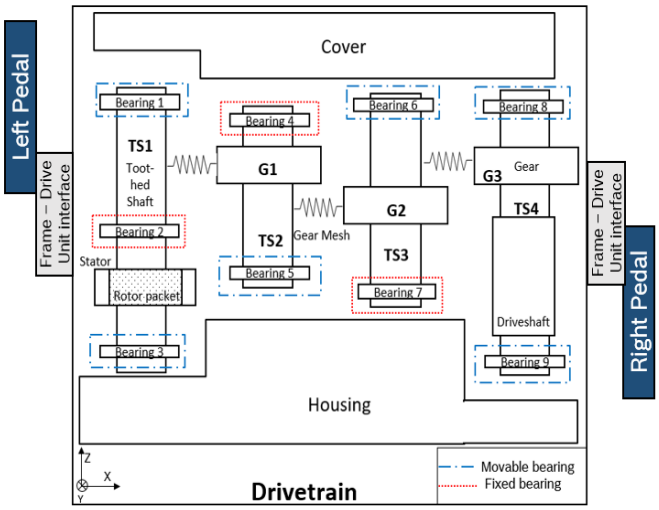


Figure 1: Architectural layout of housings, gears, toothed shafts and bearings of the drivetrain

Figure 1 describes the layout of simulated drivetrain instance to check for critical stresses, deformations and Eigen frequencies. The rotation of the toothed shafts and gears would cause transmission errors that lead to vibrations in the structure. The torque at the toothed shafts and axial load at the bearings would in turn cause the housing structures to deform. Along with the load, boundaries are also applied in the simulation template. For instance, the DOF of bearing nodes are either fixed, thus the z-axis rotation is free, or movable, thus the z-axis translation and rotation are free. Furthermore, the drivetrain is connected to pedals at the frame bracket interface; depending on the load case scenario, the model is enforced with loads at the left pedal, right pedal, shaft axles, gear toothmesh, etc.

Vibrations are influenced by structures having heavier mass or greater stiffness. Therefore, the components that contribute to the drive train vibrations must first be determined. Table 1 describes the scaled masses, Young's modulus and the yield limits of different structures available in the drivetrain. Together, these structures represent about 92 % of the total mass of the power unit. Hence, the structural validation of these components individually and in combination would help to understand the deformations and vibrations of the entire drivetrain in later stages.

The FE software Abaqus and the optimization tool Tosca from Dassault Systems are used to implement the simulation and optimization models [26].

Table 1: Modeling information of the individual components of the drivetrain (scaled)

Component	Mass (Kg)	Young's modulus (Mpa)	Plasticity limit (MPa)
Housing	0.113	0.119	0.14
Cover	0.077	0.119	0.14
Stator	0.268	0.952	0.651
TS4	0.065	0.938	1
G3	0.063	1	1
TS3	0.065	1	1
G2	0.019	1	1
TS2	0.01	1	1
G1	0.015	1	1
TS1	0.023	1	1
Driveshaft	0.044	0.933	0.94
Rotor packet	0.079	0.881	0.417
Frames	0.045	1	0.515

3.1. Static Analysis Template

This section describes the model parameters associated with simulation of the static load case. It illustrates a simulation template for analyzing the electromechanical drive for its stresses and deformations. In this scenario, the load is applied in multiple steps: first at the left pedal and later at the right pedal.

To test the mechanical strength of the power unit, it must be viewed in its most critical state, i.e., when it is subjected to the maximum forces during its use. For this reason, the bike is considered at a position with a maximum force of 1800 N on the pedals. According to the German standard DIN EN ISO 4210-8 for "bicycles-safety-technical requirements for bicycles", two cases have to be considered in order to validate the strength of a system [27]. These cases are called" uphill" and "downhill".

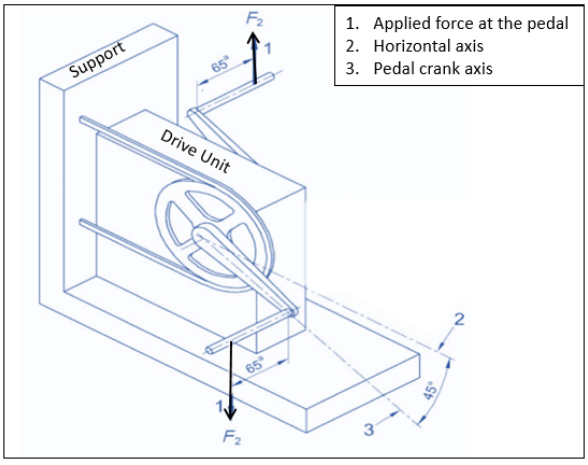


Figure 2: Schematic view of the static load case

Figure 2 corresponds to the uphill load case, where the force on the left and right pedals act in the downward and upward directions, respectively. This standard defines the position of the pedals in relation to the axis of the chain, as well as the forces and boundary conditions acting on them.

Furthermore, the Abaqus simulation model developed must comply with these standards. The boundary conditions apply to the chain and frame. In the uphill position, the simulation consists of two steps, in which a force of 1800 N is applied first to the right pedal and then to the left pedal. The mesh is then created and refined so that the simulation converges and areas with high stress are re-meshed finely for better accuracy. Table 2 describes the general information regarding the FE model.

Table 2: FE model information for static load case

Count of nodes	Count of elements	Load step 1	Load step 2
872148	525712	Right pedal 1800 N	Left pedal -1800N

A specific magnesium alloy with a yield limit of 140 MPa is used as the material for the housing structures. The fracture occurs at a plastic elongation of more than 6 %. With a safety factor of 2, the limit value at 3 % is considered further for the simulations.

The maximum stresses of both uphill and downhill simulation are comparable for their respective cases and hence do not allow the most critical case to be determined. In order to circumvent this problem, a life cycle study is performed to determine the most critical position. This is not the subject of this paper, but is carried out to find the critical load case. Uphill load case with specific orientation (not disclosed) is found out to be the worst in comparison with downhill and other orientations, therefore, it is considered for the further part of the work.

Table 3 corresponds to results obtained for the uphill simulation. It can be noted that PEEQ (equivalent plastic strain) of the two housing bodies is below the tolerable limit. The maximum stresses exceeds the yield limit. However, these are concentrated in small, highly localized areas. The highest stresses occur at the screw holes where the two parts are connected and where they are in contact with the brackets that connect them to the frame. These areas therefore need to be finely meshed [28]. The contour plots are not illustrated due to the non-disclosure agreement.

Table 3: Stress, deformation and strain energy obtained from the housing and cover

Component	Stress (MPa)	PEEQ (%)	Strain Energy (N.mm)
Housing	218.3	1.57	14210
Cover	211.2	1.32	9821

When a load is applied and the material deforms, energy is introduced into the structure. The energy introduced into the material due to the loading is referred to as the strain energy. Strain energy density gives an overview to which elements/areas of the model are contributing more in supporting the load (absorbing). As a result, an idea as to, where to strengthen the part and where to remove the material, is obtained. The strain energy is later used in optimization of housing structures, where the load-paths are identified.

3.2. Vibro-acoustic Template

The modular diagram in Figure 3 aids with the development of a correct model for curtailing the Noise Vibration and Harshness (NVH) below the threshold limits. The dynamic characteristics obtained from the vibration and acoustic module are transferred to the optimization module. Later, these values are analyzed and altered according to the changes made in the design variables.

This section gives an overview of the first step, i.e., building an simulation template, to calculate the vibration and noise. Vibrations originating at the gears and shafts need to be well understood and later contained at the desired boundaries. For this reason, modal analyses for the DU components are developed. This aids in evaluating the vibration behavior of the individual structures. The drivetrain consists of four main components: gears, motor, shafts and housings. At the development stage, essential parameters for vehicle vibration such as mass, inertia, stiffness and damping for these structures are evaluated.

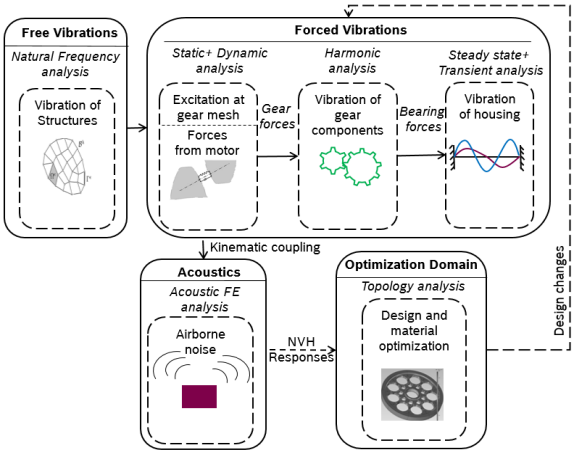


Figure 3: Flow of the FE model for dynamic simulations

Damping is the diminution, restriction or prevention of the vibrations. To achieve this, damping elements are introduced to the systems; the addition of these make the power unit more efficient and provide a better ride comfort [23]. When a damped structure, which is initially at rest, is excited with a harmonic load, it has a transient response at first which then fades out quickly. Furthermore, the structure attains a steady state which is characterized by a harmonic response with the same frequency as that of the applied load. The analysis is performed across a frequency sweep by applying the loading at a series of different frequencies and logging its response. This procedure provides solutions to the linear equations of motion when the loading is harmonic. As part of this section, several submodules of the vibro-acoustic template, such as excitations and acoustic influences, are discussed further.

3.2.1. Excitations from the Electric Motor

Electromagnetic (EM) phenomena in drive units can generate undesirable effects such as vibrations. There are, already, a few simulation methods that have been developed for understanding this effect. For instance, paper [29] from Li describes the electromagnetic simulations developed to study the characteristics of flux and EM force under different conditions using FEM, thereby it is utilized to analyze the vibration behavior of the structures. Similarly, paper [30] published by Dupont, focuses on excitations due to electromagnetic phenomena using an electromagnetic FE solver. This excitation is then projected onto the structure mesh of the stator in order to calculate the dynamic response. However, they neglected the effects of excitation due to the transmission errors. Nevertheless, through this work, the radial, tangential and axial forces arising at the stator due to the EM forces are coupled with the mechanical excitations for analyzing the vibration characteristics of the drivetrain.

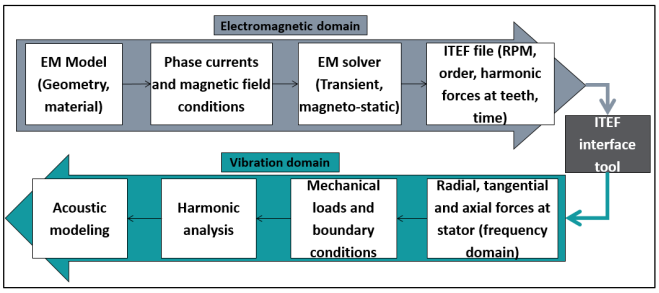


Figure 4: Flow for simulating vibration characteristics using EM forces

Figure 4 describes the flowchart for the calculation of the vibration behavior for the drivetrain based on the EM characteristics. The EM model is prepared using motor geometries and appropriate definitions, such as the phase current. Apart from the calculation of flux and field lines, the simulation also generates an ITEF (Interface for Transfer of Excitation Forces) file. This file consists of information, including RPM, orders and forces at each individual tooth, across discrete time points. The ITEF tool acts as an interface to transfer the forces onto the stator surfaces with radial, tangential and axial components along the frequency spectrum of interest. Later, these forces are used for simulating the harmonics of the drivetrain structures.

3.2.2. Excitations from the Bearings

For further analysis, the gear model with bearings and shafts is simulated. Here, the axial and radial forces at the bearings and other critical areas are calculated. The reaction forces in the simulation are logged in order to retrieve the complete profile of bearing forces. Figure 5 describes the profile of axial forces at the crankshaft after converting into the frequency domain (as an example). The other axes also have analogous profiles with their respective orders and amplitudes. These are not described in detail as the idea was to introduce the user to force amplitude vs. order graphs that are later used in the harmonic excitation step. Information about the simulation model encompassing axial forces and housing structures are explained in the papers [1, 31]. The scale on the y-axis is normalized for non-disclosure purposes.

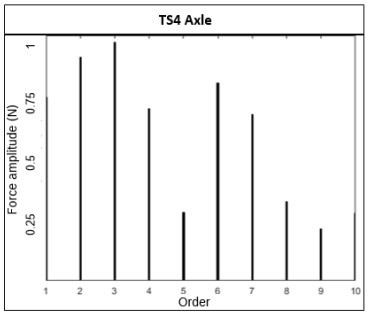


Figure 5: Forces at the TS4 axle represented in the frequency domain

3.2.3. Acoustic Mesh

The surface velocities of the housing surfaces create pressure fluctuations of the fluid around them [32]. These disturbances result in noise around the drivetrain. The acoustic pressures and intensities across different distances are estimated with the help of an acoustic FE mesh.

Acoustic elements model the propagation of the acoustic waves and are active only in dynamic analysis procedures. Figure 6 represents the schematic view of the acoustic mesh developed around the drivetrain structure. The terms n_s and n_a indicate the normal vectors of the structural and acoustic mesh, respectively. Furthermore, information corresponding to the coupling between these domains is explained by Sandberg in book [33]. The length of the acoustic elements should be able to capture the maximum frequency of interest. In (10), c, n and h correspond to the speed of sound, number of elements required to capture the wave (generally 5-8) and wavelength, respectively [31].

$$freq_{max} = \frac{c}{n \cdot h} \tag{10}$$

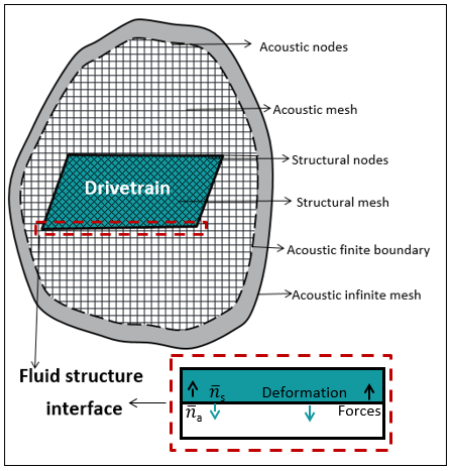


Figure 6: Structural, acoustic finite and infinite element model [20]

Exterior surfaces of the structure are tied to the inner surface of the finite acoustic mesh, hence, velocities at the structural surfaces are transferred to the acoustic nodes. Furthermore, the acoustic velocities of nodes can be altered by introducing impedances between the surfaces. Later, the acoustic finite mesh is attached to the acoustic infinite mesh for estimating the pressures at farther distances. The loss effects within the acoustic medium are modeled via volumetric drag and material impedance parameters.

Table 4: Step information for harmonic simulation

Frequency range	300 – 10000 Hz
Applied load	Axial forces (section Excitations from the bearings) and motor forces (section Excitations from the Electric Motor)
Damping factor	0.03 - 0.24 (frequency dependent) (exact values not disclosed)
Step type	Steady state dynamics - modal

3.2.4. Vibration Results

This section provides a brief overview of the acceleration magnitudes for a drive unit simulation. Table 4 indicates the step information such as the frequency range, applied load and damping constants used for calculation of harmonic responses at the housing structure.

Figure 7 describes the surface acceleration recorded at the housing structure element that has the maximum displacement for constant damping factor across the spectrum. The x-axis and y-axis represent the frequency in Hertz and amplitude in Decibels, respectively. Note that these values are scaled appropriately for non-disclosure reasons. The graph shows the influence of the damping ratio on the acceleration magnitude. For lesser damping factors, the peaks tend to be very sharp and for higher damping factors, the amplitudes are reduced and smoothed out. Therefore, the models with damping structures along the transfer path have the potential to create an impedance jump across the intersections, thereby reducing the vibrations. The horizontal line indicate the threshold limit defined using the specification data. The specification sheet including the vibration, sound limits, weight restrictions etc. are obtained from the requirements engineering team, who study the present market situation and analyze the various factors associated with customer satisfaction.

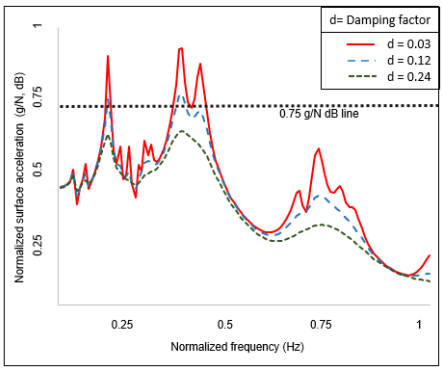


Figure 7: Acceleration simulated at a critical node with varying damping factors

Further descriptions of the results are illustrated in section 6, after describing the test setups and connected software.

4. Topology Optimization

Topology optimization starts with an initial model and determines an optimum design by modifying the properties of the material in selected areas, effectively removing elements from the analysis. The numerical methods developed to achieve structural optimization can be classified into two categories: the deterministic method and the probabilistic method. The former uses mathematical programming and a special gradient-based optimizer to perform the optimization. The latter uses heuristic algorithms which are nature-inspired and are developed based on the successful evolutionary behavior of natural systems [34, 35].

For this paper, mostly mathematical programming along with an optimal criteria approach are used in the optimization tasks. The mathematical programming performs iterative changes of the initial design for improving the objective function and fulfilling constraints in each optimization iteration, while the optimal criteria uses mathematical formulations of conditions that characterize the

optimum. The design variables are redesigned so that they fulfill the optimal criteria. The main advantages of this method are that it has faster convergence and the speed of this convergence is independent of the number of design variables. Figure 8 shows the flow of data within this optimization module for reducing the stress at critical housing areas.

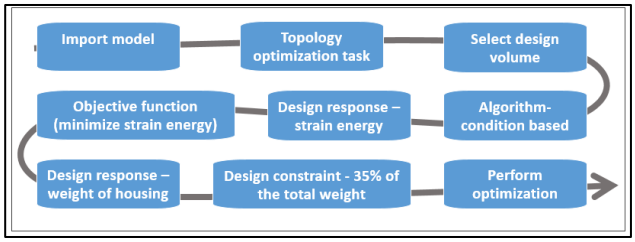


Figure 8: Optimization flow for minimizing strain energy

Table 5 describes the variables defined in the Tosca parameter file, which are used in order to optimize the drive unit for the strain energy. A design constraint with 48 % and 27 % of total weight for the housing and cover, respectively, is imposed. The area near the motor, bracket interface, other critical boundaries, etc., must not be changed. Therefore, the model is also given the frozen area for both structures. Minimization of the strain energy of both parts is considered as the objective function of the optimization task. The convergence of results is obtained at optimization iteration number 32. Later, using the Tosca GUI's Smooth module, visualization of the smoothed, optimized part and the downside of it, i.e., the removed material, is determined.

Table 5: Topology optimization definition for minimizing strain energy

Optimization task	Topology
Objective function	Minimize strain energy / Minimize compliance
Design constraint	Reduction in weight of housing and cover
Strategy	Topology sensitivity
Iteration stop	50

Figure 9 describes the reduction in strain energy through the optimization simulation. The smoothed results from the Tosca are revised in the CAD software and finally new housing and cover parts with 55 % and 31 % of the original parts are obtained.

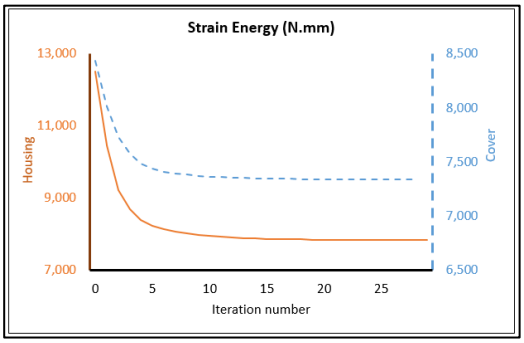


Figure 9: Strain energy vs. optimization iteration

Once the optimization task has been accomplished, new components with a mixture of the two materials can be defined. It is understood that the existing structures, i.e., the resulting elements of the optimization, have the maximum stress and strain energy, therefore, these elements were retained during the optimization. These structures are referred to as skeletons in the further sections of the work. This mechanical optimization shows that it is possible to simplify the geometry of the housing and cover whilst maintaining the stiffness of the structure. The metal skeleton can then be provided with a plastic insert in order to close and seal the drive unit.

In the next step, removed elements (negative volume) are replaced with different plastic materials to obtain the weight reduction in comparison with the original parts and, at the same time, curtail the stress limits below the thresholds.

Table 6: Maximum stress, drivetrain weight and percentage mass in comparison with original drive unit

Material	Max stress (MPa)	Drivetrain weight (gm)	Optimized drivetrain mass in %
Mg Skeleton	1530	243	47
EP43	380.6	425	82
Akulon	581.2	425	82
Delrin	361.9	467	90
Polyfort	553.4	396	76

The results in Table 6 show that the addition of plastic significantly reduces the stresses in the parts compared to the pure metal skeleton. Only a few areas of the housing have stresses that exceed the yield limit. In addition, these areas are thin and very localized. The geometry of these zones must be revised in CAD software in order to bring their maximum stress below the fracture limit.

Among the various plastics used, Delrin offers the lowest maximum stress at 361.9 MPa, while the addition of Polyfort enables the best weight savings. In fact, the magnesium skeleton supplemented with Polyfort accounts for 76 % of the mass of the original DU. EP43 is a compromise between these two cases and represents the best optimization solution. Indeed, when added to the magnesium skeleton, this allows a maximum tension of 380.6 MPa to be obtained, whilst the end part (skeleton and plastic) comprises 82 % of the original drivetrain mass. This weight saving of almost 20 % represents a potential improvement in the drivetrain. The geometry of this new structure must be revised for it to be constructed. In addition, thermal aspects must also be considered. Indeed, the structure must allow the thermal dissipation of the heat generated by the engine and transmission during the operation of the DU. The connection between metal and plastic must adapt to the different thermal expansions of the materials to ensure a perfect seal during operation.

5. Dynamic Optimization

Developing a sensible combination of objective functions and constraints is one of the most challenging steps in applying topology optimization. For the design of dynamic systems, structural vibration control is the most essential consideration. The objective of this type of optimization is to achieve the optimal material layout for the load-bearing components.

Figure 10 describes the flow of the optimization module to optimize the structural dynamics of the housing and cover, thereby improving the drivetrain. The two basic parts (cover and housing) are covered with an envelope volume (design volume). The internal forces, the electromagnetic forces of the motor and the transmission from section 3.2 are applied. The displacements of frame interface points are measured, which enables reaction forces to be calculated. The aim of the dynamic optimization is then to minimize the amplitude of the reaction forces in order to reduce the noise generated by the drive unit during its operation. For this purpose, the optimization software removes material from the design volume to form the optimal ribs. The complete design volume (from both housing and cover) weighed 1.25 kg at the beginning of the simulation. The restriction introduced via the volume constraint corresponds to a final weight corresponding to 30 % of the initial weight. This optimization is carried out in 42 iterations over 38 hours to give a final weight of 312 gm (i.e., approximately 26 % of its starting weight).

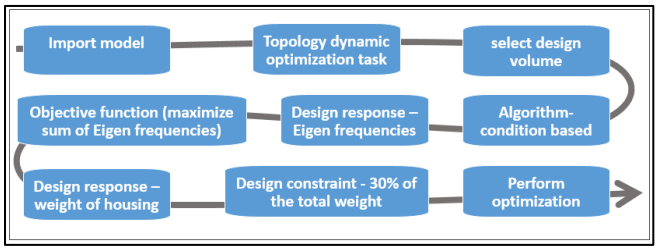


Figure 10: Optimization flow for linear dynamics

Table 7 illustrates the variables defined in the Tosca parameter file that tries to optimize the drive unit for the dynamic characteristics.

Table 7: Topology optimization definition for reducing the noise level

Optimization task	Topology
Objective function	Minimize the vibration and sound levels
Design constraint	Reduction in weight and stress of housing and cover
Strategy	Topology sensitivity
Iteration stop	80

The Frequency Response Function (FRF) is a frequency-based function used to identify the resonant frequencies, damping and mode shapes of the structure. This expresses the frequency domain relationship between an input (force) and output (acceleration) of a linear, time-invariant system. Figure 11 describes a FRF graph for the mean acceleration measured at the critical interface points. This optimization shows a decrease in the maximum acceleration by 4.2 dB. This reduction is greater at higher frequencies and reaches 12 dB compared to the initial drive unit. Overall, the acceleration of the optimized model lies well within the thresholds (limits shown in Figure 11).

Similarly, Figure 12 represents the octave spectrum simulated at 1000 mm distance from the main housing structure. This figure compares the sound pressure levels of the optimized and original drivetrain structures. Optimization shows a decrease in the sound pressure level by 6.2 dB at the frequency spectrum band with the maximum amplitude. This difference is greater at higher frequencies and reaches 13.4 dB compared to the original drivetrain. Overall, the acoustic values of the optimized model lie well within the thresholds.

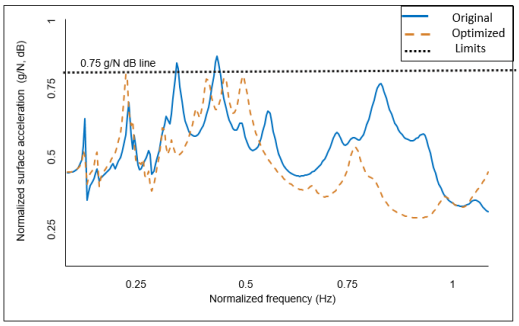


Figure 11: Vibration plot of optimized vs. original drivetrain structure

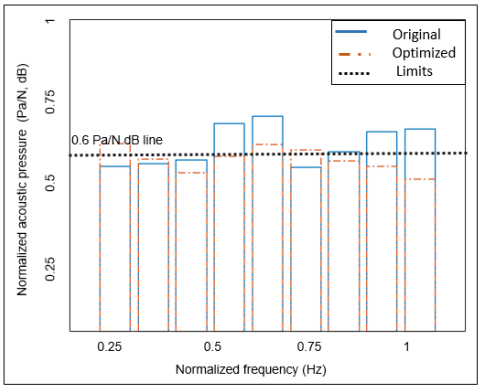


Figure 12: Octave (acoustic) plot of optimized vs. original drivetrain structure

6. Method Verification

This part of the paper describes the validation of the methodology generated using the simulation techniques in section 3.2. Initially, the experimental setup developed for a DU variant is illustrated. Later, the FE model verification is shown using the values of surface acceleration and acoustic pressure.

6.1. Experimental Setup

Experiments are performed to verify the surrounding pressure and surface velocities of the current drivetrain variant obtained from the numerical calculations. Figure 13 illustrates the schematic experimental setup of the drivetrain and corresponding measuring devices. The absorber box is used to isolate the powertrain vibrations from the surroundings. The different forms of the dashed lines indicate the surface velocities and acoustic fluctuations respectively.

RPM and motor torque profiles are considered as inputs for simulations and experimental runs. The Polytec laser vibrometer is used for the measurement of exterior vibrations of the housing and cover surfaces [36]. Acceleration sensors are mounted to obtain accelerations at precise locations. The sound pressure level is measured using the acoustic camera (from Head Acoustics) and microphones installed at a distance of 1000 mm [37]. LMS software serves as a central node to send the force reference signals and also to record the vibrations and noise [38]. Finally, after the analysis of multiple DU variants for various torque and RPM profiles, the Campbell and spectral plots are generated.

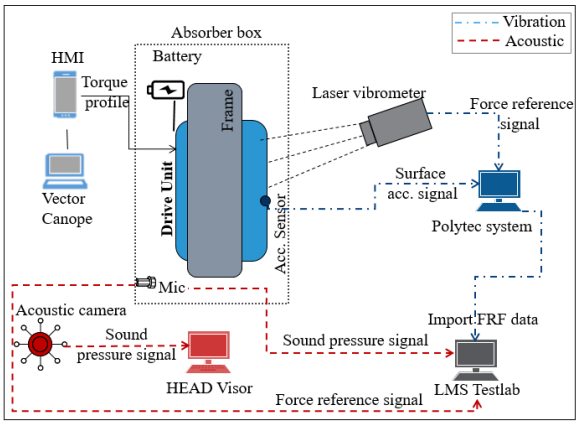


Figure 13: Experimental model for measuring noise and vibration [1]

6.2. Numerical Validation

This section attempts to establish the relationship between the experimental results and the FE values. As a part of it, the first step is to validate the model components for their natural frequencies. It aids in verifying properties like damping factors, Eigen forms and frequencies of the components (not shown in this paper). The comparison of Eigen forms between the Experiment Modal Analysis (EMA) and FE frequency analysis helped to determine the variables like loss coefficients for the spectrum of interest. This data is further used for simulating the harmonic step using the steady state dynamic solution technique. Validating the complete drivetrain at the first go can be extremely hard due to complexities associated in the system. At the start, individual structures are verified and later for the combined structures. Validating various combination of structures enabled for studying the properties of damping associated with contact.

The presence of permanent deformations can alter the dynamic characteristics. Hence, before performing the harmonic analysis of complete drivetrain, the structures are checked for plastic deformations for the static load case. In the end, the damping and material values obtained from the combined structures are used for the simulation and validation of the forced excitation (Figure 14) of the complete drivetrain. Here, an excitation with a torque (2.5 N.m), RPM (0-2500) and appropriate boundaries are used.

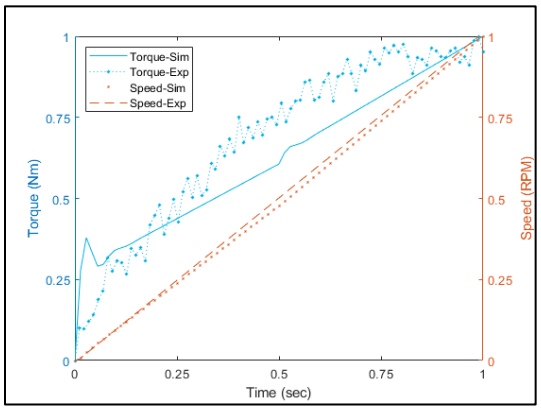


Figure 14: Torque and RPM profiles vs. time

Figure 15 shows the FRF plot of the simulation and experimental results for the vibrations of a DU variant (using the template discussed in section 3.2). The x-axis corresponds to the frequency spectrum. The y-axis represents the surface velocity of a critical node on the exterior surface of the drivetrain (g/N, dB). The numbers shown on the graph correspond to the resonance frequencies. These peaks indicate the presence of Eigen frequencies for the drivetrain components. Figure 16 describes the octave plot (1/3rd) of the acoustic quantity. The y-axis of Figure 16 represents the acoustic pressures (Pa/N, dB). The simulation and experimental values shown in Figure 15 and Figure 16 have a deviation of around 8.78 % and 4.38 % (on average), respectively.

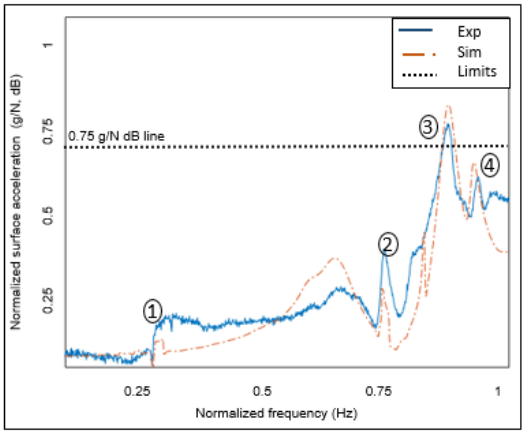


Figure 15: FRF vibration plot - Experimental vs. Simulated

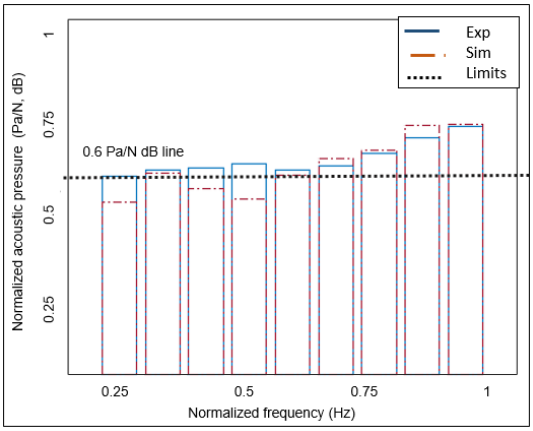


Figure 16: Sound octave spectrum - Experimental vs. Simulated

7. Conclusion

A numerical modeling method for the drivetrain with the aim of understanding the structural properties, such as stress, strain energy, compliance, surface velocities and acoustic pressure is proposed. First, a static model template with its appropriate forces and boundaries is shown. The model is then simulated for the worst load case, i.e., the uphill condition. The vibro-acoustic template introduces the user to topics related to electromagnetic influences, axial forces and structural fluid coupling. Furthermore, the harmonic forces emanating from the bearings and the motor are included into the template. Harmonic analysis is used to calculate the surface acceleration at the housings and the sound pressure in the surroundings. The next section describes two different optimized models with targets of minimizing strain energy and surface velocities. These two tasks cannot be encompassed into one single model as the solver for both the cases is different. The first scenario dealt with a static non-linear model and the second

focused on the structural dynamic characteristics. The validations are illustrated by comparing the frequency curves of simulations and experiments.

The dynamic responses described in this paper are considerably improved in comparison with the multi-body models that are illustrated in the state of the art [1]. Although the results from physical modeling approach [1] are promising, it failed in providing the information like the locations with higher vibration amplitudes in the structures. The FE method helped to realize the critical areas of the structures and to develop high-level models for calculating the acoustic reactions at precise distances. The other prior art approaches such as [6, 8, 9, 17] dealt with a single facet of the problem such as the analysis of the transmission vibrations or just a static optimization or FE modeling etc., for the traditional drive trains. The approach described in this paper aided in building a complete flow analyzing the excitations from several sources on the gears, the motor, the bearings and the housing structures for the electromechanical drives. The method also helped in connecting the static and dynamic domain to build a model with minimized deformations and vibrations simultaneously (Figure 3).

Although the simulation results obtained do not fully complement with the experiments, a substantial 78 % of results are decently fitted. The deviations between results are caused by reasons such as device calibrations, simulation errors and complexities in representing the physical significance of drivetrain parts. Improvements such as better contact, acoustic modeling in simulations and more robust test setups like anechoic chambers can aid in mitigating these differences. At the fluid structure boundary, the simulation setups transfer all the velocities from structural to acoustic domain, while in reality there is energy dissipation due to experimental conditions. Although the losses can be included using impedance boundaries in the FE, a detailed study needs to be performed to define the exact admittance and volumetric drag coefficients. Overall, the system method developed appears to have the potential to predict the NVH characteristics and can be considered for analyzing the optimized drivetrain models.

8. Outlook

Since the FE values obtained for the original drivetrain model are close enough to the experimental results, the next major step would be to build the real parts of the optimized housing and cover. It would be essential to perform validation on the improvised designs obtained from Tosca.

Other aspects such as thermal fluctuations inside drivetrain can also influence the stresses and linear dynamic characteristics of the structures. The developed simulation templates can thus be further integrated into other areas such as the thermal, control systems and the electrical system model, via heterogeneous modeling techniques.

Also, wear among the gears could affect the noise of the drive train. For example, the methods described in paper [39] can be used as a basis for understanding the dependence of the abrasion mechanism on the noise emission of the drivetrain.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Y. Kolluru, R. Doelling, L. Hedrich, "Numerical simulations of vibro acoustic behaviors related to drive train assemblies", in 2019 International Symposium on Systems Engineering (ISSE), 2019. https://doi.org/10.1109/ISSE46696.2019.8984458

[2] T. Kato, R. Mizutani, H. Matsumoto, K. Yamamoto, "Advanced technologies of traction motor for automobile", in 2013 IEEE ECCE Asia Downunder, IEEE, 147–152, 2013. https://doi.org/10.1109/ECCE-Asia.2013.6579088

[3] J. de Santiago, H. Bernho_, B. Ekergard, S. Eriksson, S. Ferhatovic, R. Waters, M. Leijon, "Electrical motor drivelines in commercial all-electric vehicles: A review", IEEE Transactions on Vehicular Technology – IEEE TRANS VEH TECHNOL, 61, 475–484, 2012. https://doi.org/10.1109/TVT.2011.2177873

[4] E. Rask, M. Duoba, H. Lohse-Busch, Recent hybrid electric vehicle trends and technologies, in 2011 IEEE Vehicle Power and Propulsion Conference, IEEE, 1–6, 2011. https://doi.org/ 10.1109/VPPC.2011.6043172

[5] J. M. Rodrıguez, R. Meneses, J. Orus, "Active vibration control for electric vehicle compliant drivetrains, in IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society", IEEE, 2590–2595, 2013. https://doi.org/ 10.1109/IECON.2013.6699539

[6] Y. Ito, S. Tomura, K. Moriya, "Vibration-reducing motor control for hybrid vehicles", RD Review of Toyota CRDL, 37–43, 2005. https://www.tytlabs.com/english/review/rev402epdf/e402_037ito.pdf, last accessed 2020/21/07

[7] K. Jensen, Optimising stress constrained structural optimisation, in 23rd International Meshing Roundtable (IMR), Elsevier, 2014. https://imr.sandia.gov/papers/abstracts/Je751.html, last accessed 2020/21/07

[8] E. Hinton, J. Sienz, fully stressed topological design of structures using an evolutionary procedure, Engineering Computations: Int J for Computer-Aided Engineering 12, 229–244, 1993. https://doi.org/10.1108/02644409510799578

[9] H. Eschenauer, A. Schumacher, T. Vietor, Decision makings for initial designs made of advanced materials, in Topology design of structures, Springer, 469–480, 1993. https://doi.org/10.1007/978-94-011-1804-0_33

[10] Y. M. Xie, G. P. Steven, "A simple evolutionary procedure for structural optimization", Computers & structures 49, 885–896, 1993. https://doi.org/10.1016/0045-7949(93)90035-C

[11] K. Suzuki, N. Kikuchi, Layout optimization using the homogenization method, in Optimization of large structural systems, Springer, 157–175, 1993. https://doi.org/10.1007/978-94-010-9577-8_7

[12] K. Mao, "An approach for powertrain gear transmission error prediction using the non-linear finite element method, Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 220, 1455–1463, 2006. https://doi.org/10.1243/09544070JAUTO251

[13] Y.-C. Chen, C.-B. Tsay, Stress analysis of a helical gear set with localized bearing contact, Finite Elements in Analysis and Design 38, 707–723, 2008. https://doi.org/10.1016/S0168-874X(01)00100-7

[14] D. A. Crolla, D. Cao, The impact of hybrid and electric powertrains on vehicle dynamics, control systems and energy regeneration, Vehicle system dynamics 50, 95–109, 2012. https://doi.org/10.1080/00423114.2012.676651

[15] R. Anderl, Simulations with NX: Kinematics, FEM, CFD, EM and Data Management. with Numerous Examples of NX 9, Hanser eLibrary, Hanser Publications, 2014.

[16] S. Iyengar, R. N. Elements of Mechanical Vibration. India, I.K. International Publishing House Pvt. Limited, 2010.

[17] J. A. Morgan, M. R. Dhulipudi, R. Y. Yakoub, A. D. Lewis, Gear mesh excitation models for assessing gear rattle and gear whine of torque transmission systems with planetary gear sets, SAE Transactions, 1780–1789, 2007. https://doi.org/10.4271/2007-01-2245

[18] H. Allik, Accuracy of infinite elements in structural acoustics modeling, The Journal of the Acoustical Society of America 77, 1985. https://doi.org/10.1121/1.2022426

[19] R.G. Grimes, J. G. Lewis, and H. D. Simon, A Shifted Block Lanczos Algorithm for Solving Sparse Symmetric Generalized Eigenproblems, SIAM Journal on Matrix Analysis and Application, 228–272, 1994. https://doi.org/10.1137/S0895479888151111

[20] R. C. Shieh, Massively parallel structural design using stochastic optimization and mixed neural net/finite element analysis methods, Computing Systems in Engineering 5, 455–467, 1994. https://doi.org/10.1016/0956-0521(94)90026-4

[21] M. P. Bendsoe, N. Kikuchi, Generating optimal topologies in structural design using a homogenization method, Computer Methods in Applied Mechanics and Engineering, 197-224, 1988. https://doi.org/10.1016/0045-7825(88)90086-2

[22] W. Desmet and D. Vandepitte, Numerical Acoustics: Finite Element Modeling for Acoustics, ISAAC 13- International Seminar on Applied Acoustics, 37-85, 2005.

[23] Y. Kolluru, R. Doelling, L. Hedrich, Multi domain modeling of NVH for electro-mechanical drives - accepted for publication, in 2020 Integrating Seamlessly NVH, 2020. https://saemobilus.sae.org/content/2020-01-1584, last accessed 2020/21/07
[24] M.P. Bendsoe and O. Sigmund, Topology optimization: Theory, methods and applications, Springer-Verlag, Berlin Heidelberg, 2003.
[25] K. Svanberg, The method of moving asymptotes - A new method for structural optimization. International Journal for Numerical Methods in Engineering, 359-373, 1987. https://doi.org/10.1002/nme.1620240207
[26] Simulia, Simulia User Assistance Guide 2018, About Abaqus and Tosca, Dassault Systems Simulia Corp, 2018. http://si0vm1337.de.bosch.com:8081/EstProdDocs2018/English/DSSIMULIA_Established.htm?show=SIMULIA_Established_FrontmatterMap/DSDocAbaqus.htm, last accessed 2020/21/07
[27] I. 4210-8:2014, Cycles - safety requirements for bicycles - part 8: Pedal and drive system test methods (iso 4210- 8:2014), Online Browsing Platform 1, 2014. https://www.iso.org/standard/59915.html, last accessed 2020/21/07
[28] P. Wriggers, Contact mechanics, Prentice Hall, New Jersey, 2007.
[29] R. Li, X. Peng, W. Wei, Electromagnetic vibration simulation of a 250 mw large hydropower generator with rotor eccentricity and rotor deformation, Energies 10, 2017. https://doi.org/10.3390/en10122155
[30] J.-B. Dupont, P. Bouvet, Multiphysics modelling to simulate the noise of an automotive electric motor, in 2012 Integrating Seamlessly NVH, 2012. https://doi.org/10.4271/2012-01-1520
[31] Y. Kolluru, R. Doelling, L. Hedrich, Numerical Analysis of the Influences of Wear on the Vibrations of Power Units - accepted for publication, in 2020 Integrating Seamlessly NVH, 1–12, 2020. https://saemobilus.sae.org/content/2020-01-1506, last accessed 2020/21/07
[32] L. Kinsler, Fundamentals of Acoustics, John Wiley & Sons Australia, Limited, 2000.
[33] G. Sandberg, R. Ohayon, Computational Aspects of Structural Acoustics and Vibration, CISM International Centre for Mechanical Sciences, Springer Vienna, 2009.
[34] R. Picelli, S. Townsend, C. Brampton, J. Norato, H. Kim, Stress based shape and topology optimization with the level set method, Computer Methods in Applied Mechanics and Engineering 329, 1-23, 2018. https://doi.org/10.1016/j.cma.2017.09.001
[35] C. Mitropoulou, Y. Fourkiotis, N. Lagaros, M. Karlaftis, Evolution Strategies Based Metaheuristics in Structural Design Optimization, 79–102, 2013. https://doi.org/B978-0-12-398364-0.00004-8
[36] P. G. Technology, Vibrometer measurements methods, 2019. https://www.polytec.com/eu/vibrometry/technology/, last accessed 2020/21/07
[37] H. acoustics, Acoustic camera / head visor(code /5000_) handbook., 2016. https://www.head-acoustics.com/eng/nvh_head-visor.htm, last accessed 2020/21/07
[38] Siemens, Lms test.lab, “Analysis and structural design manual”, Siemens Industry Software, 2015. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.372.6925&rep=rep1&type=pdf, last accessed 2020/21/07
[39] Y.Kolluru, R.Doelling, L. Hedrich, “Wear estimation using FEM and machine learning techniques”, in 2019 ECOTRIB conference, 176–178, 2019. https://ecotrib2019.oetg.at/program2.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 353-362 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Blueprint Model: An Agile-Oriented Methodology for Tackling Global Software Development Challenges

Andre Figliuolo da Cruz, Cristiano Pereira Godoy, Lanier Menezes dos Santos, Lucas Frota Marinho, Marco Santarelle Jardim, Elisangela Paiva da Silva, Cícero Augusto Pahins, Paulo Fonseca*, Felipe Taliar Giuntini

SOL department, Sidia: R & D Institute, Manaus, 69050-020,Brazil

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 23 October, 2020
Online: 20 November, 2020

Keywords:
Methodology
Agile
Scrum
Kanban
Global Software Development

ABSTRACT

In recent decades, the challenge of managing the development teams spread across different time zones has become increasingly common, raising the importance of the development of Global Software Development (GSD) techniques, in order to tackle its particular problems. This work discuss these issues in the context of Sidia, an R&D institute which implements technological solutions for global companies. The main partner of Sidia is a mobile multinational company located in overseas. The development team must cooperate with the overseas team, even though there are no overlapping working hours between both teams. Besides, the teams have a different set of skills regarding design, quality assurance, and software engineering. In order to address theses problems, we propose the Blueprint model, a Kanban and Scrum based model that supports the development of GSD systems, the allocation of tasks and teams, and the efficient communication. Finally, we discuss the aspects and lessons learned of development of project and deploy of a new model for systems development on a real-word project.

1 Introduction

Many companies adopt software development methodologies with fixed timebox periods for developing incremental deliveries, which is a strategy aligned with agile-based approaches. In these scenarios, Scrum proved to be the *de facto standard* for the most software-oriented industries, showing improved results over *waterfall* methodologies. Despite this, some groups can face difficulties when dealing with the Scrum's ceremonies in the context of Global Software Development (GSD), e.g., estimating effort during the planning sessions or even conducting daily meetings.

In this current work, we extend the definition of the Blueprint model [1], generalizing it to other scenarios and groups by furthering exploring its technical aspects and presenting add-ons to the original proposal. As described in [1] by the authors, Blueprint Model is a lightweight project management that can encourage and facilitate communication between groups in different sites when matched with the suggested group organization. In this particular work, the authors focus on reporting the main challenges when developing a project in collaboration with an external group that does not use agile methodologies, which resulted in a set of barriers to successful integration. To resolve these issues, the Blueprint Model defines its deadlines oriented by the scope of work, which leads to timeboxes of different sizes that consider the effort needed for each activity. Such an approach is different from Scrum, where *sprint* cycles have fixed size throughout the project lifecycle. In this Model, the scope of work, hereafter called *blueprint*, comprises the individual tasks required to reach the target, the necessary information needed to perform these tasks, and what is needed to integrate the work done by every group member. A blueprint defines different group layers, concerning the nature of the work performed (e.g., design group, quality assurance), and assigns the individual tasks to them. Each group layer can be further split into many other groups or even share group members. In this Model, every group must focus on a specific kind of activity since it leads to facilitated project management.

This research was developed on Sidia, an R&D institute located in Brazil, which produces and validates software for an international mobile device manufacturer. The Institute develops innovative software in a kind of areas such as machine learning, software management, and others related to mobile products. Both local and overseas groups deal with the business model integration problems and manage the Sidia's customer's standards for system and management interactions.

In order to meet its customer's demands, Sidia employs Scrum

*Corresponding author: Paulo Fonseca, Sidia: R & D Institute, Amazonas, Email: paulo.fonseca@sidia.com

in the context of a global software environment to deal with several model integration problems and manage Sidia and its customers for management interactions. In this scenario, both groups found it difficult to follow Scrum's activities and bureaucracies. In fact that the standard daily meeting system becomes arduous to follow, because, unlike Scrum recommends, groups are not multi-functional. Each group has its own set of skills, including software engineering, design, and quality assurance. Another prominent issue is to apply Scrum in large scale projects, becoming impractical with excessive management tasks and many activities to manage under a distributed scenario.

The remainder of this paper is structured as follows: Sections 2 and 3 summarizes other similar methodologies, relating them to our proposal. Section 4 shows the Blueprint approach and its implications. Section 5 shows a set of results obtained when employing our proposal on a real-world project. In Section 6 we present the conclusions and avenues for future work.

2 Background

In this section, we discuss the frameworks and methods used as references for the development of the Blueprint model, which is base on agile methodologies concepts and rules like those influenced by the values of the *agile manifest*, namely: (i) individuals and interactions over processes and tools, (ii) working software over comprehensive documentation, (ii) customer collaboration over contract negotiation, and (iv) responding to change over following a plan. As already stated, the Blueprint model shares similarities with both Scrum and Kanban.

Scrum is a structural framework created to manage work on complex products and consists of Scrum groups associated with rules, events, roles, and artifacts. Each component owned with purpose, and integration success depends on managing the relationships and interaction between them. Scrum was the primary reference when designing Blueprint.

A fundamental difference between Scrum and Blueprint is that, while the former has been used to develop software, hardware, networks of interacting function, autonomous vehicles, schools, government, and managing the operation of organizations, the latter is focused on software development. We stretch this difference in Section 4. The essence of Scrum is a small group of people while remaining highly flexible and adaptive. To achieve that, it prescribes four formal events for inspection and adaptation: (i) Sprint Planning, (ii) Daily Scrum, (iii) Sprint Review, and (iv) Sprint Retrospective. Appointed events are used in Scrum to create frequency and to reduce the meetings not defined in the regular workflow. In contrast, Blueprint is a model that reduces or eliminates the regularity of those ceremonies, once it can negatively impact some scenarios of GSD. We explore these scenarios in Section 4.4.

To task management, Blueprint employs Kanban, once it is focused on managing commitment and balancing workflow to achieve greater agility. Kanban is a method for defining, managing, and improving services that deliver knowledge work, such as professional services, creative endeavors, and the design of both physical and software products. While designing Blueprint, we make sure that it shares Kanban's same values, namely: transparency, balance, collaboration, customer focus, flow, leadership, understanding, agreement, and respect.

In summary, Blueprint deal with two typical GSD challenges:

Communication Barriers Both Scrum and Kanban define meeting and ceremonious as part of project communication, but in a GSD context, they can be challenging to coordinate, once group members and stakeholders are in different timezones.

Scope Management Although Scrum is favorable to accepting a change of scope, the typical framework implementation implies that a group is committed to the sprint scope once it begins. This scenario can cause undesirable overhead once it requires both backlog and sprint replanning.

3 Related Work

In the last two decades, agile methods increased their popularity and were established as the main software development methodologies. A large number of efforts [2–5] were focused on studying their application under different circumstances and evaluate their advantages and shortcomings. In [4] the authors conducted a systematic review of works that used Scrum guidelines in a global software development (GSD) context. The authors identified the challenges of using Scrum in such context and which are the current strategies to deal with them. The main difficulties are related to: synchronous communication, collaboration difficulties, communication bandwidth, tool support, group size and office space.

The lack of synchronized hours between the group hampers the ability of holding long meetings, such sprint planning or retrospective sessions, which can last up to four hours or more. Some groups [6, 7] handle these issues using strategies such as defining strict timeboxes for meetings' duration and executing preparation tasks in order to shorten the meeting. This approach may help teams to face GSD challenges while partially following Scrum guidelines, however, this is not possible between locations where the overlapping hours are scarce, for example, Brazil and Japan have a 12-hour offset.

Additionally, in many cases, the quality of communication is also a problem, given that distributed projects are composed by group members with diverse cultural and linguistic background, which may lead to group members not completely voicing their opinions, due to fear of being misinterpreted. A number of works [8–10] tackles these issues by proposing additional on-site meetings in multiple project phases: before starting a project, in order to reach a common understanding about the project; multiple exchange visits of group members, during the project, in order to reduce the cultural distance. These studies show that such practices are valuable to both teams as they shorten the communication gap, an important challenge in GSD. Nevertheless, these approaches have a high financial and time cost. As a way to address these issues, other approaches [11, 12] focus on enhance documentation practices in order to bridge the communication gap.

Besides these socio-cultural differences, other inherent aspects of GSD projects also cause Scrum projects to heavily rely on a diverse set of tools in order to exchange information, share group's

states and enforce Agile guidelines. In current literature, works as [13–15] used a large number of tools to achieve these goals, including enterprise wikis, shared electronic work space, global issue and bug trackers and backlog management tools, these approaches might be used in order to provide information for distributed teams in a centralized manner. However, these tools might add complexity and bureaucracy to the team operation, negatively affecting its performance.

Another common technique used in large distributed groups is to divide the group into into smaller sub-groups [6, 7, 13], according to different strategies, including: different aspects of the development group(*e.g.* design, back-end, front-end); product features and deliverable; independent modules or subsystems of the main project. This division eases group management due to the smaller gap in communication of a smaller and more focused group. Scrum may also be adapted in different ways to support the groups subdivision: a fully integrated Scrum group, where all members must participate in all meetings, with frequent meeting ensure correct and clear communication between distributed sub-groups; or separated Scrum sub-groups based on location running their own iterations while sharing key points between groups. These subdivisions tackle the management and interdependence problems in GSD because usually team members from the same aspect need to exchange information more frequently than team members from different aspects.

The majority of approaches performs only minor changes in the Scrum framework, by specifying constraints or adding features to original guidelines. Other approaches suggest more fundamental changes to Scrum guidelines or even use other Agile methodologies.

In [16], the authors investigated the impact of using a combination of Scrum and Kanban (Scrumban) on global software development and showed that challenges related to communication, cultural and strategic issues are addressed due to characteristics of Scrumban such as iterative and incremental deliveries, regular feedback ceremonies and limiting the amount of work in progress. However, there are still few challenges related to technical and security issues in GSD that are not addressed or alleviated through the use of Scrumban, which require other methodologies and tools to be mitigated.

In [17], the authors evaluated the use of Kanban on the same environment and found that some Kanban elements, such as the Kanban Board, *Inclusion Criteria*, which guarantees that all items in development deliver value to the customer, and *Reverse Items*, which enables development items to return to a previous state, are valuable to face communication challenges in GSD.

This work proposes a model called Blueprint, which is a lightweight project management that was envisioned specifically aiming global software development scenarios. This combines strengths of Scrum and Kanban, as well as the use additional tools in order to facilitate communication between groups in different sites and provide holistic project view that centralizes information for all team members. In this particular work, the authors focus on reporting the main challenges when developing a project in collaboration with an external group that did not use agile methodologies, which resulted in a set of barriers to successful integration.

4 Blueprint

4.1 Goals

The Blueprint model has been designed to deal with geographically distributed groups, handling timezone issues, which may lead to difficulties in setting up meetings, misunderstanding requirements, delay in receiving or reporting issues. Besides the timezone challenge, this model also focuses on increasing time spent on product development, parallel work, and reduce meetings overhead.

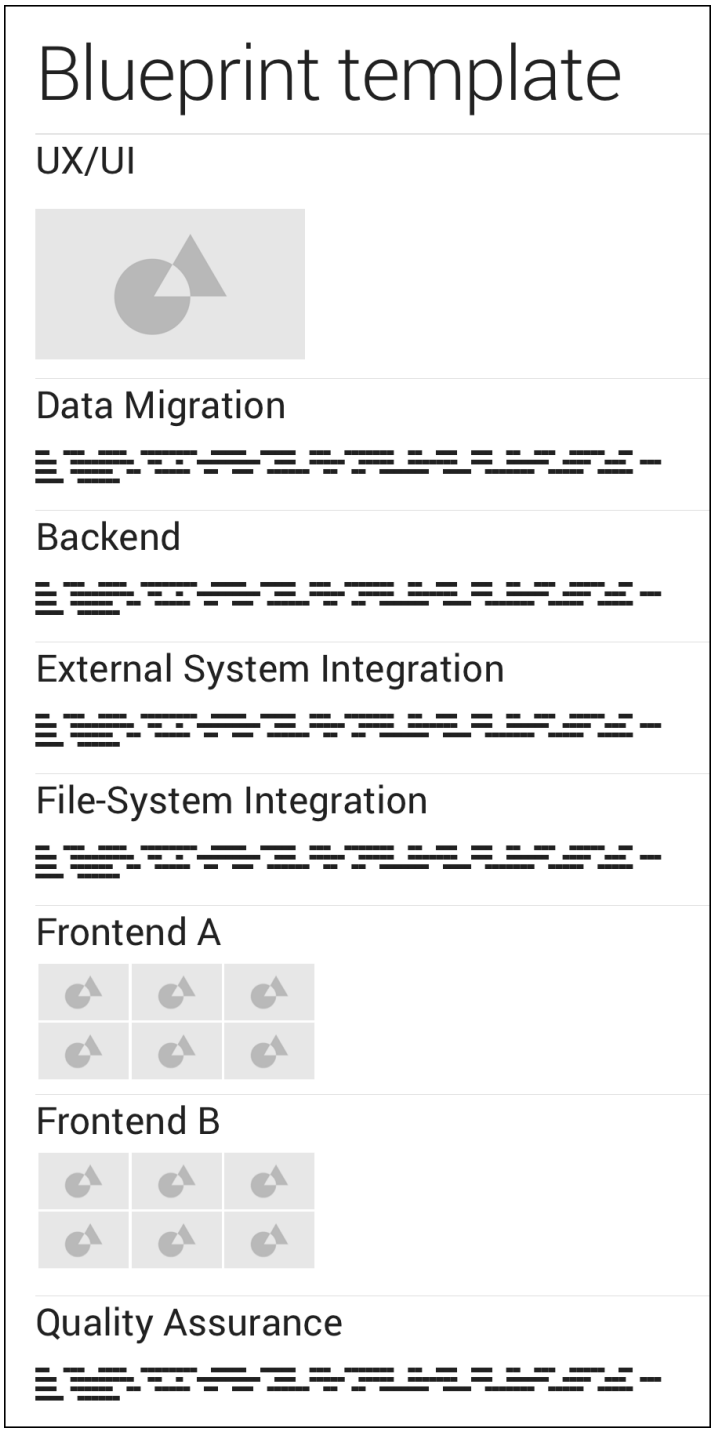


Figure 1: The *Blueprint* Layout, unifying areas *sub-teams* with information on a wiki page.

Therefore, to achieve these objectives, Blueprint seeks to simplify and improve project understanding by using visual-aid tools and helping the group prototype the main project aspects, i.e., UI and technical features. To group communication, these resources are stored in a wiki tool, available to every member so that everyone can be notified of any update.

The Blueprint model employs a hierarchical group organization of sub-groups, once it facilitates the development of unrelated requirements in parallel. Each sub-group is responsible for developing and integrating different aspects of the project into a releasable package. In contrast to other agile methodologies like Scrum, in which groups are considered self-organizing without any prior hierarchical organization [18] other than Testing, Architecture, Operations, and Business Analysis Domains, the Blueprint requires a Leader for each sub-group that is responsible for attending meetings, define requirements, and propose deadlines.

In summary, the primary purpose of Blueprint is to reduce overheads, so the model does not contain predefined which ceremonies must occur neither its regularities, which should be scheduled on demand. Blueprint suggests that each group guides itself by using a set of add-ons, e.g., a wiki tool, improving communication, and information sharing. For the coordinate task, Blueprint suggests Kanban boards, making it easy to share the information across different timezones and provide an accessible project overview.

4.2 Model

In [19], the authors describe an essential aspect of Scrum and Kanban, two popular agile methodologies used in the industry [20], is the constant expansion and improvement of its tools. A Blueprint model is a tool inspired by Scrum and Kanban that manages the project's business entities before it is developed. Blueprint introduces key adaptations to Scrum and Kanban to reduce unwanted bureaucracies and to facilitate global software development. The fundamental concepts of the Blueprint model are:

The Blueprint The concept of a blueprint is very similar to Scrum's story. While both the story and blueprint contain a brief description of the requested features, the latter contains every information needed to design, execute, evaluate, and deliver a set of features, as shown in Figure 3 1. A blueprint should be divided into meaningful sections and contain all related files and discussions to an activity, e.g., the development progress, facilitate the communication between group members and stakeholders, and lower the complexity of both maintenance and monitoring. For efficient scope management, a blueprint can be revised, merged, or divided to reflect the latest requested features and evaluate agreed scope, which should occur at regular intervals. In this document, you can find two very similar terms, the Blueprint Model and the blueprint. The first one, Blueprint Model, is directly related to the model itself. Every time the model is being addressed, it will be called as Blueprint Model, using the first letters for both words in uppercase. The other one is regarding the unity of the work of each blueprint. In this case, the first letter in lowercase refers to every blueprint created for each project, having its characteristics being explained over this document. To implement the blueprint concepts, one could use wiki tools to (i) allow discussions as recorded chats and (ii) keep a history of decisions during development. We further explore the use of wiki tools in Section 4.2.1.

Timebox and Sense of Urgency Timeboxing is a well-known time management technique employed for both project and personal management approaches, once it keeps a constant *workflow* with deliberated deadlines. The use of fixed timeboxes can be challenging, notably in projects with frequent scope changes, since it must be adopted by every group member and stakeholder, imposing difficulties in managing unexpected features requests or modifications. The Blueprint model was designed as a solution for companies and groups struggling to keep with fixed timeboxes, as suggested by agile methodologies like Scrum. In the Blueprint model, every cycle can have an independent deadline determined by the amount of work estimated by the members involved in the development, ranging from days up to weeks. Note that the total time to deliver a set of features is the sum of all activities related to a specific blueprint, similar to a Gantt chart.

Group Organization Contrary to Scrum, in which the group typically uses a flat organization, with members associated with specific roles, the Blueprint model defines a position entitle *group leader*, which is responsible for understanding the solution, define the project schedule, and evaluate deliveries. In the context of a *research and development* project, the Group Leader is also responsible for granting the paper publication. The Blueprint model organizes the group into focused sub-groups, which can be formed by specialists or multidisciplinary members. Sub-groups are advised to define a Sub-Leader to help the group leader in the project management activities, once it eases communication and offers an opportunity for every group member to express its ideas during the discussion of solutions and deadlines. We further explore the Blueprint's group organization in the following sections. Similar to other agile methodologies, the Blueprint model can be adopted by *non-coders*, e.g., testers and designers.

Contrary to the Scrum group division, Blueprint proposes a group hierarchy that can better adapt to software requirements changes during project execution. Blueprint defines its group hierarchy as *sub-groups*. The number of members per *sub-group* is not defined by Blueprint, but as an analogy to Scrum, and as discussed in [19], *sub-groups* should focus on small groups. Also, the number of *sub-groups* may change during the execution of a project.

4.2.1 Add-ons

In the following paragraphs, we present add-ons that can improve Blueprint adoption in different scenarios and groups.

Blueprint Table As explained in Section 4.2, the *blueprint table* is a management tool that should be regularly updated by the group leader to help other members realize the overall status of the project. The *tables* of Blueprint are equivalent to Kanban's boards, but instead of using *status flags* as columns, it uses the concept of *sub-group* ownership. The order of the sub-groups in the table should be defined to reflect the workflow being developed by the group, e.g., design should come first in several common situations, and quality assurance should come last in most of the scenarios.

The list of the condition is used to manage the operation of each stage described in the *blueprint table*, as shown in Table 2.

Contrary to Kanban, in which group members move tasks over the board to indicate the current status of a task, in the blueprint table, the Group Leader is responsible for moving the sub-groups status to consider the *blueprints* in the preferred order of development. The order of development must be arranged between the group, stakeholders, and Group Leader, facilitating project management. The Blueprint model defines the usage of status flags as:

- Every sub-group status should start as PENDING.
- Sub-groups without actions associated with a blueprint, assume the N/A status.
- At the point when the sub-group begins to work in a particular outline, its status should be changed to IN PROGRESS to show that not all conditions and errands are done.
- After a sub-group finishes up all tasks and conditions, the status of the sub-group ought to be set apart as DONE.
- When a sub-group concludes a blueprint, if possible, the sub-group must start working on another blueprint. If not possible, the sub-group must contact the group leader as soon as possible for resource reallocation.
- Status like *SIMILAR* and *BUG FIX* were intended to help general overview.
- SIMILAR ought to be utilized when no effort should be spent by a sub-group due to another outline. At the point when applied accurately, it can diminish modify.

Table 1: *Blueprint* suggestion of content by *subgroups*

Sub-group	Blueprint content
UX/UI Sub-Groups	▪ UX workflow annotation ▪ UI visualizations based on UX definition ▪ Business rules definition ▪ List of assets
Data Migration Sub-group	▪ Database infrastructure servers documentation ▪ Database model related to the Blueprint ▪ Migration architecture documentation ▪ Batch architecture documentation ▪ List of Scripts
Back-end Sub-group	▪ Back-end infrastructure servers documentation ▪ Back-end architecture documentation ▪ Security documentation ▪ Back-end endpoints list
External System Integration Sub-group	▪ External System infrastructure servers documentation ▪ External Integration architecture documentation ▪ External System endpoints list
File-System Integration Sub-group	▪ File-System infrastructure servers documentation ▪ File-System Integration architecture documentation ▪ File-System endpoints list
Front-end A Sub-group	▪ Front-end A infrastructure servers documentation ▪ Front-end A architecture documentation ▪ Security documentation
Front-end B Sub-group	▪ Front-end B infrastructure servers documentation ▪ Front-end B architecture documentation ▪ Security documentation
Quality Assurance Sub-group	▪ Test Cases documentation ▪ Tests Execution report ▪ Know Issue list

- BUG FIX ought to be utilized to show that a blueprint arrived at its last stage and bugs are being fixed.

Every time a new blueprint is created, abandoned or regrouped, it is advised to replan the blueprint table. The group Leader and the group must be aware of the impacts and dependencies of documentation modification during the entire project lifecycle. The same happens when remodeling sub-groups. If any sub-group is created, ceased, or regrouped, the blueprint table should be planned. In summary, a blueprint table is a tool for sub-group management that can be associated with a collection of Kanban boards for handling tasks administration.

Wiki Tools On the Blueprint model, the usage of Wiki tools is essential to organize and aggregate information on different aspects of the project, facilitating the communication between the group and stakeholders. The main points of integration are:

- The web page format makes it easy to maintain and update.
- Every blueprint should be stored as a unique wiki web page.
- The group Leader should organize all pages and ask for the group to keep every part updated.
- Every sub-group must have its own space following the same order on every blueprint page.
- The storage of technical conversations and discussions in a log format contributes to the project's maintenance.
- All related usage of blueprints in the Wiki tools used the company's infrastructure company's infrastructure and its internal Information Security policy to manage it. This means every group member was under the company's management overview and user access control rules.

Table 2: Status to control activities in the *blueprint table*.

Status	Definition
PENDING	Initial status or *sub-group* activities are on hold. It may have several causes. Descriptions should be at the Blueprint related page.
IN PROGRESS	The activities of this *Blueprint* are on development. It may contain dependencies or impediments but still in progress.
DONE	All activities are finished. Have no dependencies.
SIMILAR	When the activities of this *sub-group* are equals to the activities of the *sub-group* above. This eliminates the effort of the related activities.
BUG FIX	Fixing bugs of the *Blueprint*. The *Blueprint* is in its final status.
N/A	Not applied. It occurs when the *sub-group* have no activities for that Blueprint.

- Replanning should be easy to maintain by employing the wiki's versioning system.
- By leveraging wiki tools, groups that require necessary social distancing, home office, or are facing other adverse situations should communicate effectively by (i) integrating a Kanban board with each blueprint and (ii) using the *export* feature to share relevant information with stakeholders.

4.3 Implementing Blueprint on a Real-World Project

To show Blueprint in detail, we will use a real-world project called *The Project*, which characterized a list of *sub-groups* that can be equivalent to many webs, mobile, or embedded software projects. We first planned *The Project* with a number of *sub-groups*, which was thought to be sufficient, but at the end, the list changed to optimize group allocation, as shown in Figure 2. As showed on the Blueprint model, sub-groups can can aggregate multidisciplinary individuals or authorities with a similar profile, for example analyzers situated in Quality Assurance Sub-gathering, architects in Design Sub-gathering, engineers taking a shot at both or solely on Front end Sub-gathering and additionally Back end Sub-gathering. Note that the sub-bunch rundown can be changed by the necessities of the task to streamline bunch portion or build up another prerequisite. In *The Project*, the accompanying circumstances lead to sub-bunch rearrangement:

- To deliver the back-end features on the schedule, the sub-group was divided to help some members to work on two sets of requests.
- The client requested an alternative and entirely different UI solution from Design Sub-group that was approved to be developed, so it was decided to split the Front End Sub-group into groups *A* and *B* to allow each work on a different release.
- A new mobile solution was requested, but there is no developer with such skills in the group. Therefore, it was decided to hire new members and they will work in the entirely new Mobile Sub-group.
- After a schedule modification, the client suggested for developers to work in a new scope of migration data, and due to this, a new Migration Sub-group is created. The group Leader decided that some back-end developers and testers will work on this group part-time.

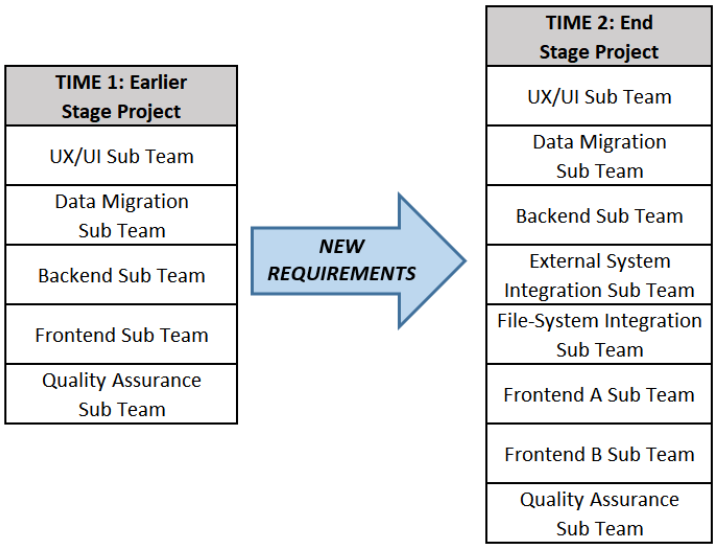


Figure 2: *Sub-groups* preparations during beginning (*time 1*) and ending (*time 2*) of *The Project* execution.

Blueprint consents to the thought that the quantity of individuals doesn't speak to quality or profitability. In the Blueprint model, each *sub-group* has a pioneer characterized as a part of mindfully, and not really work execution. Each *sub-bunch leader* is liable for the exercises and consequences of its gathering. He is additionally answerable for incorporating the errands done by the gathering into the undertaking, planning this cycle with different pioneers. All *sub-bunch leaders* react to the *group leader*, as delineated in Figure 3. This association follows the meaning of how improvement ought to be executed by the Blueprint model to permit and encourage *sub-group* communication.

After the meaning of *sub-bunch leaders*, the following stage is to execute Blueprint select and the board. This stage is a variation and conglomeration of two Scrum the board undertakings and ceremonies, in particular, *Product Backlog* and *User Stories*. All individuals are locked in to effectively take an interest in this stage during venture improvement, as it is the center of the model. As depicted in [12], the duty to characterize and keep up the Scrum's *Sprint Backlog* is in the possession of the job *Product Owner*. It is normal for the *Product Backlog* to require consistent correction and reworking. These assignments may contain business rules in the *User Stories'* depiction. A few organizations have utilized extra documentation to keep engineering definition, model plan, and framework examination refreshed by the *Product Owner*. Uniquely in contrast to Scrum, *blueprints* are business substances that are relied upon to be persistently changed, kept up, and assessed by all individuals from the task. Pioneers must have the duty to change the *blueprints*, however this assignment isn't limited to them.

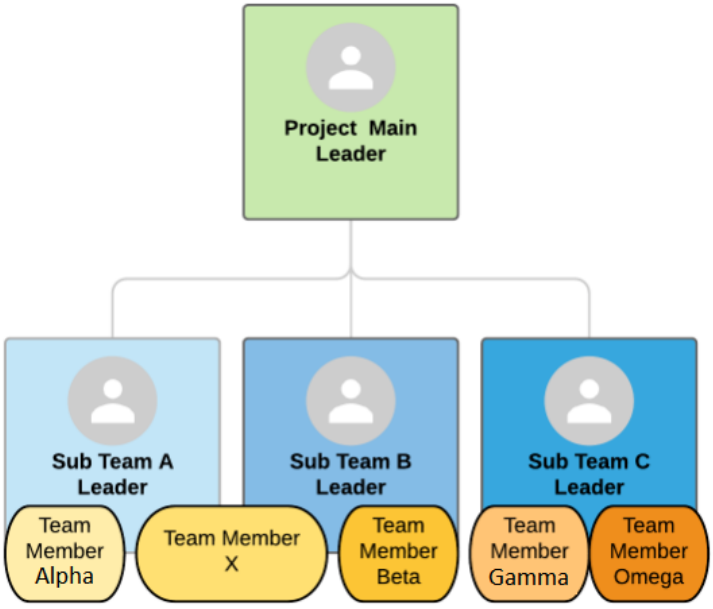


Figure 3: *Sub-group Leaders* and their relationship with the *main group Leader*.

As presented in Section 4.1, the Blueprint model was intended to help bunches in various destinations and to function admirably with distant *sub-groups*. It energizes correspondence among designers and customers by incorporating the conversation in a wiki-based website page. Each *blueprint* ought to be enlisted utilizing a wiki instrument, and its substance development relies upon each *sub-group's* work. At *The Project*, *blueprints* are wiki pages containing

all substance vital for the advancement of a module or an aspect of the undertaking. Table 2 sums up the substance of each *sub-group* obligations that was characterized in *The Project*.

Blueprints must contain all *sub-groups* related materials, as tables, diagrams or file that aggregates information to development. *Blueprints* centralizes information and communication between *sub-groups* as well as with the *group leader* and the client. Wiki tools typically allow for managing discussions between users as a recorded chat. Blueprint takes advantage of this feature to keep the history of decisions made by the group during project execution. Figure 1 illustrates a *blueprint* template that contains fields related to each *sub-group* define in *The Project*.

Blueprint maintenance is done by keeping each blueprint activity from all sub-groups up to date, which helps reflect on the project status. The usage of blueprints encourages deadlines with a focus on the agreement between sub-groups, group leaders, and stakeholders. Each sub-group deadline is a composition of all tasks planned at the blueprint, which can be replanned without impacting other blueprints. Therefore, the Blueprint model enables groups and stakeholders to control and refactor activities by proposing specific deadlines for each sub-group without re-planning activities defined.

The understanding about the hour of every movement is legitimately identified with the development of the gathering and its self-administration. After the meaning of each *blueprint* content and the principle cutoff time of *The Project*, some gathering individuals recognized the requirement for micromanagement exercises. To oblige that request, the underlying adaptation of the *blueprint table* was moved up to characterize reasonable cutoff times for all *sub-groups*. Figure 5 embodies the second form of the *The Project*'s *blueprint table*.

Blueprint encourages deadlines with a focus on *sub-group* assignments that are characterized in in agreement with the *group leader* and the client. Each *sub-group* deadline is a piece of all assignments arranged at the *blueprint*. During the development of *The Project*, bunch individuals started to demand rethinking of certain activities, and it started to impact on previously defined *blueprints*. The Blueprint empowers to gatherings and customers to control and refactor exercises by proposing specific deadlines for each *sub-group* without previously characterized.

The Blueprint model characterizes a*blueprint table* as a administration tool hat ought to be utilized to continually show the general status of an undertaking to all gathering individuals and customers. Figures 4, 5 and 6 show the *blueprint tables* during execution of *The Project*. Toward the start of *The Project*, an underlying form of the *blueprint table* was characterized by the gathering with the assistance of the customer, as observed in Figure 4. This underlying arrangement permitted all partners to concur with the interests of the task and with the substance of each *blueprint*. During the development of *The Project*, members of the *sub-groups* identified that the *SIMILAR* status helped to reduce software rework.

After different steady modifications, *The Project* characterized its last *blueprint table* as appeared in Figure 6. Note that, in examination with Figure 5, new *sub-groups* were made to oblige already impromptu prerequisites of the customer. The creation or evacuation of *sub-groups* can influence different exercises and should be all around oversaw and talked about between *sub-bunch leaders*, however never debilitated. For example, at the underlying rendition of the *blueprint table*, just one front-end gathering, called *Front-end Sub-group* was liable for all front-end exercises. At that point, when an additional necessity rose up out of the customer, another front-end *sub-group* called *Front-end B Sub-group* should have been made to build up an alternate arrangement of arrangements. The underlying front-end *sub-group* was needed to survey its exercises to help the redistribution of errands of the reestablished *blueprint table*. A portion of the gathering individuals need to chip away at bunches *A* and *B* to help extra exercises. *Front-end bunch B Sub-group* was situated at the abroad site and worked distantly with the gathering situated at Sidia. The last rundown of *blueprints* are represented in Table 3.

Table 3: List of Blueprints used in *The Project*

Blueprint Name	Example of work by blueprint
Authentication	Login page. Back-end authentication service. Front-end authentication error page.
CRUD A, B, C and D	Create, Read, Update and Delete business entity A, B, C and D. Back-end service for each. Front-end pages for each.
Extended Specification A and B	New requirement with several business workflow. Requested during the execution of the project. Replan needed.
Report A, B and C	Report generator page. Back-end service for report business model.
External Requests A and B	Unintended external integration. Requested during the execution of the project. Replan needed.
Mobile Integration and Mobile Interface	Unintended mobile version. Requested during the execution of the project. Replan needed.
Customer Service, User Overview and About	Low priority administrator pages. Gained priority due to requested changes.

The Blueprint Model as some other philosophy has its reception challenges, for instance:

- Communication hazards The Blueprint model has a great dependency on documentation process. As proposed, it is very dependent of Wiki documents in order to keep good communication between sites. This dependency combined with possible many changes requests, generates a constant necessity of Wiki documents updates. It can turn out very difficult for the group to keep the good communication under many changes requests.

- Predictability hazards If not applied as its standards by the entire group, the Blueprint model can show up as a harmful tool when it comes to deliveries and projects goals. The presented model requires management skills from each area of the group. Once the required skills are not present in the group members, the entire process can be compromised, which would impact on the completion and good quality of deliveries.

- Performance hazards Concerning software development, all sub-groups that work together have dependency of information from one another. On Blueprint model, as each sub-group is required to maintain its own Wiki page updated, the lack of update of any sub-group can generate delays, compromise

group work and even entirely block group members of working.

4.4 Use Cases

In the following paragraphs, we present some use cases in that the Blueprint model proves successful.

Global Software Development The Blueprint model was designed to support remote groups and sub-groups in different locations, since its hierarchical group organization of a Project Leader and its

Blueprint	Deadline	UI/UX	Backend	Frontend	QA
Autentication	20XX/01/10	DONE	DONE	DONE	IN PROGRESS
CRUD A	20XX/01/10	DONE	DONE	IN PROGRESS	PENDING
Report A	20XX/02/23	DONE	IN PROGRESS	PENDING	PENDING
Report B	20XX/01/16	IN PROGRESS	PENDING	PENDING	PENDING
(others)	20XX/03/15	PENDING	PENDING	PENDING	PENDING

Figure 4: Starting version of the *blueprints table*. Deadlines are characterized for *blueprints* in *The Project*.

Blueprint	UI/UX	Backend	Frontend	QA	Global Status
Autentication	20XX/01/10	20XX/01/18	20XX/01/22	20XX/01/24	20XX/01/26
	DONE	DONE	DONE	IN PROGRESS	IN PROGRESS
CRUD A	20XX/01/12	20XX/01/20	20XX/01/24		
	DONE	DONE	IN PROGRESS	PENDING	IN PROGRESS
Report A	20XX/01/14	20XX/01/22			
	DONE	IN PROGRESS	PENDING	PENDING	IN PROGRESS
Report B	20XX/01/16				
	IN PROGRESS	PENDING	PENDING	PENDING	IN PROGRESS
(others)					
	PENDING	PENDING	PENDING	PENDING	PENDING

Figure 5: Second version of the *blueprint table*. The deadlines were characterized by *sub-groups*.

ID	Blueprint	UI/UX	Backend				Frontend		QA	Global Status
			Data Migration	Backend	External migration	File integration	Frontend A	Frontend B		
1	Autentication	20XX/01/10	20XX/01/12	20XX/01/14	20XX/01/16	20XX/01/18	20XX/01/20	20XX/01/22	20XX/01/24	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE
2	CRUD A	20XX/01/12	20XX/01/14	20XX/01/16	20XX/01/18	20XX/01/20	20XX/01/22	20XX/01/24	20XX/01/25	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE	BUG FIX
3	CRUD B	20XX/01/14	20XX/01/16	20XX/01/18	20XX/01/20	20XX/01/21	20XX/01/22	20XX/01/23	20XX/01/24	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE	BUG FIX
4	CRUD C	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20	20XX/01/21	20XX/01/22	20XX/01/23	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE	BUG FIX
5	CRUD D	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20	20XX/01/21	20XX/01/22	20XX/01/23	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE	BUG FIX
6	Extended Specification A	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20	20XX/01/21	20XX/01/22	20XX/01/23	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	DONE	DONE	BUG FIX
7	Extended Specification B	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20	20XX/01/21	20XX/01/22	20XX/01/23	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	IN PROGRESS	IN PROGRESS	IN PROGRESS
8	Report A	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20	20XX/01/21	20XX/01/22	20XX/01/23	20XX/01/26
		DONE	DONE	DONE	DONE	DONE	DONE	IN PROGRESS	IN PROGRESS	IN PROGRESS
9	Report B	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20	20XX/01/21			20XX/01/26
		DONE	DONE	DONE	DONE	DONE	IN PROGRESS	PENDING	PENDING	IN PROGRESS
10	Report C	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19	20XX/01/20				20XX/01/26
		DONE	DONE	DONE	DONE	DONE	PENDING	PENDING	PENDING	IN PROGRESS
11	External Requests A	20XX/01/16	20XX/01/17	20XX/01/18	20XX/01/19					20XX/01/26
		DONE	DONE	DONE	IN PROGRESS	PENDING	PENDING	PENDING	PENDING	IN PROGRESS
12	External Requests B	20XX/01/16	20XX/01/17							20XX/01/26
		DONE	IN PROGRESS	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	IN PROGRESS
13	Mobile Integration	20XX/01/16								20XX/01/26
		DONE	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	IN PROGRESS
14	Mobile Interface	20XX/01/16								20XX/01/26
		DONE	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	IN PROGRESS
15	Customer Service	20XX/01/16								20XX/01/26
		DONE	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	IN PROGRESS
16	User Overview	20XX/01/16								20XX/01/26
		IN PROGRESS	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	IN PROGRESS
16	About									
		PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING	PENDING

Figure 6: Last version of the *blueprint table*. New *sub-groups* and *blueprints* are remembered for correlation with past tables.

respective Sub-groups facilitates daily communication, in such a way that this interaction becomes more fluid and objective for adopting a particular scope for each sub-group. In our case, we deal with a customer in a country in which the timezone is at least 13 hours ahead, making real-time communication via call or chat unfeasible. In this scenario, we use the 'wiki' to keep track of what each member of the group is planning for the next few days, what they are doing, and what was done in such a way that everyone has an overview of the project's progress, being able to follow all deadlines and features that have already been developed and those to come.

Remote Work In the face of COVID-19, several development groups and companies in different industries were forced to adopt remote work to continue their work. In this scenario, Blueprint proved to be an adequate strategy to accommodate the remote work regime, regarding the difficulty in communication between groups because from the moment employees are working from home, the practicality and speed of transmitting information are both affected. Consequently, adopting a standard of documenting the information in a centralized form and bringing the division of sub-groups, makes the dissemination of internal information through the wiki page more effective by leveraging a reliable consumption source that is validated by the client.

Scope Change Control The Blueprint also helps to manage frequently *scope changes* since unlike Scrum which suggests another iteration of a set of ceremonies when facing modified requests, such as *Project Vision*, *User Stories*, *Planning*, and others, the Blueprint offers lightweight project management that avoids these ceremonies that can cause rework and further prolong delivery forecasts. To improve on that, Blueprint proposes the idea that both the developer and client will leverage the 'wiki' as a guide for (i) decision making, (ii) discussions about new releases, and (iii) requirements alignment. By documenting information in a centralized repository, the stakeholders are able to easily monitor the requested scope changes, helping to avoid noise or loss of information from typical verbal communication.

5 Implementation Findings

The design of Blueprint was mainly oriented by the GSD challenges faced during the collaboration of remote teams on developing *The Project*, which evolved from 10 members using Scrum to a successful Blueprint case more than 20 members. The increased number of members resulted from the stakeholders' higher engagement and contribution to the process and project. Once the process started running, it allowed the team to accept more work scope, which might be understood as an increase in productivity and led to a natural team increase. Through the advancement of *The Project*, different gradual modifications were made to the Blueprint Model to energize its appropriation. This cycle included both neighborhood and far off groups to facilitate the work process, encourage worldwide programming designing, and improve appropriated correspondence. Following the summarizing of *The Project*, we notice various situations to improve the Blueprint model execution to ensuing tasks.

We found that an essential aspect of a successful Blueprint implementation is to focus on the member's sense of integration and empowerment. Regarding *The Project*, most of the individuals concurred they turned out to be more critical and made more proposals to the undertaking the executives with the appropriation of the Blueprint Model. Every part had the option to share its answers and thoughts to other people, and due to Blueprint central focuses, significant correspondence issues were survived. The primary advantage

of zeroing in on the part's feeling of reconciliation and strengthening is that the group felt inspired for monitoring Blueprint's curios, prompting improved undertaking the executives.

Another significant part of a successful usage was the group's adherence to Blueprint's rehearses. Regularly, less connected with *sub-group leaders* portrayed exercises and progress ineffectively, which can be credited to the difficulties while embracing another administration model. By the by, the individuals saw a connection between (I) the time frame pages changes, (ii) nature of these pages and (iii) commitment of the sub-pioneers. We notice a connection between diminishing the nature of Blueprint pages to diminishing code quality and last on bugs appearing.

The improvement of the communication between team and stakeholders is a crucial objective of the Model, so it is crucial to make more accessible the verification and reporting of information through a *wiki tool*. The successful implementation of the Blueprint Model enables identifying disagreements and conflicts in the Wiki and not necessarily during meetings, which enabled the resolution of these problems in private. Meeting events are not often and gathers a few members, which prevents long discussions of non-related subjects from reducing the execution time and reaching its goal as quickly as possible while the rest of the team keeps working on project tasks. After adopting the BLuePrint Model, the team had performance gains related to the planned x performed activities, related to clear and accurate communication, thus ensuring a clear understanding of the changes. The fact that did not happen using the Scrum model, due to the difficulty of managing constant changes within the sprint and the difficulty of formal meetings due to the time zone. Another improvement that is possible to observe is the quality of deliveries. In the BluePrint model, team members can maintain their specialties without the need to be multidisciplinary, and brings a significant gain in side effects, especially when it came to scope changes. Still linked to the specialty of the team members, it is also possible to see the result in the motivation of the team; it means, to work with the specialty of each one, this brings motivation and engagement in the day-to-day work, and consequently, this is reflected in product quality delivered.

Differently from other agile methodologies like Scrum, the Blueprint Model exempts ceremonies and daily meetings that need the entire team to happen, making the Model more flexible. This characteristic praises the Model because it fits well with the current pandemic scenario where the new common sense of IT company members working remotely and need to keep the flow of work to meet delivery deadlines. Some Blueprint Model positive points:

- No team ceremony: The team adopted the Blueprint Model since the beginning of our development, and we kept it when everyone on the team was doing home office. We managed to keep all deadlines using only the Wiki as a reference for the division of activities.
- Sub team division: This sub-team division was of great value for the atypical home office scenario. The team needed to maintain a very similar communication level as a live-work environment, which made the team feel more comfortable maintaining contact with companions.
- Model suited to a project scope that frequently changes: Scope change is passive to happen very commonly in any software development project. The team validated the model adherence to the common practice since every project was constantly changed. Due to the very centralized information on the Wiki, there were small relevant impacts, all highly negotiable.

6 Conclusion

In this work, we presented the extended definition of the Blueprint model, a management methodology designed to accommodate the best techniques and habits of both Scrum and Kanban to facilitate (i) global software development, (ii) groups and task allocations, and (iii) effective members communication. The straight communication between groups and the high visibility of the overall status of the projects are the main highlights of the methodology, which lead to a better scope change management with low bureaucracy and an accessible understanding of the entire development. To validate the model, we discussed its adoption on a real-world project of a large mobile-related software development company in the context of GSD. The strategic use of Blueprint's add-ons enables us to coordinate, design, develop, and test a complex set of features on a highly unstable scope scenario. Blueprint usage of UI and UX for the documentation of analysis, design, and development phases in conjunction with the lightweight cycle management tool helped to facilitate the project management, once it allows sharing responsibility between the *group Leader*, *Sub-group Leaders*, and other members in documenting and reviewing the progress of the project. More than that, the Blueprint model was improved during the execution of the project through a collaborative effort of groups located in different countries in order to improve its main benefits and to lower any adoption barriers. As future work, we plan to extend the Blueprint's add-ons to integrate other easy-to-use and well-known tools of project and task management, as well as to improve the model's adherence to other types of non-code projects.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We thank MSc. Ilsun Hong, Taehee Kim, Rafael S. Zerbini, Marcos R. G. B. Junior and Alison L. Garantizado for their valuable help during the development of this framework. This work was partially supported by Samsung Eletrônica da Amazônia Ltda., under the auspice of the informatics law no 8.387/91.

References

[1] C. P. Godoy, L. M. Santos, A. F. Cruz, R. S. Zerbini, E. P. Silva, C. A. L. Pahins, "Blueprint Model: A New Approach to Scrum Agile Methodology," in Proceedings of the 14th International Conference on Global Software Engineering, ICGSE '19, 85–89, IEEE Press, 2019, doi:10.1109/ICGSE.2019.00014.

[2] I. Zada, S. Shahzad, "Issues and implications of scrum on global software development," Bahria University Journal of Information & Communication Technologies (BUJICT), **8**(1), 2015.

[3] L. Silva, C. Santana, F. Rocha, M. Paschoalino, G. Falconieri, L. Ribeiro, R. Medeiros, S. Soares, C. Gusmão, "Applying XP to an Agile–Inexperienced Software Development Team," in International Conference on Agile Processes and Extreme Programming in Software Engineering, 114–126, Springer, 2008, doi:10.1007/978-3-540-68255-4_12.

[4] E. Hossain, M. A. Babar, H.-y. Paik, "Using scrum in global software development: a systematic literature review," in 2009 Fourth IEEE International Conference on Global Software Engineering, 175–184, IEEE, 2009, doi:10.1109/ICGSE.2009.25.

[5] J. López-Martínez, R. Juárez-Ramírez, C. Huertas, S. Jiménez, C. Guerra-García, "Problems in the adoption of agile-scrum methodologies: A systematic literature review," in 2016 4th international conference in software engineering research and innovation (conisoft), 141–148, IEEE, 2016, doi: 10.1109/CONISOFT.2016.30.

[6] M. Paasivaara, S. Durasiewicz, C. Lassenius, "Distributed agile development: Using scrum in a large project," in 2008 IEEE International Conference on Global Software Engineering, 87–95, IEEE, 2008, doi:10.1109/ICGSE.2008.38.

[7] J. Sutherland, A. Viktorov, J. Blount, N. Puntikov, "Distributed scrum: Agile project management with outsourced development teams," in 2007 40th annual Hawaii international conference on system sciences (HICSS'07), 274a–274a, IEEE, 2007, doi:10.1109/HICSS.2007.180.

[8] M. Summers, "Insights into an Agile adventure with offshore partners," in Agile 2008 Conference, 333–338, IEEE, 2008, doi:10.1109/Agile.2008.37.

[9] E. Therrien, "Overcoming the challenges of building a distributed agile organization," in Agile 2008 Conference, 368–372, IEEE, 2008, doi:10.1109/Agile.2008.9.

[10] M. Cottmeyer, "The good and bad of Agile offshore development," in Agile 2008 Conference, 362–367, IEEE, 2008, doi:10.1109/Agile.2008.18.

[11] M. Vax, S. Michaud, "Distributed Agile: Growing a practice together," in Agile 2008 Conference, 310–314, IEEE, 2008, doi:10.1109/Agile.2008.44.

[12] M. Hron, N. Obwegeser, "Scrum in practice: an overview of Scrum adaptations," in Proceedings of the 51st Hawaii International Conference on System Sciences, 2018, doi:10.24251/HICSS.2018.679.

[13] W. Williams, M. Stout, "Colossal, scattered, and chaotic (planning with a large distributed team)," in Agile 2008 Conference, 356–361, IEEE, 2008, doi:10.1109/Agile.2008.25.

[14] J. Cho, "Distributed Scrum for large-scale and mission-critical projects," AMCIS 2007 Proceedings, 235, 2007.

[15] J. Sutherland, G. Schoonheim, M. Rijk, "Fully distributed scrum: Replicating local productivity and quality with offshore teams," in 2009 42nd Hawaii International Conference on System Sciences, 1–8, IEEE, 2009, doi: 10.1109/HICSS.2009.225.

[16] A. Banijamali, M. O. Ahmad, J. Similä, M. Oivo, K. Liukkunen, et al., "Empirical investigation of scrumban in global software development," in International Conference on Model-Driven Engineering and Software Development, 229–248, Springer, 2016, doi:10.1007/978-3-319-66302-9_12.

[17] M. Tanner, M. Dauane, "The Use Of Kanban To Alleviate Collaboration And Communication Challenges Of Global Software Development," Issues in Informing Science & Information Technology, **14**, 2017, doi:10.28945/3716.

[18] S. V. Spiegler, C. Heinecke, S. Wagner, "Leadership gap in agile teams: how teams and scrum masters mature," in International Conference on Agile Software Development, 37–52, Springer, Cham, 2019, doi:10.1007/978-3-030-19034-7.

[19] H. Lei, F. Ganjeizadeh, P. K. Jayachandran, P. Ozcan, "A statistical analysis of the effects of Scrum and Kanban on software development projects," Robotics and Computer-Integrated Manufacturing, **43**, 59–67, 2017, doi: 10.1016/j.rcim.2015.12.001.

[20] M. Düchting, D. Zimmermann, K. Nebe, "Incorporating user centered requirement engineering into agile software development," in International Conference on Human-Computer Interaction, 58–67, Springer, 2007, doi: 10.1007/978-3-540-73105-4_7.

ASTES

Advances in Science, Technology and Engineering Systems Vol. 5, No. 4, 363-373 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Bilateral Communication Device for Deaf-Mute and Normal People

Raven Carlos Tabiongan*

College of Engineering, Samar State University, Catbalogan, 6700, Philippines

ARTICLE INFO

Article history:
Received: 07 May, 2020
Accepted: 09 July, 2020
Online: 28 July, 2020

Keywords:
Two-way
Communication device
Deaf-mute
Full duplex
RX-TX

ABSTRACT

Communication is a bilateral process and being understood by the person you are talking to is a must. Without the ability to talk nor hear, a person would endure such handicap. Given that hearing and speech are missing, many have ventured to open new communication methods for them through sign language. This bilateral communication device can be utilized by both non-sign language users and Deaf-mute together in a single system. Shaped as a box (8in x 8in) with two multi-touch capable displays on both ends, the contraption has several microcontrollers and touch boards within. The latter has the technology of twelve interactive capacity touch and proximity electrode pads that react when tapped, producing quick response phrases audible via speaker or earphone. These touch boards are equipped with an MP3 decoder, MIDI synthesizer, 3.5mm audio jack and a 128MB microSD card. The touch screen modules mounted on top of the microcontrollers transfer data to and from each other in real-time via receiver-transmitter (RX-TX) full duplex UART serial communication protocol. The device is lightweight weighing at about 3 lbs. The prototype device was piloted in an academic institution of special education for deaf-mute students. Participants were 75 normal and 75 Deaf-mute people aged between 18 and 30 years. The experimental results show the overall rating of the device is 90.6%. The device is designed to promote the face-to-face socialization aspect of the Deaf-mute users to the normal users and vice versa. Several third-party applications were utilized to validate the accuracy and reliability of the device thru metrics of consistency, timing and delay, data transmission, touch response, and screen refresh rates.

1. Introduction

This paper is an extension of work originally presented in the 2019 8th International Symposium on Next Generation Electronics (ISNE) [1]. Talking and hearing play a very important part of our daily lives. It helps in finishing up tasks, from the simplest to the most complex. It is inherent in humans to rely on other senses when deprived of hearing or speaking. Most people having these disabilities try to learn the art of sign language due to the urge of communicating independently. Unfortunately, being able to acquire such skills has its own sets of fallbacks, such as the inability to converse with the non-sign language, speaking, and hearing-capable people [2]. Communication is a bilateral process and being understood by the person talking to is a must [3]. Without the ability to talk nor hear, a person would be such in a handicap, socially and emotionally. Cases of these handicap conditions are commonly termed as Deaf-mute. In deaf community, the word deaf is spelled in two separate ways. The small "d", deaf represents a person's level of hearing through audiology and not being associated with other members of the deaf community whereas the capital "D", Deaf indicates the culturally Deaf people who use sign language for communication [4]. Research indicated [5] that Deaf people, especially Deaf children, have high rates of behavioral and emotional issues in relation to different methods of communication. Most people with such disabilities become introverts and resist social connectivity and face-to-face socialization. Imagine the depression and discomfort experienced when one is unable to express their thoughts, for sure frustration arises every now and then.

People rely on words and sound from the environment for them to grasp what is happening. Other alternatives to remedy the said language barrier is to acquire a sign language interpreter, but practically speaking it is not the optimal solution. An interpreter is an inconvenience basically due to the 24/7 limitation. Not everyone around can check what they are saying or is intending to

*Corresponding Author: Raven Carlos Tabiongan, Philippines, +639758694325 & raven.tabiongan@ssu.edu.ph

express. Most Deaf-mute just limit themselves to performing simple tasks to avoid irritation to others and to themselves. Persons who are Deaf-mute normally suffer when performing typical day-to-day tasks. Given that hearing and speech are missing, many have tried to open new communication medium for them such as sign language.

The above-mentioned problems can be solved by integrating a bilateral communication device capable of sending and receiving text and audio responses via full duplex serial communication protocol in real-time. Moreover, the device is equipped with an operating system that converts sign language into text, images, and audio for better communication between the Deaf-mute and the normal users [6], [7].

The main purpose of this research paper is to provide a user experience setting wherein the Deaf-mute and normal users can communicate and chat in a close-proximity, face-to-face setting. Another purpose of this system is to provide a simple and cost-effective solution that can be utilized by both Deaf-mute and non-sign language people together simultaneously in a single system.

2. Related Works

The Deaf community is not a monolithic group; it has adversity of groups which are as follows [8, 9]: (1) Hard-of-hearing people: they are neither fully deaf nor fully hearing, also known as culturally marginal people [10]. They can obtain some useful linguistic information from speech; (2) Culturally deaf people: they might belong to deaf families and use sign language as the primary source of communication. Their voice (speech clarity) may be disrupted. (3) Congenital or prelingual deaf people: they are deaf by birth or become deaf before they learn to talk and are not affiliated with Deaf culture. They might or might not use sign language-based communication; (4) Orally educated or post lingual deaf people: they have been deafened in their childhood but developed the speaking skills. (5) Late-deafened adults: they have had the opportunity to adjust their communication techniques as their progressive hearing losses.

2.1 Sensor Module Technology Approach.

Sensors and touch screen technology can be integrated in a system to bridge the communication gap between the Deaf-mute and normal people with or without knowledge in sign-language. Sharma et al. used wearable sensor gloves for detecting the hand gestures of sign language [11].

2.2 Visual Module Technology Approach.

Many vision-based technologies interventions are used to recognize the sign languages of Deaf people. For example, Soltani et al. developed a gesture-based game for Deaf-mutes by using Microsoft Kinect which recognizes the gesture command and converts it into text so that they can enjoy the interactive environment [12].Voice for the mute (VOM) system was developed to take input in the form of fingerspelling and convert into corresponding speech[13].The images of finger spelling signs are retrieved from the camera. After performing noise removal and image processing, the finger spelling signs are matched from the trained dataset. Processed signs are linked to appropriate text and convert this text into required speech. Nagori and Malode [14] proposed the communication platform by extracting images from the video and converting these images into corresponding speech. Sood and Mishra [15] presented the system that takes images of sign language as input and displays speech as output. The features used in vision-based approaches for speech processing are also used in different object recognition-based applications [16]-[22].

2.3 Product Design and Development Approach.

The device targets to improve the current way Deaf-mute communicates with the normal people, and vice versa. Thus, a bilateral communication device with the intent to promote proximity and face-to-face socialization between the parties involved is developed [1].

(1) To design and develop a portable device that enables the Deaf-mute and normal people communicate via multiple modes in a contraption that is designed for proximity and face-to-face communication.
(2) To establish a stable RX/TX full duplex serial UART communication protocol between the microcontrollers and modules within the device.
(3) To construct a sturdy packaging for the device suitable for indoor and outdoor portability and use.
(4) To test the system and device using third-party software. (Terminal Monitor, Balabolka (TTS), Arduino IDE)

a. Accuracy in transmitting data
b. Consistency in transmitting data
c. Speed (Timing and Delay)
d. Screen Refresh Rates
e. Touch Responses

(5) To test the system and device using a Five-Point Scale questionnaire given to 150 respondents:
(75 Deaf-mute and 75 normal people).

a. Value-Added
b. Aesthetic Design
c. Hardware Components
d. Built-in Features
e. User-Friendliness
f. Reliability
g. Accuracy

2.4 Mobile Application Technology Approach.

Many of the new smartphones are furnished with advanced sensors high processors, and high-resolution cameras [23]. A real-time emergency assistant “iHelp” [24] was proposed for Deaf-mute people where they can report any kind of emergency. The current location of the user is accessed through built-in GPS system in a smartphone. The information about the emergency is sent to the management through SMS and then passed on to the closest suitable rescue units, and hence the user can get rescue using iHelp. MonoVoix [25] is an Android application that also acts as a sign language interpreter. It captures the signs from a mobile phone camera and then converts them into corresponding speech. Sahaaya [26] is an Android application for Deaf-mute people. It uses sign language to communicate with normal people. The speech-to-sign and sign-to-speech technology are used. For a hearing person to interact with Deaf-mute, the text-to-speech (TTS) technology inputs the speech signal, and a corresponding

sign language video is played against that input through which the mute can easily understand. Bragg et al. [27] proposed a sound detector. The app is used to detect alert sounds and alert the deaf-mute person by vibrating and showing a popup notification.

Improvement in health care access among Deaf people possible by providing the sign language supported visual communication and implementation of communication technologies for healthcare professionals. Some of the implemented technology-based approaches for facilitating Deaf-mutes with easy-to-use services are as follows.

3. Materials and Methods

The Arduino Mega 2560 is a microcontroller board based on the ATmega2560. It has 54 digital input/output pins, which can be used as PWM outputs, 16 analog inputs, 4 UARTs hardware serial ports, a 16MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable or power it with a AC-to-DC adapter or battery. The Mega 2560 board is compatible with most shields designed.



Figure 1: Arduino Mega 2560 Microcontroller

Table 1: Arduino Mega 2560 Product Specification

Microcontroller -	ATmega2560
Input Voltage - (recommended)	7-12V
Input Voltage (limit) -	6-20V
Digital I/O Pins	54 (of which 15 provide PWM output)
Analog Input Pins -	16
DC Current per I/O Pin -	20 mA
DC Current for 3.3V Pin -	50 mA
Flash Memory -	256 KB of which 8 KB used by bootloader
SRAM -	8 KB
EEPROM -	4 KB
Clock Speed -	16 MHz
LED_BUILTIN -	13
Length -	101.52 mm
Width -	53.3 mm
Weight -	37 g

The Touch Board is a microcontroller board with dedicated capacitive touch and MP3 decoder ICs. It has a headphone socket and micro SD card holder (for file storage), as well as having 12 capacitive touch electrodes. It is based around the ATmega32U4 and runs at 16MHz from 5V. It has a micro USB connector, a JST connector for an external lithium polymer (LiPo) cell, a power switch, and a reset button. It is like the Arduino Leonardo board and can be programmed using the Arduino IDE. The ATmega32U4 can appear to a connected computer as a mouse or a keyboard, (HID) serial port (CDC) or USB MIDI device.

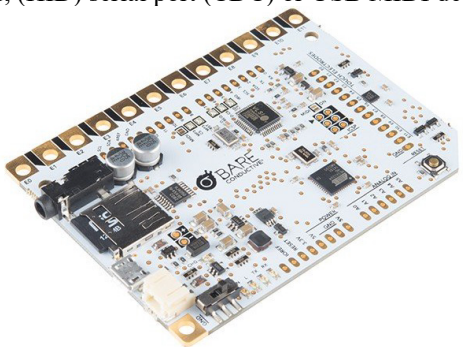


Figure 2: Bare Conductive Touch Board

Table 2: Bare Conductive Touch Board Product Specification

Microcontroller - Touch IC -	Microchip ATmega32U4 Resurgent Semiconductor MPR121
MP3 Decoder IC -	VLSI Solution VS1053b
Audio Output -	15mW into 32Ω via 3.5mm stereo socket
Removable Storage -	Up to 32GB via micro SD card
Input Voltage -	3.0V DC – 5.5V DC
Operating Voltage -	5V DC
Max. Output Current - (5V Rail)	400mA (100mA at startup)
Max. Output Current - (3.3V Rail)	300mA
LiPo Cell Connector -	2-way JST PH series – pin 1 +ve, pin 2 -ve
LiPo Charge Current -	200mA
Capacitive Touch Electrodes-	12 (of which 8 can be configured as digital I/O)
Digital I/O Pins -	20 (of which 3 are used for the MPR121 and 5 are used for the VS1053b-the latter can be unlinked via solder blobs)
PWM Channels -	7 (shared with digital I/O pins)
Analogue Input Channels -	12 (shared with digital I/O pins)
Flash Memory -	32kB (ATmega32U4) of which 4kB used by bootloader
SRAM -	2.5kB (ATmega32U4)
EEPROM -	1kB (ATmega32U4)
Clock Speed -	16MHz (ATmega32U4), 12.288MHz (VS1053b)
DC Current per I/O Pin -	40mA sink and source (ATmega32U4), 12mA source / 1.2mA sink (MPR121)
Analogue Input Resistance -	100MΩ typical (ATmega32U4)

Table 3: Bare Conductive Touch Board Input and Output Pins

Touch Electrodes - E0-E11	Connect to the MPR121 and provide capacitive touch or proximity sensing.
Serial - Pins 0 (RX) and 1 (TX)	Used to receive (RX) and transmit (TX) TTL serial data using ATmega32U4 UART.
TWI (I2C) - Pins 2 (SDA) and 3 (SCL)	Data and clock pins used to communicate with the MPR121.
IRQ - Pin 4	Used to detect interrupt events from the MPR121.
SD-CS - Pin 5	Used to select the micro SD card on the SPI bus.
D-CS - Pin 6	Used to select data input on the VS1053b.
DREQ -	Used to detect data request events from the VS1053b.
MP3-RST -	Used to reset the VS1053b.
MP3-CS -	Used to select the instruction input on the VS1053b
MIDI IN - Pin 10	Used to pass MIDI data to the VS1053b and have it behaved as a MIDI synthesizer
Headphone Output - AGND, R, L	Provide the headphone output from the VS1053b on 0.1" / 2.54mm pitch pads as an alternative to the 3.5mm socket.
External Interrupts - Pins 0, 1, 2, 3, 7	Configured to trigger an interrupt on a low value, a rising or falling edge, or a change in value.

This TFT display is big (2.8" diagonal) bright (4 white-LED backlight) and colorful (18-bit 262,000 different shades)! 240x320 pixels with individual pixel control. It has more resolution than a black and white 128x64 display. As a bonus, this display has a resistive touchscreen attached to it already. It can detect finger presses anywhere on the screen. It also includes an SPI touchscreen controller thus it only needs one additional pin to add a high-quality touchscreen controller. This display shield has a controller built into it with RAM buffering, so that almost no work is done by the microcontroller. This shield needs fewer pins than our v1 shield, in order to connect more sensors, buttons and LEDs: 5 SPI pins for the display, another pin or two for the touchscreen controller and another pin for microSD card to read images off of it.



Figure 3: TFT Touch Screen Display (2.8" diagonal)

Table 4: TFT Screen Pins

Digital #13 or ICSP SCLK -	This pin is used for the TFT, microSD and resistive touch screen data clock
Digital #12 or ICSP MISO -	This pin is used for the TFT, microSD and resistive touch screen data
Digital #11 or ICSP MOSI -	This pin is used for the TFT, microSD and resistive touch screen data.
Digital #10 -	This is the TFT CS (chip select pin). It is used by the Arduino to tell the TFT that it wants to send/receive data from the TFT only.
Digital #9 -	This is the TFT DC (data/command select) pin. It is used by the Arduino to tell the TFT whether it wants to send data or commands

Table 5: Resistive Touch Controller Pins

Digital #13 or ICSP SCLK -	This pin is used for the TFT, microSD and resistive touch screen data clock.
Digital #12 or ICSP MISO -	This pin is used for the TFT, microSD and resistive touch screen data.
Digital #11 or ICSP MOSI -	This pin is used for the TFT, microSD and resistive touch screen data.
Digital #8 -	This is the STMPE610 Resistive Touch CS (chip select pin). It is used by the Arduino to tell the Resistive controller that it wants to send/receive data from the STMPE610 only.

Table 6: Capacitive Touch Pins

SDA -	This is the I2C data pin used by the FT6206 capacitive touch controller chip. It can be shared with other I2C devices. On UNO's this pin is also known as Analog 4.
SCL -	This is the I2C clock pin used by the FT6206 capacitive touch controller chip. It can be shared with other I2C devices. On UNO's this pin is also known as Analog 5.

Table 7: MicroSD Card Pins

Digital #13 or ICSP SCLK -	This pin is used for the TFT, microSD and resistive touch screen data clock
Digital #12 or ICSP MISO -	This pin is used for the TFT, microSD and resistive touch screen data
Digital #11 or ICSP MOSI -	This pin is used for the TFT, microSD and resistive touch screen data
Digital #4 -	This is the microSD CS (chip select pin). It is used by the Arduino to tell the microSD that it wants to send/receive data from the microSD only

3.1 Power Supply Design

The two microcontrollers used in the study are the Bare Conductive Touch Board and Arduino Mega 2560 Rev3. As shown in Figure 5, both microcontrollers are directly powered. Arduino Mega 2560 Rev3 was powered by a 9V Battery with power snap for easier insertion. While the Touch Board was operated with a 3.7V 3000mAh Lithium Polymer Battery. Both power supplies are rechargeable, ensuring energy efficiency. Meanwhile, Bare Conductive Touch Board is powered via rechargeable 3.7V Li-Po Battery, which is directly inserted onto the boards LiPo socket.

The main reason why there are two designated power supplies for each microcontroller is to ensure that there is a sufficient supply independent from one another.

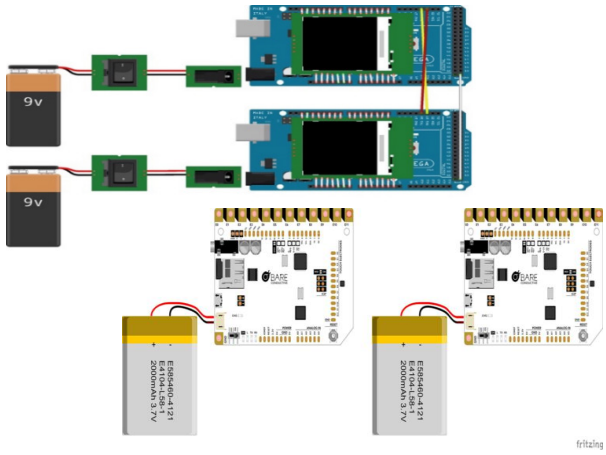


Figure 4: Illustrates how the LCD Touch Shield is mounted to Arduino Mega 2560 Rev3 with 9V power supply (top) and the Bare Conductive Touch Boards is powered thru 3.7V, 3000mah LiPo battery (bottom).

3.2 Software Application – Terminal Emulation Program

Terminal Emulation Program is a simple serial port (COM) monitoring program that is a very useful tool in debugging serial communication applications.

3.3 Software Application – Arduino Integrated Development Environment (IDE)

Figure 6 shows the real-time process happening inside the touch boards. Using the Arduino IDE 1.6.7. serial monitor, the reading indicates an accurate response, with no delay or whatsoever. From the instance that it registers a physical touch a pre-programmed audio response is played instantly. The Arduino Integrated Development Environment is a cross-platform application that is written in functions from C and C++. It is used to write and upload programs to Arduino compatible boards, but also, with the help of 3rd party cores, other vendor development boards. Furthermore, contains a text editor for writing code, a message area, a text console, a toolbar with buttons for common functions and a series of menus.

3.4 Hardware Design

The indicated connections are the I/O pins that when attached together fits in a mount position. These are the exact pin outs to where the touch screens should be connected to the microcontrollers, to execute the desired output. Basic RX-TX communication was utilized to interact and link the two pairs as shown in Figure 7.

Programming for Arduino is done on Arduino IDE which is based on C Language. It's neither completely C nor embedded C. The code written on the IDE will be passed directly to a C/C++ compiler (avr-g++).

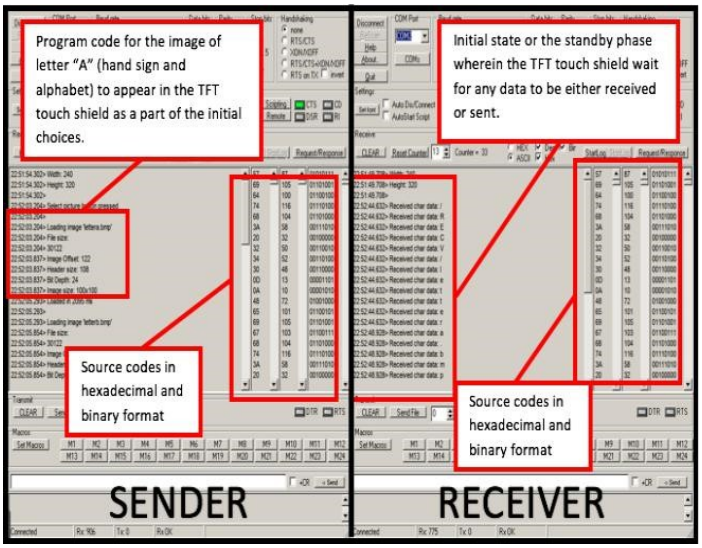


Figure 5: Actual data acquired using the Terminal Emulation Program. What is shown are the source codes in hexadecimal and binary format in real time. This was utilized to ensure accuracy and reliability in transmitting data from one shield to the other and vice-versa.

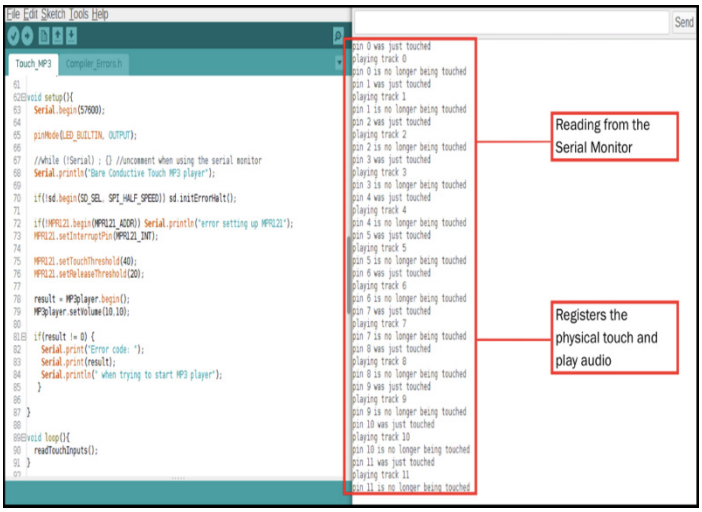
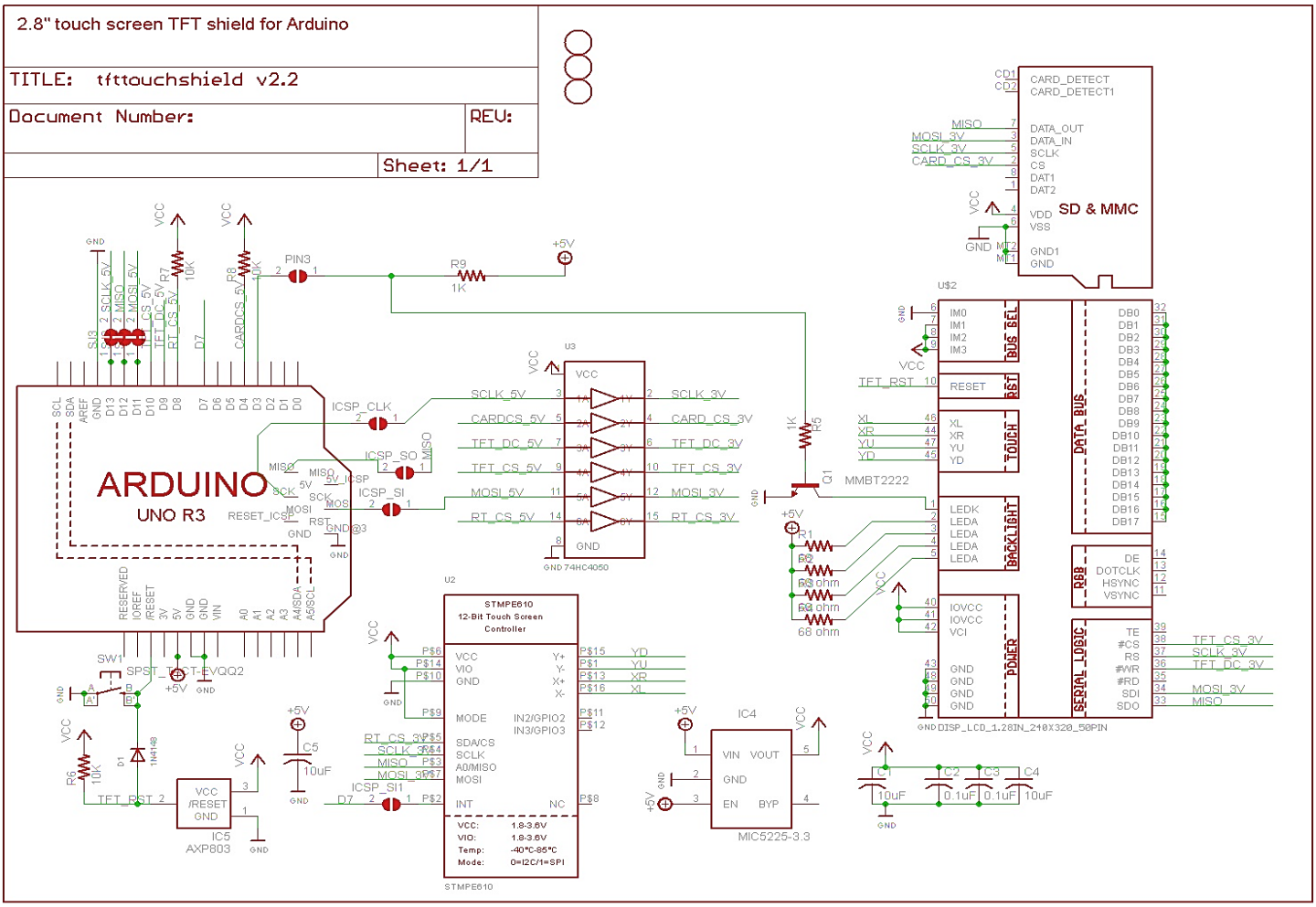


Figure 6. As indicated, for every physical tap a user made on the solder pads, it will play the preloaded phrase track. A total of twelve tracks, one for each pad.

3.5 Software Application – Balabolka Text-to-Speech (TTS)

One of the twelve preloaded phrase tracks is "*Hello! How are you? I am a Bilateral Communication Device for Hearing and speech Impaired. A communication device that pertains to people suffering from hearing and speaking disabilities most referred to as deaf, mute, and deaf-mute individuals. I am capable of being used both indoor and outdoor environment. I am dedicated to bridge the language gap between the impaired subjects to those of the normal, talking, non-sign language people. Have a good day.*"



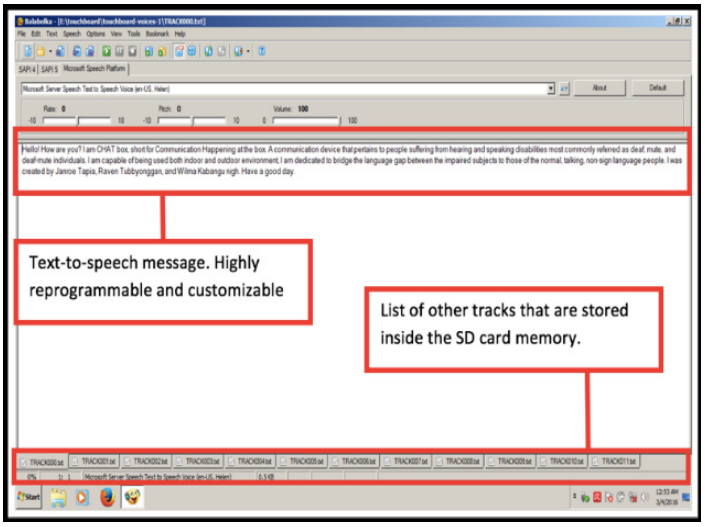
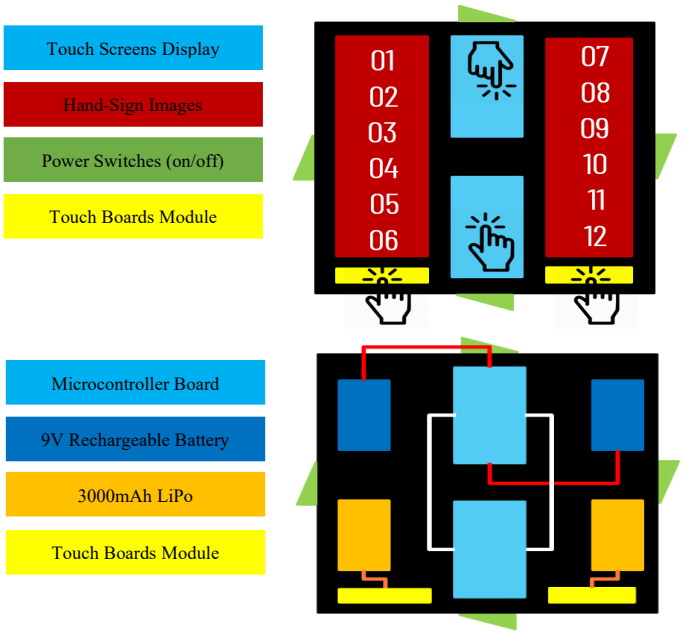


Figure 8: Actual data seen using a third-party program named Balabolka. What shown is the raw text file to be imported as .mp3 format to the micro SD Card which will then be inserted to the Touch Board's compartment.

The system is composed of two microcontroller and two touch boards. It can receive and send sign language images in .bmp format via touch screens through the microcontrollers' RX-TX serial communication. The system comprises four microcontrollers, one electric paint, two earphones, two rechargeable 9V batteries and two 3.7V lithium-ion rechargeable batteries with 3000mAh capacity and a twin barrel battery charger. (1) A surface consists of twelve common words and/or phrases based from the 1000 most common phrases used); (2) Two touch screens consisting of preloaded standard English alphabet and phrases for the non-sign language users. The impaired subjects have two options either to tap the surface with the signed image of their choice, correspondingly the Touch Board will read the message out loud via built-in speaker or to communicate using the touch screens like texting.

It is a bilateral communication device comprising of the following processes and components: (1) two microcontrollers, two touch boards, configured to act as with powerful processing power with a built-in Li-Po and Li-Ion battery charger. Standard audio jack for speakers and earphones. Composed of twelve electrode solder pads which produce audio when touched or tapped: (2) Innovative User Interface. Two microcontrollers with two mounted screen display module and touch boards Solder Pads; (3) Stable UART Communication. An RX-TX serial connection was utilized to send and receive data from one touch screens to another: (4) On

Top Guide. A surface consisting of twelve hand sign images with their corresponding translations; (5) Text-to-Speech Capable. A third-party text-to-speech program (TTS) called Balabolka, enabling the converted audio in an .mp3 format to be exported within the Touch Board's SD card memory; (6) Lightweight and Sturdy Casing. A lightweight packaging made from stainless alloy was constructed to form a rigid box for indoor and outdoor usage; (7) Accessories Integrated. Two earphones attached on the two touch boards respectively for audio feature and One twin barrel battery charger: (8) Long-lasting Battery Consumption. Two rechargeable 9V 220mAh battery supply for the microcontrollers; Two rechargeable 3.7V 3000mAh li-ion battery; (9) Convenient On & Off Preferences. Two separate rocker switches (on/off) states for the touch screens; Two separate built-in switches (on/off) states found in the touch boards for interface purposes; (10) Non-Water-Resistant. Any nearby water reservoir could damage the device when exposed to.

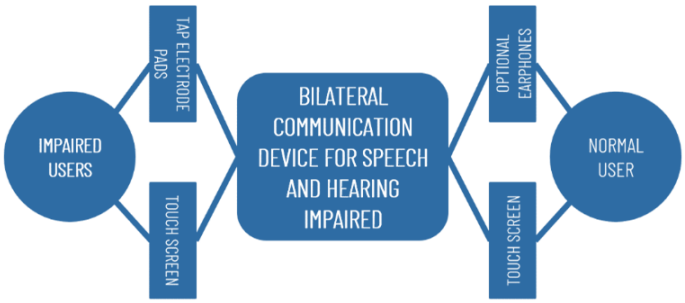


5. Perspective Renderings of the Device

Based on design it has an eight by eight (8x8) inch aluminum casing, with four differently sized compartments dedicated for the two 2.8-inch touch shields comprising the thirty percent (30%), two touch boards comprising the ten percent (10%), and display of hand sign images with English phrase translation comprising the sixty percent (60%) of the space. Additionally, there are four (4) separate power switches, one for each microcontrollers and boards, respectively. The device is approximately weighed at around 2.8 lbs. to 3.0 lbs.

6. Proposed Methodology

Distress caused by a verbal barrier which hinders the subjects from being able to communicate freely and easily to sign language incapable individuals and vice-versa is the focal point of concern this project aims to resolve. This communication device consists of two major parts: (1) a surface consists of two Bare Conductive Touch Boards preloaded with the twelve phrases and words with its corresponding display of hand sign images converted to its English translation and (2) two TFT Touch Shields mounted in Arduino Mega 2560 Rev3 respectively, preloaded with the standard English alphabet and American Sign Language symbols. The Deaf-mute have two options either to tap the desired electrode pads on the touch board or to communicate using the TFT touch shield like a multi-touch screen. In this regard, we proposed and developed a system that converts and interprets American sign-language (ASL) hand-sign image to its corresponding English translation and vice versa. Aside from that, it can also produce speech or audio responses via interactive capacity touch and proximity electrode pads. The bilateral communication device promotes two-way communication between Deaf-mute and normal person in a close. The normal person refers to the one who has no hearing or vocal impairment or disability. The main features of the prototype's operating system are listed below.



Users communicate with one another using the device and its peripherals.

6.1 Normal to Deaf-Mute Person Communication.

This module takes text message of a normal person as an input and outputs an image file in a .bmp format, that displays sign language images for the Deaf-mute person. Each image file is given tags and indexing. The steps of normal to Deaf-mute person communication are as follows:

- The application takes text of normal person as an input.
- The application converts the text message of a normal person into an image by using the text-to-image conversion program.
- The program matches the text to any of the image tags and index associated with the file and displays the corresponding sign for the Deaf-mute.

6.2 Deaf-Mute to Normal Person Communication.

Not everyone has knowledge of sign language. The Deaf-mute have two options either to tap the desired electrode pads preloaded with the most commonly used English words and phrases located on the touch board or to communicate using the TFT touch shield like a multi-touch screen, swiping and tapping hand-sign and illustrative images of English words in alphabetical order by default. A search option is also integrated to the system to find images on the screen faster.

6.3 System Flowchart of the Operating System.

It is straightforward bilateral communication device with simple, yet sharp responses made possible through the layer of process it undergoes. As shown in Figure 11, the device starts with tapping

(choosing the words), followed by chatting (forming syntax and semantics), then a process of sending or receiving (decision-making), after that two alternative options follow either to ignore (no) or to reply (yes). To ignore means ending the system flow while replying resumes the entire process creating a loop from tapping to responding, until the user ends it. The reply stage involves the sending and receiving process which constitutes conversion of text message to picture message (normal user to Deaf-mute) or conversion of picture message to text message (Deaf-mute to normal user).

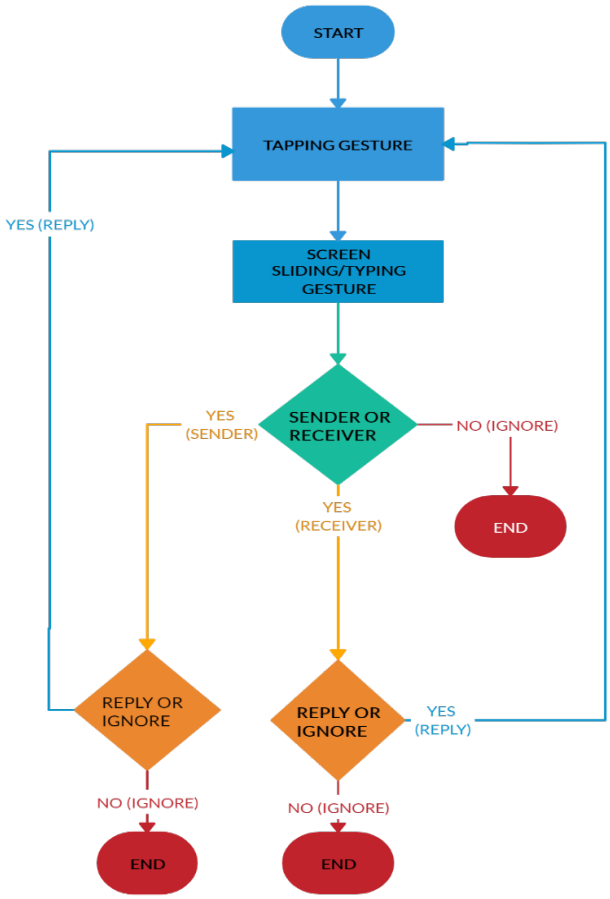


Figure 11: Presents the System Flowchart: it shows the internal logic of the entire system, thus providing the backbone of the how the project works (start and end)'

6.4 Framing.

Framing process is used to split the pre-emphasizedaudio files into short segments. The Bare Conductive Touch Boards consists of capacitive and proximity electrodes that is responsive to touch. The said board also stores audio files via microSD card. The audio files is represented by N frame samples and the interframe distance or frameshift is M ($M < N$). In the program, the frame sample size (N) = 256 and frameshift (M) = 100. The frame size and frameshift in milliseconds are calculated as follows:

$$\text{FrameSize(ms)} = f_n = \frac{1}{N*M} = 25.6\text{milliseconds} \quad (1)$$

FrameShift = 10 milliseconds

6.5 Statistical Treatment of Data.

Weighted Mean is used for calculating the weighted mean statistics for the given data set. If all the values are equal, then the weighted mean is equal to the arithmetic mean. It is a kind of average, wherein instead of each data point contributing equally to the final mean, some data points contribute more weight than the others. If all the weights are equal, then the weighted mean equals the arithmetic mean.

Formula:

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i w_i)}{\sum_{i=1}^{n}(w_i)} \quad (2)$$

where,

Σ = the sum of
w = the weights
x = the value

7. Experimental Results and Discussion

7.1 Experimental Setup

This communication device consists of two major parts: (1) a surface consists of two Bare Conductive Touch Boards preloaded with the twelve phrases and words with its corresponding display of hand sign images converted to its English translation and (2) two TFT Touch Shields mounted in Arduino Mega 2560 Rev3 respectively, preloaded with the standard English alphabet and American Sign Language symbols. The Deaf-mute have two options either to tap the desired electrode pads on the touch board or to communicate using the TFT touch shield like a multi-touch screen. The device is powered via two separate power supply. Rechargeable 9V DC supply charges up the Arduino microcontrollers and a different rechargeable 3.7V, 3000mAh LiPo powers up the touch boards. Sufficient power supply is vital to ensure the device functionality and features work.

7.2 Message Conversion and Transmission Phase

The working operating system of the bilateral communication device. (a) Write/Text Message Mode allows normal user to type-in English words and sent it to Deaf-mute. (b) Picture Message Mode converts the previously sent text message, into its translation in hand-sign images. (c) Picture Message Mode allows the Deaf-mute to select sign language image sets and sent to normal user. (d) Write/Text Message converts the previously sent picture message, into its English word translation. The operating system treats each image stored in the memory card of touch screen displays, as indexed then call each .bmp image file once its match word is chosen. The same protocol works the other way around from text messages to picture message.

7.3 Qualitative Feedback

Researchers formalized questionnaire survey to evaluate the overall performance device through several criteria. The demographics of the respondents includes fifty percent (50%) non sign language users aged 18 to 30 years old and fifty percent (50%) Deaf-mute aged 18 to 30 years old. The researcher divides the population into strata based off the participant's racial demographics. Here, there are two strata, one for each of the two racial categories. Using this information, a disproportional stratified sample was further implemented. The basis for rating are the following parameters: (1) Value Added – The device was advantageous to the speech and hearing impaired people; (2) Aesthetic Design – The packaging of the device prove to be stable and sturdy; (3) Hardware Components - device's hardware is properly placed and well organized; (4) Built-in Features – The device has innovative user interfaces like the touch shield making it more interesting; (5) User-Friendliness - The device is easy to use, very convenient, and innovative; (6) Reliability – The data sent and received by the device are consistent and correct; and (7) Accuracy – The data sent and received by the device are exact and true on what was intended to.

7.4 Application Interface

Initial bootup of the screens from both ends of the device differs from each other. Normal users and Deaf-mute have their designated position to where they should be at. Display screen of the Deaf-mute contains picture message tab showing the multiple sign language images with English translation at the bottom, in grid view. These images can be a word, a letter, or a phrase. These are arranged alphabetically, in which an additional search bar is added for faster image searches. Meanwhile, display screen of the normal users contains a standard QWERTY keyboard layout. Both screens have sent button, wherein the conversion and transmission of data is initialized and completed once data is received from one end to the other

In this example (a), the message contains the word "hello". From this point, the normal user will tap "Send Message", then this will be sent and converted to the corresponding hand sign images on the second TFT Touch Shield for the Deaf-mute users. Accordingly, the previously sent "hello" word was converted into hand sign images as shown at (b).

In this example (c), the image that has the word "all done" was tapped. This is how the Deaf-mute will communicate with the normal user. Consequently, the image that has the word "all done" was received by the TFT Touch Shields from the normal user in an English-based text form (d).

7.5 Questionnaire Design

Corresponding interpretation includes: Excellent (4.51-5.00); Very Good (3.51-4.00); Average (2.51-3.50); Fair (1.51-2.50); Poor (1.00-1.50); Very Poor (0.00-0.99). Likert surveys are also quick, efficient, and inexpensive methods for data collection. They have high versatility and can be sent out fast. In answering the questionnaire, the respondents are group in pairs of normal users and Deaf-mute, for a total of 75 pairs.

The respondents rated the device based on the following criteria:

a. Value Added
b. Aesthetic Design
c. Hardware Components
d. Built-in Features
e. User-Friendliness
f. Reliability
g. Accuracy

Test cases involved:

a. Accuracy in transmitting data
(back and forth sending and receiving 10 times)
b. Consistency in transmitting data
(back and forth sending and receiving 10 times)
c. Speed (timing and delay in seconds)
d. Screen Refresh and Transitions Rates (min. 2 seconds)
e. Touch Responsiveness (min. 2 seconds)
f. User Experience (easy-to-use)
g. User Interface (easy-to-learn)

Table 8: Respondents' Scale

Range	Interpretation
5	Excellent (4.51-5.00)
4	Very Good (3.51-4.50)
3	Average (2.51-3.50)
2	Fair (1.51-2.50)
1	Poor (1.00-1.50)
0	Very Poor (0.00-0.99)

Table 9: Tallied Results based on the Respondents' Scale

Metrics	Users' Feedbacks						Weighted Mean	Interpretation
	0	1	2	3	4	5		
Value Added	0	130	10	10	0	0	4.80	Excellent
Aesthetic Design	0	50	100	0	0	0	4.33	Very Good
Hardware Components	0	70	80	0	0	0	4.47	Very Good
Built-in Features	0	90	40	20	0	0	4.47	Very Good
User-Friendliness	0	100	40	0	10	0	4.53	Excellent
Reliability	0	100	30	20	0	0	4.53	Excellent
Accuracy	0	110	20	20	0	0	4.60	Excellent
Overall Rating							4.53	Excellent

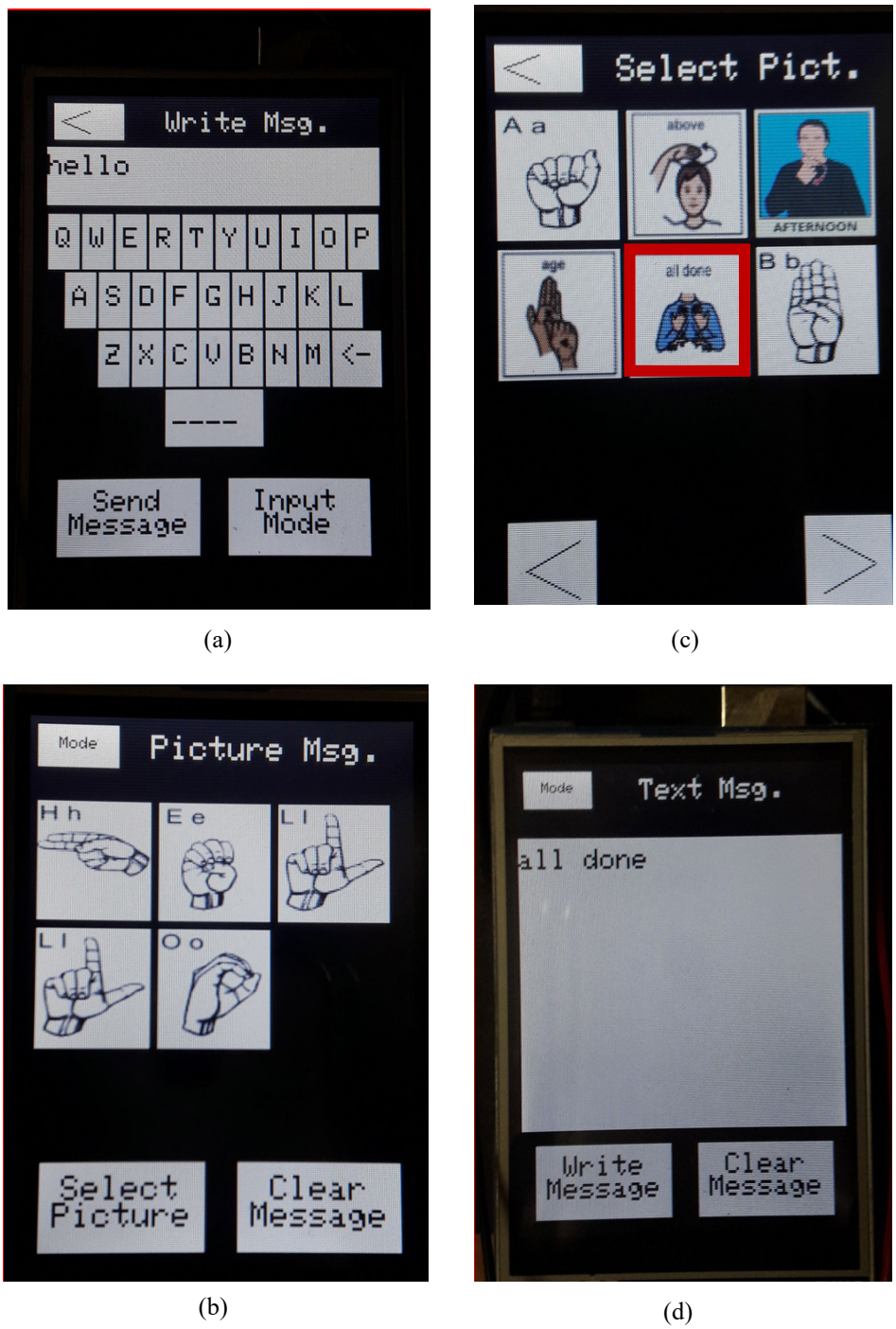


Figure 12: The working operating system of the bilateral communication device. (a) Write/Text Message Mode allows normal user to type-in English words and sent it to Deaf-mute. (b) Picture Message Mode converts the previously sent text message, into its translation in hand-sign images. (c) Picture Message Mode allows the Deaf-mute to select sign language image sets and sent to normal user. (d) Write/Text Message converts the previously sent picture message, into its English word translation.

8. Conclusion

Communication is a bilateral process and being understood by the person you are talking to is a must. Without the ability to talk nor hear, a person would endure such handicap. Given that hearing and speech are missing, many have ventured to open new communication methods for them through sign language. This bilateral communication device can be utilized by both non-sign language users and Deaf-mute together in a single system. Shaped as a box (8in x 8in) with two multi-touch capable displays on both ends, the contraption has several microcontrollers and touch boards within. The latter has the technology of twelve interactive capacity touch and proximity electrode pads that react when tapped, producing quick response phrases audible via speaker or earphone. These touch boards are equipped with an MP3 decoder, MIDI synthesizer, 3.5mm audio jack and a 128MB microSD card. The touch screen modules mounted on top of the microcontrollers transfer data to and from each other in real-time via receiver-transmitter (RX-TX) full duplex UART serial communication protocol. The device is lightweight weighing at about 3 lbs. The prototype device was piloted in an academic institution of special

education for deaf-mute students. Participants were 75 normal and 75 Deaf-mute people aged between 18 and 30 years. The experimental results show the overall rating of the device is 90.6%. The device is designed to promote the face-to-face socialization aspect of the Deaf-mute users to the normal users and vice versa. Several third-party applications were utilized to validate the accuracy and reliability of the device thru metrics of consistency, timing and delay, data transmission, touch response, and screen refresh rates.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The proponent would like to acknowledge the Samar State University particularly the College of Engineering and the Center for Engineering, Science and Technology Innovation for the continuous support, resources, and unwavering encouragements in order to achieve milestones and deliverables set for this research undertaking.

References

[1] R. Tabiongan, "Chatbox: two-way schemed communication device for speech and hearing impairment," in 2019 8th International Symposium on Next Generation Electronics (ISNE), Zhengzhou China, CN, 2019. doi: 10.1109/ISNE.2019.8896577

[2] T. Humphries and J. Humphries, "Deaf in the time of the cochlea," Journal of Deaf Studies and Deaf Education, 16(2), 53-163, 2010. doi: 10.1093/deafen/enq054

[3] V. Kimmelman, "List of sign languages and their abbreviations," Information Structure in Sign Languages, 23-26, 2019. doi: 10.1515/9781501510045-206

[4] I.W. Leigh and J.F. Andrews, "Deaf people and society: psychological, sociological and educational perspectives", Psychology Press, 2016. doi: 10.4324/9781315473819

[5] P. Vostanis, M. Hayes, M. Du Feu, and J. Warren, "Detection of behavioural and emotional problems in deaf children and adolescents: Comparison of two rating scales," Child: Care, Health and Development, 23(3), 233–246, 1997. doi: 10.1111/j.1365-2214.1997.tb00966.x

[6] J. Gugenheimer, K. Plaumann, F. Schaub et al., "The impact of assistive technology on communication quality between deaf and hearing individuals," in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, 669–682, March 2017. doi: 10.1145/2998181.2998203

[7] S. Kaur, and K.S. Dhinda, "Design and Development of Android Based Mobile Application for Specially Abled People". Wireless Pers Commun 111, 2353-2367, 2020. doi: 10.1007/s11277-019-06990-y

[8] J.L. PrayandI. K. Jordan, "The deaf community and culture at a crossroads: Issues and challenges," Journal of Social Work in Disability and Rehabilitation, 9, 2, 168–193, 2010. doi: 10.1080/1536710X.2010.493486

[9] S. Barnett, "Communication with deaf and hard-of-hearing people: A guide for medical education," Academic Medicine: Journal of the Association of American Medical Colleges, 77(7), 694–700, 2002. doi: 10.1097/00001888-200207000-00009

[10] E. Rothe and A. Pumariega, "Culture, identity, and mental health: conclusions and future directions" Oxford Medicine Online, February 2020. doi: 10.1093/med/9780190661700.001.0001

[11] M.V. Sharma, N.V. Kumar, S.C. Masaguppi, S. Mn, and D.R. Ambika, "Virtual talk for deaf, mute, blind and normal humans," in Proceedings of the 2013 1st Texas Instruments India Educators' Conference, TIIEC 2013, 316–320, April 2013. doi: 10.1109/TIIEC.2013.63

[12] F. Soltani, F. Eskandari, and S. Golestan,"Developing a gesture based game for deaf/mute people Using microsoft kinect," in Proceedings of the 2012 6th International Conference on Complex, Intelligent, and Software Intensive Systems, ICCISIS 2012, 491– 495, July 2012. doi: 10.1109/CISIS.2012.55

[13] A. K. Tripathy, D. Jadhav, S. A. Barreto, D. Rasquinha, and S.S. Mathew, "Voice for the mute", in Proceedings of the 2015 International Conference on Technologies for Sustainable Development, ICTSD 2015, February 2015. doi: 10.1109/ICTSD.2015.7095846

[14] N. P. Nagori and V. Malode, "Communication Interface for Deaf-Mute People using Microsoft Kinect," in Proceedings of the 1st International Conference on Automatic Control and Dynamic Optimization Techniques, ICACDOT2016, 640–644, September 2016. doi: 10.1109/ICACDOT.2016.7877664

[15] A. Sood and A. Mishra, "AAWAAZ: A communication system for deaf and dumb," in Proceedings of the 5th International Conference on Reliability, Infocom Technologies and Optimization, ICRITO 2016, 620–624, September 2016. doi: 10.1109/ICRITO.2016.7785029

[16] M. Yousuf, Z. Mehmood, H.A. Habib et al., "A Novel Technique Based on Visual Words Fusion Analysis of Sparse Features for Effective Content-Based Image Retrieval," Mathematical Problems in Engineering, 2018, Article ID 2134395, 13 pages, 2018. doi: 10.1155/2018/2134395

[17] D. Zhong and I. Defee, "Visual retrieval based on combination of histograms of ac block patterns and block neighborhood," International Conference on Image Processing, Atlanta, GA, 2006, 1481-1484, 2006. doi: 10.1109/ICIP.2006.312562

[18] Z. Mehmood, T. Mahmood, and M. A. Javid, "Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine," Applied Intelligence, 48(1), 166–181, 2018. doi: 10.1007/s10489-017-0957-5

[19] N. Ali, K.B. Bajwa, R. Sablatnig, and Z. Mehmood, "Image retrieval by addition of spatial information based on histograms of triangular regions," Computers & Electrical Engineering, 54, 539–550, 2016. doi: 10.1016/j.compeleceng.2016.04.002

[20] Z. Mehmood, S.M. Anwar, N. Ali, H.A. Habib, and M. Rashid, "A Novel image retrieval based on a combination of local and global histograms of visual words," Mathematical Problems in Engineering, 2016, Article ID 8217250, 1-12, 2016. doi: 10.1155/2016/8217250

[21] Z. Mehmood, F. Abbas, T. Mahmood, M.A. Javid, A. Rehman, and T. Nawaz, "Content-based image retrieval based on visual words fusion versus features fusion of local and global features," Arabian Journal for Science and Engineering, 1–20, 2018. doi: 10.1007/s13369-018-3062-0

[22] S. Jabeen ,Z. Mehmood, T. Mahmood, T. Saba, A. Rehman, and M. T. Mahmood, "An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model," PLoS ONE, 13, 4, e0194526, 2018.doi: 10.1371/journal.pone.0194526

[23] S. Ghanem, C. Conly, and V. Athitsos, "A Survey on Sign Language Recognition Using Smartphones," in Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, 171–176, ACM, June 2017. doi: 10.1145/3056540.3056549

[24] L.B. Chen, C.W. Tsai, W.J. Chang, Y. M. Cheng, and K.S.M. Li, "A real-time mobile emergency assistance system for helping deaf-mute people/elderly singletons," in Proceedings of the IEEE International Conference on Consumer Electronics, ICCE 2016, 45-46, January 2016. doi: 10.1109/icce.2016.7430516

[25] R. Kamat, A. Danoji, A. Dhage, P. Puranik, and S. Sengupta, "MonVoix-An Android Application for the acoustically challenged people," Journal of Communications Technology, Electronics and Computer Science, 8, 24–28, 2016. doi: 10.22385/jctecs.v8i0.123

[26] A. Kaur and P. Madan, "Sahaaya: Gesture Recognising System to Provide Effective Communication for Specially-Abled People," *2012 IEEE Global Humanitarian Technology Conference*, Seattle, WA, 2012, 93-97, doi: 10.1109/GHTC.2012.65

[27] D. Bragg, N. Huynh, and R. E. Ladner, "A personalizable mobile sound detector app design for deaf and hard-of-hearing users," in Proceedings of the 18th International ACMSIG ACCESS Conference on Computers and Accessibility, ASSETS2016, 3–13, 2016. doi: 10.1145/2982142.2982171

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 376-381 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Semiclassical Theory for Bacteria Motility Under External Electric Fields and Interactions with Nanodevices

Huber Nieto-Chaupis*

Universidad Autónoma del Perú, Panamericana Sur Km. 16.3, Villa el Salvador, Lima, Perú

ARTICLE INFO

Article history:
Received: 18 June, 2020
Accepted: 02 September, 2020
Online: 21 September, 2020

Keywords:
Bacteria
Classical Electrodynamics
Probability Amplitude

ABSTRACT

The accurate identification and characterization of microbiological species is a must that allows us to design engineered pharmacology to tackle down the diversity of diseases derived of them. This paper presents a study about the usage of both classical dynamical and electrodynamics in conjunction to the Feynman's path integral to describe as well as identify the displacement of bacteria in closed spaces. Our methodology consists in the in the usage of the usage of probability amplitude to investigate the theoretical motility of bacteria can be disturbed through electrical interactions from the fact that them contain ions in their biochemical composition.Our study yields that bacteria might exhibit a pattern of probabilities as function of space and time. It would be advantageous for an engineered nanodevice that would sense them uniquely through electrical interactions.

1 Introduction

It is well-known that some type of bacteria might be lethal in humans [1] moreover in the shortest times, so that their understanding with respect to physical and chemical interactions constitutes a must. Nowadays, notable advances to tackle down diseases caused by bacteria and virus have demanded to engineer precise pharmacology [2]. Furthermore, with the expected arrival of nanomedicine, one expects that the so-called nanodevices [3] can achieve an efficient task to defeat any intrusion of microbiological compound in humans. Thus, an accurate knowledge of bacteria's properties and biochemical characteristics turns out to be critic for the application of an accurate nano pharmacology, with the lowest cost-benefits fraction [4].

In this paper, theory of bacteria's response to electrical interactions derived from both classical electrodynamics and quantum mechanics path integrals is treated. Essentially, particular attention is paid on the probability amplitude. Commonly it is written as [5],

$$\mathcal{A} = \langle z_b|\mathbf{U}(t_b, t_a)|z_a\rangle\,.$$

with $\mathbf{U}(t_b, t_a)$ the quantum mechanics evolution operator,

$$\mathbf{U}(t_b, t_a) = \mathrm{Exp}\left[-i\left(\frac{\mathbf{P}^2}{2\mathrm{M}\hbar} + \frac{Q_J\mathbf{V}(\mathbf{R})}{|\mathbf{R}-\mathbf{R}_J|}\right)\right].$$

Although this equation is a pure quantum mechanics description of a compound electrically charged, the employment of distances at the order of nanometer leads us to implement this approach by which quantum effects cannot be discarded. Thus this paper has a target to present a scheme to identify bacteria through the measurement of the probability as function of time and spatial displacement. Because bacteria have a well defined ionic composition, it make them sensitive to electrical interactions. Therefore bacteria can be disturbed by electric fields and potentials, as consequence their dynamics is strongly affected. In this manner, chemotaxis and quorum among them are modified. Therefore, bacteria would exhibit patterns in according to their internal ionic composition. With this information, then is plausible their identification. The usage of quantum mechanics appears as a closed-form tool that would describe perturbed paths of bacteria. Thus, one can ask about the probability that an aggregation of bacteria is disabled in their main functions in presence of external electric fields. This paper has as purpose to answer this question.

The rest of this paper is as follows: in second section all the classic physics that would model bacteria and their spatial dynamics is presented. In third section, the implementation of the probability amplitude with the gained knowledge from the classical analysis is done. In fourth section, the implementation of integer-order Bessel's functions is done. This has as objective to extract the possible probabilities while a single bacterium is trapped in a closed space. Finally, conclusion of paper is presented.

*Corresponding Author: Huber Nieto-Chaupis, Email: hubernietochaupis@gmail.com

2 Bacteria in Classical Physics

As studied by Bechhold at 1904, bacteria are carriers of negative charges. At 1932 Leo Olitzki studied bacterial migration of cells under various values of pH [6]. After at 1932, Burke and Gibson analyzed electric reactions of bacteria observing spatial translation to the positive pole and demonstrating the negative charge composition of bacteria under fixed electrodynamic situations [7]. Therefore, it is possible to extract an exact dynamics to understand the response and behavior of bacteria or any microbiological aggregation in a scenario of electrical exposure.
Thus, by knowing the electric potential, the entire dynamics is derived from the Newton's equation given by: $\vec{\mathbf{F}} = Mvdv/ds = -Q\vec{\nabla}[\Phi(x,y,z) + \phi(x,y,z)]$ with $\phi(x,y,z)$ a noise that emerges in macro-biological scenarios. However in isolated environments $\phi(x,y,z) \approx 0$.

2.1 The Classical Dynamics Approach

Consider the Poisson's equation defined by an electric potential $\mathbf{U}(\mathbf{r})$ and charge volumetric density ρ written as:

$$\nabla^2\mathbf{U}(\mathbf{r}) = -\frac{\rho}{\epsilon_0} \tag{1}$$

and ϵ_0 the electrical permittivity. Then after an integration in both members of Eq.(1),

$$\nabla \int (\nabla \mathbf{U}(\mathbf{r}))\, dV = -\frac{1}{\epsilon_0}\int \rho(\mathbf{r}) dV$$
$$\int \nabla\cdot\mathbf{E} dV = \frac{1}{\epsilon_0}\int \rho(\mathbf{r}) dV \Rightarrow \mathbf{E}\int d\mathbf{A} = \frac{Q}{\epsilon_0} \Rightarrow,$$

with $\mathbf{E}$ the electric field one can arrive to:

$$\mathbf{E} = \frac{Q}{\epsilon_0 \int d\mathbf{A}}, \tag{2}$$

where the well-known Gauss's law is used. In this manner the electric force $\mathbf{F} = q\mathbf{E}$, onto a single bacterium of charge q due to the presence of an external total charge Q is written as $\mathbf{F} = q\mathbf{E} = \frac{qQ}{\epsilon_0\int d\mathbf{A}} \Rightarrow \int vdv = \int \frac{2qQ}{M\epsilon_0\int d\mathbf{A}} d\mathbf{s}$ yielding the displacement:

$$R(t) = \int_0^T dt \sqrt{\int \frac{2qQ}{M\epsilon_0\int d\mathbf{A}} d\mathbf{s}} = T\sqrt{\frac{2qQ}{M\epsilon_0}\int \frac{d\mathbf{s}}{\int d\mathbf{A}}}, \tag{3}$$

clearly the charge Q supposed to be the one that acts onto the bacterium is a finite sum of independent charges. Thus one gets:

$$R(t) = T\sqrt{\frac{2q\sum_j^J Q_j}{M\epsilon_0}\int \frac{d\mathbf{s}(t)}{\int d\mathbf{A}}}, \tag{4}$$

in the same logic, the displacement done by a compound of N bacteria due to the presence of Q charges is written as:

$$R(t) = T\sqrt{\frac{2\sum_n^N q_n \sum_j^J Q_j}{M\epsilon_0}\int \frac{d\mathbf{s}(t)}{\int d\mathbf{A}}}. \tag{5}$$

If all charges have same magnitude, then $\sum_n^N q_n = Nq$ and $\sum_j^J Q_j = JQ$. A toy model can be established from Eq.(5) in the sense that the net displacement for N bacteria and J external charges can be written as:

$$R(t) = T\sqrt{NJ}\sqrt{\frac{2qQ}{M\epsilon_0}\int \frac{d\mathbf{s}(t)}{\int d\mathbf{A}}}. \tag{6}$$

Integrations inside the square root can be carry out in a straightforward manner in according to the choice of coordinate system. For instance a choice is the cylindric system allows us to write $\int d\mathbf{A}$=$2\pi L\mathbf{r}(t)$. On the other hand while the electric force acts onto the radial direction then $d\mathbf{s}(t) \to d\mathbf{r}(t)$ therefore one obtains that:

$$R(t) = T\sqrt{NJ}\sqrt{\frac{2qQ}{2\pi\epsilon_0 ML}\int \frac{d\mathbf{r}(t)}{\mathbf{r}(t)}}$$
$$= T\sqrt{NJ}\sqrt{\frac{2qQ}{2\pi\epsilon_0 ML}\mathrm{Log}\left[\frac{\mathbf{r}(t)}{\mathbf{r}(0)}\right]}. \tag{7}$$

Clearly the classic physics establishes that the displacement is dictated by the inverse of the root square of product of ML. In this manner, when $ML >> NJ$ the net displacement might be negligible.

2.2 The Electrodynamics Classical Approach

A rigorous treatment about the dynamics of any ionic compound in an electrical scenario demands to use exact solutions of either Poisson or Laplace equations. By following the picture that a single bacterium is enclosed inside a cylindric system, it is possible to derive a closed-form formulation of the bacteria displacement. In fact, consider a charged object inside a finite cylindric shape, then the electric potential becomes the Jackson's potential [8] and it is written as:

$$\mathbf{U}(\mathbf{x},\mathbf{x}') = \frac{2Q}{\pi\epsilon_0 LR^2}\sum_{m,k,n}^{\infty} \mathrm{e}^{im(\phi-\phi')}\sin\left(\frac{k\pi z}{L}\right)\sin\left(\frac{k\pi z'}{L}\right)\times$$
$$\times\frac{J_m\left(\frac{x_{mn}r}{R}\right)J_m\left(\frac{x_{mn}r'}{R}\right)}{\left[\left(\frac{x_{mn}}{R}\right)^2+\left(\frac{k\pi}{L}\right)^2\right]J_{m+1}^2(x_{mn})}. \tag{8}$$

Thus, one can see that actually this full solution is the product of the individual solutions of the involved coordinates, so that one can state that the solution is separable as: $\mathbf{U}(\mathbf{x},\mathbf{x}') = \Phi(\phi,\phi')Z(z,z')R(\rho,\rho')$. While one can adopt the radial solution, it is explicitly dependent on the integer numbers m,k,n.
In this manner, the solution along the z coordinate appears as the one that exhibits tractability from the mathematical point of view, and that fits with the physics about the motility and spatial displacements of bacteria. Subsequently, for a fixed radius and polar angle, the Jackson's potential can be written in its simplified form as:

$$\mathbf{U}(\mathbf{z},\mathbf{z}') = \frac{2Q}{\pi\epsilon_0 LR^2}\sum_{k}^{\infty}\sin\left(\frac{k\pi z}{L}\right)\sin\left(\frac{k\pi z'}{L}\right). \tag{9}$$

With this one can estimate the electric field from $\mathbf{E} = -\nabla \mathbf{U}(\mathbf{z}, \mathbf{z}')$ created along the longitudinal coordinate turns out to be:

$$\mathbf{E} = -\hat{\mathbf{k}}\frac{\partial}{\partial z'}\mathbf{U}(\mathbf{z}, \mathbf{z}') = \frac{2kQ}{\epsilon_0 L^2 R^2}\sum_k^{\infty} \sin\left(\frac{k\pi z}{L}\right)\cos\left(\frac{k\pi z'}{L}\right)\hat{\mathbf{k}}. \quad (10)$$

When the difference $|z - z'| << 1$ the electric field depends on this so that the identity sin(2x)=2 sin(x)cos(x) is used. Thus, a single bacterium containing an ion of charge q inside the cylinder feels this field and it is going to experience a net electric force given by:

$$\mathbf{F}(\mathbf{z}) = q\mathbf{E}(\mathbf{z}) = \frac{-kqQ}{\epsilon_0 L^2 R^2}\sum_k^{\infty} \sin\left(2\frac{k\pi z}{L}\right)\hat{\mathbf{k}}. \quad (11)$$

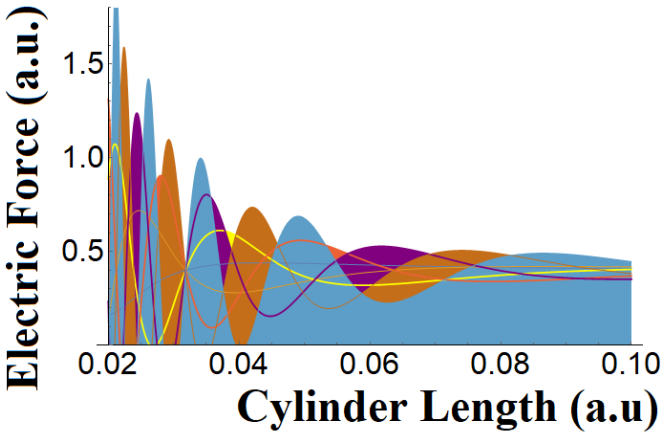


Figure 1: The intensity of electric force $\|\mathbf{F}(\mathbf{z})\|$ from Eq.(11) in the S.I system on a single bacterium as function of the length of cylinder L for values of k from 1 to 7. A noise field of 40% is attained.

3 The Probability Amplitude of Bacteria Displacement

3.1 Motivation

The prospective implementation of nano technologies through nano devices, would demand to employ accurate theoretical formulations that would have to encompass for example the nano biological scenarios. Under this view, emerges quantum mechanics as a robust theoretical frame that would describe the interactions through the language of probabilities.

In fact, commonly in quantum mechanics the probability amplitude $\mathcal{A}$ requires of the exact knowledge of the Green's function or the function that dictates the space-time propagation. For example, the propagation of a massive particle from the point t_a, z_a to t_b, z_b is given by:

$$\mathcal{A} = \langle z_b|\mathbf{U}(t_b, t_a)|z_a\rangle\,. \quad (12)$$

In the case of a free particle, the evolution operator is explicitly given by:

$$\mathbf{U}(t_b, t_a) = \mathrm{Exp}\left[\frac{iE(t_a - t_b)}{\hbar}\right] = \mathrm{Exp}\left[\frac{ip^2(t_a - t_b)}{2M\hbar}\right]. \quad (13)$$

In praxis one needs to employ the representation of momentum given by the completeness relation $\int dp\,|p\rangle\langle p| = \mathrm{I}$ on gets:

$$\mathcal{A} = \frac{1}{2\pi\hbar}\int dp \mathrm{Exp}\left[-i\left(\frac{p^2(t_b - t_a)}{2M\hbar} + \frac{p(z_a - z_b)}{\hbar}\right)\right]. \quad (14)$$

In order to evaluate the integral, it is required the explicit form of the initial and final states at the momentum representation. Thus, these states can be written as $\mathrm{Exp}(-i\frac{pz_a}{\hbar})$ and $\mathrm{Exp}(i\frac{pz_b}{\hbar})$ respectively with normalization constants for both states given by $\frac{1}{\sqrt{2\pi\hbar}}$, with the causal condition $t_a > t_b$ then the integration turns out to be in a straightforward manner resulting in (the reader can check that the square completion "trick" is used.):

$$\mathcal{A} = \sqrt{\frac{M}{i2\pi\hbar(t_a - t_b)}}\mathrm{Exp}\left[\frac{iM(z_b - z_a)^2}{2\hbar(t_b - t_a)}\right]. \quad (15)$$

Clearly Eq.(15) requires of an additional physics that by the which one can describe the displacement of bacteria in presence of external fields. Therefore, Eq.(15) can model all those trajectories with those intensities that are negligible and cannot to affect dynamics.

3.2 The Path Integral Scheme

The scheme of path integral invented by Feynman [9] inside the framework of quantum electrodynamics [10] emerged as an independent methodology to find probability amplitudes of elementary processes.

For instance, the case subject to a negligible external potential one could be treating the free particle containing kinetic energy. Therefore, the probability amplitude contains an infinite number of paths for a spatial displacement initialized at z_b, t_b arriving to z_a, t_a. In order to calculate the amplitude probability the employed time for carrying out such motility demands to assume that the whole employed time has N+1 slices in the form: $t_\ell = \epsilon\ell + t_a$ with the integer ℓ running from $\ell = 0, 1, ..., N, N+1$ so that the concrete problem consists in to evaluate the product of integrals from the general formulation:

$$< z_b, t_b|z_a, t_a >= \lim_{N\to\infty}\frac{1}{\sqrt{2i\pi\epsilon/m}}\times$$
$$\times\prod_{i=1}^{N}\int_0^L \frac{dz_i}{\sqrt{2\pi\epsilon/m}}\exp\left[i\sum_{i=1}^{N+1}\frac{m}{2\epsilon}(z - z_{i-1})^2\right]. \quad (16)$$

It should be noted that the kinematic part is denoted by the free-particle Lagrangian

$$\mathcal{L} \approx \sum_{i=1}^{N+1}\frac{m}{2\epsilon}(z - z_{i-1})^2\,. \quad (17)$$

Turning back on the bacteria displacement, one can apply the method given above to attack the problem of bacteria motility when it is subject to external potential. In fact, Janke and Kleinert [11] have demonstrated that the path integral of a free particle between two space points inside a box is actually the product of the well-known wave functions when these closed-form solutions of the Schrödinger's equation. Thus, Janke and Kleinert have found that

$< z_b, t_b|z_a, t_a >$ turns out to be a closed-form expression characterized by having a complex exponential, and reads:

$$< z_b, t_b|z_a, t_a >= \frac{2}{L}\sum_{q=1}^{\infty}\sin(\frac{\pi q z_b}{L})\sin(\frac{\pi q z_a}{L}) \text{Exp}\left[-i\frac{(\pi q)^2(t_b-t_a)}{2mL^2}\right]. \quad (18)$$

It is interesting to note that of product of two sin functions matches well to right-side of Eq.(9). In this manner, because the presence of the wave functions: $\Psi(\beta z_a) = \sqrt{\frac{2}{L}}\text{Sin}(\beta z_a)$ and $\Psi(\beta z_b) = \sqrt{\frac{2}{L}}\text{Sin}(\beta z_b)$, with $\beta = \frac{q\pi}{L}$, one can interpret that are solutions for a free particle in a cylinder. Furthermore, Eq.(18) can be also written in a simplified manner as:

$$< z_b, t_b|z_a, t_a >= \sum_{q=1}^{\infty}\Psi(\beta z_b)\Psi(\beta z_a)\text{Exp}\left[-i\frac{\beta^2(t_b-t_a)}{2m}\right], \quad (19)$$

thus for small displacements, $z_a \approx z_b$ so that Eq.(19) can be written in a compact form as:

$$< z_b, t_b|z_a, t_a >= \sum_{q=1}^{\infty}|\Psi(\beta z_a)|^2\,\text{Exp}\left[-i\frac{\beta^2(t_b-t_a)}{2m}\right], \quad (20)$$

exhibiting the square of wave function denoting the probability density. With the definition of $\gamma = \frac{2Q}{\pi\epsilon_0 L R^2}$ the Jackson's potential from Eq.(9) can be rewritten as:

$$\mathbf{U}(\mathbf{z_b}, \mathbf{z_a}) = \gamma\sum_{q}^{\infty}\sin(\beta z_b)\sin(\beta z_a). \quad (21)$$

Certainly, Q denotes the charge that generates the Jackson's potential inside and outside of cylinder of length L and radius a. Therefore, the quantum mechanics paths are contained inside the cylinder. It is noteworthy to remark that the geometry "seen" by the charge Q is a finite cylinder.

3.3 Bacteria in a Cylindrical Geometry

Consider now a negatively charged bacteria under circumstances as given in [**?**] with electric properties as studied and simulated in [**?**] with charge Q_B that can move along the surface of cylinder, then the force per unit of charge attained to a single bacterium can be derived in a straightforward manner as:

$$\frac{\mathbf{F}}{Q_B} = \mathbf{E} = \nabla\mathbf{U} = \frac{Q}{4\pi\epsilon a^2}\sum_{q=1}^{\infty}\text{Exp}\left(-i\frac{t\beta^2}{2M}\right)\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right). \quad (22)$$

where the electric potential has been approximated to the probability amplitude. Under this assumption it is clear that Eq.(22) differs notable from the pure classic force. One can note that the quantity M and Q_B becomes the mass and charge of a single bacterium. The infinite sum gives account of the quantum paths. Aside, the term $\int < z_b, t_b|z_a, t_a > dt$ exhibits the character probabilistic since $< z_b, t_b|z_a, t_a >$ is understood as the density of probability.
It is interesting to note that from Eq.(22) the basic Coulomb force between the bacteria and test charge separated by a distance a each other can be derived, yielding:

$$\mathbf{F} = \frac{QQ_B}{4\pi\epsilon a^2}\sum_{q=1}^{\infty}\text{Exp}\left(-i\frac{t\beta^2}{2M}\right)\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right),$$

and the classic Coulomb law is recovered yielding:

$$\mathbf{F}_{\text{CL}} = \frac{Q_B Q}{4\pi\epsilon a^2} \quad (23)$$

that would indicate that bacterium and a charge Q (as nano device for instance) are experiencing a Coulomb force each other along the quantum path (as seen at Fig.2) with a separation given by a distance a the cylinder radius.
In other words the single bacterium in a cylinder with $a << L$ is attained to be on the cylinder surface. In this manner, the charge Q generates a Coulomb-like force to a single bacterium.
In addition, the gradient of the probability $\int < z_b, t_b|z_a, t_a > dt$ provides the stochastic component to the dynamics of bacteria in presence of an external charge Q.

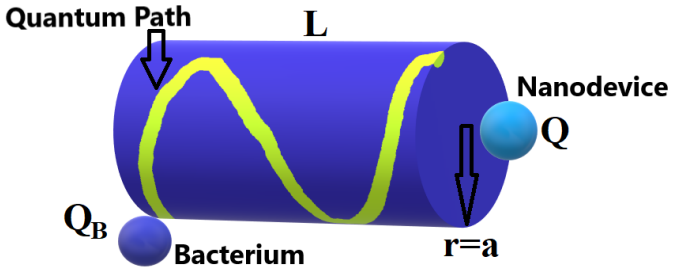


Figure 2: Sketch of the quantum path along the external area of cylinder indicating the constant radial distance between the bacterium and nanodevice.

The resulting hybrid force or semi-classical force might opt only of the real part with respect to the complex exponential, so that one gets that:

$$\mathbf{F} = \mathbf{F}_{\text{CL}}\sum_{q=1}^{\infty}\text{Cos}\left(\frac{t\beta^2}{2M}\right)\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right). \quad (24)$$

Below in Fig.3 this force is plotted. For this end the quantity $\int < z_b, t_b|z_a, t_a > dt$ is assumed to be proportional to αz that means that the probability to find the bacteria inside the cylindric geometry increases with the longitudinal distance z. Despite of this, one can see that the electric force decreases with the bacterial mass or aggregation [12].

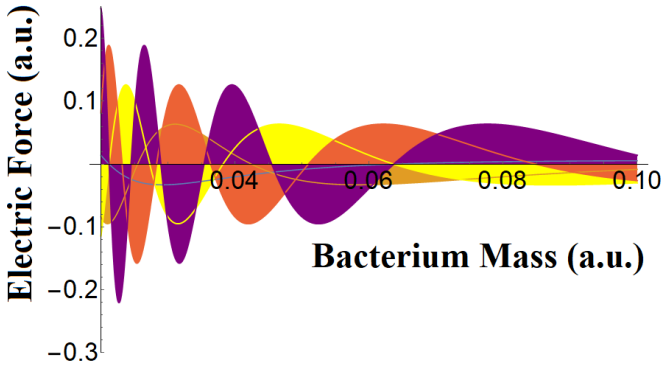


Figure 3: The electric force at the S.I as function of bacteria mass in according to Eq.(24). The force appears to be extenuated in up to a 40% of their top values as seen for a mass of 0.08 expressed in arbitrary units.

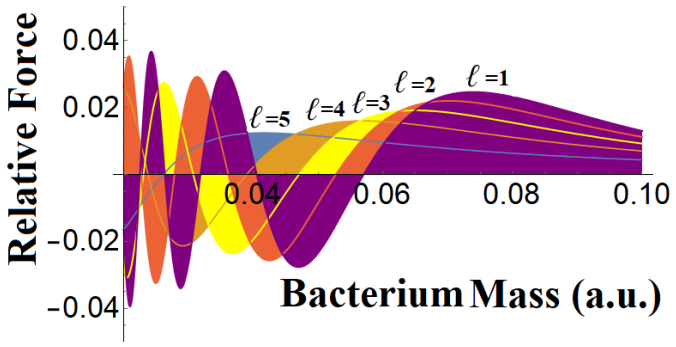


Figure 4: The relative electric force from Eq.(26) in the S.I up to for the first 5 orders of the Bessel function.

4 Bessel Shapes of Bacterium Paths

Turning back to Eq.(24), attention is paid on the complex exponential,

$$\mathbf{F} = \mathbf{F}_{\rm CL}\sum_{q=1}^{\infty}{\rm Exp}\left(-i\frac{t\beta^2}{2M}\right)\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right). \tag{25}$$

by which the usage of the Bessel expansion appears to be feasible in this analysis through the implementation of

$${\rm Exp}\left(-i\frac{tq^2\pi^2}{2ML^2}\right) = {\rm Exp}\left(-i\frac{t}{2M}\sin\lambda\right) \tag{26}$$

where the change $\sin(\lambda) = \frac{q^2\pi^2}{L^2}$, therefore the direct application of the expansion then reads as:

$${\rm Exp}\left(-i\frac{t}{2M}\sin\lambda\right) = \sum_{\ell} J_\ell\left(\frac{t}{2M}\right){\rm Exp}(-i\ell\lambda), \tag{27}$$

so that the relative force reads:

$$\frac{\mathbf{F}}{\mathbf{F}_{\rm CL}} = \sum_{\ell} J_\ell\left(\frac{t}{2M}\right)\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right). \tag{28}$$

In Fig.4, Eq.(28) the relative force was plotted up to for 5 orders of the Bessel function. For this end $\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right)$ turned out to be proportional to $\mathrm{Cos}(z_b)$. The dominant distribution is for $\ell = 1$. In contrast to Fig.3, the relative force does not decreases for large values of bacterium mass. The fact that one can see small values, then it might correlated to the dominance of the classic of Coulomb force against the quantum description. It is fully coherent for large values of bacteria's mass that exhibits a small relative force indicating that that for large masses the Coulomb force governs the electrical interactions.

In order to include the probabilistic concept in Eq.(28) the relative force can be understood as the gradient of quantum mechanics probability in the sense of:

$$\left|\frac{\mathbf{F}}{\mathbf{F}_{\rm CL}}\right|_\ell = J_\ell\left(\frac{t}{2M}\right)\nabla\mathcal{P}(z_b) \tag{29}$$

with $\mathcal{P}(z_b)$ the probability of finding a single bacterium at z_b.

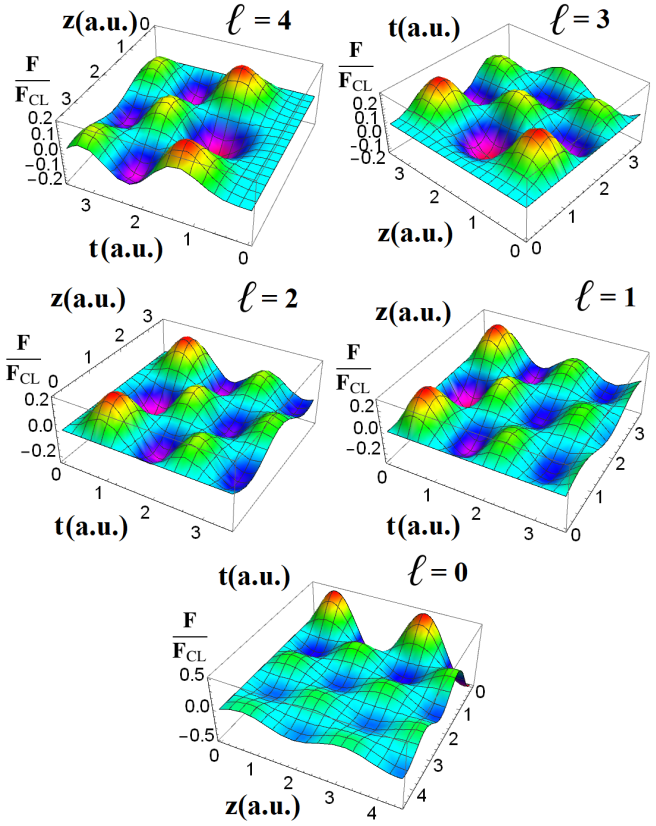


Figure 5: The relative electric force from bottom to top for the orders ℓ=0-4 as function of space and time (expressed in arbitrary units) for a fixed value of bacterium mas $M = 1$.

4.1 Probabilities as Electric Patterns of Bacteria

In Fig.5 up to 5 different 3D distribution of the Eq.28 at a fixed mass, and as function of the z coordinate and time are plotted.

Since the relative force is proportional to the gradient of probability $\nabla\left(\int < z_b, t_b|z_a, t_a > dt\right)$, then Fig.5 can be seen as the curves of probabilities.
In this sense the peaks that all surfaces are exhibiting would correspond to the best scenarios by which one can measure the mass leading to the identification of a particular specie. In bottom panel the case when ℓ=0 is shown. One can see that at $t = 0$ up to 2 maximum values for $z = 0.5$ and 3.0 are exhibited by the 3D curve. One clearly can see that the approximation of Eq.(28) might give a better interpretation of this curve in the sense that for $\ell = 0$ results in $\left|\frac{\mathbf{F}}{\mathbf{F}_{\rm CL}}\right|_\ell \nabla\mathcal{P}(z_b)$ exhibiting the important weight of the quantum mechanics probability that would be the cause for the apparition of peaks. In middle right panel, the relative force for $\ell = 1$ displays again two peaks (both curves in color red) for t=0.1, while at the left side the case for $\ell = 2$ again is exhibited two peaks for t=0.2. Therefore, the order of the Bessel function might be understood as the cause of a slight shift at time with a reduced relative force. This is analogue to the case of the wave packet whose amplitude is minimized with time. Therefore the relative force has actually a quantum behavior as consequence of the classical and quantum engagement through the Jackson's potential. However, the degradation of these surfaces of probabilities are balanced with the apparition of minor peaks as seen at left and right plots. In top left plot, the case when $\ell = 3$ is plotted. Here there are two peaks at $t = 1.5$ at $z = 1$ and $z = 3.5$. It is noteworthy as done in previous case, one can see the apparition of minor peaks. Again, the variation as to up and down of the relative force is entirely related to the quantum behavior. The drastic case is seen at the left panel where $\ell = 4$ whose peaks appear to be rather reduced in contrast to previous cases. In effect, at $t = 2$ and $z = 2.5$ is seen a hole resulting in a null probability and the electric interaction with a nano device [13] inside the cylinder is either lost or not enough to measure the tracking of bacterium under the assumption that **there multiple paths that bacterium would opt while is inside of a closed geometry.**

5 Conclusion

In this paper, the analysis of any microscopic biological specie such as a single bacterium or virus containing a ionic aggregation inside, has been treated through classical and quantum methodologies. The results, have shown that the path integral formulation might enclose the classical electrodynamics formulation dictated by the Jackson's potential. Thus, the relative force was defined as the ratio of the quantum mechanics over the Coulomb-like under the assumption that an engineered nano device [14] coexist inside the cylinder. The 3D plots (done with the package Wolfram [15] for figures 1-5) for different orders of Bessel function have exhibited that the bacterium acquires a pure quantum mechanics behavior. Further studies to understand this quantum and classical correspondence in a more general view are required to support the diverse transitions between the classical and quantum mechanics in nano systems subject to classical fields.

References

[1] A. F. Boggs, "Antibacterial resistance: who pays?" Expert Opinion on Therapeutic Patents, **13(8)**, 1107–1112, 2005, doi:10.1517/13543776.13.8.1107.

[2] L. Good, "Antisense antibacterials," Expert Opinion on Therapeutic Patents, **12(8)**, 1173–1179, 2002, doi:10.1517/13543776.12.8.1173.

[3] H. Nieto-Chaupis, "Macrophage-Like Nanorobots To Anticipate Bacterial Dynamics," in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 0873–08789, 2019, doi:10.1109/CCWC.2019.8666506.

[4] H. Nieto-Chaupis, "Bacteria Nano Communications Described by a Machine Learning Theory Based on the Feynman Path Integral," in 2019 IEEE 53th Annual Conference on Information Sciences and Systems (CISS), 1–4, 2019, doi:10.1109/CISS.2019.8692893.

[5] J. J. Sakurai, Modern Quantum Mechanics, Chapter II Section 2.5, Addison-Wesley, 2010.

[6] L. Olitzki, "Electrical Charge of Bacterial Antigens," The Journal of Immunology, **4**, 251–256, 1932, doi:10.1186/s11671-019-3005-z.

[7] V. Burke, F. O. Gibson, "The Gram Reaction and the Electric Charge of Bacteria," The Journal of Immunology, **2**, 211–214, 1932, doi:10.1186/s11671-019-3005-z.

[8] J. D. Jackson, Classical Electrodynamics, Third Edition, Chapter II, Problem 3.13, Wiley, 1999.

[9] R. P. Feynman, "Space-Time Approach to Quantum Electrodynamics," Physical Review, **76**, 769, 1949, doi:10.1103/PhysRev.76.769.

[10] R. P. Feynman, "Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction," Physical Review, **80**, 440, 1950, doi:10.1103/PhysRev.80.440.

[11] W. Janke, H. Kleinert, "Summing Paths for a Particle in a Box," Lettere Nuovo Cimento, **25**, 297–300, 1979, doi:10.1007/BF02776259.

[12] M. M.-H. Hasnaa Fatehi, M. T. Figge, "Modelling cellular aggregation induced by chemotaxis and phototaxis," Mathematical Medicine and Biology: A Journal of the IMA, 373–384, doi:10.1093/imammb/dqq002.

[13] Y. W. Mingyu Wang, Z. Yang, "Cooperation Method of Symmetrically Distributed Multi-Nanorobotic Manipulators Inside SEM for Nanodevice Constructing," in 2018 IEEE 13th Annual International Conference on Nano/Micro Engineered and Molecular Systems (NEMS), 543–548, 2018, doi:10.1109/NEMS.2018.855.6984.

[14] Q. Xia, "Memristive nanodevices: CMOS compatibility and novel applications," in 2016 IEEE 18th Mediterranean Electrotechnical Conference (MELECON), 1–4, 2016, doi:10.1109/MELCON.2016.7495316.

[15] "Wolfram Mathematica https://www.wolfram.com," .

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 391-397 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Comparison of Gaze Points Among Viewing Conditions (Resolution, Display Size, Viewer Position) During Video Viewing

Miho Shinohara*, Yusuke Nosaka, Reiko Koyama, Riko Nakagawa, Takuya Sarugaku, Mitsuho Yamada

Graduate School of Information and Telecommunication Engineering, Tokai University, Tokyo, 108-0074, Japan

ARTICLE INFO

Article history:
Received: 16 July, 2020
Accepted: 01 November, 2020
Online: 20 November, 2020

Keywords:
Ultra-High Resolution
Average Gaze Position
Viewing Position

ABSTRACT

Ultra-high resolution broadcasting such as 4K and 8K are becoming more popular for at-home use. The position of the viewer's mean gaze point when they are looking at a larger display depends on the video clip, viewing position, and viewing distance. The International Telecommunication Union provided the document BT-2022; as the standard viewing conditions for the subjective evaluations of flat-panel TV displays. However, it is rare that a television is viewed at home under standard viewing conditions. In most homes, a TV of any size is watched from the viewer's desired position according to the layout size of the room hosting the TV. No study has been conducted to determine how the movement of a viewer's line of sight changes when a display with different resolution is viewed from a position different from the standard viewing condition. If the gaze position when viewing TV differs due to the difference in resolution, the size of the display, and the difference in observation position, it affects the method of TV production. This would also affect consumers considering the resolution, installation location, and viewing location of a TV. Here, we clarified that the mean gaze position and its standard deviation are almost the same viewing conditions (i.e., resolution, display size, and viewing position) are changed, even when the same contents are viewed. Our findings demonstrate that consumers could buy a display of their preferred resolution and size according to their room layout.

1. Introduction

In Japan, high-definition broadcasting with a resolution of 2K has become widespread, and the public use of ultra-high-definition 4K and 8K broadcasting began on December 1, 2018. In the International Telecommunication Union (ITU-R) document 'BT-2022' ('General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays'), the standard viewing distance for a television's display is defined as the maximum distance at which a person with a visual acuity of 1.0 can see the pixel structure [1]. Since the current HDTV broadcasting has approx. 2 million pixels at 1920×1080, the standard viewing distance when watching HDTV images is 3H (H; Display height). In contrast, 4K broadcasting has ~ 8 million pixels at 3840×2160, and that of 8K broadcasting has ~ 33 million pixels at 7680×4320.

It was reported that when the resolution becomes higher, the pixel structure cannot be seen even when one is approaching it, and thus it is possible to view from a closer distance and enjoy a wide-angle image [2]. The ITU-R stipulates that a viewer should watch from the center of the front of the screen even when approaching the screen, but at each home, the viewing position differ depending on the room layout, the number of viewers, the size of the TV purchased, and more. The display is frequently viewed from sites that are not at the center of the screen. According to a survey of TV viewing positions, the positions spread horizontally ±60° and vertically ±30°, and 37% of viewers look at the display from the right side of the TV, only 3% from the center, and 60% from the left side [3]. When watching a movie on a large screen, the impression of the movie varies greatly depending on the position of the viewer's seating, such as the center of the screen, the left and right edges of the front seats, and the rearmost seats. However, no research has examined how the position of the content where a viewer views on a TV display changes depending on the position of the viewer's seating. If the content is viewed differently depending on the seating position, the producers of televised content must consider the effects of the viewing position when producing the content. Consumers seeking a TV to purchase also need to consider where to put the TV and furniture.

*Corresponding Author: Miho Shinohara, 0cjnm010@mail.u-tokai.ac.jp

A study of the differences in the viewing of contents depending on the TV's resolution compared NTSC; analog color system with (525 scanning lines) and high-definition (1,125 scanning lines) [4]. In the low-resolution NTSC, the viewer's gaze points are distributed in the center of the screen because multiple cameras are switched in order to provide the camera shot that the content's director wants to show in the center. In high-definition, the entire screen is displayed, and the viewers' gaze points are spread over the entire screen. Several research groups have verified the differences in viewing distance by using parameters such as screen size, display brightness, resolution, and video content [5-8]. In those studies, the contents were viewed from the center position of the display, and the viewing position was not changed in the horizontal or vertical direction. In addition, an NTSC or phase alternating line (PAL) system was used (not higher-definition 4K), and the results cannot always be applied to a 4K or 8K display. In the present study, using 4K high-definition images that have never been used, we examined the differences in viewers' lines of sight when they were watching television from a distance that is not the standard viewing distance and from various positions (right, center or left). Viewing 4K or 8K ultra-high-definition TV with large-screen images from a wide field of view at a short distance has become established in for home use. As a result, viewers are more likely to see at various viewing positions, not limited to the center of the screen, as seen in movie theaters, but there is no study that examined the effect of viewing position on viewing behavior from eye movements. Therefore, our present findings will provide meaningful data for both viewers and video producers to know whether or not the observation position affects the gaze position.

The purpose of our research is to objectively evaluate the psychological effects of changes in display resolution, size, and viewing position on viewing behavior from eye movements. For that purpose, two kinds of experiments were conducted.

In Experiment 1, we analyzed the influence of the display resolution on the viewing positions of individuals watching three video clips on the display. We showed the subjects a video clip whose resolution was 4K (3840×2160), 2K (1920×1080) or 1K (960×540) at the viewing distance of 0.75H. In Experiment 2, we analyzed the influence of the size of the screen on the viewing position. We showed the subjects a video clip whose screen size was 1/1, 1/2 or 1/4 at 0.75 H as the viewing distance [9].

2. Experimental Method

2.1. Experimental equipment

When viewing a 4K image at the viewing distance of 1.5H, the field of view is 60°. The range that a human being can see by using only saccades, which is a high-speed eye movement, is up to 15° [10]. Head movements are used together with eye movements to exceed the 15° range. It was reported that head movements contributed twice as much as the eye movements when moving the line of sight to fixate at one point [11]. It is also known that in the ratio of head movement to the line of sight, the head share becomes 60%-80% when the line of sight moves outside a 30° range [12, 13]. We thus surmised that in this experiment, it was necessary to detect changes in the line of sight that included the addition of the subjects' head movements to their eye movements. We used the line-of-sight measuring device developed in our laboratory [14]. The accuracy of the horizontal direction was approx. 1.5° to -1.7° at ±30° corresponding to the visual angle of the display edge at 1.5H. The accuracy of the perpendicular direction was approx. 2.6° to -1.4° at ±18°, corresponding to the visual angle of the display edge at 1.5H [4]. We measured eye movements (EMR-8B; Nac Image Technology, Tokyo) and head movements (3DM-GX4; Microstrain, Williston, VT, USA). Data were recorded at 60 Hz. In the experiment, the subjects looked at the display in a free posture without being fixed on the head-chin rest of the experimental set-up.

2.2. Viewing conditions

The viewing distance of the display was fixed at 1.5H, and the subject was allowed to view the display from three positions: the center (display center), the right (1/4 to the right from the display center), and the left (1/4 to the left from the display center) as shown in Figure 1. The standard viewing distance for image-quality evaluations is not 1.5H when the image resolution is 2K or 1K, or when the screen size is 1/2 or 1/4: however, in a most households, the room layout is prioritized, and the viewing distance often does not change even if the screen resolution or size changes. We therefore fixed the viewing distance. Since the subjects' head movements were measured simultaneously with their eye movements, the subject's head movements were not restricted and the subject was instructed to watch the TV in a natural posture. The subjects watch videos with three types of resolution, three screen sizes, and three viewing positions, but each subject was prevented from seeing the same image more than once. This experiment was conducted according to BT-2022 and followed the Ethical Provisions of Tokai University's guidelines for "Research using humans."

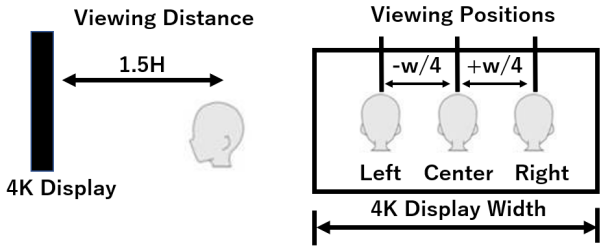


Figure 1: Viewing conditions.

2.3. Laboratory environment

The video clips were presented to a 4K projector (VPL-VW 500ES; Sony, Tokyo) using a personal computer (PC) as the signal source. The height of the screen was 147 cm, the width of the screen was 262 cm, and the screen brightness was 0.04 cd/m^2 at the minimum and 104 cd/m^2 at the maximum. The illuminance in the room was 69.7 Lux. The layout of the laboratory, the procedure of the experiment, and the selection of subjects were based on BT-2022.

2.4. Changes in gaze position due to the difference in resolution (Experiment 1)

The details of the video clips used in Experiment 1 are shown in Table 1 and in Figures 2-4 (scenes used for video clips 1-3). The 4K video clips consisted mainly of natural scenes (mountains, flowers, seas) and were downloaded from a free video clips material website [15]. The 4K video clips were recorded in the

60p/XAVCS format, which is also used for broadcasting, and have low compression. Each video clip length was 40 sec. The series of 4K video clips was edited to a length of approx. 4 min with the use of nonlinear editing software (VegasPro13; Sony). The 2K and 1K video clips were down-converted from 4K video clips using the same nonlinear editing software. A still video showing the fixation point in the center was inserted before and after each video clip presented to the subject, so that it could be recognized even if the calibration of the eye movement was misaligned. The subjects were 25 college students (10 males and 15 females) in their twenties, seven of whom had the visual acuity score of >1.0: the other 18 had the visual acuity score of 0.7 (including corrected vision). We analyzed the data of 21 of the 25 subjects (seven males, and 14 females) because four subjects had a large number of measurement errors during the experiment. No subject viewed the same image more than once. Each subject was randomly determined which image to view at which resolution and viewing position.

Table 1: Video clips used in Experiment 1

Video clip	Time	Resolution
Video clip 1: River, ship or bridge	4 min 14 sec	4K, 2K and 1K
Video clip 2: Cherry blossoms or Mt. Fuji	4 min 2 sec	4K, 2K and 1K
Video clip 3: Town or animal	4 min 9 sec	4K, 2K and 1K



Figure 2: Scenes used for video clip 1.



Figure 3: Scenes used for video clip 2.



Figure 4: Scenes used for video clip 3.

2.5. *Changes in gaze position due to the difference in screen size (Experiment2)*

The details of the video clips used in Experiment 2 are shown in Table 2 and Figures 5-7 (scenes used for video clips 1-3). The video clips were edited as in Experiment 1, and 1/4-size and 1/2-size video clips were generated by reducing the number of pixels from 4K video clips using the nonlinear editing software. The subjects were 24 college students (nine males and 15 females) in their twenties, eight of whom had a visual acuity score of >1.0: the other 16 had a visual acuity score of 0.7 (including corrected vision).

Nineteen of the 24 subjects also participated in Experiment 1. There were no measurement errors in the data of all 24 subjects, and all were used for analysis. No subject viewed the same image more than once. Each subject was randomly determined which image to view at which screen size and viewing position.

Table 2: Table 2. Video clips used in Experiment 2

Video clip	Time	Screen size
Video clip 1: Flowers	4 min 7 sec	1/4, 1/2, 1/1 size of 4K
Video clip 2: Okinawa	4 min 2 sec	1/4, 1/2, 1/1 size of 4K
Video clip 3: Yokohama	3 min 14 sec	1/4, 1/2, 1/1 size of 4K



Figure 5: Scenes used for video clip 1.



Figure 6: Scenes used for video clip 2.



Figure 7: Scenes used for video clip 3.

3. Results

The distribution of the gaze points of each subject was analyzed while the subject viewed the video clips. We used 10 deg/s as the threshold speed for extracting the gaze point from the gaze movement [16]. The horizontal and vertical components of gaze were calculated by the following formulas.

Horizontal components of Gaze = Horizontal components of Eye movement + Horizontal components of Head rotation

Vertical components of Gaze = Vertical components of Eye movement + Vertical components of Head rotation

The mean gaze position and its standard deviation (std. dev.) during viewing have often been used to compare the gaze behaviors of individual watching video clips [4]. We thus calculated the mean gaze position and its std. dev. from the remaining 4 minutes of data excluding the 1 minute immediately after the presentation of the video clips, at which point the data were not stable.

The mean gaze point position and the std. dev. of the mean gaze point distribution showed the same tendencies in all three video clips, and we therefore analyzed the data of the three video clips together. We performed a two-way analysis of variance (ANOVA) to analyze the results. The level of significance was set at 10%.

3.1. Results of Experiment 1: Change in gaze position due to the difference in display resolution (Experiment 1)

During the measurement of a person's gaze, there may be many blinks, or the eye movement sensor may shift, causing an error in eye movement. Of the 21 subjects test in Experiment 1, the seven subjects who had almost no measurement error during the measurements for all the 1K, 2K, and 4K video clips were used for the analysis in order to perform statistical analysis with all the accurate data, because, it is extremely difficult to perform all experiments completely without measurement errors due to the nature of eye movement measurement. Figure 8 shows the averages and standard deviations of the mean gaze point positions of these seven subjects. In the figure, the center of the screen is 0°, and the angles displayed on the horizontal and vertical axes represent the viewing angles used by the subject. Figure 9 shows the mean gaze positions obtained by summarizing the results of the three viewing positions for each resolution.

The mean gaze position at the left and right viewing positions was around the center of the screen, and the average gaze position at the central viewing position was slightly left of the center of the screen regardless of the resolution, as shown in Figure 8. In the summary of the results of the three viewing positions, the mean gaze position is located at the center of the screen regardless of the differences in resolution, as shown in Figure 9. These results suggest that there was a slight difference in the mean gaze position depending on the viewing position, but it was not due to the difference in display's resolution.

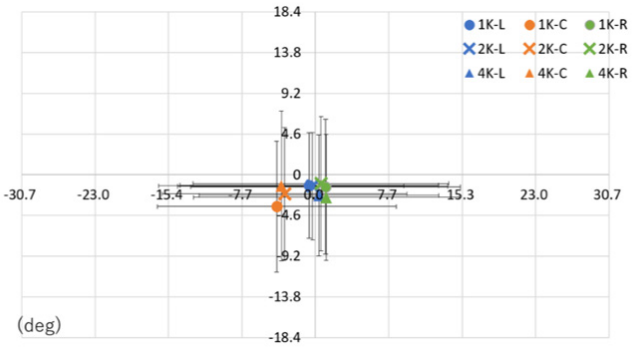


Figure 8: Averages of the mean gaze position and std. dev. (resolution and viewing position)

The results of the two-way ANOVA for mean gaze position with resolution and viewing position as variables are illustrated in Figure 10. No significant differences were observed for resolution, but a multiple comparison analysis revealed a significant difference in viewing position in the right-central and left-central positions. No significant differences were detected in the std. dev. of mean gaze position as shown in Figure 11; however, there was a slight but significant difference in viewing positions at the 10% level of significance.

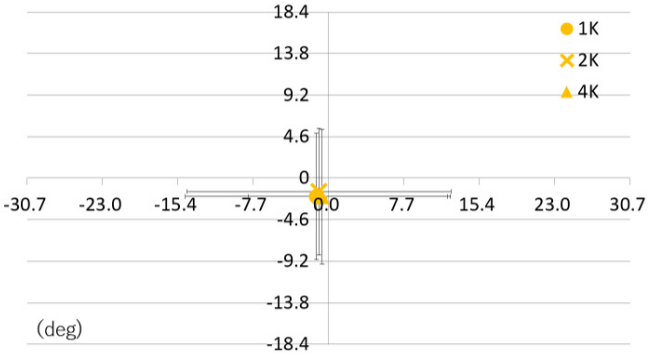


Figure 9: Averages of the mean gaze position and std. dev. (resolution).

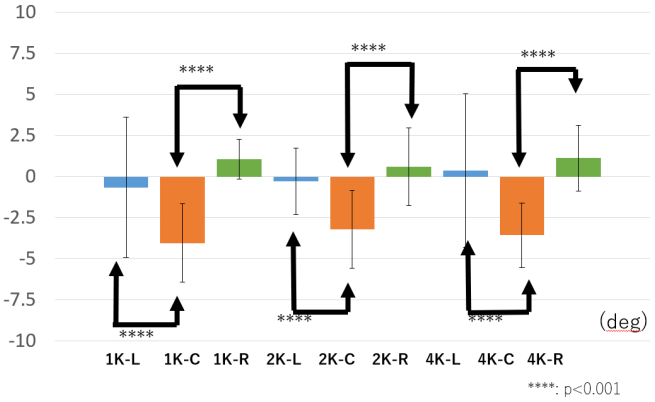


Figure 10: Results of the two-way ANOVA for mean gaze position (resolution).

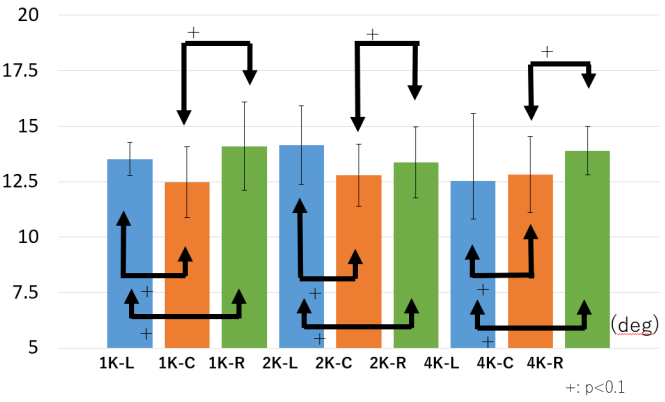


Figure 11: Results of the two-way ANOVA for the std. dev. of mean gaze position (resolution).

3.2. Results of Experiment 2: Change in gaze position due to the difference in screen size (Experiment 2)

In Experiment 2, of the 24 subjects, eight who had almost no measurement error during the measurements for all of the 1K, 2K,

and 4K video clips were used for the analysis to perform the same accurate statistical analysis as in Experiment 1. The average mean gaze position of these eight subjects who viewed 1/4-, 1/2- and 1/1-size video clips at each viewing position is shown in Figure 12. Figure 13 provides the mean gaze positions obtained by summarizing the results of the three viewing positions for each resolution. In this experiment, since only the screen size was changed without changing the viewing distance, the angle of viewing the screen expected from the subjects would be 1/4 and 1/2 for the 1/4 and 1/2 sizes, respectively. Therefore, to examine whether each subject gazed at the same place in the display and had the same gaze range, we normalized these graphs to the 4K 1/1 size.

The results for a 1/4-size image (Fig. 12) indicated that the subjects viewed the area to the right of the center screen from the left and central viewing positions, and they viewed the screen center from the right viewing position. The results for a 1/2-size image (Fig. 12) revealed that the subjects viewed the area to the right of the center screen from the right and central viewing positions, and they viewed the lower right of the screen from the left viewing position. The results for a 1/1-size image (Fig. 12) demonstrated that the subjects viewed the area to the right of the center screen from the right and central viewing positions, and they viewed the screen center from the left viewing position. The results shown in Figure 12 suggest that the mean gaze position depending on the viewing position tended to change slightly when the size of the screen becomes small. According to Figure 13, the average of the results for all subjects and all viewing position regardless of the screen size shows that the mean gaze position falls in the area to the right of the center screen.

The results of the two-way ANOVA for mean gaze position are illustrated in Figure 14, and those of the two-way ANOVA for the standard deviation of the mean gaze position are shown in Figure 15. No significant differences in the size of the screen or in viewing position were revealed by the two-way ANOVA. Although the mean gaze position tended to change slightly depending on the viewing position when the screen size and viewing position changed, the mean gaze position and the standard deviation of the mean gaze position did not show significant differences.

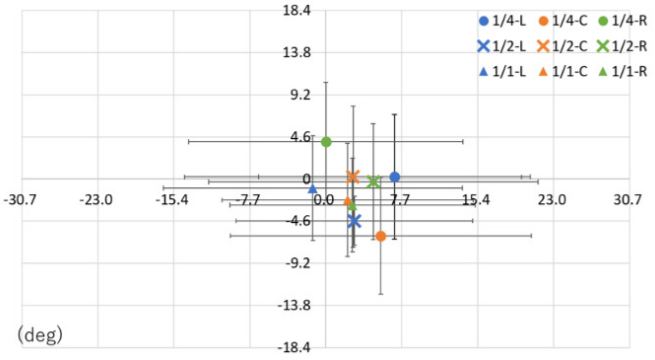


Figure 12: Averages of the mean gaze position and std. dev. (screen size and viewing position).

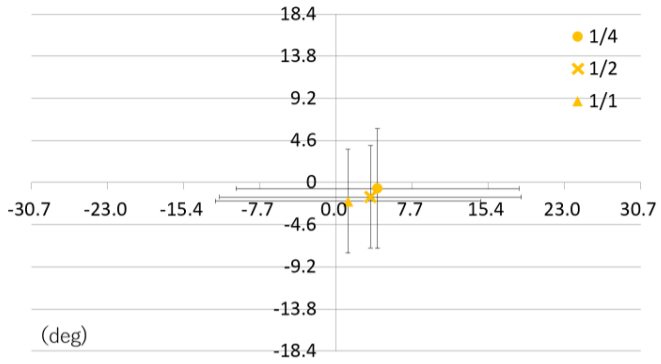


Figure 13: Averages of the mean gaze position and std. dev. (screen size).

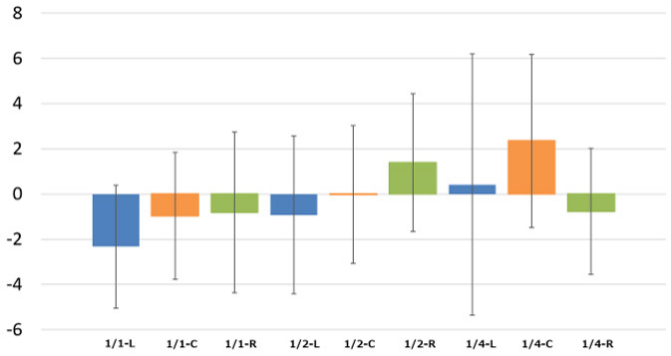


Figure 14: Results of the two-way ANOVA for mean gaze position (screen size).

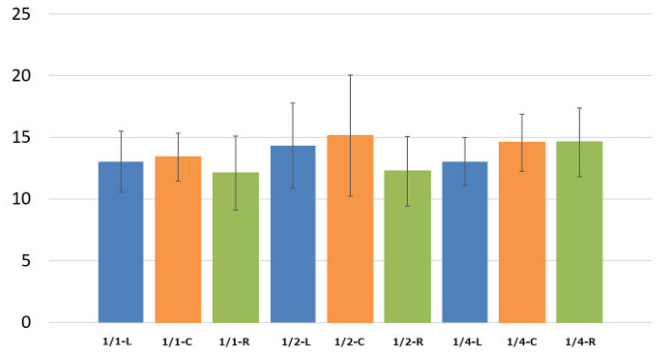


Figure 15: Results of the two-way ANOVA for the std. dev. of the mean gaze position (screen size).

4. Discussion

There were no significant differences in the mean gaze position or the standard deviation of the average gaze position due to differences in resolution, even when the viewing position was changed. Similarly, even when the viewing position was changed, no significant differences were found in the average gaze position or the standard deviation of the mean gaze position due to differences in screen size.

The tendency for the initial gaze position to fall at the center of the image when looking at a still image is called center bias. Photographer bias and viewing strategies have been suggested to be major factors in center bias [7]. Photographer bias is the practice of placing an object or the actor in whom the photographer is most interested near the center of the screen. With respect to viewing

strategies, people usually turn their eyes to the center of the image, having learned through frequent experience with photographer bias that objects that attract attention are generally arranged around the center of the screen. We believe that our present results in which the subjects watched the screen center regardless of differences in resolution and the size of the screen demonstrate a tendency similar to center bias; that is, while center bias is defined as a tendency to look to the center of a still image, our subjects looked to the center of video clips, rather than still images.

5. Conclusion

Broadcasting in 4K and 8K ultra-high-definition video has started, and viewers can enjoy high-definition, wide-angle video near the display. The conditions for subjectively evaluating a display are strictly defined by the ITU-R. Subjective evaluations of displays by broadcasters and manufacturers are performed under these conditions. However, in the home where the display is actually located, the display is viewed at a free position according to the size of the TV, the layout of the room, and structure of the household. For example, when watching a movie in a large-screen movie theater, the viewers experience a different sense of presence and emotion after viewing from the seats at the left and right ends of the front row, the seats at the center of the movie theater, and the seats at the rear. Television broadcasting is also expected to be viewed from various positions in the home by viewing a high-definition, wide-field display. If the impression produced by the image differs depending on the viewing position, it is necessary for the consumers as well as the image producers and the display manufacturers to take this into consideration.

However, the evaluations of displays are performed according to the ITU-R and are rarely performed under the different conditions. Our present experiments show that the mean gaze position during viewing depends on the video clips, the viewing position, and the viewing distance, with the size and resolution of the display fixed and the viewing position as a parameter [17]. This research was conducted with a constant display size and resolution, but in reality, displays with various resolutions and sizes are found in various homes, and it has been necessary to investigate the effects of resolution and size. Therefore, in Experiment 1, we used the resolution and the viewing position as parameters, and in Experiment 2, the screen size and the viewing position were used as parameters, and the mean gaze position and its standard deviation were analyzed. In analyses of the gaze behavior of still images such as photographs, the average gaze position and the variance of gaze points are examined by heat maps and other methods, and there is a photographer bias that places the part that the photographer wants to show to the viewer in the center [18]. It is thus known that there is a center bias in which viewers empirically tend to see the center of the photograph [19].

In addition, regarding evaluations of the viewing of TV images, the mean gaze position and its standard deviation are often used to compare gaze behaviors while subjects watch video clips [4, 17], and we thus applied the mean gaze position and its standard deviation to the analysis of viewing behavior when the resolution, screen size, and viewing position were changed. When 4K-resolution content is viewed from 1.5H, the field of view seen from the eye is 60°. It was reported that the range of visual information that can be captured only by eye movements in daily activities is at most 15° [10]. We therefore, measured the line of sight (head movement + eye movement) with the subjects in a free posture without fixing the head in the device that we developed that can measure head movement and eye movement at the same time [14].

In Experiment 1, the resolution was changed to 1K, 2K, and 4K, and the mean gaze position and its standard deviation were obtained by changing the viewing position using the same-size display. The difference in the mean gaze position was seen by the difference in the viewing position. Notably, the mean gaze position was different between the central viewing position and the left and right viewing positions. There was a slight change in the standard deviation of the mean gaze position due to the difference in the viewing position. However, the changes in the mean gaze position and its standard deviation were slight.

In Experiment 2, the display size was changed to 1/4, 1/2, and 1/1, and the mean gaze position and its standard deviation were obtained by changing the viewing position using the same resolution. The mean gaze position and its standard deviation did not change even when the screen size was changed.

With the introduction of 4K and 8K broadcasting, broadcasting media and video contents have diversified. However, in each home, displays with various resolutions and sizes are being viewed at any viewing position. Our present results demonstrated that when the same content was viewed at different resolutions, screen sizes, and viewing positions, there was almost no change in the mean gaze position, which is one of the viewing behaviors. This means that video producers can engage in program production without needing to being aware of the resolution of the display to be viewed, the screen size, or the viewing position. In addition, consumers can purchase a display according to the size and layout of the room and their preferences, and watch their television from any position.

References

[1] Recommendation BT.2022 https://www.itu.int/rec/R-REC-BT.2022/en
[2] Y. Shishikui, “Super Hi-Vision R\&D,” NHK STRL R\&D, **137**, 4-9, January, 2013.
[3] S. Kubota, A. Shimada, S. Okada, Y. Nakamura, E. Kido, “Television Viewing Conditions at Home,” The Journal of The Institute of Image Information and Television Engineers, **60**(4), 597-603, 2006, doi: 10.3169/mta.7.112.
[4] M. Yamada, T. Fukuda, “Quantitative Evaluation of Eye Movements as Judged by Sight-Line Displacements,” SMPTE Journal, **95**(12), 1230-1241 1986, doi: 10.5594/J04053.
[5] N. Narita, M. Kanazawa, F. Okano, “Optimum Screen Size and Viewing Distance for Viewing Ultra High-Definition and Wide-Screen Images,” The Journal of The Institute of Image Information and Television Engineers, **55**(5), 773-780, May, 2001, doi: 10.3169/itej.55.773.
[6] M. Arbito, “Studies of the Infkuence of Display Size and Picture Brightness on the Preferred Viewing Distance for HDTV Program,” SMPTE J., **103**(8), 517-522, August, 1994, doi: 10.5594/J06461.
[7] A.M. Lund, “The Influence of Video Image Size and Resolution on Viewing-Distance Preferences,” SMPTE J., **120**(5), 406-415, May, 1993, doi: 10.5594/J15915.
[8] S. Kubota, K. Kishimoto, S. Goshi, S. Imai, Y. Igarashi, T. Matsumoto, H. Haga, T. Nakatsue, Y. Mano, Y. Kobayashi, “Preferred Viewing Distance for High Definition Television LCDs,” The Journal of The Institute of Image Information and Television Engineers, **65**(8), 1215-1220, August, 2011, doi: 10.3169/itej.65.1215.
[9] R. Koyama, R. Nakagawa, M. Suganuma, S. Mochiduki, M. Yamada, “The

Effect of Viewing Position by Changing Display Resolution and Display Size," IMQA 2018, 103-109, September, 2018.
[10] A.T. Bahill, D. Adler, L. Stark, "Most naturally occurring human saccades have magnitudes of 15 deg or less," Invest. Ophthalmol., **14**(6), 468-469, 1975.
[11] B. Biguer, C. Prablance, M. Jeannerod, "The contribution of coordinated eye and head movements in hand pointing accuracy," Experimental Brain Research, **55**(3), 462-469, 1984, doi: 10.1007/BF00235277.
[12] M.A. Gresty, "Coordination of head and eye movements to fixate continuous and intermittent targets," Vision Research, **14**(6), 395-403, 1974, doi: 10.1016/0042-6989(74)90238-7.
[13] G.R. Barns, "Vestibulo-ocular function during co-ordinated head and eye movements to acquire visual targets," The Journal of Physiology, **287**, 121-147, 1979, doi: 10.1113/jphysiol.1979.sp012650.
[14] H. Takahira, K. Kikuchi, M. Yamada, "A System for Measuring Gaze Movement and Hand Movement Simultaneously for Hand-Held Devices," The Institute of Electronics, Information and Communication Engineers Trans. Commun., **E98-B**(1), 51-61, January, 2015, doi: 10.1587/transcom.E98.B.51.
[15] "Royalty Free 4K HD Footage (XAVC S format).", http://footage3.opensp c2.org/HDTV/ footage/4K/60p/XAVCS/., May 10, 2018.
[16] R. Fukuda, M. Sakuma, E. Nakamura, T. Fukuda, "An Experimental Consideration on the Definition of a Fixation Point," The Japanese Journal of Ergonomics, **32**(4), 197-204, 1996, doi: 10.5100/jje.32.197.
[17] H. Takahira, S. Mochiduki, M. Yamada, "Analysis of Gaze Movement When Changing Viewing Distance and Viewing Position," The IEICE Transactions on Information and Systems (Japanese Edition), **J99-D**(3), 293-306, March, 2016, doi: 10.14923/transinfj.2015PDP0005.
[18] P. H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, L. Itti, "Quantifying Center Bias of Observers in Free Viewing of Dynamic Natural Xcenes," Journal of Vision, **9**(7-(4)), 1-16, 2009, doi: 10.1167/9.7.4.
[19] D. J. Parkhurst, E. Nieber, "Scene content selected by active vision," Spatial Vision, **16**(2), 125-154, 2003, doi: 10.1163/15685680360511645.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 393-400 (2021)

www.astesj.com

Special issue on Multidisciplinary Innovation in Engineering Science and Technology

ASTESJ

ISSN: 2415-6698

Environmental Entrepreneurship as an Innovation Catalyst for Social Change: A Systematic Review as a Basis for Future Research

Carol Dineo Diale[1,*], Mukondeleli Grace Kanakana-Katumba[2], Rendani Wilson Maladzhi[2]

[1]*Department of Psychology, Industrial Psychology, Rhodes University, Grahamstown, 6139, South Africa*

[2]*Department of Mechanical and Industrial Engineering, University of South Africa, Johannesburg, 0002, South Africa*

ARTICLEINFO

Article history:
Received: 09 November, 2020
Accepted: 13 January, 2021
Online: 22 January, 2021

Keywords:
Environmental entrepreneurship
System dynamics
Innovation

ABSTRACT

There are pressures to adopt sustainable behaviour more so in generating profits and benefiting the society to accelerate green efforts through a green framework. The overarching goal of the paper is premised through various works of literature, building the ecosystem the elements highlighted by most researchers in the field of environmental entrepreneurship. The various models reviewed consists of generic incubators and entrepreneurship, and societal and environmental factors. Environmental entrepreneurship is often used interchangeably with concepts such as green entrepreneurship and ecopreneurship which under-researched globally, with non-existent efforts on the applicability and modelling of key environmental entrepreneurship within a specific context utilising the system dynamics approach. In order to assess the environmental entrepreneurship ecosystem, the authors adopted a system dynamic approach to determine key variables that enable the development of the system. A literature review was conducted, and of the 135 articles reviewed, n=92 peer-reviewed articles met the criteria that the researchers set. Some of the results emanating from a systematic review are environmental policy, green skills, financial and non-financial support, societal and behavioural factors, environmental agility, ethics and governance, and access to markets. The theoretical results are simulated using system dynamics modelling. Due to limited research on the above-mentioned topic, a possible impacting variable (Exogenous variables) was broadened to add value to, and have an impact on, the study. Upon reviewing the above-mentioned models, the framework emerged signalling elements to be simulated in the system dynamics model, which were then theoretically contextualised for the South African context. The theoretical virtual system dynamic model forming part of the framework will be tested and validated in the next study. The applicability of the theoretical ecosystem to South African context as well as future recommendations are provided in the study.

1. Introduction

This paper is an extension of "Green Entrepreneurship Model Utilising the System Dynamics Approach: A Review" presented at the IEEM 2019 Conference in Macau [1]. The conference proceeding argued that Green entrepreneurship as a technology is an area that has not been explored fully, let alone its application the system dynamics spectrum. Furthermore, the paper argued that empirical scholarly research in entrepreneurship within Green industries is at infancy stages in South Africa, and little research work has been done. In [2], the author conducted a study focusing on green entrepreneurs in the South African context. While [3-5] concentrated on the international markets. However, a complete ecosystem surrounding environmental entrepreneurship, and system dynamics remains a gap under consideration. The existing research is centred around a single factor of green entrepreneurship [1], [3-5], with less focus on finding possible multiple solutions to minimise climatic challenges. In this paper, the focus is on environmental entrepreneurship, particularly with the aim of contributing to social norms and finding innovative

*Corresponding Author: Dineo Diale, c.diale@ru.ac.za

approaches to utilising system dynamics for modelling endogenous and exogenous variables of the system. In the current paper, the focus of environmental entrepreneurship is within the spectrum of key drivers of environmental entrepreneurship, social change and behavioural factors. The latter focus is deliberate in creating access to services for the community through empowerment, development, and awareness, and ultimately laying a theoretical foundation on how environmental entrepreneurship can be utilised. These efforts may accelerate the sustainable development goals as environmental entrepreneurs are always expected by societies to be concerned with the environmental factors [6]. Some of the literature defined environmental entrepreneurship by using phrases such as eco-entrepreneurship, ecopreneurship, and ecological entrepreneurship [7-9].

Green entrepreneurship and ecopreneurship is minimally included in the current paper. This is due to the fact that the authors are advocating for the concept of green entrepreneurship, ecopreneurship and environmental entrepreneurship to be defined and dealt with separately so as to have an impact and foster the implementation thereof. It is notable that the latter terms are somewhat similar. However, there is a slight difference between the concepts, whereby, Green entrepreneurship focuses on exploiting markets for profits and saving the planet by leveraging green thematic areas [5]; ecopreneurship has a distinctive feature of ensuring efficiency and effectiveness [10]; and environmental entrepreneurship has a distinctive feature of social, behavioural norms to develop and create awareness [11, 12]. Environmental entrepreneurship can be defined as the concept of radiating adoption of sustainable behaviour and social entrepreneurship, whilst benefiting the society [11].

The current paper focuses on environmental entrepreneurship by contributing of utilising the system dynamics approach to understand the behaviour of the system and its interaction with sub-systems. A further contribution that the current paper hopes to add is to create a theoretical opportunity of the contribution of environmental entrepreneurship to accelerate the sustainable development goals. The current study illustrates how environmental entrepreneurship contributes to sustainable development, social and behavioural factors affecting environmental entrepreneurship in facilitating access to the service, and fostering environmental entrepreneurship as a form of social change. Social dynamics and sustainable development need to be borne in mind when fully implementing green initiatives to support environmental entrepreneurship. Environmental entrepreneurship, as a complex phenomenon, needs to be viewed from different angles to ensure that solutions are gained from multiple sources at the same time. The underlying goal of the paper is premised through various works such as generic business incubators, environmental factors, and societal norms capitalising on social entrepreneurship to build the ecosystem. From the various models the authors discovered key drivers to environmental entrepreneurship, social and behavioural factors as part of the framework. The theoretical ecosystem consists of key drivers, social and behavioural factors, as well as the theoretical framework contextualised in the South African context. The subsequent sections discuss the research questions methodology, and future recommendations.

2. Research questions

The research questions were identified through the research gaps [13] that were identified while reviewing literature and where possible research contributions can be made setting future research agendas. The research questions are as follows:

What are the theoretical key drivers enabling environmental entrepreneurship as a form of social change?

What are theoretical social and behavioural factors affecting environmental entrepreneurship in facilitating access to the service?

How can theoretical key drivers be applied in formulating a South African context framework?

3. Research Background

3.1. Environmental entrepreneurship for social change

Environmental entrepreneurs need to have an agenda of creating and formulating sustainable and green products and services, and at the same time, meet social, economic and environmental objectives [8], [14-16]. It is essential to focus on social dynamics and sustainable development by fully implementing green initiatives. Goal 5, 'gender equality', of the sustainable development goals can be fulfilled by promoting women in environmental entrepreneurship, and Goal 8, 'inclusive and sustainable economic growth, full and productive employment and decent work for all' [17], can be fulfilled by contributing to a better quality of life and serving as a social change catalyst. Studies have shown that women are more likely to venture into environmental entrepreneurship and are more inclined to uplift communities through social entrepreneurship [11, 18].

Research or publication on environmental entrepreneurship as a contribution to sustainable development goals and its interrelatedness or linkages to the system dynamics in the South African context is minimal. Environmental challenges are due to humans not adopting sustainable efforts [19]. The current authors assert that efforts need to be in the form of environmental psychology to minimise challenges due to human activity. Environmental entrepreneurship focuses heavily on the social sphere, transforming communities and the need to adopt agility for the viability of environmental entrepreneurship. In most instances, environmental entrepreneurship forms the foundation of green entrepreneurship by advocating new or innovative ways to save the planet with green products and services [20].

Worldwide, there is minimal research on women as environmental entrepreneurs. In [21], the author holds a strong view that developing women in green cooperatives, especially in rural areas, may alleviate poverty and promote sustainable rural development. This may ultimately contribute to creating green jobs [22]. There has been some effort in tracking environmental entrepreneurship from the early 1990s [23]. Another social factor that has not received much attention is popularising female entrepreneurs who embark on environmental entrepreneurship. There is lack of research on the female identity in relation to their roles as environmental entrepreneurs, which will need to be taken

into perspective and how they optimise or create a new identity in transitioning into a green economy. In [17], the authors shed some light in proving the motive of women embarking in environmental entrepreneurship is to uplift the society.

The current authors are of the opinion that environmental entrepreneurship can assist in accelerating the sustainable development goals, particularly goal 1: zero poverty, goal 5: gender equality, and goal 8: inclusive and sustainable economic growth, full and productive employment and decent work for all.

3.2. The innovative contribution of environmental entrepreneurship utilising system dynamics modelling

The innovative contribution of environmental entrepreneurship as a social change is in the form of system dynamics modelling. The system dynamics approach models the key drivers and factors affecting environmental entrepreneurship as well as social and behavioural factors that create a new perspective of a theoretical entrepreneurship framework. Environmental entrepreneurship, as a complex phenomenon, needs to be viewed from different angles to ensure that solutions are obtained from multiple sources at the same time. System dynamics is used to conceptualise the causal loop relationships or stocks and flows, with the intention of identifying variables that affect the ecosystem feeding into the whole system [24-26]. "System dynamics is the combination of theory, method and philosophy in analysing the behaviour of the system [26]." Furthermore, system dynamics focuses on a feedback loop and nonlinear relationships [25]. System dynamics is viewed as the process or feedback relationship between different parts to understand complexity and behaviour throughout the lifespan, or the process of the system [27]. The need to utilise the complex system to have holistic thinking serves as a research agenda [28]. Therefore, the authors of the current paper further contribute to the system dynamics approach or modelling by looking at environmental entrepreneurship as a complex factor.

4. Methodology

The research methodology followed was in the form of systematic review. The rationale for using a systematic review as a methodology afforded the authors access to the various methodologies and outcomes within the areas of environmental entrepreneurship and various societal norms. This enabled the researchers to formulate a framework with triangulation of efforts from previous researchers. Furthermore, the rationale for systematic review is to explore the area under investigation, simulate theoretically, the key drivers, social and behavioural support factors in quest to create a future research agenda free of bias. The systematic review as a methodology was beneficial, especially during the COVID-19 pandemic, which has affected the entire world, as interviews or questionnaires would not have succeeded during these challenging times. The use of online data collection still poses challenges due to social inequalities and the majority of people not having access to emails or not being able to afford online video platforms in South Africa. The systematic review focused on peer-reviewed articles utilising search engines such as Google Scholar, Sage (Journal of Entrepreneurship Theory and Practice), the Journal of Cleaner Production, Science Direct and the Journal of Sustainability. Grey resources such as newspapers media alerts were excluded when conducting a systematic review. A theoretical review was conducted using a literature review of which n=92 sources (peer-reviewed articles) out of 135 articles, met the criteria that the researchers set for the ecosystem. Duplication of results was removed from the 135 articles. The current study is qualitative in nature and thematic content analysis was used to analyse the results. According to [29], the most preferred and effective way to analyse a qualitative study is to use thematic content analysis. Content thematic analysis attempts to summarise comments into meaningful categories [29]. According to [30] content analysis is a general term to analyse text. It is the systematic coding pillar used for exploring a large amount of textual information to determine trends and patterns [30]. Furthermore, content thematic analysis enables the researcher to condense a broad and clear description of a factor. Themes that have emerged from a review of the literature are environmental entrepreneurship (in terms of key drivers and support), and social and behavioural factors regarding environmental entrepreneurship. The search criteria were environmental entrepreneurship, and system dynamic social factors regarding environmental entrepreneurship, with a refined search from 2010-2020. The software, Vensim, was utilised to simulate the virtual abstract system dynamics model. The title, introduction, abstract and findings sections were used to select the relevant articles and discard other articles. The findings are summarised in Table 1.

5. Findings

The identification of environmental key drivers was affected by conducting an entrepreneurship ecosystem. In order for sustainable entrepreneurship to flourish, awareness of environmental challenges and how they currently affect the nation needs to be set as a research agenda. The literature review and themes such as environmental entrepreneurship (in terms of key drivers and support), and social and behavioural factors regarding environmental entrepreneurship, served as a form of reference. The latter enabled the researchers to manipulate the key variables or drivers for environmental entrepreneurship to be popularised. Environmental entrepreneurship needs to be analysed using small and large-scale criteria and weighed up against the weakness and strengths in order for individuals to learn from and serve, as best practices.

The themes and summary of the afore-mentioned findings are explained in the Table 1. The ecosystem of environmental entrepreneurship in this paper comprises of key drivers in a form of financial and non-financial support, social and behavioural factors as depicted in table 1.

The key drivers such as funding, credit, capital financing, investing in environmental entrepreneurship, and crowdfunding, can be classified under green finance mechanisms in support of environmental entrepreneurship with social orientation [9, 29], [30-33], [35, 36], [44, 45]. Funding of enterprises or ventures is always a challenge, more specifically financing environmental entrepreneurship. Green proofing as advocated by [43] gives rise to a consultancy option, which the current authors deduce that the programme can be utilised to facilitate and manage green change dynamics. Some of the pillars supporting environmental entrepreneurship are infrastructure, skills and education, funding,

(environmental entrepreneurship incubators) products, markets affordability and political will [12, 29, 30, 33, 43]. However, the latter financial key drivers cannot operate in isolation and therefore, may need non-financial support in the form of behavioural change. With key drivers illustrated above, the next theme leads us to behavioural and social factors to strengthen the ecosystem of environmental entrepreneurship

Table 1: Ecosystem for environmental entrepreneurship

Key drivers of environmental entrepreneurship	**Social and behavioural factors**	**Authors' contribution to the environmental entrepreneurship (South African context)**	**Common factors from various models**
Strategic support mechanism [29]	Modelling of human behaviour	Ambassadors or society representatives from various cultural groups, tribal kings, ward councillors and volunteers	Strategic alliance
Environmental policy [12, 30]	Societal support, Alertness, motivation, Altruism [32]	Curriculum review and service-learning	Collaboration, wellbeing, alliance and governance
Tax exemptions and tax breaks, financial mechanism, cost effective products and services [31,32]	Crowdfunding, impact of social norms, development agencies [35-38]	Memorandum of understanding, exchange programmes, and international linkages	Financial and Non-financial support
Access to markets, legal support [12; 30,33,34]	Environmental psychology Environmental consciousness, collective Environmental entrepreneurship and green behaviour, Pro environmental behaviour [39-44]	Green change dynamics programme	Green consumer behaviour and adoption of green behaviour

The behavioural and social factors are premised within environmental agility innovation, transformation, motivation, ecopreneurial management, profiling of environmental entrepreneurship, development agencies crowdfunding, environmental policy, legal framework business incubators commitment and green skills to govern the rights of environmental entrepreneurs [28-30], [34-38], [45-54]. More entrepreneurs from small scale enterprises are environmentally conscious [55]. Literature suggests that in order for desired change to take place, social context plays a huge role, drawing from the work of [56, 57]. Society can play a huge role in praising and supporting environmental innovation, whereas entrepreneurs could ensure that they exploit the markets and ensure they offer affordable products and services to the communities at large. The support from government is needed to contribute to the subsidies and waiving of some costs relating to environmental entrepreneurship, as part of the key drivers [57,58]. Networks, non-governmental stakeholders, financial structural support systems, tax exemptions and tax breaks are some of the key drivers that can support environmental entrepreneurship [29, 59]. The authors are proposing that skills development and human capital need to be the cornerstone of accelerating environmental entrepreneurship through an environmental entrepreneurship talent pipeline. The motives of environmental entrepreneurship have been shown that education is a factor. It has been proved that if individuals have higher levels of education, they tend to embark on environmental entrepreneurship [11].

The system dynamics approach is in the form of modelling the key drivers and factors affecting environmental entrepreneurship as explained above, as well as social and behavioural factors creating a new perspective of entrepreneurship. The modelling of the afore-mentioned results can be depicted using the system dynamics model in Figure 1.

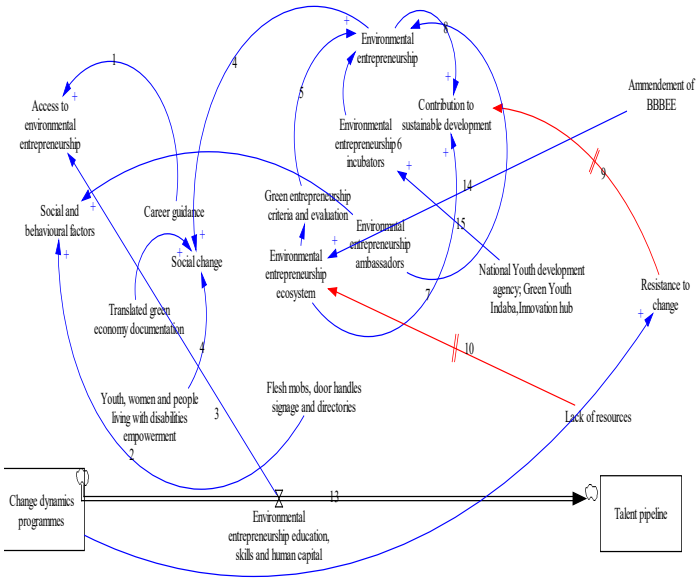


Figure 1: Virtual system dynamics model of environmental entrepreneurship

The system dynamics model (Figure 1) illustrates the importance of the interdependency of multiple variables or factors within the environmental entrepreneurship spectrum. The (+) sign represents a positive relationship and the (–) sign represents a negative relationship. Numbers depict each loop, while delays and threats are represented by a red line with two strokes. The system dynamics model consists of stocks and flows. Loop number 1 represents accessibility services that are influenced by career guidance, and the stock is represented by environmental entrepreneurship education, skills and human capital (loops 3 and 13).

Loop 2 represents an order for social and behavioural factors to be understood and incorporated for social change; and societies can be reached through flesh mobs, door handle signage, directories and ambassadors. Empowering youth, women and people living with disabilities about how to enter the environmental markets, contribute to a social change. Social change then may accelerate environmental entrepreneurship. For

environmental entrepreneurship to gain momentum, the criteria and evaluation system needs to be in place (loop 5).

Furthermore, contribution to environmental entrepreneurship can gain momentum because of the ambassadors, environmental entrepreneurship incubators (loop 6) and centres such as the National Youth Development Agency, Green Youth Indaba, Innovation Hub (loop 15) and emerging environmental entrepreneurship non-profit and profit organisations. Additionally, business incubators are influenced by financial and non-financial mechanisms as exogenous variables.

Environmental entrepreneurship serves as catalyst and contribution to sustainable development positively demarcated by loop 8 and loop 7. However, resistance to change and a lack of resources causes delays and threats to the ecosystem represented by loops 9 and 10. At the same time, there are inputs from the environment, which is BBBEE or legislation represented by a loop 14, which contribute positively to the environmental entrepreneurship ecosystem.

Stocks and flows are indicated by number loop 13. Currently, this research proposes two stocks namely change dynamics and talent pipeline, which are influenced and accelerated by environmental entrepreneurship, education and skills. At the same time, environmental change dynamics can be used to minimise resistance to change.

6. Discussions of environmental entrepreneurship key drivers and social factors

6.1. Key drivers of environmental entrepreneurship

Previous research suggests that creating a clientele, market, and support for the business are important pillars for a start-up to flourish [26]. One important factor that will determine the success or failure of a business is customers and customers' attitudes, which is often overlooked. Creation of the market and funding, or any support for green entrepreneurship, appears to be taboo and an unheard-of practice, therefore, it can be deduced that it needs to start with creating awareness, and unpacking human behaviour as an action. The process needs to be simplified for communities to understand the green economy phenomenon and from there, gradually improve to investigating if there is capability to optimise resources and opportunities for business gain and profit generation and saving the planet for future generations.

There is lack of investors which poses challenges and threats to the system. Investors in environmental entrepreneurship, especially in these sectors, are cautious investors, where scepticism and reluctance arise when it comes to investing to environmental entrepreneurship enterprises. In [33], the author serves to close the gap of financing by introducing crowdfunding. Crowdfunding can be explained as the sourcing of funds through other mechanisms within the lenses of cultural and social norms and creating initiatives online to raise funds and finance. The initiatives to enable transition and a green economy are to ensure ambassadors or champions in different sectors are available. These sectors include the private, public, tertiary, and schooling sectors, and are to be included from the planning phase to the implementation phase. However, society at large or communities are bypassed, and this may add to further challenges and delays in the adoption of new behaviour such as the environmental entrepreneurship. In order to capitalise on collective efficacy in transitioning to a green economy, society, and stakeholders, need to form part of the movement. Collective efficacy can be defined as a group of people coming together to remedy a challenge (as in this instance, the environmental challenges), and have the confidence and capabilities to embark on the movement [41].

The need for curriculum review and inclusion of environmental entrepreneurship in South African basic and higher education needs to be considered to ensure the sustainability of environmental entrepreneurship. In [43], the author echoes the latter idea of curriculum inclusion of environmental entrepreneurship in schools. The green skills documentation and lessons learnt could be drawn from parts of Europe in regulating the economic factor of green skills for the country's economy [59]. The SEED programme is another support mechanism that supports sustainability entrepreneurship [60-62].

6.2. Societal and behavioural factors

There is an action of social responsibilities and environmental policies to be modified concurrently of which the researchers believe that environmental entrepreneurship with the lens of system dynamics could be incorporated onto policies to view the challenge as a system. A study conducted in Bosnia asserts that there is role ambiguity among government stakeholders and educational curricula regarding green entrepreneurship [63]. The latter argument can then be strengthened, and a clear framework provided where relevant stakeholders have stipulated the role and support required to accelerate the agency of adoption of environmental entrepreneurship.

6.3. Threats to the environmental entrepreneurship system

Through the analysis of literature, there have been efforts to investigate green entrepreneurship from international perspectives. Researchers then grouped existing literature to identify barriers to environmental entrepreneurship. Some of the internal barriers were a lack of resources and adequate knowledge, poor implementation of policies and strategies, and attitudes and an organisational culture that discourage environmental management [34]. Furthermore, external barriers that were identified were, difficulty in obtaining certification for environmental practices and systems, economic uncertainty, a lack of legal and institutional frameworks, specialised support and networks among small businesses and the creation of green jobs, and non-compliance with policies. The lens in regard to level of education, support services for incubation, entrepreneurial confidence, business climate, job creation, and clientele may be the solution to the above-mentioned challenges.

6.4. Contextualisation to the South African context

The authors believe that South Africa can implement strong coordination in the form of a memorandum of understanding with the countries that have developed environmental entrepreneurship. This will allow exchange programmes within the SMEs in diverse environmental entrepreneurship, particularly targeting Youth, Black people, women and people living with disabilities.

To broaden this perspective, Broad-based Black Economic Empowerment (BBBEE) needs to also include additional lenses of facilitating environmental entrepreneurship, and points should be allocated in the form of tendering systems in diverse environmental entrepreneurship. To promote gender equality, goal 5 of the sustainable development goals, environmental entrepreneurship can be spearheaded through women empowerment and be integrated into the BBBEE codes to ensure the economic and social standing of women in the South African context. Furthermore, to facilitate the access of services, the scope of career guidance can be expanded to include environmental entrepreneurship as a transformative process. The authors offer recommendations to accelerate access to markets and provide social contexts that are to be embedded in environmental entrepreneurship. In order to further accelerate access to markets, environmental policy needs to specifically stipulate what the rights of environmental entrepreneurs are and how they can access markets [12, 42].

The environmental entrepreneurship pipeline needs to be in place and included in policies, that especially target the youth. The National Youth Development Agency could spearhead funding and include the criteria to promote entrepreneurship in environmentally friendly products, services, consulting or distribution. The National Youth Development Agency could work hand in hand with Green Youth Indaba and an Innovation Hub where several youth programmes are housed, to ensure the sustainability of entrepreneurship in targeting the youth. Environmental change dynamics should be implemented whereby green readiness of change and managing fear of the unknown within Green initiatives could be developed, and awareness and training initiatives with environmental enterprises are instilled into society to contribute to the level of confidence in sustainable energy enterprise sectors. Green skills can be incorporated into the South African Skills Development Act to facilitate an initiative regarding green skills training and development. Change dynamics programmes such as resistance to change and sustainable development goals in a developing country can be introduced. The existing change dynamics programmes can be used as a foundation to design and implement the ecosystem. The South African Skills Development Act could be amended to facilitate the initiative of training and development, and the keeping of training and development records to aid sustainable development goals, and green skills for monitoring of grants and incentives.

7. Future recommendation

Action is needed with regards to social responsibilities. Further action is for environmental policies to incorporate environmental entrepreneurship. The authors further contribute that in order for environmental challenges to be minimised and environmental entrepreneurship to flourish, human behaviour needs to be modelled through environmental Psychology in terms of social psychology, behaviour, motivation, cultural and gender factors, personality traits, and attitudes towards a green economy and environmental entrepreneurship. To further aid environmental entrepreneurship as a social change, the development and focus of research and support needs to be given to women.

8. Conclusions

The current paper explored various works of literature such as key drivers from business incubators, social and behavioural factors, environmental entrepreneurship, and building the theoretical ecosystem of the elements highlighted by most researchers in the field of environmental entrepreneurship. Environmental entrepreneurship, as a complex phenomenon, needs to be viewed from different angles to ensure that solutions are gained from multiple sources at the same time.

The discovery made during the review of literature is that an ecosystem of environmental entrepreneurship can be broadly clustered within key drivers (support and accelerators), and social and behavioural factors that are contextualised in the South African context. The latter broad factors can further be explained through the accessibility of services that are influenced by career guidance, environmental entrepreneurship education skills and human capital. In order for social and behavioural factors to be understood and incorporated for social change, societies can be reached through flesh mobs, door handle signage, directories and ambassadors.

The empowerment of youth, women and people living with disabilities about how to enter the sphere of environmental entrepreneurship needs to be the cornerstone and serve as an accelerator to the system. Therefore, environmental entrepreneurship serves as catalyst and contributes to sustainable development, theoretically. As a way forward, the above hypothesis needs to be tested and validated to build a systems dynamic model taking foundation from the systematic review conducted by the current authors. Furthermore, to try and bridge the gap, more interdisciplinary qualifications need to be encouraged and a call for industrial psychologists to collaborate with experts in the fields of industrial engineering, environmental science and economics, about meeting sustainable development goals, should be undertaken.

Conflict of Interest

The authors declare no conflict of interest

Acknowledgment

The authors would like to thank the anonymous reviewers of the IEEE IEEM conference proceedings, which ultimately led to the extension of conference proceedings and thereby, the journal publication. The authors would also like to thank the reviewers and organisers of the first annual French-South Africa innovation days held in December 2019, which also contributed to the publication of the current paper.

References

[1] C.D. Diale, G. Kanakana-Katumba & R.W. Maladzhi. Green Entrepreneurship Model Utilising the System Dynamics Approach: A Review. In 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 384-389. (IEEE), 2019, December. doi.org/10.1109/IEEM44572.2019.8978804.

[2] C. Mukonza. Analysis of Factors Influencing Green Entrepreneurship in South Africa, 2016.

[3] F. Farinelli, M. Bottini, S. Akkoyunlu, P. Aerni & P. Green entrepreneurship: the missing link towards a greener economy. Atdf Journal, **8**(3/4), 42-48, 2011.

[4] M. Schaper. Understanding the green entrepreneur. In Making Ecopreneurs, 27-40. Routledge, 2016.

[5] P.N. Vaidya & D.V. Honagannavar. Green entrepreneurship towards sustainable environment. International Journal of commerce and management **3**(1), 88-91, 2017.

[6] J.C. Allen & S. Malin. Green entrepreneurship: a method for managing natural resources? Society and natural resources, **21**(9), 828-844, 2008. doi.org/10.1080/08941920701612917.

[7] A. Kuckertz & M. Wagner. The influence of sustainability orientation on entrepreneurial intentions – Investigating the role of business experience. Journal of Business Venturing, **25**(5), 524-539, 2010. doi.org/10.1016/j.jbusvent.2009.09.001.

[8] M. Lenox & J.G. York. Environmental entrepreneurship. The Oxford handbook of business and natural environment, 70-92, 2011.doi.org/10.1093/oxfordhb/9780199584451.003.0004.

[9] E. Mieszajkina. Ecological entrepreneurship and sustainable development. Problemy Ekorozwoju–Problems of Sustainable Development, **12** (1), 163-171. 2016.

[10] L. Walley, D. Taylor & K. Greig. Beyond the visionary champion: Testing a typology of green entrepreneurs. Making ecopreneurs: developing sustainable entrepreneurship, **2**, 59-74, 2010.

[11] J. Hörisch, J. Kollat & S.A. Brieger. What influences environmental entrepreneurship? A multilevel analysis of the determinants of entrepreneurs' environmental orientation. Small Business Economics, **48**(1), 47-69, 2017.

[12] L.E. Huggins. Environmental entrepreneurship: markets meet the environment in unexpected places. Edward Elgar Publishing, 2013.

[13] J.W. Creswell & J.D. Creswell. Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications, 2017.

[14] J. Kirkwood & S. Walton. What motivates ecopreneurs to start businesses? International Journal of Entrepreneurial Behavior & Research, **16**(3), 204-228, 2010. doi.org/10.1108/13552551011042799.

[15] T. McEwen. Ecopreneurship as a solution to environmental problems: implications for college level entrepreneurship education. International Journal of Academic Research in Business and Social Sciences, **3**(5) 264, 2013.

[16] D. Vasilevska & B. Rivza. Green Entrepreneurship as a Factor of Sustainable Economic Development in Baltic States. International Multidisciplinary Scientific GeoConference: SGEM, **18**(5.3), 423-430, 2018. doi.org/10.5593/sgem2018/5.3/S28.054.

[17] A.R. Nunes, K. Lee & T. O'Riordan. The importance of an integrating framework for achieving the Sustainable Development Goals: the example of health and well-being. BMJ global health, **1**(3) e000068, 2016. doi.org/10.1136/bmjgh-2016-000068.

[18] D.M. Hechavarria, A. Ingram., R. Justo & S. Terjesen. Are women more likely to pursue social and environmental entrepreneurship? In Global Women's Entrepreneurship Research. Edward Elgar Publishing, 2012. doi.org/10.4337/9781849804752.00016.

[19] G. Nhamo. Green economy readiness in South Africa: A focus on the national sphere of government. International Journal of African Renaissance Studies-Multi-, Inter-and Transdisciplinarity, **8**(1), 115-142, 2013. doi.org/10.1080/18186874.2013.834628.

[20] V. Harini. & D.T. Meenakshi. Green Entrepreneurship Alternative (Business) Solution to Save Environment. Asia Pacific Journal of Management & Entrepreneurship Research, **1** (3), 79, 2012.

[21] S.E. Sanyang & W.C. Huang. Green cooperatives: A strategic approach developing women's entrepreneurship in the Asian and Pacific region. World Journal of Agricultural Sciences, **4**(6), 674-683, 2008.

[22] M. Drăgoi, I.E. Iamandi, S. Munteanu, R. Ciobanu & R. Lădaru. Incentives for developing resilient agritourism entrepreneurship in rural communities in Romania in a European context. Sustainability, **9**(12), 2205, 2017. doi.org/10.3390/su9122205.

[23] D. Holt. Where are they now? Tracking the longitudinal evolution of environmental businesses from the 1990s. Business Strategy and the Environment, **20**(4), 238-250, 2011. doi.org/10.1002/bse.697.

[24] D.H. Meadows. Thinking in systems: A primer. Chelsea green publishing, 2008.

[25] J.W. Forrester. Some basic concepts in system dynamics. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 9, 2009.

[26] R. Teplov, J. Väätänen & D. Podmetina. A System Dynamic Approach to Modelling the Endogenous and Exogenous Determinants of the Entrepreneurship Process. Journal of Innovation Management, **4**(2), 68-95, 2016. doi.org/10.24840/2183-0606_004.002_0005

[27] N. Abdelkafi & K. Täuscher. Business models for sustainability from a system dynamics perspective. Organization & Environment, **29**(1), 74-96, 2016. doi.org/10.1177/1086026615592930.

[28] K. Schaefer, P.D. Corner & K. Kearins. Social, environmental and sustainable entrepreneurship research: what is needed for sustainability-as-flourishing? Organization & environment, **28(4),** 394-413, 2015. doi.org/10.1177/1086026615621111.

[29] D. Waddell, A. Creed, T.G. Cummings & C.G. Worley. Organisational change: Development and transformation. Cengage AU, 2019

[30] M. Vaismoradi, H.Turunen & T. Bondas. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing & health sciences, **15**(3), 398-405, 2013. doi.org/10.1111/nhs.12048.

[31] P.B. Zamfir. (2014). Supporting Green Entrepreneurship in Romania: Imperative of Sustainable Development. Romanian Econ. Bus. Rev, **9**, 35-44, 2014.

[32] N. Thompson, K. Kiefer & J.G. York. Distinctions not dichotomies: Exploring social, sustainable, and environmental entrepreneurship. In Social and sustainable entrepreneurship. Emerald Group Publishing Limited, 2011. doi.org/10.1108/S1074-7540(2011)0000013012.

[33] A.V. Barkov & Y.S. Grishina. Methodological approaches to theoretical construction of the model of legal support for environmental entrepreneurship. Civil Law 4, 3-7, 2018. doi.org/10.33397/2619-0559-2019-1-1-115-134.

[34] O.A. Yakovleva. Environmental entrepreneurship (legal aspect). Legal Concept= Pravovaya Paradigma, **17**(2), 2018.

[35] J. Hörisch. Crowdfunding for environmental ventures: an empirical analysis of the influence of environmental orientation on the success of crowdfunding initiatives. Journal of cleaner production, 107, 636-645, 2015. doi.org/10.1016/j.jclepro.2015.05.046.

[36] F. Massa-Saluzzo & L. Toschi. The Impact of Local Social Norms on Access to Finance: The Case of Environmental Entrepreneurship. In Academy of Management Proceedings 2019, (1), 14820. Briarcliff Manor, NY 10510: Academy of Management, 2019, July. doi.org/10.5465/AMBPP.2019.38.

[37] E.I. Nikolaou, D. Ierapetritis & K.P. Tsagarakis. An evaluation of the prospects of green entrepreneurship development using a SWOT analysis. International Journal of Sustainable Development & World Ecology, **18**(1), 1-16, 2011. doi.org/10.1080/13504509.2011.543565.

[38] C. Rodgers. Sustainable entrepreneurship in SMEs: a case study analysis. Corporate Social Responsibility and Environmental Management, **17**(3), 125-132, 2010. doi.org/10.1002/csr.223.

[39] S. Schaltegger & M. Wagner. Sustainable entrepreneurship and sustainability innovation: categories and interactions. Business strategy and the environment, **20**(4), 222-237, 2011. doi.org/10.1002/bse.682.

[40] S. Schaltegger. A framework and typology of ecopreneurship: Leading bioneers and environmental managers to ecopreneurship. In Making Ecopreneurs pp. 95-114. Routledge, 2016.

[41] S. Cheema, B. Afsar, B.M. Al-Ghazali & A. Maqsoom. How employee's perceived corporate social responsibility affects employee's pro-environmental behaviour? The influence of organizational identification, corporate entrepreneurship, and environmental consciousness. Corporate Social Responsibility and Environmental Management, **27**(2), 616-629, 2020. doi.org/10.1002/csr.1826.

[42] J.P. Doh, P. Tashman & M.H. Benischke. (2019). Adapting to grand environmental challenges through collective entrepreneurship. Academy of management perspectives, **33**(4), 450-468, 2019. doi.org/10.5465/amp.2017.0056.

[43] I. Fritsche, E. Jonas & T. Kessler. Collective reactions to threat: Implications for intergroup conflict and for solving societal crises. Social Issues and Policy Review, **5**(1), 101-136, 2011. doi.org/10.1111/j.1751-2409.2011.01027.x.

[44] W.R. Meek, D.F. Pacheco & J.G. York. The impact of social norms on entrepreneurial action: Evidence from the environmental entrepreneurship context. Journal of Business Venturing, **25**(5), 493-509, 2010. doi.org/10.1016/j.jbusvent.2009.09.007.

[45] K. Foster, J. Jelen & A. Scott. (2010). Teaching Environmental Entrepreneurship at an Urban University: Greenproofing. Metropolitan Universities, **21**(1), 73-87, 2010.

[46] R. Ranjan. Deriving double dividends through linking payments for ecosystem services to environmental entrepreneurship: The case of the invasive weed Lantana camara. Ecological Economics, **164**, 106380, 2019. doi.org/10.1016/j.ecolecon.2019.106380.

[47] R. Antolin-Lopez, J. Martinez-del-Rio & J. J. Cespedes-Lorente. Environmental entrepreneurship as a multi-component and dynamic construct: Duality of goals, environmental agency, and environmental value creation. Business Ethics: A European Review, **28**(4), 407-422, 2019. doi.org/10.1111/beer.12229.

[48] A. Domańska, B. Żukowska & R. Zajkowski. Green entrepreneurship as a connector among social, environmental and economic pillars of sustainable

development. Why some countries are more agile? Problemy Ekorozwoju, **13**(2), 2018.
[49] A. de Bruin & K. Lewis. Little acorns in action: green entrepreneurship and New Zealand micro-enterprises. Making ecopreneurs: Developing sustainable entrepreneurship, 95-107, 2010.
[50] J. L. Thompson & J. Scott. Environmental entrepreneurship: The sustainability challenge, 2010.
[51] J.K. Hall, G.A. Daneke & M.J. Lenox. Sustainable development and entrepreneurship: Past contributions and future directions. Journal of business venturing, **25**(5), 439-448, 2010. doi.org/10.1016/j.jbusvent.2010.01.002.
[52] J.M. Scott. & J. Thompson. Making ecopreneurs: developing sustainable entrepreneurship. International Journal of Entrepreneurial Behavior & Research, 2012. doi.org/10.1111/joms.12214.
[53] G.D. Markman, M. Russo, G.T. Lumpkin, P.D. Jennings & J. Mair. Entrepreneurship as a platform for pursuing multiple goals: A special issue on sustainability, ethics, and entrepreneurship. Journal of Management Studies, **53**(5), 673-694, 2016.
[54] J. Thompson. Incredible Edible–social and environmental entrepreneurship in the era of the "Big Society". Social Enterprise Journal, 2012.
[55] M. Schaper. Understanding the green. Making ecopreneurs: Developing sustainable entrepreneurship, **7**, 2010.
[56] K. Hockerts & R. Wüstenhagen. Greening Goliaths versus emerging Davids—Theorizing about the role of incumbents and new entrants in sustainable entrepreneurship. Journal of business venturing, **25**(5), 481-492, 2010. doi.org/10.1016/j.jbusvent.2009.07.005.
[57] J.A. Schumpeter. "Economic theory and entrepreneurial history", in Clemence, R.V. (Ed.), Essays on Entrepreneurship, Innovations, Business Cycles, and the Evolution of Capitalism, 7th ed., Transaction Publishers, New Brunsick, NJ, 254-72, 2004. doi.org/10.4324/9781351311489-21.
[58] S. Choudhary. & N. Patil. Green entrepreneurship: role of entrepreneurs in energy economics in Nepal. Annual Research Journal of Symbiosis Centre for Management Studies, **3**(1), 166-175, 2015.
[59] M. Loock. Going beyond best technology and lowest price: on renewable energy investors' preference for service-driven business models. Energy Policy **40**, 21–27, 2012. doi.org/10.1016/j.enpol.2010.06.059.
[60] Y.D. Uslu, Y. Hancıoğlu & E. Demir. Applicability to green entrepreneurship in Turkey: A situation analysis. Procedia-Social and Behavioral Sciences, 195, 1238-1245, 2015. doi.org/10.1016/j.sbspro.2015.06.266.
[61] OECD/Cedefop, Greener Skills and Jobs, OECD Green Growth Studies, OECD Publishing. doi.org/10.1787/9789264208704-en, (2014).
[62] H. Creech, L. Paas, G.H. Gabriel, V. Voora, C. Hybsier & H. Marquard. Small-scale social-environmental enterprises in the green economy: supporting grassroots innovation. Development in Practice, **24**(3), 366-378, 2014. doi.org/10.1080/09614524.2014.899561.
[63] I. Silajdžić, S.M. Kurtagić & B. Vučijak. Green entrepreneurship in transition economies: a case study of Bosnia and Herzegovina. Journal of Cleaner Production, **88**, 376-384, 2015. doi.org/10.1016/j.jclepro.2014.07.004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 402-410 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Enhancing Decision Making Capabilities in Humanitarian Logistics by Integrating Serious Gaming and Computer Modelling

Za'aba Bin Abdul Rahim[1,*], Giuseppe Timperio[1], Robert de Souza[1], Linda William[2]

[1]*The Logistics Institute Asia Pacific, National University of Singapore, 21 Heng Mui Keng Terrace #04-01, Singapore 119613*

[2]*School of Informatics & IT, Temasek Polytechnic, 21 Tampines Avenue 1, Singapore 529757*

ARTICLE INFO

Article history:
Received: 30 January, 2020
Accepted: 25 May, 2020
Online: 11 June, 2020

Keywords:
Supply Chain Management
Humanitarian Logistics
Serious Gaming
Computer Modelling
Enhanced Decision Making

ABSTRACT

The field of humanitarian logistics has in recent times gained an increasing attention from both academics and practitioners communities alike. Although various research groups have addressed theoretical and technical developments in humanitarian logistics using conventional research tools, applied research appears to be often dependent on practitioners' inputs. This paper is an attempt to fill the existing gaps between academic research and practitioners' needs and proposes an integrated framework that consists of serious games and computer modelling. The serious games component aims to raise awareness on humanitarian logistics issues as well as provide a platform to facilitate the acquisition of inputs from humanitarian practitioners. Based on these inputs, a computer model will be developed. To test the framework, a real-life case study about the prepositioning of strategic stockpiles in Indonesia, one of the countries with the highest disaster risk exposure on a global scale, was used. Findings of this work highlight the role of serious games as risk-free environments for players to design strategies enhancing disaster preparedness in conjunction with broadly used research methodologies such as computer modelling.

1. Introduction

This paper is an extension of work originally presented in 2018 Winter Simulation Conference (WSC), Gothenburg, Sweden by de Souza, et al. by [1]

ReliefWeb, a specialized digital service of the UN Office for the Coordination of Humanitarian Affairs (OCHA), reported that between the years 1994 and 2013, more than 6,500 natural calamities occurred globally. On average, 218 million lives were affected, and 1.35 million deaths were recorded annually [2]. In the year 2017 alone, for instance, more than 300 catastrophes occurred, of which more than half were natural [3]. These disasters claimed more than 11,000 people's lives, and it's estimated that they have caused an economic loss of over 300 billion US dollars. It is almost two times as many as the USD 180 billion economic losses estimated for the precedent year, and well above the average USD 190 billion monetary losses for the period 2005-2015 [3]. In addition, the combination of climate change, political and social crises, and progressive urbanization of disaster-prone areas seem to push even further the frequency and scale of humanitarian crises globally.

Therefore, in a context of increased number and scale of humanitarian crises in conjunction with reduced resources, there is a need for greater efficiency in the deployment of relief assets. This is particularly relevant for logistics (purchasing, transportation, and storage), that is by far the biggest contributor to humanitarian expenditures (60-80%) [4].

In humanitarian context, logistics can be defined as the activity of "planning, implementing and controlling the efficient, cost-effective flow of and storage of goods and materials as well as related information, from point of origin to point of consumption for the purpose of alleviating the suffering of vulnerable people" [5]. Compared to the commercial context, humanitarian logisticians are required to be cost-effective, agile, and time responsive.

Generally, disaster relief operations are run in a highly chaotic environment where the needs of affected population are in a fast and continuous development, and resources such as

*Za'aba Bin Abdul Rahim, Email: zaaba@nus.edu.sg

supplies, people, funds, and technologies are scarce [6] [7]. Hence, the logistics processes and disaster relief chains need to be streamlined to speed up the response time for the pressing need of critical supplies. Given these requirements, humanitarian supply chains require to be designed proactively, well ahead of the time when disasters occurr [8] [9].

To address supply chain issues, computer models are conventionally used. The main benefit of computer models is that they allow to replicate (to a certain extend) real-life operations in virtual environments. In turns, this allows to identify possible solutions, to test their feasibility in a virtual environment, and to mitigate the risk of making costly mistakes prior to the implementation in a real-world setting [10]. Amongst others, computer simulation appears to be the most popular decision support tool for supply chain design decisions as well as for the assessment of supply chain policies. A number of simulation techniques were used by other scholars to address research issues in the humanitarian logistics space including System Dynamics (SD) [11] [12], Discrete Event Simulation (DES) [13] [14] and Agent Based Simulation (ABS) [15] [16].

However, practical experience suggests that while the aforementioned research techniques are indeed relevant, practitioners' inputs are also a key component for research inquiries related with humanitarian logistics. In fact, while methodologies such as network optimization or simulation can be used to generalize and extend the humanitarian logistics preparedness through scenario-based analysis, it appears that to date there is no clear method facilitating the acquisition of inputs from practitioners.

In another front, there has also been an increase in the use of serious games to increase awareness and provide an avenue for wider audience to learn and experience humanitarian logistics, as well as to facilitate the acquisition of experiential inputs from practitioners. Serious games have been around for many years and have recently gained increased popularity [17]. Playing games increases interactivity among players and provide time-flexibility to pace their learning in their own time [18]. Serious games in humanitarian logistics provide a safe environment for the players to plan, implement and evaluate their humanitarian logistics strategies without the dangers of real-world consequences. It provides different sets of scenarios or environments compared to computer simulation models [19].

In this paper, we focus on developing a decision support framework integrating the use of serious games and computer models to address complex supply chain issues in the domain of disaster relief. We apply a role-playing simulation-based board game, titled THINKLog: Humanitarian Logistics Gameplay to gather inputs from practitioners. THINKLog is an interactive board game designed that can be scaled to include different scenarios for logistics and supply chain management including for humanitarian logistics. THINKLog is a learning framework where different scenarios can be generated by combining different concepts and challenges in logistics. One of the scenarios that is available is the Humanitarian Logistics Gameplay. This gameplay helps to introduce the importance and complexity of supply chain management and logistics in humanitarian context and deepen the understanding of humanitarian logistics for the players. The gameplay focuses on the warehouse location (Preparedness), inventory management and deliveries (Response).

This game has been played both locally (Singapore) and regionally (Southeast Asia). In this paper, the results from one of the game workshops are presented as an evidence that the game provided an interactive method to apply and experiment with supply chain concepts. The workshop was conducted with a official operating in disaster preparedness and response from multiple ASEAN countries at the ASEAN Coordinating Centre for Humanitarian Assistance on Disaster Management (AHA) Centre Executive (ACE) Programme in November 2018 in Jakarta, Indonesia.

Using the gathered inputs from the practitioners during and after the game session, a computer model is developed. As a case study, we applied it to a real-life case study about prepositioning of life-saving kits in Indonesia. Subsequently, a simulation model using AnyLogistix software was developed.

The rest of the manuscript is organized as follows. Section 2 debates the academic literature previously developed in this domain and focuses its attention on simulation-based serious games applied to supply chain and logistics and computer models in the niche of stockpile prepositioning. Section 3 describes the problem statement and methodology used in this paper. Section 4 describes the THINKLog – Humanitarian Logistics Gameplay game, a real-life simulation model and key findings of both. Section 5 summarizes the key findings of this work alongside its limitations and future extensions.

2. Literature Review

In this section, we review relevant academic literature in the areas of simulation-based serious games and computer simulation applied to the domain of humanitarian logistics and stockpile prepositioning.

2.1. Simulation-Based Serious Gaming

Recently, there has been an upward trend towards the use of serious games in creating awareness in specific areas and domains. They are widely used in areas like education [20], military [21], healthcare [22] and city planning [23]. Games help players grasp the learning objectives by providing hands-on engaging and motivating experiences. This manuscript will emphasize humanitarian simulation-based games.

Simulation-based games embed the process of hypothesizing, probing, and reflecting upon the simulated world within the game to promote learning and create awareness [24]. By using specific storylines in the game, players try to solve a particular task or challenge to learn specific concepts. Players are able to perform repeated experimentations and understand the consequences of their decisions, thus promoting players' incentive and learning transfer to the real-life context. [19]. Some other studies have even shown that these kind of games advanced players' awareness and understanding of scientific knowledge [19] [25].

There is a fair number of supply chain simulation-based games available in the market that are used as a learning tool to teach certain topics of supply chain management. In this paper, we looked into simulation-based games with humanitarian

storylines. Two of such games are AFTERSHOCK [26] and STOP DISASTERS! [27]. AFTERSHOCK is a board game where players assumes roles of different agencies involved in the response of an emergency and its early phase of the recovery. STOP DISASTERS! is an online game whereby the player is needs to prepare for the impending disaster by improving the infrastructure in order to save the lives.

These games do offer a certain level of experiential experience or learning to the players. However, they do not provide an avenue to elicit current issues, especially faced by humanitarian logistics domain experts, practitioners and decision makers. There may be valuable information/data that may be captured, analyse and derived from the games that are not necessarily captured.

Table 1 shows the summary and comparison of the two aforementioned games against our proposed one namely THINKLog.

Table 1. Games Comparison

	STOP DISASTERS!	AFTERSHOCK	THINKLog (humanitarian logistics version)
Topic	Pre-disaster buildings capabilities and defences	Interagency cooperation in response to emergency and early recovery phase of a humanitarian crisis	Logistics and facility location identification in humanitarian context
Type of game	Simulation	Simulation	Simulation
Objective	To build the defence of an existing structures for the population in face of an upcoming disaster	To address the urgent humanitarian needs of the local population.	To build a warehouse based in the concept of Multi Criteria Decision Making (MCDM) to deliver an uninterrupted and continuous supply of the required life-saving kits to the disaster area
Type of tasks	Operation activities/tasks	Operation activities/tasks	Strategic planning (in preparation stage) and operation activities/tasks (in response stage)
Playing mode	Computerized	Board game	Board game with digital companion app

2.2. Computer Modelling in Humanitarian Logistics

Logistics system design with focus on storage and transportation systems have received significant attention in the domain of commercial supply chains [28], [29]. Nevertheless, the importance of humanitarian logistics applications has been recognized only in recent times. One of the areas receiving the greatest attention in the space of humanitarian logistics is stockpile prepositioning and network design.

Balcik & Beamon (2008) are considered pioneers in this field. Their study analyses the facility location problem in consideration of constraints such as budget and capacity of facilities. The overarching goal of their work is to optimize the flow of relief supplies all the way to disaster affected zones. McCoy & Brandeau [30] looked at storage and shipping policies and the main outcomes of their work include effective distribution strategies and trade-off between stock size and service level. Roh, et al. [31] analyzed the supply prepositioning location problem from both macro and micro perspectives using inputs from high-level decision-makers and operational experts.

Simulation literature studies humanitarian logistics problems using different methods including system dynamics, discrete-event simulation, and agent-based modeling.

System dynamics (SD) is a modelling technique often used to evaluate the nonlinear behavior of complex systems, Basic elements of SD modelling are known as stocks, flows, internal feedback loops, table functions and time delays [32]. One of the first applications of SD in the humanitarian logistics space was by Besiou et al. [11]. Their work applies SD to the issue of fleet management. Costa et al. [12] instead used SD with the aim of addressing the challenge of coordination in disaster response.

Discrete-event simulation (DES) is a modelling technique used to sequence discrete events over time. Once a particular event occurs, the state of the system will change. No changes will occur between two consecutive events [33]. Iakovou, et al. [13] developed a DES model to assess the impact of emergency sourcing on supply chain performances. Noreña, et al. [14] used DES to understand the robustness of the logistics of medical supply in Bogotá (Colombia).

Agent-based modeling (ABM) refers to a series of computer simulation techniques whereby the actions of autonomous pre-defined agents are simulated [15]. Horner & Widener [16] used an integrated approach ABM and GIS to foresee the damage caused by heavy storms on the transportation network in use case of Florida City.

2.3. Research Gaps

Despite the significant body of literature available in the domain of stockpile prepositioning, to our knowledge it appears that most research groups have tackled humanitarian logistics issues, and facility location problems in particular, from a purely conceptual standpoint with limited understanding of practitioners' viewpoints. Although these approaches provide fine solutions to this class of problems, it is also evident that there is a need for supplementing them with practitioner perspectives. It is intended that, decision-makers will be able to gain insights on the following matter:

- How to leverage on serious games to acquire inputs from decision makers and practitioners?
- How to integrate serious-games and computer models to assure the robustness of the computer model and its relevance to the practice?

2.4. The Contribution of The Current Research

This paper demonstrates the applicability of the proposed framework for network design in the use case of Indonesia. The work presented in this manuscript will apply THINKLog and simulation to demonstrate the relevance of the proposed solution approach. The contribution of our study is on the integration of serious games and computer modelling with the aim of generating insights into the domain of network design using practitioner's inputs.

3. Problem Statement and Methodology

Indonesia is the 14th largest nation by size, spans across three time zones and counts over 260 million people living across more than seventeen thousand islands [9]. The country sits on the edges of four tectonic plates and is part of the so called "Ring of Fire". According to Statista [34], between 1900 and 2016, 113 earthquakes were recorded in the Indonesian archipelago, which claimed nearly 200,000 human lives. According to the Centre for Research and Epidemiology of Disasters, Indonesia ranks in the 'top 5 countries most frequently hit by natural disasters' after China, India, the Philippines, and the United States [35]

3.1. Problem Statement

When disasters hit in remote areas of the archipelago, existing response capacities are significantly stretched. Despite the efforts of the national authorities to boost inter-island and intra-island connectivity through infrastructure development [36], the existence of ad-hoc logistics hubs to store those critical inventories to be deployed in times of emergency, would be highly beneficial for the effective provision of humanitarian assistance. Hence, the case study will focus on the pre-positioning of relief supplies at strategic locations across Indonesia to improve country's logistics capabilities to respond to natural disasters.

3.2. Methodology

Because the problem statement requires the embedding of inputs from practitioners, decision makers, and humanitarian logistics domain experts, the integration of serious games for the collection of inputs, and computer simulation as an analytical closed form of analysis is proposed.

In particular, to address the identified problem statement, the following multi-phase framework encompassing two main steps was conceptualized and tested:

(1) Phase 1. Socializing the problem at hand with practitioners, decision makers, and humanitarian logistics domain experts, and gathering of inputs using serious games.

(2) Phase 2. Building the computer model and running scenario-based analysis for the identification of the solution to the problem at hand.

Figure 1 is an outline of the proposed framework which was tested in the use case of stockpile prepositioning in the use case of Indonesia using THINKLog game in integration and computer simulation in AnyLogistix.

4. THINKLOG Game

4.1. THINKLOG Game, An Overview

ThinkLog is an expandable interactive logistics and supply chain management board game that was designed specifically to help players in learning about SCM concepts. The game aims to complement teaching and learning activities in classrooms and workshops by introducing SCM concepts through roleplaying and simulation. The initial design of the game was for a pure board game play: main board, demand cards, gameplay/rules and game master [37]. This game also includes a digital companion application [38]. With the inclusion of this app, more SCM concepts and different scenarios of game play can be added, such as the Humanitarian Logistics game play.

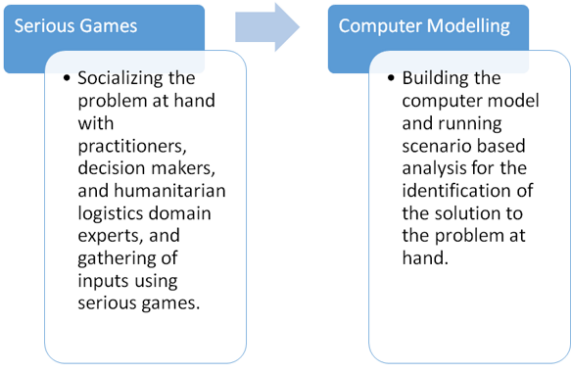


Figure 1. Proposed Framework



Figure 2: THINK:Log game components

4.2. Humanitarian Logistics Gameplay

The main learning objectives for this gameplay are to introduce the notions of Multi Criteria Decision Making (MCDM), and particularly its applications in warehouse location selection problems for disaster preparedness, while providing a high level

view on the challenges related with information flow and coordination in logistics and supply chain management.

MCDM, that is a structured framework able to analyze decisions problems involving complex objectives concepts, is adapted in this gameplay to help the identification of possible locations for the warehouses [9], [39], [40].

There are two stages in this gameplay - Preparation and Response. In the Preparation, the players input their weightage to the criteria defined in the MCDM framework directly through the companion app. Figure 3 shows the input screen for the players. There are 5 criteria for consideration, including Distance, Congestion, Cost, Coverage and Risk. Figure 4 shows one of the criteria and its definition. The companion app will then generate recommended locations based on the weightage as input by the player. In the second stage namely Response, players will have to deliver the requested relief items to the affected population while taking into consideration the possible random events that may occur and disrupt their distribution operations. Examples of events include congestions and /or further calamities that may impede their logistics plan. Figure 5 shows the basic game flow of this game play.

Players assume the role of a Humanitarian Agency Officer in charge of designing and coordinating an uninterrupted and continuous flow of relief items to disaster affected areas. The player has to strategize the most optimal way to deliver such relief items. In the event where the player fails to deliver the required number of items within the stipulated time, the player will 'earn' a failure token as a penalty. The player with the least number of failure tokens wins the game.

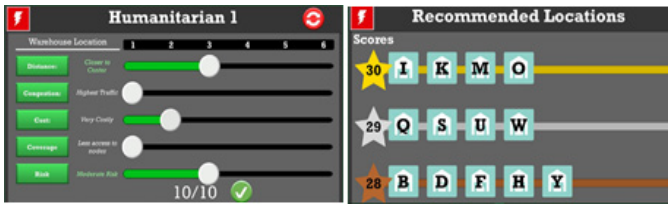


Figure 3: Screenshots of the Warehouse Location selection screen using the digital companion application



Figure 4: Sample screenshot of the MCDM criteria in the digital companion application

4.3. Game Experience

In November 2018, we conducted a workshop using the THINKLog game in Jakarta, Indonesia. This workshop was part of the ASEAN Coordinating Centre for Humanitarian Assistance on Disaster Management (AHA) Centre Executive (ACE) Programme conducted by the ASEAN Coordinating Centre for Humanitarian Assistance (AHA Centre). The exercise involved 16 senior government officials operating in disaster preparedness and response from multiple ASEAN countries, whom had limited knowledge in humanitarian logistics.

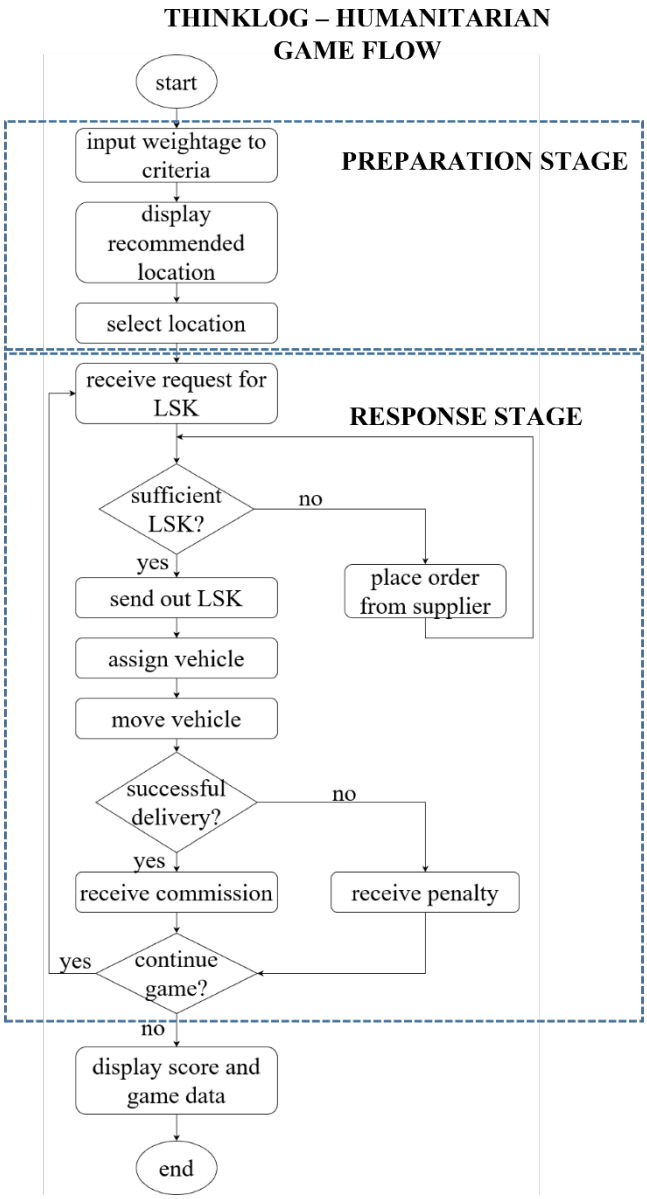


Figure 5: Basic game flow of THINKLog Humanitarian play

We conducted a brief presentation on humanitarian logistics and related supply chain concepts prior to the game play session. After the game play, participants were asked to fill up feedback forms. We then used those data to evaluate the impact of the game to increase the participants' understanding about humanitarian logistics and the overall game experience. We also used the workshop as a platform to gather inputs from the participants to be used in the simulation model.

The session allowed us to gather two separate sets of inputs. The first set is to evaluate participants' game experiences and it used a questionnaire of 14 questions. The sample of questions are shown in Table 2 below.

Table 2: Sample of feedback questions

0: Strongly disagree 1: Disagree 2: Neutral 3: Agree 4: Strongly agree

Your Opinion	Rate (0-4)
1. I feel that the whole session is interesting	0 – 1 – 2 – 3 – 4
2. I would like to play this game again in future	0 – 1 – 2 – 3 – 4
3. I encounter difficulties in understanding the game rules	0 – 1 – 2 – 3 – 4
4. I know exactly what I need to do in the game	0 – 1 – 2 – 3 – 4
5. I feel that every steps/decisions made in this game has meaningful meaning	0 – 1 – 2 – 3 – 4
6. I feel very involved in this game	0 – 1 – 2 – 3 – 4
7. Interaction with other players motivates me to understand this game better	0 – 1 – 2 – 3 – 4
8. Overall session is too long/boring	0 – 1 – 2 – 3 – 4
9. The game motivates me to ask questions/discuss	0 – 1 – 2 – 3 – 4
10. I obtained interesting information through this session	0 – 1 – 2 – 3 – 4
11. I am interested in the content provided in this game	0 – 1 – 2 – 3 – 4
12. I did not learn anything from this session	0 – 1 – 2 – 3 – 4
13. The Companion App helps me understand the game	0 – 1 – 2 – 3 – 4
14. The Companion App helps me learn from the game	0 – 1 – 2 – 3 – 4

Participants were asked to provide answers using the 5-point Likert scale with 1 as strongly disagrees and 5 as strongly agree. There were sixteen valid responses received. The positive experience's average score is above 3 and the negative experience's is below 2, as shown in Figure 6. We may conclude that the participants had a good overall game experience.

The second feedback asked the participants to list down the key learning points that they have gathered from the session. We then compared the responses with the game's intended learning objectives. We received ten valid responses. We extracted the responses and identified the topics that they mentioned in their responses. The topics and the connection between the topics are shown in Figure 7.

From the analyzed feedback forms, we identified three main topics, namely: Plan, Warehouse and Location. Three out of six responses (3/6) that mentioned about Plan also mentioned about Warehouse and Location. It can be further breakdown to important considerations for planning the warehouse location, such as distance, congestion, cost, coverage and risk.

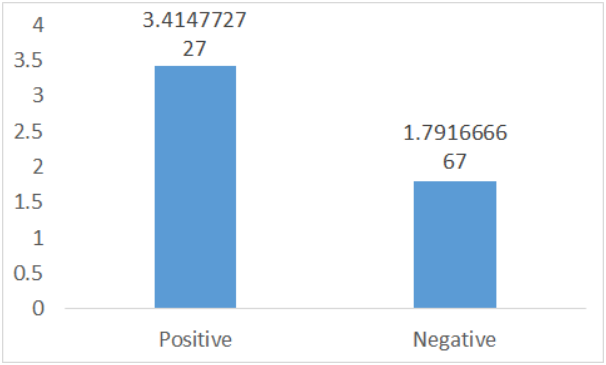


Figure 6: Overall experience

With the results, we are able to conclude that:

- The participants' learning outcome is aligned with the game's learning objectives.
- The game exposes the participants to the humanitarian logistics complexity.
- The feedback confirmed the need of embedding key elements such as distance (closeness to affected areas), congestion (the lower, the better), cost (the lower, the better), coverage (closer to highly populated disaster prone areas) and risk (facility to be located in a disaster free zone).

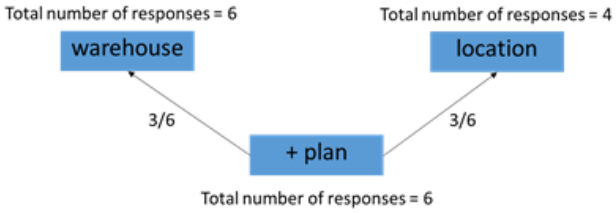


Figure 7: Learning Outcome

In addition, we also recorded the practitioners' inputs during and after the game session. Their inputs include the critical criteria to use for determining the location of a warehouse, and the importance of each criterion.

The feedback shows that the game enhances the players' awareness to the importance of decision making, especially in facility location. Players spent a substantial amount of time inputting the weightage for the criteria, deliberating which criteria should be of a higher importance. As the game provides 5 selected decision criteria for the MCDM, the players are more eager to

experiment these using real-data simulation model that incorporates even more criteria for consideration, which will also involve the framework of selecting those criteria. Table 3 shows the criteria that was selected for the THINKLog game as compared to the criteria used for the model in Section 4.4. [41]

During the game play session, it was also noted that most, if not all, players placed higher weightage to Risk and Access to affected zones criteria. All players have the same number of vehicles (2 trucks) and warehouse capacity (30 LSK) in the game. As the game has only 1 disaster area (demand point), this may be sufficient. Whereas in real-life situation, there may be more than 1 demand point at any given point of time during a disaster.

Also, for this game play configuration, we intentionally excluded the transportation cost that may occur to deliver the LSK to the disaster area, for ease of play purpose. As such, the player may not be able to optimize its relief operation.

Table 3: The criteria used in THINKLog game

Criteria	Description used in THINKLog game	THINKLog game
Coverage	Refers to the number of nodes connected.	✓
Access to affected zones	Refers to the time taken from when the order is placed till the order is delivered. The closer one is to the affected area, the shorter lead time it will have,	✓
Risk	Refers to the chance of a disaster striking the node.	✓
Access to infrastructure	Not used in the game	×
Access to corridor	Not used in the game	×
Congestion	Refers to the frequency of traffic congestions happening around that area	✓
Cost	Refers to the maintenance cost of a warehouse in a node. The area around the downtown costs higher than those at the outskirt	✓
National development plan (NDP)	Not used in the game	×

Section 4.4 describes the model we have developed to help locate the most appropriate alternative locations for the warehouse (facilities) to serve the different demand points, taking into consideration the transportation cost.

4.4. Simulation Model for Humanitarian Logistics

Based on the practitioners' input and the game session's outcomes of the game play, we developed a computer simulation model to delve the understanding on network design for stockpile prepositioning in the use case of Indonesia. The aim of this model was to determine, out of 9 alternative locations, which were the most appropriate 6 sites to use for the positioning of emergency response facilities.

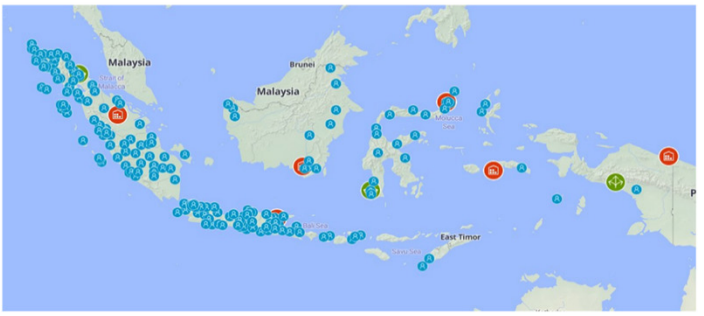
There were 186 demand points identified across the archipelago. Each demand points is served by a single facility. The model's overarching objective is to identify the possible locations for the facilities to serve these demand points, while keeping the transportation cost at the minimum. We used the actual distance and land fuel cost in this model to calculate the transportation cost between the facility and demand points. For those demand points without land connections (e.g. small islands), we set the fuel cost to extremely high values as to replicate intermodal transportation.

The results of the model are summarized in Figure 8 while the entire case study can be found in the paper by The Logistics Institute – Asia Pacific [42]. Figure 8(a) and 8(b) show possible network configurations identified by the computer model whereas figure 8(c) shows the network structure prospected by discussions with practitioners.

By inputting these 3 alternative network configurations into the model, findings show that the network structure identified using network optimization has the potential to reduce the transportation cost by 15%.



8a) Optimum configuration from Simulation tool



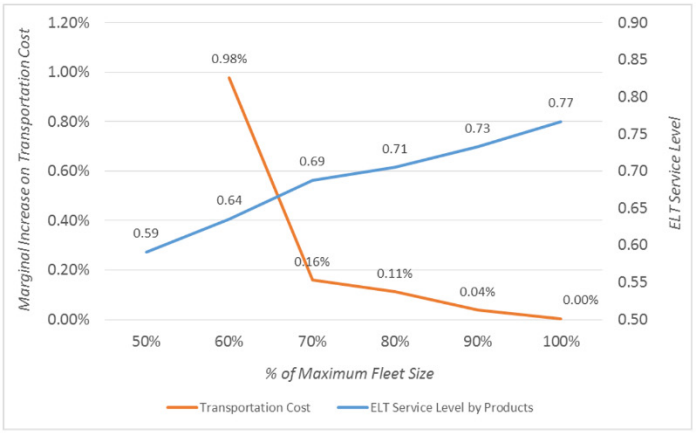
8b) Second best configuration from Simulation tool



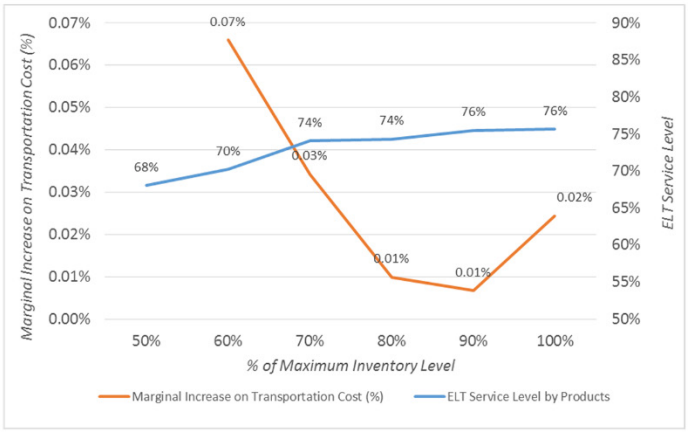
8c) Configuration from preliminary discussions with several experts in Indonesia

Figure 8: Disaster Relief Model

Subsequently, we stress-tested the network configuration upon two key logistics parameters namely fleet size and inventory. Figure 9 shows the results of the computer simulation model in regard to service level at when fleet size and inventory are set at different values.



9 a) Impact of Fleet size on Transportation Cost and service level



9b) Impact of Inventory Level on Transportation Cost and service level

Figure 9: Simulation Result for Fine-tuning the Network Configuration on Two Parameters

5. Conclusion

In this paper, the focus was on integrating serious games and computer modelling as a novel way to raise awareness on humanitarian logistics issues as well as to facilitate the acquisition of inputs from humanitarian practitioners. A real-case simulation model was developed with the aim of selecting the most suitable locations for establishing a network of facilities for prepositioning stocks of life-saving goods in Indonesia. We also had preliminary discussions with several experts on a prospected network configuration. When we compared these two configurations, the simulation model configurations allow for an estimated cost saving of 15%.

A role-based board game, THINKLog, is proposed as an easy-to-use tool to raise awareness in humanitarian logistics for non-technical personnel. The game has two stages, the Preparation and Response stage. The feedbacks gathered from the game play session shows that the participants have a good overall game experience with a substantial positive influence on the players. This is supported by an evaluation of the learning objectives by the participants whom reported that they were able to grasp them during the game play.

This work has few limitations. First, the sample size of senior government officials operating in disaster preparedness and response from multiple ASEAN nations is fairly small. Playing the game with a larger group of practitioners would allow to gather deeper insights on the game play and inputs to the computer model. Secondly, on the simulation model, the dataset on small and medium scale disasters is confined to the biennium 2014-2015. An extension of this database with the inclusion of a greater number of disasters would provide a more accurate estimation of demand.

For future extension, we see an opportunity to add more immersive game features to the game, like augmented reality (AR). This may enhance the game's visualization and overall gameplay experience. Secondly, we would like to evaluate the THINKLOG game using a broader group of individuals with different age groups, educational and professional backgrounds, and level of understanding of supply chain and humanitarian logistics concepts. This would enable us to gather a more comprehensive view of game's ability to raise awareness.

Conflict of Interest

The authors declare no conflict of interest related to this paper.

Acknowledgment

The THINKLog board game was first developed as part of "Temasek Foundation International –National University of Singapore Urban Transportation Management Programme in Indonesia" programme. It is supported by Temasek Foundation International and Coordinating Ministry for Economic Affairs of Indonesia.

References

[1] R. D. de Souza, L. William, G. Timperio and Z. B. A. Rahim, "Simulation M And Simulation-Based Serious Gaming In Humanitarian Logistics," in *Winter Simulation Conference* , 2018.

[2] ReliefWeb, "The human cost of natural disasters 2015: a global perspective," 6 3 2015. [Online]. Available: https://reliefweb.int/report/world/human-cost-natural-disasters-2015-global-perspective. [Accessed 5 12 2017].

[3] Swiss Re Institute, "Natural catastrophes and man-made disasters in 2017: a year of record-breaking losses," Swiss Re Institute, 2018.

[4] L. Van Wassenhove, "Humanitarian Aid Logistics: Supply Chain Management in High Gear," *Journal of Operational Research Society,* vol. 57, no. 5, pp. 475-489, 2006.

[5] A. S. Thomas and L. R. Kopczak, "Fritz Institute," 2005. [Online]. Available: http://www.fritzinstitute.org/pdfs/whitepaper/fromlogisticsto.pdf. [Accessed 25 March 2016].

[6] B. Balcik and M. Beamon, "Facility Location in humanitarian relief," *International Journal of Logistics: Research and Applications,* vol. 11, no. 2, pp. 101-121, 2008.

[7] Y. Nahleh, A. Kumar and F. Daver, "Facility Location Problem in Emergency," *International Journal of Mechanical, Aereospace, Mechatronic, and Manufacturing Engineering,* vol. 7, no. 10, pp. 2113-2118, 2013.

[8] C. L'Hermitte, B. Brooks, M. Bowles and P. Tatham, "Investigating the strategic antecedents of agility in humanitarian logistics," *Disasters,* vol. 41, no. 4, p. 672–695, 2016.

[9] G. Timperio, G. Panchal, A. Samvedi, M. Goh and R. De Souza, "Decision support framework for location selection and disaster relief," *Journal of Humanitarian Logistics and Supply Chain Management,* vol. 7, no. 3, pp. 222-245, 2017.

[10] C. Thierry, G. Bel and A. Thomas, "Supply Chain Management Simulation: An Overview," in *Supply Chain Management Simulation*, Wiley-ISTE; 1 edition , 2008, pp. 1-39.

[11] M. Besiou, O. Stapleton and L. N. Van Wassenhove, "System dynamics for humanitarian operations," *Journal of Humanitarian Logistics and Supply Chain Management,* vol. 1, no. 1, pp. 78-103, 2011.

[12] O. Costa, J. Santos, M. Martins and U. Yoshizaki, "A system dynamics analysis of humanitarian logistics coordination," Delft, 2015.

[13] E. Iakovou, D. Vlachos, C. Keramydas and D. Partsch, "Dual sourcing for mitigating humanitarian supply chain disruptions," *Journal of Humanitarian Logistics and Supply Chain Management,* vol. 4, no. 2, pp. 245-264, 2014.

[14] D. Noreña, R. Akhavan-Tabatabaei and L. Yamín, "Using discrete event simulation to evaluate the logistics of medical attention during the relief operations in an earthquake in Bogota," Phoenix, AZ, USA, 2011.

[15] V. Grimm and S. F. . Railsback, "Individual-based Modeling and Ecology," 17 May 2004. [Online]. Available: http://www2.humboldt.edu/ecomodel/documents/Grimm-Railsback05.pdf. [Accessed 20 August 2017].

[16] M. W. Horner and M. J. Widener, "The effects of transportation network failure on people's accessibility to hurricane disaster relief goods: a modeling approach and application to a Florida case study," *Natural Hazards,* vol. 59, no. 3, p. 1619–1634, 2011.

[17] U. Ritterfeld, M. Cody and P. Vorderer, "Serious games: Mechanisms and effects," *Internationa Journal of Gaming and Computer-Mediated Simulations,* vol. 1, no. 3, pp. 89-94, 2009.

[18] W. William, "Current trends in educational technology research: The study of learning environments," *Educational Psychology Review,,* vol. 14, no. 3, p. 331–351, 2002.

[19] H.-T. Hou, "Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration," *Computers in Human Behavior,* vol. 48, no. 1, pp. 424-435, 2015.

[20] M. Graafland, J. M. Schraagen and P. Schijven. M., "Systematic review of serious games for medical education and surgical skills training," *British Journal of Surgery,* vol. 99, no. 10, pp. 1322-1330, 2012.

[21] C. Lim and H. Jung, "A study on the military Serious Game," *Advanced Science and Technology Letters,* vol. 39, pp. 73-77, 2013.

[22] M. A. Garcia-Ruiz, J. Tashiro, B. Kapralos and M. V. Martin, "Crouching Tangents, Hidden Danger: Assessing Development of Dangerous Misconceptions within Serious Games for Healthcare Education," in *Virtual Immersive and 3D Learning Spaces: Emerging Technologies and Trends*, Kansas State University, USA, IGI Global, 2011 , p. 412.

[23] A. Gómez-Rodríguez, J. González-Moreno, D. Ramos-Valcárcel and L. Vázquez-López, "Modeling serious games using AOSE methodologies.," in *11th International Conference on Intelligent Systems Design and Applications*, 2011.

[24] J. Hamari, D. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke and T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Computers in Human Behavior,* vol. 54, no. 1, pp. 170-179, 2016.

[25] M. Ma, A. Oikonomou and L. Jain, Innovations in Serious Games for Future Learning. In Serious Games and Edutainment Applications, London: Springer, 2011.

[26] PAXsims, "AFTERSHOCK," 4 2015. [Online]. Available: https://paxsims.wordpress.com/aftershock/.

[27] UN Office for Disaster Risk Reduction, "STOP DISASTERS!," 2019. [Online]. Available: https://www.stopdisastersgame.org/stop_disasters/.

[28] S. Conceição, S. Pedrosa, A. Neto, M. Vinagre and E. Wolff, "The facility location problem in the steel industry: a case study in Latin America," *Production Planning & Control,* vol. 23, no. 1, pp. 26-46, 2010.

[29] H. Steenhuis and E. De Bruijn, "Production Planning & Control," *Assessing manufacturing location,* vol. 15, no. 8, pp. 786-795, 2007.

[30] J. McCoy and M. Brandeau, "Efficient stockpiling and shipping policies for humanitarian relief: UNHCR's inventory challenge," *Operations Research-Spektrum,* vol. 33, no. 3, pp. 673-698, 2011.

[31] S. Roh, S. Pettit, I. Harris and A. Beresford, "The pre-positioning of warehouses at regional and local levels for a humanitarian relief organisation," *International Journal of Production Economics,* vol. 170, no. Part B, pp. 616-628, 2015.

[32] System Dynamics Society, "Introduction to System Dynamics," 2010. [Online]. Available: http://lm.systemdynamics.org/what-is-s/. [Accessed 9 January 2018].

[33] S. Robinson, Simulation: The Practice of Model Development and Use, 1st ed., Chichester: John Wiley & Sons Ltd, 2004.

[34] Statista, "Countries with the most earthquake fatalities 1900-2016," 2017. [Online]. Available: https://www.statista.com/statistics/269649/earthquake-deaths-by-country/. [Accessed 5 January 2018].

[35] AIR, "Five Countries Most Frequently Hit by Natural Disasters," 24 September 2015. [Online]. Available: http://www.air-worldwide.com/Blog/Five-Countries-Most-Frequently-Hit-by-Natural-Disasters/. [Accessed 21 February 2018].

[36] Indonesia-investments, "Indonesia-investments," 6 une 2017. [Online]. Available: https://www.indonesia-investments.com/news/todays-headlines/business-investment-climate-of-indonesia-improving-competitiveness/item7878. [Accessed 20 February 2018].

[37] Lindawati, E. Nugroho, R. Fredericco, Z. B. A. Rahim and R. de Souza, "ThinkLog: Interactive learning for supply chain management," in *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2017.

[38] L. William, A. Rahim, R. de Souza, E. Nugroho and R. Fredericco, "Extendable Board Game to Facilitate Learning in Supply Chain Management," *Science, Technology and Engineering Systems Journal,* vol. 3, no. 4, pp. 99-111, 2018.

[39] M. Köksalan, J. Wallenius and S. Zionts, "An early history of multiple criteria decision making," *Journal of Multi-Criteria Decision Analysis,* vol. 1, no. 2, pp. 87-94, 2013.

[40] J. Brans, P. Vincke and B. Mareschal, "How to select and how to rank projects: The PROMETHEE method," *European Journal of Operational Research,* vol. 24, no. 2, pp. 228-238, 1986.

[41] The Logistics Institute - Asia Pacific, "How to identify the most appropriate locations for establishing an efficient network of emergency facilities? – A Discussion Paper," The Logistics Institute - Asia Pacific, Singapore, 2016.

[42] The Logistics Institute - Asia Pacific, "Integrated Decision Support Framework for Enhancing Disaster Preparedness: A Pilot Application in Indonesia," The Logistics Institute - Asia Pacific, Singapore, 2018.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 408-415 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

To the Question of Multi-Criteria Optimization of Aircraft Components in Order to Optimize its Life Cycle

Sergey Alekseevich Serebryansky*, Alexander Vladimirovich Barabanov

Department Aircraft Design and Certification, Moscow Aviation Institute (National Research University), Moscow, 125993, Russia

ARTICLE INFO

Article history:
Received: 20 August, 2020
Accepted: 21 October, 2020
Online: 20 November, 2020

Keywords:
Complex Methodology
Optimization of The Design Cycle
Aerodynamic Drag
Rational Parameters
Nose of The Aircraft

ABSTRACT

The paper examines the impact on the product life cycle of reducing the duration of the development stages of aviation technology through the use of multi-criteria optimization. The subject of the study is the nose of a supersonic aircraft and the process of linking it. This example describes in detail the application of a comprehensive design methodology for aircraft components using top-level criteria as a way to significantly reduce the development time and the risks of obtaining negative test results when entering a full-scale sample.

1. Introduction

The great Russian aircraft designer Pavel Osipovich Sukhoi said: "We must not stop at what we have achieved, but must constantly move forward. After all, what was new and advanced yesterday will only be satisfactory tomorrow." This statement is still relevant until now because to take leadership positions in aviation, engineering, Economics, and business, you must constantly work ahead of time and your competitors. The current trend in the field of mechanical engineering is a significant (from 10-15 to 5-8 years) reduction in the development time of new aircraft systems due to the use of mathematical modeling, high unification and the use of scientific and technical reserve. The use of complex design techniques is one of the elements based on scientific and technical groundwork and allows you to reduce the design time.

2. Aircraft product life cycle

Within the lifecycle of the aircraft as a product, the aircraft goes through the following stages (gates):

Figure 1 shows the typical life cycle of an aircraft product, which is divided into internal and external design stages. The focus of the current work is optimization of internal design stages, so we will analyze them in more detail.

It is worth noting that key technical decisions are made at the stages of the preliminary project and preliminary technical design. The work is carried out by a relatively small number of OKB employees, the minimum amount of financial resources from the project budget is used, and the maximum number of technical decisions is made as shown in Figure 2. According to the results of these stages, the appearance of the designed aircraft is fixed.

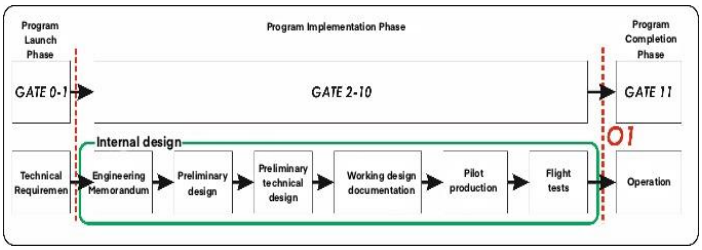


Figure 1: The typical lifecycle of the aviation system

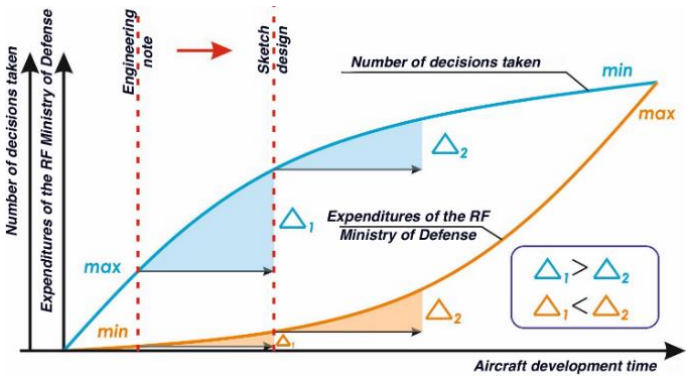


Figure 2: Dependence of the number of decisions made and the customer's costs on the project implementation time

*Corresponding Author: Sergey Alekseevich Serebryansky, s-s-alex@mail.ru

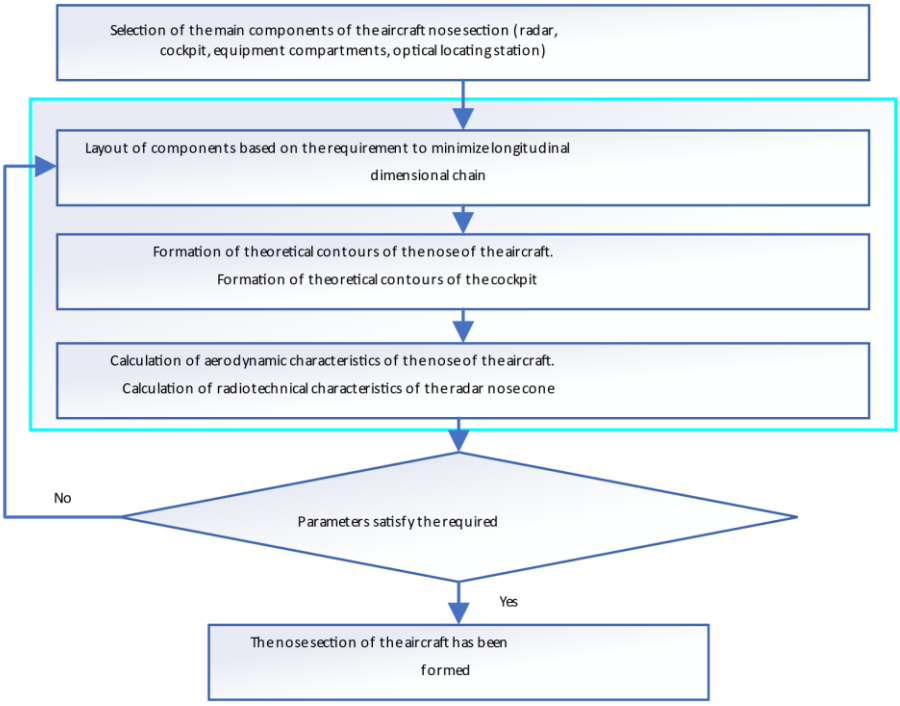


Figure 3: Typical algorithm for developing the nose of the aircraft

Thus, the stages of project initiation are the most effective in terms of cost and number of technical decisions:

$$\left.\begin{matrix}\Delta_i^{expenses}\\ \Delta_i^{time}\end{matrix}\right\} \rightarrow \min(opt)$$

However, these stages have the greatest technical risks, and errors in the technical decisions made may become fatal for the further fate of the project.

As can be seen from Figure 2, the initial stages of project implementation are the most effective in terms of the ratio of the customer's costs to the number of decisions made. Simultaneously, the technical decisions taken at these stages determine the further technical appearance of the project and can significantly affect its results.

Figure 3 shows an algorithm for linking the nose of the aircraft, consisting of sequential operations and, in case of errors, having an iterative character.

To complete the formation of the theoretical contour of the head part of the aircraft it must validate the obtained electronic and aerodynamic characteristics of the selected form of the radio waves fairing requirements to ensure the characteristics of the radar. Meanwhile, the speed and range characteristics of the aircraft will depend on the aerodynamic characteristics in the future, and the potential capabilities of the locator will depend on the radio technology. Thus, such a seemingly insignificant thing as the shape of the nose of an aircraft can determine its key characteristics of combat use. Moreover, the shape of the nose part of the minimum resistance and the shape of the maximum radar-transparency of a thing are incompatible because the best radio-technical characteristics are usually of small elongation nosecone (the ratio of the square of the length of the nosecone to the area of the initial cross-section), such a shape allows you to ensure the uniformity of the distribution of the angles of incidence of the electromagnetic wave front on the surface of the nosecone wall, the best aerodynamic characteristics on the contrary have the body with maximum elongation, it allows reducing significantly the resistance at supersonic flight speeds, see Figure 4:

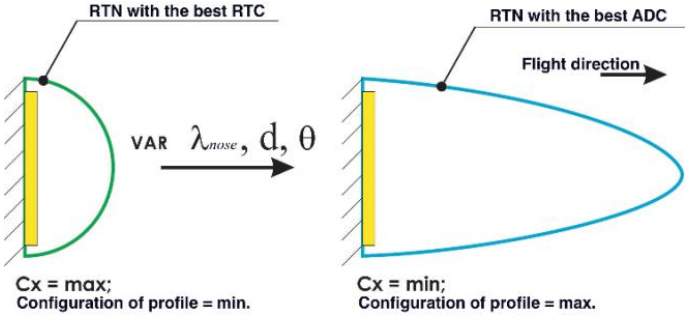


Figure 4: Comparison of nosecone shapes of the aircraft

From Figure 4 we can conclude that the final shape of the nosecone will be a compromise between the two shown in the figure. In this case, the wrong form at this stage may lead to the need to repeat the whole cycle of flight tests in case of its change after the construction of prototypes (stages of pilot production and flight tests).

3. Optimization of the aircraft life cycle

The scientific novelty of the work consists in carrying out multi-criteria optimization of the geometric parameters of the studied RPO using the Russian-made LOGOS software product and in terms of functions, which is an analog of the FLUENT product. The results of mathematical modeling and evaluation of aerodynamic characteristics have been validated with existing empirical methods, and the software product itself is widely used in Russian industrial projects.

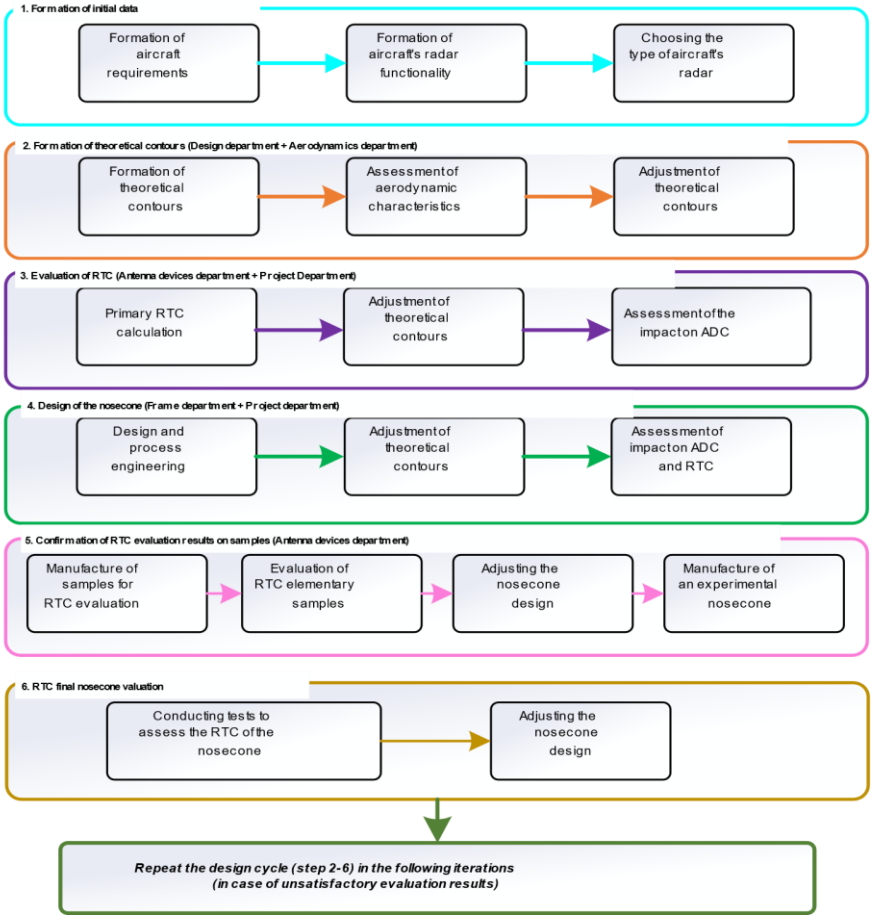


Figure 5: The existing algorithm for the development of the fairing

Reducing the timing of development work can be achieved by reducing the development and testing of the aircraft complex. Concurrently, correct and timely technical decisions made at the development stage will also reduce the time required for ground and flight tests. Let's look at an example of optimizing the development stages of a nose radio-transparent fairing (RPO). The current state of the RPO design algorithm is shown in Figure 5.

After analyzing Figure 4, we can conclude that the design of the radio-transparent fairing according to the existing methodology does not meet the requirements for the development of new aircraft systems. To meet the customer's requirements for the development of new AK in 5-8 years, it is necessary to significantly optimize this algorithm. To do this, you must: To optimize this algorithm, it is necessary:

1) In accordance with [1], [2] select a comprehensive criterion for evaluating the characteristics of RPO radar, both in terms of radio engineering (RTC) and aerodynamic (ADC) characteristics;

2) Form a research algorithm, in which to establish the relationship between the parameters of the shape of the fairing and its ADC and RTC.

3) Based on the results of the study, formulate technical recommendations for choosing rational parameters of the fairing shape, depending on the initial data and technical requirements.

As a complex criterion of the upper level (1), having analyzed the scientific and technical background in the field of research of radio technical characteristics (RTC) [3] and aerodynamic characteristics (ADC) [4], [5], [6], it is possible to take the ratio of the criterion of the first level of aerodynamic characteristics to the criterion of the first level of radio engineering characteristics:

$$X^{*} = ArgMax\left(\frac{ArgMax(RTC)}{ArgMin(ADC)}\right) \quad (1)$$

In this case, we will choose as the criterion of the first level for RTC nosecone its transmission coefficient (TC)—a dimensionless value equal to the ratio of the radiation flow through the medium (nosecone) to the radiation flow to its surface. The first level criterion for ADC is its drag coefficient (Cx), a dimensionless value reflecting how the body (nosecone) reacts with the environment (air).

According to this criterion, it is necessary to conduct research on possible options for the geometric appearance of the fairing and choose rational parameters that meet all the requirements. The algorithm for conducting these studies is shown in Figure 6.

Based on the results of the above studies, the method for selecting the geometric parameters of the fairing will allow optimizing the stages of RPO development. Adjusted according to the results of the optimization algorithm for the design of the fairing is presented in Figure 7.

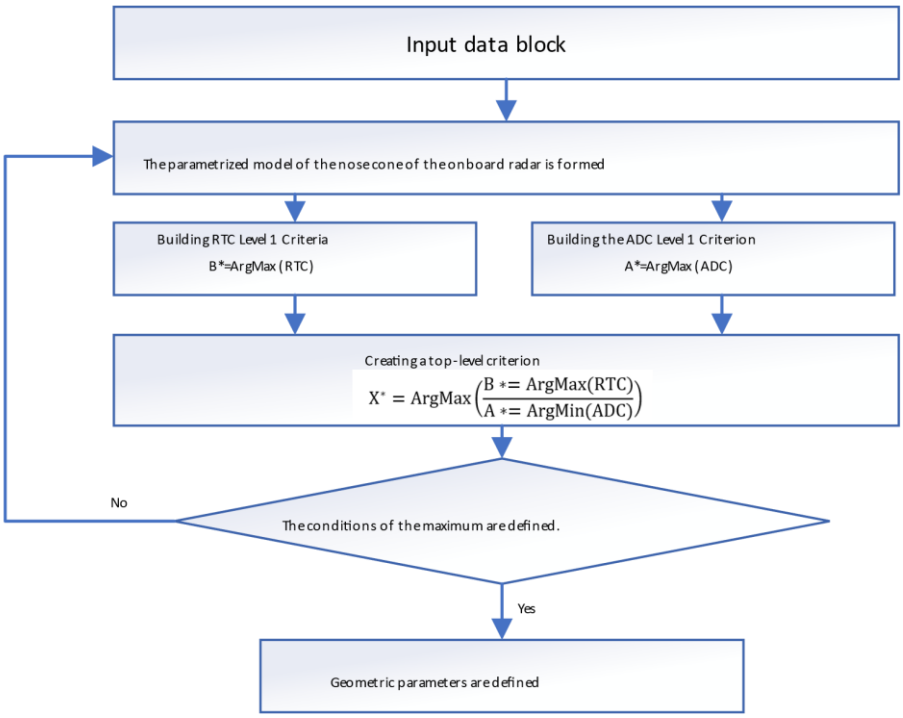


Figure 6: Algorithm for studying geometric parameters of RPO

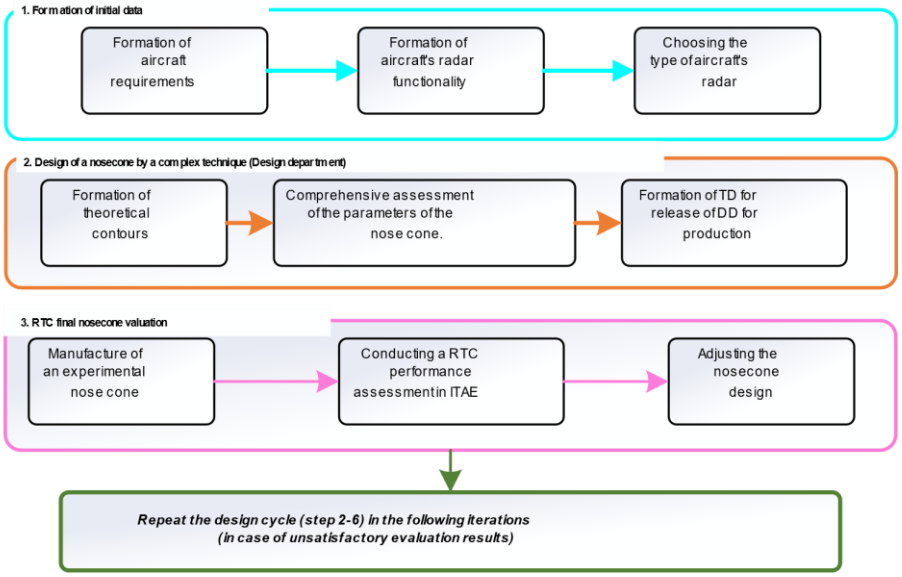


Figure 7: Adjusted the algorithm design of the fairing

Thus, the development of such methods for designing aircraft components can significantly reduce the number of design stages, as well as the involvement of company personnel, which in turn reduces the time and cost of product development.

4. Forming an optimal fairing with minimal drag

A typical example of multi-criteria optimization of an aircraft component to significantly reduce the product life cycle is the optimization of the geometric parameters of the radio-transparent fairing in order to obtain a body of minimal resistance. For this optimization, you must:

Firstly, to choose a complex criterion for evaluation of aerodynamic characteristics;

Secondly, to make up an algorithm for solving this problem and determine a method for evaluating aerodynamic characteristics that will have the required level of sensitivity to changes in key geometric parameters and the required accuracy;

Third, using the selected algorithm, to conduct research and, using a complex criterion, select rational geometrical parameters of the nose section of the aircraft for a given flight mode.

4.1. Choosing a comprehensive criterion

As noted earlier in this work, the issue of selecting rational geometric parameters for the nose section of the aircraft for the supersonic cruise mode is under consideration $M \sim 1.5$. To select the main criterion of aerodynamic characteristics of the nose

section, it is necessary to analyze the existing aerodynamic characteristics of the aircraft components as a whole.

According to [7], there are three main criteria for the aerodynamic characteristics of an object: C_x – aerodynamic resistance coefficient, C_y – lifting factor, Δmz_0 – component contribution to the longitudinal static torque of the aircraft as a whole:

1) Coefficient C_{xnose} as shown in Part 1. The introduction is one of the essential parts of the aerodynamic drag coefficient of the pressure of the aircraft as a whole $C_{xfpressure}$ in particular for $M > 1$.

2) In doing so C_{ynose} does not make a significant contribution to is not more than 0.1% of the total C_y as a whole.

3) It should be noted that the nose part of the aircraft can be shaped as a body rotation, and as a body of complex shape in the cross section. So for the body of rotation the contribution to Δmz_0 the aircraft as a whole tends to 0, and for a complex body it requires deep analysis and has a significant impact due to the large removal of the aerodynamic focus of the fore part of the aircraft from the center of the masses, which leads to losses on balancing and reduction of fuel efficiency of the aircraft in cruising mode of flight. Because for the current work is considered a cruise flight of the aircraft in the mode of $M \sim 1.5$ let's look at the fore part in the shape of a rotational body.

Thus, it is possible to draw an unequivocal conclusion that the criterion of a choice of rational geometrical parameters of a nose part of an airplane is the minimum value of the aerodynamic resistance coefficient, depending on basically geometrical options. This criterion can be written mathematically as follows (see (2)):

$$A^* = ArgMin\,(C_{x\,nose}(\lambda_{nose}, d, \theta)). \quad (2)$$

4.2. Selection of performance evaluation method

At the present time a significant scientific and technical reserve has been accumulated in the field of research of aerodynamic characteristics of bodies of various shapes and geometry, the main branch institute and center of competence of aerodynamics in the Russian Federation is the Central Aerodynamic Institute named after Professor N.E. Zhukovsky (TsAGI). Based on TsAGI's extensive work to determine the aerodynamic performance of various components of the aircraft, empirical dependencies of the AP on a few geometric parameters were obtained. These methods and dependencies are presented in [7] and are used to determine the AP (C_x and C_y) of the main components of the aircraft (wing, fuselage, plumage) and build integral AP of the aircraft. The empirical approach is widely used by the country's leading aviation and missile companies.

However, because these dependencies have been obtained from numerous wind tunnel experiments on wind tunnel models, there are certain assumptions in the wind tunnel estimates due to model manufacturing tolerances and measurement errors.

The trend formed in the XXI century of constant toughening of requirements to newly developed aircraft complexes (ACs) by the State Customer, related to ensuring competitive edge over foreign ACs, requires minimization of design reserves in the development of AKs and a more accurate assessment of the AP of both components and the aircraft as a whole.

In recent years, the main trend and element of product life cycle management in the aircraft industry is the digitalization of the life cycle stages in a single information space, including various types of tests [8] and the creation of electronic twins of the aircraft, including aerodynamic, through the application of mathematical simulation of geometry flow by finite element method [9]. The models received by results of similar digital tests are verified and possess necessary sensitivity to change of geometrical characteristics, are shown as algorithm in Figure 2, also allow to receive result with high accuracy and can be parametrized that essentially reduces terms of research.

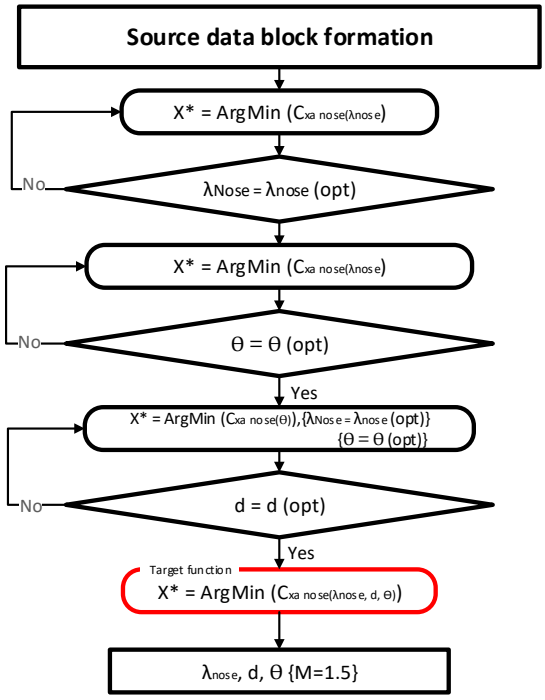


Figure 8: Algorithm for solving the problem of determining the rational parameters of the aircraft fuselage nose section

4.3. Problem solving algorithm

For effective solution of the problem within the framework of this work it is necessary to apply methods of mathematical modeling with the use of parametrized mathematical models of the surface, where the main geometrical parameters of the fore-body indicated earlier are used as variables of parametrization.

Main calculation parameters are accepted as a block of input data: altitude and speed parameters of flight, values of angles of attack, roll and slide, as well as atmospheric parameters at a selected altitude and speed.

The problem is solved iteratively, and at the first stage is the optimal value of the parameter of elongation of the nose part of the aircraft (λ_{nose}) as a most comprehensive parameter. Next, for an optimum elongation value, the optimum angle of tangent to the forming curve is selected (θ) and the final step is to determine the optimal value of the discriminant in the formation curve. (d).

It should be noted that for each stage a mathematical model of the investigated surface is built, and the range of parameter values can change based on the conditions of ensuring the smoothness of the fore-body surface (absence of double curvature) [10], [11].

4.4. Selecting the number of settlement cases

Calculated flight conditions are shown in Table 1. During the calculation the atmospheric parameters are accepted in accordance with the international standard atmosphere.

Table 1: Calculated conditions for solving the problem

	Parameter	Value
1	Estimated number of M flight	1.5 M
2	Estimated flight altitude	11 000 m
3	Calculated Flight Angle of Attack (α)	2^0
4	Calculated sliding angle (β)	0^0
5	Calculated roll angle (γ)	0^0
6	Atmospheric parameters	In accordance with the ISA

The calculation is performed in the selected calculation conditions for each of the values of the parametrization variables and in accordance with the algorithm presented in Part 4. The range of values of the parameterization variables and the step are shown in Table 2.

Table 2: Value range of variables

Parameterization parameter values								
λ_{nose}			$\theta_{,0}$			d		
min	max	step	min	max	step	min	max	step
2	5	1 (0.5 from 2 to 3)	20	45	5	0.5	0.8	0.1

The total number of settlement cases for the calculation is determined by formula (3):

$$N = n_\lambda + n_\theta + n_d, \quad (3)$$

Where n_λ – fuselage fore-nose extension parameter number, n_θ – number of angle values of the tangent curvature parameter, n_d – discriminant parameter number of the constitutive curve.

Thus, to solve the problem it is necessary to make 15 calculation experiments.

In accordance with the research algorithm (see Figure 8), a cycle of computational experiments to determine the dependence of values has been performed in this work $C_{x\,nose}$ the nose of the aircraft from its extension (λ_{nose}), the results of calculations are recorded in Table 3, the dependence of parameters is shown in Figure 9.

Table 3: Results of the dependency calculation C_{xnose} on λ_{nose} .

	θ	d	λ_{nose}	$C_{x\,nose}$
1	25^0	0.5	2	0.21
2			2.5	0.15
3			3	0.11
4			4	0.06
5			5	0.04

As can be seen from Table 3 and Figure 6, there is an inverse relationship between λ_{nose} and C_{xnose} for the nose section of the aircraft, so that the body with elongation has minimal aerodynamic resistance $\lambda_{nose} = 5$.

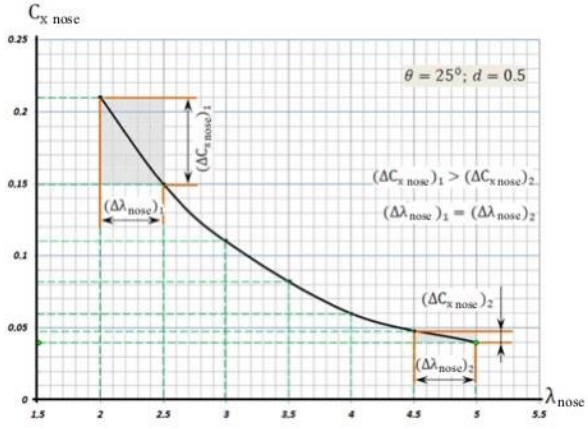


Figure 9: Dependency C_{xnose} on λ_{nose}

Taking into account these results in accordance with the algorithm (see Figure 8), we have estimated the change in the value of C_{xnose} change angle θ . In this case, to increase the sensitivity of the model to changes in the angle of tangent in the geometric model of the object surface, a spherical blunting of the spout with the $r = 15\,mm$ in this case the influence of the coefficient appears for the model $C_{xattenuation}$ in case of blunting, which is for M = 1.5 $C_{xattenuation} = 0.68$, and for $r = 15\,mm$ $C_{xattenuation} \sim 0.22$. The results of calculations are given in Table 4, the dependence is shown in Figure 10.

Table 4: Results of the dependency calculation $C_{x\,nose}$ on θ .

	λ_{nose}	d	$\theta_{,0}$	$C_{x\,nose}$
1	5	0.5	20	0.062547
2			25	0.062555
3			30	0.062565
4			35	0.062575
5			40	0.062587
6			45	0.062598

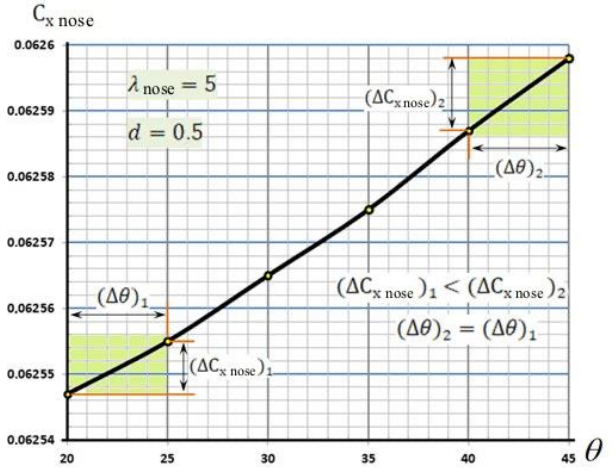


Figure 10: Dependency $C_{x\,nose}$ at θ

As can be seen from Table 4 and Figure 10, the aerodynamic resistance of the nose of the aircraft increases when the angle of tangent to the forming nose part of the aircraft. However, the difference between $C_{x\,nose}$ for $\theta = 20^{O}$ and $C_{x\,nose}$ for $\theta = 45^{O}$ at $\lambda_{nose} = 5$ not exceeding 1%. Based on these results, it can be concluded that the angle θ does not significantly affect the aerodynamic resistance of the nose section of the aircraft at $M = 1.5$. For the next phase of the study, we will accept $\theta = 20^{O}$ as a parameter with minimal aerodynamic resistance.

The next step in the study of the dependence of aerodynamic characteristics of the nose section of the aircraft on its geometrical parameters is to determine the dependence of $C_{x\,nose}$ from the curvature of the constituent – d. To determine the effect of curvature on the $C_{x\,nose}$ aircraft, a comparative analysis of the conical nose AP and the parabolic nose (at d=0.5) in determining the dependence of the $C_{x\,nose}$ on λ_{nose}. The results of this analysis are listed in Table 5, the dependencies are shown in Figure 11.

Table 5: Comparison $C_{x\,nose}$ at d.

	λ_{nose}	$C_{x\,parabola}$	$C_{x\,cone}$
1	2	0.21	0.23
2	2.5	0.15	0.17
3	3	0.11	0.13
4	4	0.06	0.08
5	5	0.04	0.059

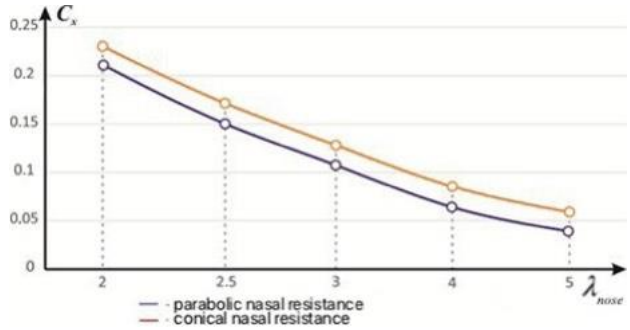


Figure 11: Comparison of nasal AP with conical and parabolic forming parts

As can be seen from Table 4 and Figure 8, both for conical and parabolic forming dependencies obtained in Figure 6 are fair. However, for the parabolic formative level $C_{x\,nose}$ lower than conical for all values λ_{nose}. Based on these results, it can be concluded that the nose section of the aircraft with parabolic formation has the lowest aerodynamic resistance.

To determine the rational value of the curvature forming within the framework of this work the dependence analysis has been performed $C_{x\,nose}$ at d for the nose section with the previously selected rational parameters λ_{nose} and θ, the results of this assessment are shown in Table 5, the dependence is shown in Figure 12.

Table 6: Dependency $C_{x\,nose}$ at d parabolic forming.

	λ_{nose}	θ	d	$C_{x\,nose}$
1	5	20^0	0.5	0.062547
2			0.6	0.062551
3			0.7	0.062558
4			0.8	0.062564

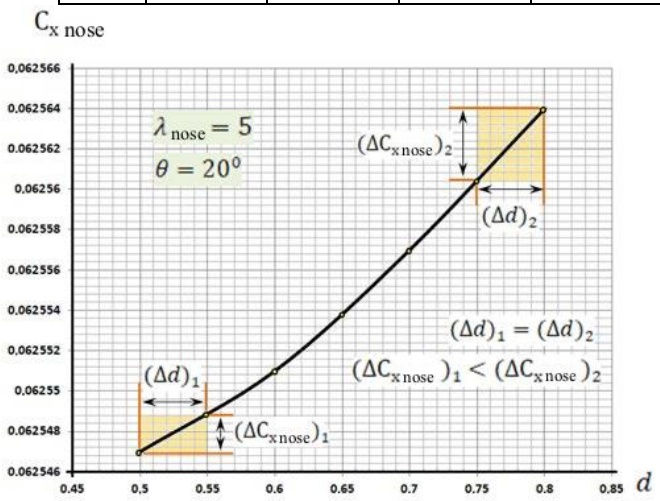


Figure 12: dependency $C_{x\,nose}$ at d

As can be seen from Table 5 and Figure 12, when the curvature of the aircraft nose increases, the aerodynamic resistance of the nose part increases. However, the difference between $C_{x\,nose}$ at $d = 0.5$ and $C_{x\,nose}$ for $d = 0.8$ at $\lambda_{nose} = 5$ and $\theta = 20^{O}$ not more than 1%. Based on these results, it can be concluded that parameter d does not significantly affect the aerodynamic resistance of the nose section of the aircraft at $M = 1.5$. As a rational value d at $M = 1.5$ take d =0.5 as a parameter with minimal aerodynamic resistance.

4.5. Optimal parameters

Within the framework of the work performed by the methods of mathematical modeling the dependences of aerodynamic resistance of the nose part of the aircraft on the main geometrical parameters varying in the selected range were obtained for supersonic cruise of the aircraft at $M = 1.5$. The following conclusions can be drawn from these studies:

1. Parameter of elongation of the nose of the aircraft λ_{nose} has a significant impact on the aerodynamic characteristics of the nose section of the aircraft. Dependence $C_{x\,nose}$ from λ_{nose} is reversed: when the parameter value is increased λ_{nose} level $C_{x\,nose}$ is decreasing. Reasonable value λ_{nose} to ensure minimum aerodynamic resistance in the selected flight mode is $\lambda_{nose} = 5$.
2. The parameter of the angle of tangent curvature to the forming nose part of the aircraft θ does not have a significant impact on the aerodynamic characteristics of the nose section of the aircraft when $M = 1.5$. Dependency $C_{x\,nose}$ for θ is direct: when the parameter value is increased θ level is increasing $C_{x\,nose}$. Rational importance θ to ensure minimum aerodynamic resistance in the selected flight mode is $\theta = 20^{o}$.
3. The nose section of the aircraft with a parabolic formation has a lower level of aerodynamic resistance than the nose section with a conical formation 3.

4. The parameter of curvature of the forming nose part of the aircraft d does not have a significant impact on the aerodynamic characteristics of the nose part of the aircraft at $M = 1.5$. Dependency $C_{x\,nose}$ from d has a direct character: when the value of parameter d increases, the level $C_{x\,nose}$ increases. The rational value d for minimum aerodynamic resistance in the selected flight mode is $d = 0.5$.

Thus, based on the conclusions (1–4) can be concluded that the main geometric parameter of the nose of the aircraft, which determines the level of its aerodynamic characteristics, for $M = 1.5$ is an extension of the nose of the aircraft λ_{nose}.

The nose section of the aircraft with rational geometrical parameters has minimal aerodynamic resistance: $\lambda_{nose} = 5$, $\theta = 20^{\circ}$, $d = 0.5$.

5. Conclusion

The main conclusion from the materials presented in this paper is that optimization of the development cycles of aircraft components leads to optimization of the product life cycle as a whole. The considered example of forming an optimization basis for choosing rational geometric parameters of RPO clearly showed a significant reduction in the RPO design cycle due to the use of complex techniques.

In this work, a total of about 90 computational experiments were performed, with an average labor intensity (if there is free computing power) of ~5 h/h. Thus, using the results of this work in the design of a new supersonic aircraft will save ~450 h/h of research work, which is ~2.5 months of work for one specialist of the organization.

The introduction of such an integrated approach in all industries and for all types of products will make it possible to make a new leap in the field of mechanical engineering in the shortest possible time. [12].

References

[1] E.S. Wentzel, Operation research, Moscow: Science, 1988.

[2] A.V. Barabanov, S. A. Serebryansky, "Approach to determining the rational shape of a supersonic aircraft using mathematical modeling methods", Vestnik SURGU, 2020

[3] V.A. Kaplun, Radar domes, Moscow: Sov. Radio, 1974.

[4] L.G. Artamonova, A.V. Kuznetsov, N.N. Pesetskaya, Verification of aerodynamic characteristics of the aircraft. Moscow, MAI, 2010

[5] G.S. Byushgens, Aerodynamics and flight dynamics of trunk aircraft, TSAGI, 1995.

[6] G.S. Byushgens, Aerodynamics, stability and control of supersonic aircraft, Phismatlit, 1998.

[7] G.A. Kolesnikov, V.K. Markov, A.A. Michailyuk, "Aerodynamics of aircraft". Moscow, Machinostroenie, 1993

[8] A.V. Barabanov, S.A. Serebryansky, "Algorithm for the formation of rational geometric shape of the radio-transparent nose radar fairing of the aircraft," Engineering J.: Science and innovation, **1**, 2020, doi: 10.18698/2308-6033-2020-1-1948

[9] S.A. Serebryanskii, A.V. Barabanov, "To the Question of Optimizing Product Life Cycle STAGES," Proc. of 12th International Conference Management of Large-Scale System Development, MLSD, doi: 10.1109/MLSD.2019.8911045, 2019

[10] D.D. Kulikov, V.S. Babanin, "Creating a parametric part model in the CAD system environment", Sci. and Tech. J. of Information Technologies, Mechanics and Optics, **4** (74), 2011.

[11] P.A. Koporushkin, A.S. Partin, D.V. Kurennov, "Algorithm for calculating parametrized objects in the system of geometric modeling of machine-building products", Automation and modern technologies, 2008.

[12] S. A. Serebryansky, A.V. Barabanov, "An approach to determining the rational shape of the nose of a supersonic aircraft using mathematical modeling methods", Bulletin of SUSU, mechanical engineering, **20(**3), 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 411-418 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Framework for Measuring Workforce Agility: Fuzzy Logic Approach Applied in a Moroccan Manufacturing Company

Fadoua Tamtam*, Amina Tourabi

National School of Applied Sciences, Systems Engineering and Decision Support Laboratory, 80000, Morocco

ARTICLE INFO

Article history:
Received: 03 May, 2020
Accepted: 22 May, 2020
Online: 11 June, 2020

Keywords:
Workforce agility
Fuzzy logic
Agile enablers

ABSTRACT

In today's Moroccan business environment, companies need to implement organization agility by developing an agile workforce that is able to deal with the environment volatility. Thus, the agile workforce concept has been appeared as a necessary and sufficient condition to achieve agility. Focusing on agile enablers influencing workforce agility is an important area but currently there is limited literature available. Acknowledging its importance, we continued our literature exploration in order to identify the enablers of workforce agility. Then, we describe a list of four enablers with different criteria and attributes. This paper further proposed fuzzy logic approach to evaluate different measures of the workforce agility. The results suggest that engagement, knowledge sharing, acceptance of changes and self-motivation are the most important attributes of agile workforce. Apart from that, different agile workforce attributes need to be improved in order to achieve the extremely agile level of the workforce.

1. Introduction

This paper is an extension of work originally presented in 4th World Conference on Complex Systems (WCCS) [1].

For many years, unexpected and dynamic changes [2] represent a common reality facing many organizations from different sizes and sectors. This volatility makes traditional approaches useless for fulfilling current organizational goals. Under this pressure [3], enterprises need to implement agility in order to achieve profitability [4]. The agility means the ability to quickly respond and adapt to volatile market environments. According to the different definitions of "agility" proposed in the literature [2], agility includes different competitive criteria like speed, flexibility, innovation, adaptability, proactivity, quality, productivity, profitability, customization, and knowledge [3] which help to focus on products and services driven by customer instead of those driven by the company [5].

The successful implementation of agility requires adapting all enterprise elements such as goals, technology and people to the unexpected changes [2]. Thus, agility concept has been extended to cover different organization areas. Within this paper [5], the term workforce agility appeared as the employee ability to adapt and evolve quickly by providing innovative solutions to different problems during any phase of the company project/program. Therefore, agile workforce needs to exhibit sufficient skills [6] which have been implicitly mentioned in Zhang and Sharifi [7] definition: Agile workforce is knowledge worker [3] with a broad vision and who is able to deal with environmental turbulence by capturing the advantageous side of these changes [8]. The same definition was presented by others researchers like Breu, et al. [5] who defined the concept as environmental responsiveness to market volatility [9], also Patil and Suresh [6] defined it as the ability to respond to customer needs and uncertain changes within the stipulated time.

From the literature review, the agile workforce is an essential facet of the overall agility level of the organization [3] since it allows to achieve different organizational benefits [5]. However, there was less focus on the theoretical and empirical validation of workforce agility enablers [5]. The main aim of the present paper is to identify the crucial factors influencing the workforce agility [6]. The following section develops a list of four workforce agility enablers by identifying their criteria and attributes from the literature [5].

2. Literature review: Enablers of the agile workforce

After reviewing the literature, four enablers have been identified as important factors playing a crucial role in improving the workforce agility.

* Fadoua Tamtam, Email: fadoua.tamtam@gmail.com

Individually, to be agile, the workforce needs to be adaptable or flexible. Then, an agile workforce needs to be proactive [6], which means self-anticipating the activities influencing positively the changes [2]. Also, workforce with innovative behavior is an agile workforce who can identify the need for a new product/service/process/technology or improving those existing.

From what was mentioned previously, the workforce handles multiple tasks or programs simultaneously, which means the workforce needs to focus on the most important ones for his work. If the working conditions are stressful, the workforce should be resilient. Other behaviors should be demonstrated by the agile workforce as getting knowledge of marketplace, business environment, organization operations and future priorities. These behaviors are grouped under the name business orientation. Also, it is important for agile workforce to achieve multiple competencies as those related to management, business process change, technical [6], information technology or software which were identified by Breu et al. [5,6]. Other researchers have identified responsiveness and intelligence as the main attributes of agility. Also, highly motivated and informed workforce is beneficial for the organization success.

Collectively, agile workforce needs to cooperate and share knowledge with internal or external groups which ensures fluid information, communication and knowledge flow across these groups which is important for agile team [6].

Table 1 summarizes enablers, criteria and attributes discussed in the workforce agility literature [5].

Table 1: Workforce agility enablers (Adapted from [2,3,6])

Workforce agility enablers	**Workforce agility criteria**	**Workforce agility attributes**
Workforce status (E_1)	Adaptability/ Flexibility **(E_{11})**	Simultaneously work on multiple work assignments within program or across different groups in organization (E_{111})
		Move quickly from role to role, to new tasks and responsibilities (E_{112})
		Engage and disengage often and easily with others with a singular focus on task accomplishment as per requirement (E_{113})
		Having interpersonal and cultural adaptability (E_{114})
		Acquire skills of professional flexibility (E_{115})
		Constant and quick learning of new skills, technologies, and procedures (E_{116})
		Acceptance rate of the workforce to altered working time or work locations (E_{117})
	Competency/ Self-awareness **(E_{12})**	Developing new skills or competencies within a short span (E_{121})
		Developing new skills in business process change (E_{122})
		Developing new management skills (E_{123})
		Developing new technical skills (E_{124})
		Developing new skills in information technology (E_{125})
		Developing new software skills (E_{126})
		Comprehension of new ideas, knowledge, or technologies (E_{127})
		Creativity and innovation in problem-solving (E_{128})
Workforce autonomy (E_2)	Proactivity **(E_{21})**	Dynamically explore new opportunities (E_{211})
		Identify and anticipate problems related to change (E_{212})
		Accomplish the promising goals (E_{213})
		Personal initiative (E_{214})
		Solution of the change related problems (E_{215})

	Innovation **(E_{22})**	Continuously work towards gaining proficiency in multiple competency areas (E_{221})
		Share the gained knowledge with its partners and collaborators (E_{222})
		Exploring into new markets (E_{223})
		Generates new ideas to tackle the unidentified change requests made by customers (E_{224})
		Generates new ideas to come out of the ambiguous situations faster (E_{225})
	Resiliency **(E_{23})**	Taking calculated risks (E_{231})
		Coming out comfortably from ambiguous situations quickly (E_{232})
		Positive attitude to change, to new ideas and technology (E_{233})
		Tolerance to stressful and uncertain situations (E_{234})
	Intelligence/ Responsiveness [6] **(E_{24})**	Quick response to changing customer needs and market conditions (E_{241})
		Sensing changes (E_{242})
		Execute things smoothly (E_{243})
	Quickness/ Speed **(E_{25})**	Shorter transition or recovery time (E_{251})
		Faster completion time (E_{252})
		Products or services delivery speed (E_{253})
		Problem-solving speed (E_{254})
Workforce job (E_3)	Focus **(E_{31})**	Set priorities while handling multiple programs (E_{311})
		Set priorities to drive towards solutions (E_{312})
		Demonstrate a strong sense of urgency to deliver a set goal (E_{313})
		Demonstrate right focus (E_{314})
	Business Orientation **(E_{32})**	Get knowledge of marketplace (E_{321})
		Get knowledge of business environment (E_{322})
		Get knowledge of the organization's operations (E_{323})
		Get knowledge of future priorities (E_{324})
		Get aligned with the organizational values (E_{325})
	Informative **(E_{33})**	Serious information seekers (E_{331})
		Keep informed in order to achieve the objectives or clarify problems (E_{332})
		Personally undertake research, analysis or investigation (E_{333})
		Use contacts or information networks to obtain useful information about technologies and processes related to programs they are working on (E_{334})
Workforce involvement (E_4)	Collaboration **(E_{41})**	Ability to collaborate with other teams, functions and organizations (E_{411})
		Avoid duplication of efforts (E_{412})
		Use decision-makers from different domains (E_{413})
		Smooth flow of knowledge and information across the boundaries of groups (E_{414})
		Willingness to enter unexpected collaborations (E_{415})

		Tolerance to different or new opinions of people from other disciplines (E_{416})
		Tolerance to different or new approach in collaboration (E_{417})
		Ease of communication (E_{418})
		Autonomy in collaboration (E_{419})
	Self-motivation/ Ownership **(E_{42})**	Self-motivation for seeking solutions to change issues (E_{421})
		Workforce needs possibility of growth (E_{422})
		Workforce needs recognition (E_{423})
		Workforce needs advancement (E_{424})
		Workforce needs technical supervision (E_{425})
		Pursue goals in the face of setbacks (E_{426})

3. Fuzzy logic approach to measure workforce agility of a Moroccan company

Acknowledging the importance of workforce agility, it has been assessed through several means such as [10] fuzzy logic, exploratory methodology, descriptive statistics, mathematical modeling and discriminate analysis [11].

Due to the ill-defined and vague indicators which exist within agility assessment [12], the decision-makers are unable to make a significant measurement when available information is scarce [13]. Thus, fuzzy logic has been widely used to handle real-life problems which are subjective, vague, and imprecise in nature [14]. These phenomena can be assessed only by linguistic values rather than numerical terms [13]. Using fuzzy concepts, each linguistic term can be associated with membership function [12]. When it is easy to determine an exact membership function we can use only type-1 fuzzy logic system, which means in our circumstances we don't need to evaluate the workforce agility measures by type-2 fuzzy logic system.

4. Fuzzy logic application

4.1. About the company

The proposed assessment has been done for an industrial company (hereafter referred as Indus_Comp) at southern part of Morocco [15]. Indus_Comp was started in 1999 and it designs and develops products dedicated to individual homes and professional projects. Currently Indus_Comp faces the challenges of modernization and competition and it should be able to offer an unlimited choice of products and service. In order to sustain in this agile environment, Indus_Comp needs to identify the weaker enablers [12] which can affect its competitive positioning.

In this context, and according to the hierarchical composition of the company, a questionnaire has been distributed to five decision-makers of Indus_Comp. The collected data are provided below [15].

4.2. Results and discussion

Each decision-maker has been instructed to express the suitable linguistic variables (Table 2) in terms of importance weight and performance rating against each agile workforce attribute. The linguistic scale has been furnished in Table 3. Then, these linguistic variables have been converted into fuzzy intervals (Table 4). The next step is to calculate the aggregated importance weights and aggregated performance ratings of the agile workforce criteria and enablers (Table 5). Finally, we obtained the fuzzy workforce agility index and fuzzy performance importance index of all the attributes (FPII) [15]. A detailed description of fuzzy logic application is given as follows [16]:

- Step 1: Identification of agile workforce enablers, criteria and attributes [17]: From the literature, four enablers, twelve criteria and sixty-four attributes were identified (Table 1).
- Step 2: Collection of performance rating and importance weights of agile workforce attributes: Five decision-makers (DM1, DM2,…, DM5) from Indus_Comp were asked to provide the required detail [15] in terms of linguistic variables (Table 2).

Table 2: Fuzzy calculation of agile workforce enablers

Agile Workforce	Importance weight					Performance rating				
	DM_1	DM_2	DM_3	DM_4	DM_5	DM_1	DM_2	DM_3	DM_4	DM_5
E_{111}	M	H	M	VL	VL	G	F	F	F	VP
E_{112}	H	H	FH	FH	H	G	F	F	F	VG
E_{113}	M	H	FL	M	M	VG	E	F	G	G
E_{114}	M	M	M	M	M	F	F	F	G	G
E_{115}	H	M	FH	M	FH	F	F	G	VG	VG
E_{116}	H	FH	M	H	FH	P	F	P	G	F
E_{117}	FH	H	FH	M	FH	W	W	P	F	P
E_{121}	H	H	M	FH	H	W	W	VP	F	F
E_{122}	M	H	H	FH	H	G	VG	E	E	G
E_{123}	VH	H	H	VH	H	G	F	G	G	VG
E_{124}	H	FH	H	FH	FH	E	E	G	F	E
E_{125}	H	FH	H	FH	H	G	G	G	VG	G
E_{126}	FH	M	FH	M	FH	F	G	F	F	G
E_{127}	H	M	M	M	H	G	G	G	VG	F
E_{128}	VH	H	H	VH	VH	VP	F	P	P	G
E_{211}	H	FH	M	FH	H	W	F	G	F	G
E_{212}	H	H	H	FH	H	G	VG	G	F	F

E_{213}	FH	M	FH	FH	M	W	W	F	G	VP
E_{214}	FH	FH	FH	FH	FH	VP	P	P	F	P
E_{215}	VH	H	H	VH	H	F	G	E	VG	E
E_{221}	H	FH	H	FH	FH	VP	P	W	F	F
E_{222}	FL	M	FL	FL	M	VG	F	E	F	G
E_{223}	H	FH	H	M	FH	W	VP	G	E	G
E_{224}	H	M	M	FH	H	VP	F	F	W	E
E_{225}	FH	H	H	H	FH	F	G	E	VG	G
E_{231}	H	FL	H	M	H	F	F	VG	F	G
E_{232}	M	M	FH	H	FL	G	E	F	G	F
E_{233}	H	VH	FH	FL	FL	E	VG	G	E	E
E_{234}	VH	M	VH	H	VH	VP	P	P	W	G
E_{241}	H	FH	VH	H	H	VP	G	E	G	G
E_{242}	H	FH	M	H	M	F	VG	VG	G	VG
E_{243}	FH	FH	FH	FH	FH	VP	VG	F	G	F
E_{251}	FH	M	FH	FH	M	W	W	VP	F	VP
E_{252}	H	FH	H	FH	M	G	F	G	VG	E
E_{253}	H	M	FH	H	H	W	W	G	F	G
E_{254}	FH	M	H	FH	FH	VP	P	W	P	F
E_{311}	H	FH	H	FH	M	G	F	P	VP	G
E_{312}	FH	FH	H	FH	H	P	W	F	G	P
E_{313}	H	H	H	H	H	W	W	F	F	P
E_{314}	VH	H	VH	VH	H	F	G	P	P	P
E_{321}	FH	H	FH	M	FH	E	VG	G	E	F
E_{322}	H	H	H	H	H	E	E	VG	F	F
E_{323}	FH	M	H	H	H	P	F	F	G	VG
E_{324}	H	FH	H	H	M	E	VG	G	F	G
E_{325}	H	H	H	H	H	E	F	G	F	F
E_{331}	H	M	M	M	M	W	W	F	G	F
E_{332}	FH	FH	H	H	FH	W	P	VP	G	F
E_{333}	H	FH	FH	H	H	E	E	E	G	E
E_{334}	H	M	FH	M	H	VP	P	F	G	F
E_{411}	H	FH	FH	FH	M	F	G	P	F	G
E_{412}	H	FH	H	FH	M	F	E	VG	G	F
E_{413}	H	FH	FH	H	H	G	F	F	E	G
E_{414}	H	H	FH	H	M	VP	F	F	F	F
E_{415}	FH	M	M	FH	FH	E	G	G	F	F
E_{416}	H	FH	H	FH	FH	G	E	F	F	VG
E_{417}	M	M	M	FH	FH	W	W	W	F	P
E_{418}	H	M	FH	FH	H	E	E	E	E	VG
E_{419}	H	FH	H	H	H	E	VG	F	P	F
E_{421}	M	H	M	M	FH	VG	G	F	E	E
E_{422}	FH	FH	M	M	H	W	G	P	F	F
E_{423}	H	M	M	M	FH	G	F	G	E	F
E_{424}	FH	H	H	H	H	G	F	F	F	VG
E_{425}	M	H	FH	H	H	E	G	F	P	VP
E_{426}	FH	H	M	H	FH	W	VP	P	VP	F

- Step 3: Matching the linguistic terms with the corresponding fuzzy intervals [16,17]: Table 3 presents linguistic terms and its appropriate fuzzy numbers for weight and ratings assignment [15] chosen from literature [10,16] .

Table 3: Fuzzy intervals for approximating linguistic terms (Adapted from [16])

Importance Weight		Performance Rating	
Linguistic variable	**Fuzzy number**	**Linguistic variable**	**Fuzzy number**
Very Low (VL)	(0, 0.05, 0.15)	Worst (W)	(0, 0.5, 1.5)
Low (L)	(0.1, 0.2, 0.3)	Very Poor (VP)	(1, 2, 3)
Fairly Low (FL)	(0.2, 0.35, 0.5)	Poor (P)	(2, 3.5, 5)
Medium (M)	(0.3, 0.5, 0.7)	Fair (F)	(3, 5, 7)
Fairly High (FH)	(0.5, 0.65, 0.8)	Good (G)	(5, 6.5, 8)
High (H)	(0.7, 0.8, 0.9)	Very Good (VG)	(7, 8, 9)
Very High (VH)	(0.85, 0.95, 1.0)	Excellent (E)	(8.5, 9.5, 10)

- Step 4: Application of fuzzy calculations on the weight and rating of each agile attribute, criteria and enabler. First, we calculate the average fuzzy weight and average fuzzy performance rating of each agile attribute [10] as shown in the following example.

E_{111} average fuzzy weight = [M+H+M+VL+VL]/5 = (0.3, 0.5, 0.7)/5, (0.7, 0.8, 0.9)/5, (0.3, 0.5, 0.7)/5, (0, 0.05, 0.15)/5, (0, 0.05, 0.15)/5 = (0.26, 0.38, 0.52)

E_{111} average fuzzy performance rating = [G+F+F+F+VP]/5 = (5, 6.5, 8)/5, (3, 5, 7)/5, (3, 5, 7)/5, (3, 5, 7)/5, (1, 2, 3)/5 = (3.0, 4.7, 6.4)

Then, we calculate the rating of each agile criterion [10]. An example of this calculation is demonstrated below.

Example: Rating of the criterion E_{11}=

[(3.0, 4.7, 6.4) ⊗ (0.26, 0.38, 0.52) ⊕ (4.2, 5.9, 7.6) ⊗ (0.62, 0.74, 0.86) ⊕ (5.7, 7.1, 8.4) ⊗ (0.36, 0.53, 0.70) ⊕ (3.8, 5.6, 7.4) ⊗ (0.3, 0.5, 0.7) ⊕ (5.0, 6.5, 8.0) ⊗ (0.46, 0.62, 0.78) ⊕ (3.0, 4.7, 6.4) ⊗ (0.54, 0.68, 0.82) ⊕ (1.4, 2.6, 4.0) ⊗ (0.50, 0.65, 0.80)] / [(0.26, 0.38, 0.52) ⊕ (0.62, 0.74, 0.86) ⊕ (0.36, 0.53, 0.70) ⊕ (0.3, 0.5, 0.7) ⊕ (0.46, 0.62, 0.78) ⊕ (0.54, 0.68, 0.82) ⊕ (0.50, 0.65, 0.80)] = (3.68, 5.27, 6.87)

Table 4: Fuzzy calculation of the criteria rating

Agile workforce	**Agile workforce attributes**	**Average fuzzy performance rating**	**Average fuzzy weight**	**Criteria rating**
E_{11}	E_{111}	(3.0, 4.7, 6.4)	(0.26, 0.38, 0.52)	(3.68, 5.27, 6.87)
	E_{112}	(4.2, 5.9, 7.6)	(0.62, 0.74, 0.86)	
	E_{113}	(5.7, 7.1, 8.4)	(0.36, 0.53, 0.70)	
	E_{114}	(3.8, 5.6, 7.4)	(0.3, 0.5, 0.7)	
	E_{115}	(5.0, 6.5, 8.0)	(0.46, 0.62, 0.78)	
	E_{116}	(3.0, 4.7, 6.4)	(0.54, 0.68, 0.82)	
	E_{117}	(1.4, 2.6, 4.0)	(0.50, 0.65, 0.80)	
E_{12}	E_{121}	(1.4, 2.6, 4.0)	(0.58, 0.71, 0.84)	(4.54, 5.97, 7.37)
	E_{122}	(6.8, 8.0, 9.0)	(0.58, 0.71, 0.84)	
	E_{123}	(5.0, 6.5, 8.0)	(0.76, 0.86, 0.94)	
	E_{124}	(6.7, 8.0, 9.0)	(0.58, 0.71, 0.84)	
	E_{125}	(5.4, 6.8, 8.2)	(0.62, 0.74, 0.86)	
	E_{126}	(3.8, 5.6, 7.4)	(0.42, 0.59, 0.76)	
	E_{127}	(5.0, 6.5, 8.0)	(0.46, 0.62, 0.78)	
	E_{128}	(2.6, 4.1, 5.6)	(0.79, 0.89, 0.96)	
E_{21}	E_{211}	(3.2, 4.7, 6.3)	(0.54, 0.68, 0.82)	(2.46, 3.58 4.79)
	E_{212}	(4.6, 6.2, 7.8)	(0.66, 0.77, 0.88)	
	E_{213}	(1.8, 2.9, 4.2)	(0.42, 0.59, 0.76)	
	E_{214}	(2.0, 3.5, 5.0)	(0.50, 0.65, 0.80)	
	E_{215}	(6.4, 7.7, 8.8)	(0.76, 0.86, 0.94)	
E_{22}	E_{221}	(1.8, 3.2, 4.7)	(0.58, 0.71, 0.84)	(3.83, 5.19, 6.53)
	E_{222}	(5.3, 6.8, 8.2)	(0.24, 0.41, 0.58)	
	E_{223}	(3.9, 5.0, 6.1)	(0.54, 0.68, 0.82)	
	E_{224}	(3.1, 4.4, 5.7)	(0.50, 0.65, 0.80)	
	E_{225}	(5.7, 7.1, 8.4)	(0.62, 0.74, 0.86)	
E_{23}	E_{231}	(4.2, 5.9, 7.6)	(0.52, 0.65, 0.78)	(4.49, 5.90, 7.25)
	E_{232}	(4.9, 6.5, 8.0)	(0.40, 0.56, 0.72)	
	E_{233}	(7.5, 8.6, 9.4)	(0.49, 0.62, 0.74)	
	E_{234}	(2.0, 3.2, 4.5)	(0.71, 0.83, 0.92)	
E_{24}	E_{241}	(4.9, 6.2, 7.4)	(0.69, 0.80, 0.90)	(4.84, 6.20, 7.53)
	E_{242}	(5.8, 7.1, 8.4)	(0.50, 0.65, 0.80)	
	E_{243}	(3.8, 5.3, 6.8)	(0.50, 0.65, 0.80)	
E_{25}	E_{251}	(1.0, 2.0, 3.2)	(0.42, 0.59, 0.76)	(2.85, 4.03, 5.32)
	E_{252}	(5.7, 7.1, 8.4)	(0.54, 0.68, 0.82)	
	E_{253}	(2.6, 3.8, 5.2)	(0.58, 0.71, 0.84)	
	E_{254}	(1.6, 2.9, 4.3)	(0.50, 0.65, 0.80)	
E_{31}	E_{311}	(3.2, 4.7, 6.2)	(0.54, 0.68, 0.82)	

	E_{312}	(2.4, 3.8, 5.3)	(0.58, 0.71, 0.84)	(2.47, 3.94, 5.47)
	E_{313}	(1.6, 2.9, 4.4)	(0.7, 0.8, 0.9)	
	E_{314}	(2.8, 4.4, 6.0)	(0.79, 0.89, 0.96)	
E_{32}	E_{321}	(6.4, 7.7, 8.8)	(0.50, 0.65, 0.80)	(5.29, 6.78, 8.15)
	E_{322}	(6.0, 7.4, 8.6)	(0.7, 0.8, 0.9)	
	E_{323}	(4.0, 5.6, 7.2)	(0.58, 0.71, 0.84)	
	E_{324}	(5.7, 7.1, 8.4)	(0.58, 0.71, 0.84)	
	E_{325}	(4.5, 6.2, 7.8)	(0.7, 0.8, 0.9)	
E_{33}	E_{331}	(2.2, 3.5, 5.0)	(0.38, 0.56, 0.74)	(4.01, 5.22, 6.44)
	E_{332}	(2.2, 3.5, 4.9)	(0.58, 0.71, 0.84)	
	E_{333}	(7.8, 8.9, 9.6)	(0.62, 0.74, 0.86)	
	E_{334}	(2.8, 4.4, 6.0)	(0.50, 0.65, 0.80)	
E_{41}	E_{411}	(3.6, 5.3, 7.0)	(0.50, 0.65, 0.80)	(4.62, 6.04, 7.41)
	E_{412}	(5.3, 6.8, 8.2)	(0.54, 0.68, 0.82)	
	E_{413}	(4.9, 6.5, 8.0)	(0.62, 0.74, 0.86)	
	E_{414}	(2.6, 4.4, 6.2)	(0.58, 0.71, 0.84)	
	E_{415}	(4.9, 6.5, 8.0)	(0.42, 0.59, 0.76)	
	E_{416}	(5.3, 6.8, 8.2)	(0.58, 0.71, 0.84)	
	E_{417}	(1.0, 2.0, 3.3)	(0.38, 0.56, 0.74)	
	E_{418}	(8.2, 9.2, 9.8)	(0.54, 0.68, 0.82)	
	E_{419}	(4.7, 6.2, 7.6)	(0.66, 0.77, 0.88)	
E_{42}	E_{421}	(6.4, 7.7, 8.8)	(0.42, 0.59, 0.76)	(3.81, 5.30, 6.74)
	E_{422}	(2.6, 4.1, 5.7)	(0.46, 0.62, 0.78)	
	E_{423}	(4.9, 6.5, 8.0)	(0.42, 0.59, 0.76)	
	E_{424}	(4.2, 5.9, 7.6)	(0.66, 0.77, 0.88)	
	E_{425}	(3.9, 5.3, 6.6)	(0.58, 0.71, 0.84)	
	E_{426}	(1.4, 2.6, 3.9)	(0.54, 0.68, 0.82)	

After that, we calculate the rating of each agile enabler by using the same calculation method (Table 5). For example, the rating of the enabler E_1 is calculated as

E_1= [(3.68, 5.27, 6.87) ⊗ (0.5, 0.65, 0.8) ⊕ (4.54, 5.97, 7.37) ⊗ (0.7, 0.8, 0.9)] / [(0.5, 0.65, 0.8) ⊕ (0.7, 0.8, 0.9)] = (4.18, 5.66, 7.13)

Table 5: Fuzzy calculation of the enablers rating

Agile workforce enablers	Agile workforce criteria	Criteria rating	Fuzzy importance weight of the agile workforce criteria	Enabler rating	Fuzzy importance weight of the agile workforce enablers
E_1	E_{11}	(3.68, 5.27, 6.87)	(0.5, 0.65, 0.8)	(4.18, 5.66, 7.13)	(0.5, 0.65, 0.8)
	E_{12}	(4.54, 5.97, 7.37)	(0.7, 0.8, 0.9)		
E_2	E_{21}	(4.54, 5.97, 7.37)	(0.5, 0.65, 0.8)	(4.11, 5.46, 6.80)	(0.5, 0.65, 0.8)
	E_{22}	(3.83, 5.19, 6.53)	(0.5, 0.65, 0.8)		
	E_{23}	(4.49, 5.90, 7.25)	(0.5, 0.65, 0.8)		
	E_{24}	(4.84, 6.20, 7.53)	(0.5, 0.65, 0.8)		
	E_{25}	(2.85, 4.03, 5.32)	(0.5, 0.65, 0.8)		
E_3	E_{31}	(2.47, 3.94, 5.47)	(0.5, 0.65, 0.8)	(4.08, 5.42, 6.74)	(0.5, 0.65, 0.8)
	E_{32}	(5.29, 6.78, 8.15)	(0.7, 0.8, 0.9)		
	E_{33}	(4.01, 5.22, 6.44)	(0.5, 0.65, 0.8)		
E_4	E_{41}	(4.62, 6.04, 7.41)	(0.5, 0.65, 0.8)	(4.21, 5.67, 7.07)	(0.5, 0.65, 0.8)
	E_{42}	(3.81, 5.30, 6.74)	(0.5, 0.65, 0.8)		

- Step 5: Calculation of the fuzzy workforce agility index of Indus_Comp [17]: Using the same calculation method, we obtained

Fuzzy workforce agility index = [(4.18, 5.66, 7.13) ⊗ (0.5, 0.65, 0.8) ⊕ (4.11, 5.46, 6.80) ⊗ (0.5, 0.65, 0.8) ⊕ (4.08, 5.42, 6.74) ⊗ (0.5, 0.65, 0.8) ⊕ (4.21, 5.67, 7.07) ⊗ (0.5, 0.65, 0.8)] / [(0.5, 0.65, 0.8) ⊕ (0.5, 0.65, 0.8) ⊕ (0.5, 0.65, 0.8) ⊕ (0.5, 0.65, 0.8)] = (4.14, 5.55, 6.93)

- Step 6: Matching the fuzzy workforce agility index with the appropriate linguistic label [16,17]: From the literature, we selected five linguistic labels (slowly agile, fairly agile, agile, very agile, extremely agile) (Table 6) [16]. Then, we convert the index into the suitable label using Euclidean distance method [10]. This method calculate the distance from the index to each agility level as shown below [16]:

Table 6: Linguistic labels and the corresponding fuzzy intervals (Adapted from [16])

Level of agility	Fuzzy intervals
Slowly Agile	(0, 1.5, 3)
Fairly Agile	(1.5, 3, 4.5)
Agile	(3.5 5 6.5)
Very Agile	(5.5, 7, 8.5)
Extremely Agile	(7, 8.5, 10)

D (FAI, Slowly Agile) = $\{(4.14-0)^2 + (5.55-1.5)^2 + (6.93-3)^2\}^{1/2}$ = 7.00

D (FAI, Fairly Agile) = $\{(4.14-1.5)^2 + (5.55-3)^2 + (6.93-4.5)^2\}^{1/2}$ = 4.40

D (FAI, Agile) = $\{(4.14-3.5)^2 + (5.55-5.0)^2 + (6.93-6.5)^2\}^{1/2}$ = 0.95

D (FAI, Very Agile) = $\{(4.14-5.5)^2 + (5.55-7)^2 + (6.93-8.5)^2\}^{1/2}$ = 2.53

D (FAI, Extremely Agile) = $\{(4.14-7)^2 + (5.55-8.5)^2 + (6.93-10)^2\}^{1/2}$ = 5.13

The minimum distance between the index and the linguistic label was obtained for the "agile" label [10], as it is showed in figure 1.

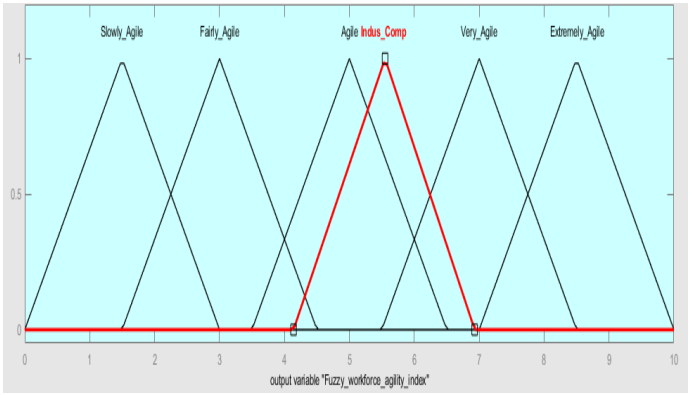


Figure 1: Fuzzy workforce agility index of Indus-Comp

- Step 7: Calculation of fuzzy performance importance index (FPII) and its score [10]: In order to improve the

agility level, we calculate the FPII of each agile workforce attribute. Then, the score of each FPII represents its contribution to the agility of workforce. The score of the 64 FPIIs are presented in table 7. If the score is weak it means that the attribute doesn't contribute to the overall workforce agility [16] and vice versa.

An example of $FPII_{111}$ and its score are calculated as:

$FPII_{111} = [(1, 1, 1) – (0.26, 0.38, 0.52)] \otimes (3.0, 4.7, 6.4) = (2.22, 2.91, 3.07)$

Score of $FPII_{111} = (2.22 + 4 \times 2.91 + 3.07) / 6 = 2.82$

Table 7: FPII and ranking score of the agile workforce attributes

Agile workforce attributes	FPII	Ranking score
E_{111}	(2.22, 2.91, 3.07)	2.82
E_{112}	(1.60, 1.53, 1.06)	1.46
E_{113}	(3.65, 3.34, 2.52)	3.25
E_{114}	(2.66, 2.80, 2.22)	2.68
E_{115}	(2.70, 2.47, 1.76)	2.39
E_{116}	(1.38, 1.50, 1.15)	1.42
E_{117}	(0.70, 0.91, 0.80)	**0.86**
E_{121}	(0.59, 0.75, 0.64)	**0.70**
E_{122}	(2.86, 2.32, 1.44)	2.26
E_{123}	(1.20, 0.91, 0.48)	**0.89**
E_{124}	(2.81, 2.32, 1.44)	2.25
E_{125}	(2.05, 1.77, 1.15)	1.71
E_{126}	(2.20, 2.30, 1.78)	2.20
E_{127}	(2.70, 2.47, 1.76)	2.39
E_{128}	(0.55, 0.45, 0.22)	**0.43**
E_{211}	(1.47, 1.50, 1.13)	1.43
E_{212}	(1.56, 1.43, 0.94)	1.37
E_{213}	(1.04, 1.19 1.01)	1.13
E_{214}	(1.00, 1.22, 1.00)	1.15
E_{215}	(1.54, 1.08, 0.53)	**1.06**
E_{221}	(0.76, 0.93, 0.75)	**0.87**
E_{222}	(4.03, 4.01, 3.44)	3.92
E_{223}	(1.79, 1.60, 1.10)	1.55
E_{224}	(1.55, 1.54, 1.14)	1.47
E_{225}	(2.17, 1.85, 1.18)	1.79
E_{231}	(2.02, 2.06, 1.67)	1.99
E_{232}	(2.94, 2.86, 2.24)	2.77
E_{233}	(3.82, 3.27, 2.44)	3.22
E_{234}	(0.58, 0.54, 0.36)	**0.52**
E_{241}	(1.52, 1.24, 0.74)	1.20
E_{242}	(2.90, 2.48, 1.68)	2.42
E_{243}	(1.90, 1.85, 1.36)	1.78
E_{251}	(0.58, 0.82, 0.77)	**0.77**
E_{252}	(2.62, 2.27, 1.51)	2.20
E_{253}	(1.09, 1.10, 0.83)	**1.05**
E_{254}	(0.80, 1.01, 0.86)	**0.95**
E_{311}	(1.47, 1.50, 1.12)	1.43
E_{312}	(1.01, 1.10, 0.85)	**1.04**
E_{313}	(0.48, 0.58, 0.44)	**0.54**
E_{314}	(0.59, 0.48, 0.24)	**0.46**
E_{321}	(3.20, 2.69, 1.76)	2.62
E_{322}	(1.80, 1.48, 0.86)	1.43
E_{323}	(1.68, 1.62, 1.15)	1.55
E_{324}	(2.39, 2.06, 1.34)	1.99
E_{325}	(1.35, 1.24, 0.78)	1.18
E_{331}	(1.36, 1.54, 1.30)	1.47
E_{332}	(0.92, 1.01, 0.78)	**0.96**
E_{333}	(2.96, 2.31, 1.34)	2.26
E_{334}	(1.40, 1.54, 1.20)	1.46
E_{411}	(1.80, 1.85, 1.40)	1.77
E_{412}	(2.44, 2.18, 1.48)	2.11
E_{413}	(1.86, 1.69, 1.12)	1.62
E_{414}	(1.09, 1.28, 0.99)	1.20
E_{415}	(2.84, 2.66, 1.92)	2.57
E_{416}	(2.23, 1.97, 1.31)	1.90
E_{417}	(0.62, 0.88, 0.86)	**0.83**
E_{418}	(3.77, 2.94, 1.76)	2.88
E_{419}	(1.60, 1.43, 0.91)	1.37
E_{421}	(3.71, 3.15, 2.11)	3.07
E_{422}	(1.40, 1.56, 1.25)	1.48
E_{423}	(2.84, 2.66, 1.92)	2.57
E_{424}	(1.43, 1.36, 0.91)	1.30
E_{425}	(1.64, 1.54, 1.06)	1.48
E_{426}	(0.64, 0.83, 0.70)	**0.78**

After calculating the scores of the 64 FPIIs, the Indus_Comp management was consulted to decide the appropriate threshold in order to rank the attributes. Subsequently, threshold 1.1 was identified as the management scale [10,16], and sixteen attributes have a lower score than 1.1 (Table 7):

- Acceptance rate of the workforce to altered working time or work locations ;
- Tolerance to different or new approach in collaboration ;
- Creativity and innovation in problem-solving [3].
- Developing new skills or competencies within a short span ;
- Developing new management skills ;
- Continuously work towards gaining proficiency in multiple competency areas ;
- Shorter transition or recovery time ;
- Products or services delivery speed ;
- Problem-solving speed ;
- Set priorities to drive towards solutions ;
- Demonstrate a strong sense of urgency to deliver a set goal ;
- Demonstrate right focus ;
- Keep informed in order to achieve the objectives or clarify problems ;
- Pursue goals in the face of setbacks [6].
- Solution of the change related problems ;
- Tolerance to stressful and uncertain situations [2].

5. Conclusions

Our purpose was evaluating workforce agility. In the theoretical part we presented a conceptual model including four enablers, namely workforce status, workforce autonomy, workforce job and workforce involvement. In the empirical part we used the fuzzy logic. First, we collected data from decision-makers, then selected the appropriate fuzzy numbers for interpreting the linguistic variables, and calculated the fuzzy workforce agility index. Based on Euclidean distance computation, the workforce of Indus_Comp was agile. However,

by calculating FPII and ranking different scores we identified the attributes needing some improvement so that the workforce of Indus_Comp could be extremely agile.

The contribution of this work is to help the companies to evaluate their existing level of the workforce agility by using the fuzzy logic. This method has take into account the ambiguity of each workforce agility enabler in order to calculate the overall workforce agility index. Then, the company could compare this index with that of competitors in order to identify its workforce attributes strengths and weaknesses.

Despite the above benefits, there are some limitations: In our case, the fuzzy logic depends mainly on the decision-maker only (one parameter) which excludes other employees perception who will have more detail on the technological or technical workforce attributes. Thus, taking into account other parameters can give a better tuning of results. Also, this method needs to be computerized in order to reduce time and errors that may be done in the calculation [15] or even to integrate the results provided by the fuzzy logic with other computing schemes like QFD (Quality Function Deployment), TISM (Total Interpretive Structural Modeling)…etc. For further research, this study should be conducted in different organizations of different sizes and sectors in order to carry out a comparative study. Apart from this, our conceptual model for workforce agility evaluation could also be extended in order to include other workforce aspects.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors acknowledge the financial support of the National Centre for Scientific and Technical Research (CNRST) under the Excellence Research Scholarships Program.

References

[1] F. Tamtam, A. Tourabi, "Agile capabilities in Moroccan companies: Criteria and practices," in 2019 4th World Conference on Complex Systems (WCCS), Ouarzazate, Morocco, 2019. https://doi: 10.1109/ICoCS.2019.8930721

[2] B. Sherehiy, W. Karwowski, "The relationship between work organization and workforce agility in small manufacturing enterprises," International Journal of Industrial Ergonomics, **44**(3), 466–473, 2014. https:// doi.org/10.1016/j.ergon.2014.01.002

[3] R. Qin, D. A. Nembhard, "Workforce agility in operations management," Surveys in Operations Research and Management Science, **20**(2), 55–69, 2015. https://doi.org/10.1016/j.sorms.2015.11.001

[4] R. Qin, D. A. Nembhard, "Workforce agility for stochastically diffused conditions—A real options perspective," International Journal of Production Economics, **125**(2), 324–334, 2010. https://doi.org/10.1016/j.ijpe.2010.01.006

[5] K. Breu, C. J. Hemingway, M. Strathern, D. Bridger, "Workforce agility: the new employee strategy for the knowledge economy," Journal of Information Technology, **17**(1), 21–31, 2002. https://doi.org/10.1080/02683960110132070

[6] M. Patil, M. Suresh, "Modelling the enablers of workforce agility in IoT projects: a TISM approach," Global Journal of Flexible Systems Management, **20**(2), 157–175, 2019. https://doi.org/10.1007/s40171-019-00208-7

[7] Z. Zhang, H. Sharifi, "A methodology for achieving agility in manufacturing organisations," International Journal of Operations & Production Management, **20**(4), 496–513, 2000. https://doi.org/10.1108/01443570010314818.

[8] A. Muduli, "Exploring the facilitators and mediators of workforce agility: an empirical study," Management Research Review, **39**(12), 1567–1586, 2016. https://doi.org/10.1108/MRR-10-2015-0236

[9] C. L. Bosco, "The relationship between environmental turbulence, workforce agility and patient outcomes," Ph.D Thesis, The University of Arizona, 2007.

[10] M. Suresh, R. Patri, "Agility assessment using fuzzy logic approach: a case of healthcare dispensary," BMC Health Services Research, **17**(1), 394, 2017. https://doi.org/10.1186/s12913-017-2332-y

[11] S. Alavi, D. Abd. Wahab, N. Muhamad, B. Arbab Shirani, "Organic structure and organisational learning as the main antecedents of workforce agility," International Journal of Production Research, **52**(21), 6273–6295, 2014. https://doi.org/10.1080/00207543.2014.919420

[12] S. Vinodh, S. R. Devadasan, B. Vasudeva Reddy, K. Ravichand, "Agility index measurement using multi-grade fuzzy approach integrated in a 20 criteria agile model," International Journal of Production Research, **48**(23), 7159–7176, 2010, https://doi.org/10.1080/00207540903354419

[13] C.-S. Tsai, C.-W. Chen, C.-T. Li, "Align agile drivers, capabilities and providers to achieve agility: a fuzzy-logic QFD approach," in Supply Chain, V. Kordic, Ed. I-Tech Education and Publishing, 2008.

[14] C. K. M. Lee, C. T. Y. Ru, C. L. Yeung, K. L. Choy, W. H. Ip, "Analyze the healthcare service requirement using fuzzy QFD," Computers in Industry, **74**, 1–15, 2015. https://doi.org/10.1016/j.compind.2015.08.005

[15] S. S. K. Sahu, "Agility evaluation in fuzzy environment," BTech Thesis, National Institute of Technology, 2013.

[16] C.-T. Lin, H. Chiu, Y.-H. Tseng, "Agility evaluation using fuzzy logic," International Journal of Production Economics, **101**(2), 353–368, 2006. https://doi.org/10.1016/j.ijpe.2005.01.011

[17] M. Deksnys, "Organizational agility in high growth companies," Ph.D Thesis, Mykolas Romeris University, 2018.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 414-425 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Hybrid Model for Coronary Heart Disease Prediction in Thai Population

Chalinee Partanapat*, Chuleerat Jaruskulchai, Chanankorn Jandaeng

Management of Information Technology, School of Informatics, Walailak University, Nakhon Si Thammarat 80161, Thailand

ARTICLE INFO

Article history:
Received: 14 July, 2020
Accepted: 08 September, 2020
Online: 21 September, 2020

Keywords:
Data Mining
Feature Selection
Coronary Heart Disease
Socioeconomic Status
Rediscretization on Clinical Values

ABSTRACT

The ability to verify the critical risk factors related to an effective diagnosis is very crucial for improving accuracy on coronary heart disease prediction. The objective of this research is to find the best predictive model for coronary heart disease diagnosis. Three approaches are set up to achieve the goals (1) investigating the classifier algorithms that are most suitable for the Thai heart disease dataset in this study (2) exploring features analyzed to be the significant risk factors in the predictive model, both major risk factors, and socioeconomic status and (3) rediscretizing the predefined clinical values on certain major risk factors. In order to achieving the optimal model before incorporating with feature selection process, several classifier approaches are conducted in this experiment. The study shows that the most effective classifiers ranked from the highest accuracy are Support Vector Machine, Naïve Bayes, Decision Tree, and Multi Layer Perceptron. Support Vector Machine produces the highest accuracy of 88.18%, with respect to both major risk factors and socioeconomic factors. Moreover, when adjusted thirteen major risk factors and five socioeconomic factors altogether, the accuracy is proved to be better than conducting each one alone. To investigate the better predictive performance of our study, feature selection methods of both filter and wrapper groups are employed with exploring the hybrid models to identify the most relevant features for Thai coronary heart disease. Relief Attribute Evaluation with Bayes Theorem is proved to be the best one with the accuracy of 92.59%, classified by SVM. To prove the accuracy enhancement, we perform rediscretization model on predefined medical values to examine different physical and personalized information of each person which can be incurred the coronary heart disease in different situation. The findings found that equal-depth rediscretization values on 7 major risk factors as Obesity, Hypertension, age, LDL, HDL, Fasting Blood Sugar, and Triglyceride, influences and improves with the better accuracy than predefined values of 95.50% classified by SVM. Thus, this finding shows that the proposed technique definitely outperforms predefined values from medical field.

1. Introduction

Coronary Heart Disease (CHD) is a chronic and fatal disease that is appeared to become the major health problem in Thailand. The most critical major risk factors causing such disease include Hypertension, HDL, LDL, Triglyceride, or diabetes. Base on literature reviews, many researchers attempted to utilize data mining techniques to enhance prediction ability on heart disease, which it was mostly based only on the major risk factors (MRF). Nevertheless, the factors related to socio-economic parts, called as socioeconomic status (SES), can also be influenced as one of crucial factors affecting on coronary heart disease. These factors can be found in various developing countries as well as in Thailand. The SES, like age and sex, can be mentioned after empirical demonstration that it can be a powerful predictor discovering the heart disease for Thai people [1]. Commonly, the high death rates from heart disease happened in Thailand can be caused by stress, hard works, and others. But even low SES can recently be found as influential factors related to heart disease problems for Thai patient [2]. In addition, the factors such as high blood pressure level, obesity, and smoking, lower education level, occupation as well as unhealthy life-styles can also affect heart disease [3]. Although previous researchers conducted experiment SES for heart disease prediction, they merely used statistical

*Corresponding Author: Chalinee Partanapat, Department of MIT, WU, Nakhon Si Thammarat 80161, Thailand, Email: joepooh14@gmail.com

methods to derive the study result. However, there have been few studies utilizing the data mining techniques. Thus, we need to find out which major or biological risk factors and SES factors that are the critical risk factors for Thai population. This study applied data mining techniques task as supervised machine learning methods via various feature selection models. Moreover, the rediscretization approaches are also used to find patterns that are hidden in the datasets to explore which risk factors can be the significant ones in prediction coronary heart disease for Thai population. In the last phase of this study, we proposed rediscretized methods for transforming the predefined medical values to novel rediscretized values to investigate the better accuracy for Thai heart study. The purposes of this research are to (1) explore the critical major and SES risk factors analyzed to be critical for Thai population (2) to rediscretize the predefined medical values on some of major risk factors.

2. Literature Review

For the last few years, many researchers have been attempting to employ advanced data mining technique approaches for prediction in medical field. In this study, the literature review is categorized into 3 paradigms. Firstly, explore heart disease prediction applying SES independently. The predictive model by applying Logistic Regression methods with Linear Regression using Statistical Software as SAS was conducted [4]. The result implied that gender and age are correlated with income levels and education, which is critical for incurring Coronary Heart Disease. Furthermore, the correlation between Risk of All-Cause Cardiovascular Mortality and Socioeconomic Status was performed [5]. The authors conducted the experiment by applying Twenty-Three risk factors. They employed association among socioeconomic status to evaluate the performance by Cox proportional hazard models using SAS Statistical Software. This research explored the correlation between lower social-economic status (SES) and worse conditions of health were directly related to the heart disease. The result revealed that each of the 23 socioeconomic status conducted was critically correlated, especially when assigning age with income.

Secondly, classification is applied in this study in order to classify Thai people who are healthy and who can be taken in to the consideration of heart disease. In [6], the authors developed a research on Effective Heart Disease Prediction by employing feature selection as association algorithm. They used Maximal Frequent Itemset Algorithm (MAFIA) classified by Non Linear Integrals to obtain the effective result. Further study of a predictive methodology on heart disease patient diagnosis based on decision tree J48 and Bagging Algorithms was presented [7]. Feature selection algorithms as Information Gain, GINI, and Gain Ratio were employed generated as ranking of most significant features. The highest performance is assessed on the basis of accuracy as 84.1%. Additionally, heart disease prediction applying j48, Bayes Net, SMO, and Multilayer perception with 10-fold cross validation using WEKA software was proposed [8]. SMO produced the highest accuracy of 89%. The heart disease prediction focusing on non- linear classification algorithm has also been developed [9]. They conducted with big data tool as Hadoop Distributed File System (HDFS) along with SVM for optimizing the set of feature. The research compared the performance of different data mining techniques. SVM in parallel fashion was executed as the best performance of prediction with 85% of accuracy.

Finally, discretization has been developed as applying for the first step in classification process in data mining approaches. In [10], the authors developed a discretization method and Information Gain on Naïve Bayes to diagnose heart disease. The authors used the heart disease dataset recruited from the UCI dataset composing of 270 instances and 14 features. The experiment was conducted with 10 fold cross validation. The results of the application showed that the accuracy of Naïve Bayes classifier was 85.18% while Naïve Bayes with discretization and information gain obtained the increased accuracy of 85.56%. Further study was conducted on the experiment of Heart Disease Prediction proposing different types of Decision along with discretization [11]. They evaluated the performance of alternative of decision trees by calculating sensitivity, specificity, and accuracy. The result showed that nine voting Equal Frequency Discritization with Gini Index Decision Tree produced the best accuracy with 85.3%, when compared with other methods without discretization.

3. Major Contribution of the Work

This study has the usefully applicable contribution both for data mining and medical field as (1) the effective performance of a hybrid model for coronary heart disease prediction for Thai population, which explores novel strategy achieved by combining feature selection approaches for choosing the most relevant attributes and rediscretization model, can be utilized for public health personnel as a primary prevention tool; and (2) a rediscretization model of this heart disease prediction system can be applied to the other diseases that have the same risk factors as stroke [12].

4. Proposed Method

This study proposed the hybrid models of feature selection methods and rediscretization on predetermined clinical values of major risk factors to prove the better performance on predictive accuracy on heart coronary disease. We firstly conducted hybrid of feature selection models to explore the most relevant attributes or risk factors that strongly related to the coronary heart disease. In the final of the study, the rediscretization on medical values was proposed to investigate the enhanced performance on prediction and to prove the strongly robust risk factors that can be the significant indicators for coronary heart disease in Thai population. In this phase, the stage of the proposed method includes data gathering, research framework, proposed hybrid models of prediction, and experiment process. Then, the algorithms of feature selectors chosen to develop the model for measuring the performance are described. Finally, the results from both hybrid predictive model and rediscretization model are compared for observing the prediction accuracy and are analyzed in discussions.

4.1. Dataset Collection

The dataset used in this work was recruited from Ramathibodi Hospital, Bangkok, Thailand. This Thai heart disease dataset

consisted of 20,000 records, composing of 11,240 male and 8,760 female. Additionally, this data set contains 1,205 record with CHD. There are two types of attributes, first is 13 major risk factors (attributes) and the other type is 5 SES. Detail values of each attribute and their corresponding predefined values by medical field are listed in Table1 and Table 2. The full dataset of participants were conducted in this research to explore the best fit classifiers for dataset in this study. We then performed the analysis on the dataset to determine the relation between baseline major risk factors and SES to incur CHD. The 13 major risk factors applied for the experiment are age, gender, HDL, LDL, Fasting Blood Sugar, Hypertension, Exercise, Smoking, Alcohol intake, Diabetes, Obesity, Hereditary, and Triglyceride. The 5 socioeconomic status used in the experiment are income, education, Occupation, Jobs, and Living Conditions.

The 13 major risk factors (attributes), which have the corresponding predefined values of medical field are listed in Table 1.

Table 1: Description of each attribute for major risk factors

No.	Attributes	Descriptions	Values
1	Gender	Male or Female	0: Female 1: Male
2	Age	Age in years	Continuous
3	HDL	Good cholesterol (mg/dl)	0: >= 40 1: < 40
4	LDL	Bad cholesterol (mg/dl)	0: <= 130 1: >130
5	fbs	Fasting Blood Sugar (mg/dl)	0: <= 120 1: >120
6	hp	Hypertension	0: No 1: Yes
7	Exercise	Physical Activity	0: No Exercise 1: Less Than 3d/wk 2: Greater or Equal than 3d/wk
8	Smoke	Smoking Habit	0: Non-Smoking 1: Smoking
9	Alcoholic	Alcoholic Habit	0: Non-Drinker 1: Drinker
10	Diabetes	Diabetes Mellitus	0: No 1: Yes
11	Obesity	Measured By BMI	0: No 1: Yes
12	Hereditary	Family Member Diagnosed with HD	0: No 1: Yes
13	Tg	Triglyceride (mg/dl)	0: 0 – 200 1: >200

The additional five risk factors related to socioeconomic status are employed to assess in the prediction analysis in order to derive improved result as well as effective prediction. Moreover, the result generated from the study can provide the identification on whether correlated with the major risk factors to cause the coronary heart disease. The factors concerning socioeconomic status and their representing values are detailed in Table 2.

4.2. Experiment settings

In this work, WEKA was used as a tool for preprocessing and classifying CHD dataset. For preprocessing step, Weka was facilitated on feature selection, both traditional and designed hybrid model, and was employed for discretization. Finally, classifier algorithms were used for prediction analysis.

Three experiments were conducted for this study. The experiment performance was assessed by Confusion Matrix, which it was the standard metrics to measure the accuracy. In the meantime, the performances derived from the classifiers were also compared in order to see the different views of the results.

The experiment that had been done in the first time had the objective to identify the classification algorithms that can effectively investigate the most relevant dataset of Thai heart disease. In this stage, the study incrementally studied and separately evaluated 13 factors related to biological (MRF) perspectives and 5 factors about the socioeconomic status (SES). Next step, both MRF and SES were assessed together. The predictive performance as accuracy in this designed experiment was examined the effective measurement between conducting MRF and SES separately and adjusting both of them. Lastly, the comparison for both scenarios on the derived result was also conducted.

Table 2: Description of each attribute of socioeconomic risk factors

No.	Attributes	Descriptions	Values
14	Income	Income levels (THB)	0 = <=10,000 1 = >10,000 - 20,000 2 = >20,000 - 50,000 3 = >=50,000
15	Edu	Education Levels	0 = Secondary 1 = Vocational 2 = Bachelor 3 = Higher than Bachelor
16	Occupation	Occupation Levels	0 = Manager 1 = Intermediate 2 = Self-Employed 3 = Technical
17	Jobs	Employment status	0 = Unemployed 1 = Employed
18	liv. cond.	Living condition	0 = Rented 1 = Owner

The second experiment focused on applying various feature selection models to CHD dataset. This experiment was designed to evaluate the efficiency of classification by examining performance of filter and wrapper approaches, including proposed hybrid models pioneered from combined feature selector methods to improve the classification accuracy. In this section, 10 filter algorithms and wrapper approach were applied. Furthermore, the hybrid models generated from combined feature selectors were developed. All of filter algorithms were classified by 4 classifiers and the accuracy of each were compared. For wrapper approaches, and the hybrid ones, were proposed to investigate the

enhancement of predictive performance. Feature selection methods employed in this study are as followings:-

- Correlation Based Feature Selection (CFS): Typically, CFS algorithm evaluates strong features subsets containing features that are highly correlated with the class [13]. We mainly performed both CFS and CFS with other three feature selection approaches in this study. 18 risk factors were used to identify the robust and relevant factors. The factors that executed the value highly related to the class selected by CFS were age, gender, LDL, hypertension, smoking, diabetes, obesity, triglyceride, income, and education.
- Bayes Theorem: based on statistical technique describing the usage of the Theory of conditional probability, which assume with possible causes for a given observed event. It can be computed from the knowledge that has already been occurred. It identifies the relation between event and probability [14]. Applying Bayes' theorem, assume that one would like to investigate whether the prospective participants have a chance to get the heart disease by knowing their triglyceride. Triglyceride is the critical causes of CHD, the participants' triglyceride information can be determined the probability of CHD occurrence.
- CFS and Bayes Theorem: A novel hybrid feature selection model generated from CFS with Bayes Theorem combination was proposed. The features selected by CFS algorithm will be assigned to Bay's theorem for conclusively better reduction. Correspondingly, the insignificant features will be removed. Bay's theorem selects the feature with the highest conditional probability. In this study, CFS with Bayes Theorem is proved to give the better accuracy than the single selector approach.
- Chi-squared attribute evaluation: A Chi-squared feature selection is normally applied to test the independence of two variables or features. The aim of this feature selection is to select the features which are significantly dependent on the other feature or the response. When two features are significantly different, the chi-square value will be smaller, and it can be implied that features are not related to each other as it is not dependent on the response [15]. In this study, we focused on exploring the critical risk factors that are correlated with CHD for Thai people.
- Random Tree: Algorithms used for selecting features by randomized method. It works effectively on large dataset as having the greatest solving problem in impurity, while reducing time for training model and achieving accuracy improvement [16]. This selector algorithm produced the highest accuracy when combined with CFS to determine the critical risk factors of Thai CHD.
- Info Gain Attribute Evaluation: An algorithm proposed for feature ranking and feature reduction evaluated by Entropy, which characterizes the impurity of a collection of examples [17]. In this study, the risk factors selected by this method are reduced to 8 factors with higher accuracy than without feature selection.
- SVM Attribute Evaluation: This approach evaluates the features, which are ranked with respect to the square of coefficients as linear using SVM classification algorithm [18].
- Wrapper Subset Evaluation: A learning algorithm that are applied to evaluate the feature subsets and use the predictive model to score them due to the accuracy to detect the possible relationships between variables [19]. In this study, we applied wrapper approach including Genetic Algorithm and Sequential Search to solve the problem of performing of intensive calculation.

The third experiment was developed to assess and to compare the new rediscretized values with the predefined medical values. In this experiment, we proposed rediscretization model on 7 major risk factors (MRF) that have continuous and predefined categorical values. The new rediscretized values are automated binning by WEKA, starting from 2 to 10 bins, before feeding to the best feature selector from the previous experiment. Lastly, the accuracy of rediscretized values of each risk factor will be compared with the corresponding predefined medical values.

4.3. Designed Framework

The framework designed for this research was proposed focused on hybrid feature selection approaches for predicting coronary heart disease, which assembled numerous methods of attribute selectors via classification task to enhance the accuracy of prediction. The various hybrid feature selection models were combined from two different attribute selectors. All of these hybrid models, proposed to prove the improvement in accuracy, were generated into prospective classifiers to compare the predictive performance. Rediscretization was conducted at the last step to investigate the highest accuracy. The steps of designed framework were detailed and portrayed in Figure 1.

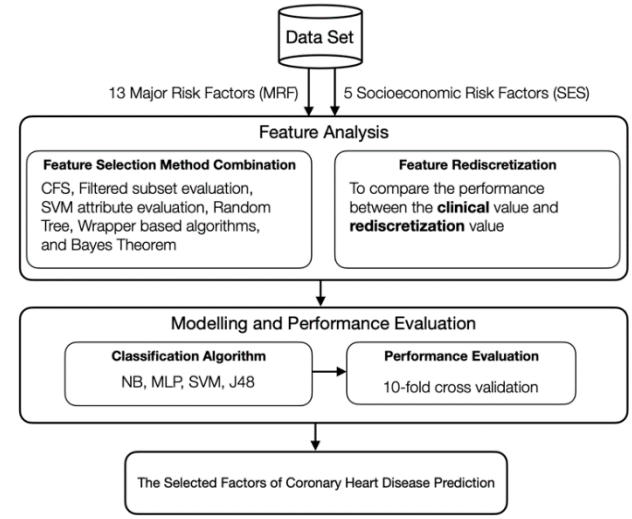


Figure 1: Proposed Framework of Predictive Model for CHD

In the phase of feature analysis, the In the phase of feature analysis, the research employed 18 feature selection models, including 10 filter approaches, wrapper models, and hybrid models in order to attain the significant attributes. The number of relevant or critical attributes drawn by these feature selector approaches would be fed into 4 classifiers, consisting of Naïve Bayes [20], Decision Tree J48 [21], Multilayer Perceptron (MLP), and Support Vector Machine (SVM) [22]. These 4 classifiers were proved to be the best ones resulting from the first study stage. Accordingly, the performance evaluation of different classification algorithms for each feature selector approach was compared and analyzed. Nevertheless, several significant MRF derived from attribute selection methods related to lifestyle such as no exercise, eating discipline, or personalization were

discovered to have the different effect on each individual. Hence, the study employed rediscretization model for predefined clinical values regarding MRF in order to investigate the improved predictive performance.

4.4. Experimental Results

The experimental result was divided into 3 parts corresponding to the experimental set up, which can be demonstrated as followings

4.4.1. Experiment 1

Before feature selection, we conducted 6 classifiers to find the best fit for our dataset, which were Naïve Bayes, K-Nearest Neighbor (KNN), J48, Neural Network (NN), MLP, and SVM, and we found that KNN and NN gave the accuracy less than 80% for every type of risk factors. Finally, only 4 classifiers were selected, due to predictive performance higher than 80%. Also the performance evaluation as accuracy was compared on 5 factors of socioeconomic status (SES) with 13 major risk factors (MRF). The experimental result obtained clearly revealed that that the entire 13 MRF produced much better accuracy performance than only socioeconomic factors. Classification algorithm as SVM obviously executed as the highest classifier compared with the other three. Accordingly, predicting heart disease considering only SES in Thai people, the predictive performance cannot be accepted due to low accuracy. The comparison of predictive performance of these four classifiers between all MRF and SES factors are shown in Table 3.

Table 3: Comparison of accuracy between all major risk factors (MRF) and all socioeconomic risk factors (SES)

Classifiers	Accuracy (%)		
	MRF	SES	MRF+SES
Naïve Bayes	86.70	47.33	86.14
J48	83.17	43.33	84.88
MLP	82.50	54.68	83.66
SVM	86.88	56.78	**88.18**

From Table 3, the result showed that after adjusting all MRF and SES together, the predictive performance is a little better than that of MRF itself. It indicates that when considered Socioeconomic status (SES) as a predictor of coronary heart disease independently of the major risk factors as gender, age, diabetes, hypertension, smoking, physical activity, cholesterol, and body weight, SES itself is not a powerful determinant for health problem even it is inversely associated with risk for CHD. For example, low education or low income can be proved to have a chance of occurrence of CHD but not directly contributed to CHD without analyzing the major risk factors together. Accordingly, this finding found that consideration of SES with major risk factors would address more CHD risk. The performance of all 18 risk factors and each type of risk factors are compared and are evaluated by four classifiers. The comparison of results is represented in table 3.

According to Table 3, the result confirmed that adjusting major risk factors with SES provides the better accuracy. The above result was performed before process of feature selection, SVM classifier produced the best accuracy of 88.18%. The result examined the relationship between major risk factors and SES that when these two types of risk factors were assessed together, they affect the accuracy more than considered SES alone. Low education was defined as a low SES. If considered with major risk factors, it is strongly related with smoking [23].

4.4.2. Experiment 2

After achieving the performance evaluation before feature selection, the number of irrelevant attributes was minimized by exploring four feature selector approaches, including CFS, Filtered Subset Evaluation, SVM Attribute Evaluation, and Random Tree respectively. The critical relevant attributes obtained from each attribute selector approach were presented in Table 4.

Based on the results given in Table 4, among 4 feature selector approaches, Random Tree produced the minimum number of risk factors, which decreased to 7 factors. The factors including gender, age, smoking, LDL, and education were then accounted as the critical factors influencing CHD incurrence for Thai people. Accordingly, the finding discovered that all classifiers produced the higher accuracy after removing irrelevant attributes.

Subsequently, we proposed three hybrid feature selection approaches, which were developed from combining CFS with Filtered subset evaluation, CFS with SVM attribute evaluation, and CFS with Random Tree. Accordingly, the study proved that the three proposed hybrid models produced the better predictive accuracy. The relevant attributes or the critical risk factors achieved from three proposed hybrid feature selection models were shown in Table 5.

From the result given in Table 5, the best performance measurement yielded to SVM classifier with the highest accuracy of 90.55%, selected by CFS with Random Tree attribute selector approaches.

As a result displayed in Table 4 and Table 5, we observed the finding and found that the combination of two feature selection approaches called hybrid models produced the better accuracy than applying a single feature selector. Furthermore, Education which is the SES factors is appeared in every selector approach, both a single selector method and hybrid models. Thus, the result in this phase can be implied that low education level is considered as a critical risk factor related to SES and has the direct effect on CHD in Thai people. Consequently, Thai people with less education have more chance to raise CHD.

The findings indicated that classifiers give the improved accuracy after utilizing the hybrid feature selector approaches. The crucial factors including age, gender, LDL, smoking and low education can raise CHD of Thai people. Additionally, diabetes was found that it can also have an effect on Thai heart disease related with age, sex, cholesterol and smoking.

After results of hybrid feature selectors conducted in filter group were shown in Table 5, there were other feature selection in another group as wrapper subset evaluation of wrapper group was also implemented. Herewith, other five feature selections including wrapper subset evaluation (WSE), information gain attribute evaluation (IGAE), qui-squared attribute evaluation (CAE), gain ratio attribute evaluation (GRAE), one attribute evaluation (OAE), and relief attribute evaluation (RAE) had been employed as they were accepted to investigate the enhanced accuracy of prediction of factors affecting CHD occurrence in Thailand. The reduced selected factors are displayed in the Table 6.

Table 4: Number of selected attributes by each attribute selection method and attributes selected

Selection Method	Selected Attributes	Accuracy (%)				
		Naïve Bayes	J48	MLP	SVM	Average
CFS subset evaluation	10 (1, 2, 4, 6, 8, 10, 11, 13, 14, 15)	87.70	85.14	84.74	88.00	85.14
Filtered subset evaluation	8 (1, 2, 3, 6, 8, 10, 13, 15)	87.14	83.60	80.88	85.60	83.38
SVM attribute evaluation	8 (1, 2, 5, 8, 10, 14, 15)	88.00	85.74	85.50	88.74	85.94
Random Tree	7 (1, 2, 3, 4, 6, 8, 15)	88.18	86.00	87.11	**89.66**	**86.59**

Table 5: Selected features by CFS with Filtered subset evaluation, CFS with SVM attribute evaluation, and CFS with Random Tree

Selection Method	Selected Attributed	Accuracy (%)				
		Naïve Bayes	J48	MLP	SVM	Average
CFS	10 (1, 2, 4, 6, 8, 10, 11, 13, 14, 15)	87.70	85.14	84.74	88.00	85.14
CFS + Filtered subset evaluation	8(1, 2, 4, 6, 8, 10, 13, 15)	88.18	85.62	88.62	86.00	85.63
CFS + SVM attribute evaluation	7(1, 2, 4, 5, 8, 10, 15)	86.18	88.91	88.14	89.91	87.12
CFS + Random Tree	6(1, 2, 4, 8, 10, 15)	88.92	89.14	89.71	**90.55**	88.39

Table 6: The selected attributes by attribute selection methods

Selection Method	Selected Attributes	Accuracy (%)				
		Naïve Bayes	J48	MLP	SVM	Average
Wrapper Subset Evaluation	7(1, 2, 4, 8, 10, 15)	87.50	85.82	84.86	88.54	86.93
Info Gain Attribute Evaluation	8(1, 2, 4, 5, 8, 10, 13, 15)	87.00	85.18	84.66	87.74	86.15
Chi-squared Attribute Evaluation	9(1, 2, 3, 4, 5, 6, 8, 10, 13, 15)	83.70	81.66	82.88	85.50	83.44
Gain Ratio Attribute Evaluation	8(1, 2, 4, 6, 8, 10, 13, 15)	87.00	84.70	85.18	87.50	86.10
One Attribute Evaluation	8(1, 2, 4, 5, 8, 10, 13, 15)	87.50	85.18	86.88	88.00	86.89
Relief Attribute Evaluation	7(1, 2, 4, 8, 10, 15)	88.89	85.97	87.50	**89.50**	**87.97**

Table 6 displayed the result of selected attributes from feature selection algorithms mentioned above. There are 5 crucial biological factors as age, sex, LDL, diabetes, smoking, and socioeconomic factors as education accepted as the crucial factors as picked in every selector approach. This study implies that gender, age, LDL, diabetes, smoking and levels of education are the critically related factors causing CHD for Thai population.

In addition, the study revealed that Triglyceride was appeared to show the impact on prediction and could have the critical effects on Thai Heart Disease, associated with age, sex, cholesterol and smoking and diabetes.

The selected attributes were picked into classification algorithms to investigate the prediction accuracy. The accuracy of related factors towards CHD can be seen in the Table 6, demonstrating comparative results.

From Table 6, after performing the hybrid feature selector approaches, classifiers produced the enhanced accuracy. Relief attribute evaluation method (RAE) classified by the SVM produced best accuracy as of 89.50%, indicating that age, sex, LDL, smoking, diabetes, and low education can affect CHD occurrence. Further analysis on the study result from six methods indicated that the diabetes can be one of crucial factors influencing coronary heart disease for Thais.

After using feature selection from filter group, we attempted to use other wrapper feature selections to study the related factors influencing CHD and then compare its performance evaluation with feature selectors in filter group. The studied wrapper models are then conducted and explained. First, the wrapper algorithms that have been conducted together with Baysian classifier were calculated to work out on average accuracies. Furthermore, the other two wrapper models as Forward Selection (FS) wrapped methodologies were experimented on feature subsets to assess the prediction performance about the coronary heart disease factors. The first one uses the Best First Search (BFS) which is the search algorithm that explores the graph by expanding the promising node with the best score [24]. The second one we generate the subset of features using the Sequential Floating Forward Search

(SFFS) [25]. The principle of such technique is to add one or more attributes progressively. For SFFS, after each step forward, backward steps would be applied to improve the efficiency of wrapper selector. The proposed wrapper with GA approach is also compared with four classifiers.

The finding suggested that gender, age, LDL, diabetes, smoking and education are the critical factors as they were selected in every feature selection method, and Triglyceride was also significantly involved as the major cause of CHD.

The list of features or risk factors for heart disease dataset generated by each of wrapper approach, which are Genetic Algorithm [26], Best First Search, Sequential Forward Floating Search, and without feature selection, is shown in Table 7.

Table 7: The Selected Features by different wrapper Techniques

	Feature	Selected Features					
		GA				BFS	SFFS
		NB	SVM	MLP	J48	NB	NB
1	Gender	✓	✓	✓	✓	✓	✓
2	Age	✓	✓	✓	✓	✓	✓
3	HDL	✓	✓		✓		
4	LDL	✓		✓		✓	✓
5	FBS		✓	✓		✓	
6	Hypertension		✓	✓	✓		
7	Exercise	✓		✓			
8	Smoke	✓	✓		✓	✓	✓
9	Alcoholic	✓	✓	✓			
10	Diabetes	✓	✓	✓	✓		
11	Obesity		✓				
12	Hrd.	✓	✓				
13	Tg	✓	✓		✓		
14	Income	✓	✓	✓			
15	Edu		✓	✓	✓	✓	✓
16	Ocp.						
17	Jobs		✓			✓	✓
18	Liv.Cond						

As the consequence of employing the wrapper feature selection method, the Table 8 displays the results related to the accuracy that has been derived by implementing 10-fold cross validation. Applying GA wrapper generated with BN yields the highest effective prediction accuracy as of 89.50%, which is accounted as the most accurate value compared to other feature selection methods. It is interpreted that the GA wrapper operated with Genetic Algorithm classified by Naïve Bayes can explore the most relevant and strongest features. In addition, the accuracy of classification algorithms with different wrapper feature selection are also mentioned and displayed in the Table 8.

As the result shown from wrapper feature selection group in Table 8, it can be again confirmed that the GA wrapper with BN classifier can produce the highest predictive accuracy of CHD as 89.50%. After investigating with wrapper methods, the other filter feature selectors were applied to reduce the irrelevant risk factors.

After that, the six-inconsequence hybrid models with Bayes theorem, which include CFS, Info Gain Attribute Evaluation, Chi-squared Attribute Evaluation, Gain Ratio Attribute Evaluation, One Attribute Evaluation and Relief Attribute Evaluation, have been experimented and their results obtained the improved predictive accuracy. The attributes attained from these selection methods are shown in the Table 9.

Table 8: The performance evaluation of wrapper-based feature selection algorithms

Wrapper Method	Accuracy (%)			
	Naïve Bayes	J48	MLP	SVM
GA Wrapper	**89.50**	88.82	85.60	84.55
BFS Wrapper	87.50	84.53	83.50	84.18
SFFS Wrapper	88.18	87.17	81.89	83.50
Without FS	86.50	84.50	80.88	80.12

Referring to the hybrid feature selection models in Table 9, the numbers of attributes deducted after applying Chi-squared attribute evaluation is 9, One attribute evaluation is 8, Gain Ratio attribute evaluation is 8, and Relief attribute evaluation is 7, respectively. These selected risk factors are processed into Bayes Theorem. The CFS method with Bayes Theorem generates 6 attributes. The info gain attribute evaluation method with Bayes theorem produces 7 attributes. The chi-squared attribute evaluation method with Bayes Theorem gives 8 attributes. The gain ratio attribute evaluation method with Bayes Theorem generates 7 attributes. The one attribute evaluation method with Bayes Theorem executes 7 attributes. And lastly, the relief attribute evaluation method with Bayes Theorem produces 6 attributes. The results also indicated that critical factors influencing CHD for Thai people can count on gender, age, LDL, smoking, and education. Nevertheless, this chosen dataset achieved from the hybrid models is summarily proved in accuracy improvement classified by four classification algorithms. The comparison of classifiers accuracy with deducted attributes is presented in Table 9.

As a result obtained in Table 9, when adjusting Bayes with these six feature selectors, the predictive accuracy is absolutely proved an improved performance. Relief attribute evaluation (RAE) with the attribute selector Bayes Theorem classified by SVM yields the highest accuracy of 92.59%. The attributes including age, sex, LDL, smoking, and low education are counted as the significant factors affecting the coronary heart disease for Thai people. Beyond this, the diabetes is revealed to be one of the critical factors conclusively associated with age, sex, cholesterol and smoking raising the CHD for Thai population.

4.4.3. Experiment 3

In the last step of our study, according to the values of some risk factors are continuous values or have the predefined range of value from the medical theory, which is not well suitable for nowadays lifestyle of Thai people. For example, attribute such as age, is shown in the clinical value as continuous, which is inappropriate for predictive analysis in medical domain and have no meanings to be the input for heart disease prediction. In this section, we propose the rediscretization method to improve the accuracy on predictive model for CHD of prediction. The proposed methodology involved systematically testing discretization techniques.

Discretization is one of the preprocessing techniques that is used for transforming the different continuous attributes, variables to discrete or nominal attributes [27]. In addition,

Table 9: The selected attributes by hybrid selection methods **with Bayes Theorem**

Selection Method With Bayes Theorem	Selected Attributed	Accuracy (%)				
		Naïve Bayes	J48	MLP	SVM	Average
CFS	6(1, 2, 4, 8, 10, 15)	89.88	86.60	87.11	**90.94**	88.63
Info Gain Attribute Evaluation	7(1, 2, 4, 8, 10, 13, 15)	88.50	86.88	87.50	89.88	88.19
Chi-squared Attribute Evaluation	8(1, 2, 4, 5, 8, 10, 13, 15)	85.92	83.29	84.07	87.50	85.20
Gain Ratio Attribute Evaluation	7(1, 2, 4, 8, 10, 13, 15)	88.00	85.88	86.50	88.67	87.26
One Attribute Evaluation	7(1, 2, 3, 4, 6, 8, 15)	88.18	86.50	87.88	89.50	88.02
Relief Attribute Evaluation	6(1, 2, 4, 8, 10, 15)	90.50	86.87	88.50	**92.59**	89.62

Table 10: Comparison of accuracy between Equal Width Binning and Equal Depth Binning with Age Attribute

No. of Bins	Accuracy (in%)	
	Equal Width Binning	Equal Frequency Binning
2	89.50	88.18
3	**91.00**	**95.33**
4	90.55	94.00
5	88.50	93.66
6	87.18	90.28
7	85.33	87.33
8	84.10	85.50
9	81.18	83.70
10	79.88	81.28

Table 11: Discretization on continuous and some of numeric factors of 7 risk factors

No.	Attributes	Descriptions	Clinical Value	Discretized Value
2	Age	Age (years)	Continuous	0: 0 – 37 1: 38 – 50 2: >= 51
3	HDL	Good cholesterol (mg/dl)	0: >= 40 1: < 40	0: >= 60 1: 50 – 59 2: 40 - 49 3: < 40
4	LDL	Bad cholesterol (mg/dl)	0: <= 130 1: >130	0: <= 100 1: 100 – 130 2: 130 – 160 3: >= 160
5	fbs	Fasting Blood Sugar (mg/dl)	0: <= 120 1: > 120	0: <= 125 1: >125
6	hp	Hypertension (mm/hg)	0: No 1: Yes	0: <= 125 1: 121 – 150 2: > 150
11	Obesity	Measured By BMI	0: No 1: Yes	0: 19-22 1: 23-25 2: 26-30 3: > 30
13	Tg	Triglyceride (mg/dl)	0: 0 – 200 1: > 200	0: 0 – 130 1: > 130-179 2: ≥180 – 259 3: > 259

several classifiers entail that some would work better on discretized or binarized data as the valuable meanings in the new range of values.

At the earlier stage of this experiment, we performed the simplest unsupervised discretization methods on prospective risk factors based on equal width and equal frequency discretization. Hence, the methodology of the research work is equal width binning and equal frequency binning to investigate which one produced the better accuracy for our heart disease dataset. The primarily results showed that equal frequency gave the higher accuracy than equal width. Thus, in this study, the equal depth (Equal Frequency) discretization was selected to conduct the experiment to help improving the predictive performance. To confirm the result, attribute "Age" was exemplified. The equal width method divided the range into the user-defined number of approximately equal size of intervals.

The equal frequency method divides the values into k groups and each group contains approximately the same number of training values [28]. The new discretized value of each factor will be automated binning. The number of bins was generated by WEKA tool. We tried 9 numbers of bins starting from 2 bins to 10 bins. Table 10 contains the measurement of accuracy regarding to the number of bins generated.

This study specifies the parameter by the number of bins and how many values should be included in each bin or each interval. According to the result shown in Table 10, Equal Frequency discretization approach gave the better accuracy and the bin that achieved the highest accuracy was selected to perform rediscretization on the final experiment. Rediscretization on medical values designed to 7 attributes or risk factors is shown in Table 11.

According to the Table 11, first four major risk factors, which are age, HDL, LDL, and fasting blood sugar has been rediscretized. The rediscretized values of attribute "Age" was categorized as followings: the range of 0-37 years old contains 6,132 people, which is accounted for 31%. The range of 38-50 years old contains 8,258 people, which is accounted for 41%. The range of age more than 50 years old contains 5,610 people, which is accounted for 28%. This range of rediscretized values of attribute "Age" gives the best accuracy when compared with any other number of bins. We tried 2 to 10 bins of discretization and the number of 3 bins produced the best performance. Thus, the rediscretized values of 3 bins were selected for the study. The attribute "HDL" was rediscretized into 4 ranges. The range of more than or equal to 60 mg./dl contains 3,232 people, which is accounted for 16%. The range of 50-59 mg./dl contains 5,162 people, which is accounted for 26%. The range of 41-49 mg./dl contains 6,221 people, which is accounted for 31%. The range of less than or equal to 40 mg./dl contains 5385 people, which is accounted for 27%. Similar to attribute age, we tried rediscretization from 2 bins to 10 bins and the selected 4 bins gave the best accuracy. The attribute "LDL" was categorized into 4 bins of new range of values. The range of less than or equal to 100 mg./dl contains 3,657 people, which is accounted for 18%. The range of 101-130 mg./dl contains 4,953 people, which is accounted for 25%. The range of 131-160 mg./dl contains 5,745 people, which is accounted for 29%. The range of less than or equal to 40 mg./dl contains 5,645 people, which is accounted for 28%. Similar to attribute age, we tried rediscretization from 2 bins to 10 bins and the selected 4 bins produced the best performance evaluation. The attribute "Fasting Blood Sugar" was rediscretized to the new range of values as 2 bins. We selected this bin as the best accuracy. The first range is less than or equal to 125 mg/dl, which involves 12,240 people. Another range is more than 125 mg/dl, which involves 7,760 people.

Furthermore, as the experiment conducted in the phase of feature selection, hypertension is one of the critical risk factors for incurring the chance of occurrence for CHD. Thus, this study applied additional risk factor as hypertension for rediscretization. Many previous researches defined and divided the range of systolic blood pressure (SBP) in different range of values due to the purpose and the main idea of their researches. Generally, Hypertension was categorized according to blood pressure readings and the examination by physician is usually performed under the multivariate period of time [29]. Hypertension is considered related to several factors, such as age, body mass index, blood pressure, smoking, family history, and physical inactivity [30]. In our study, the blood pressure determination used for experiment is Systolic Blood Pressure (SBP), and the SBP that we obtained is categorized in to class "Yes" and "No", and the rediscretized values from predefined ones, which has been chosen as the best accuracy from the bins explored from 2 to 10 bins. The rediscretized values selected in the new range can be categorized in to 3 bins, which are less than or equal to 120 mm/hg, 120 – 150 mm/hg, and more than 150 mm/hg. Systolic blood pressure with our participant's characteristics is shown in Table 12.

Table 12: Characteristics of Participants with rediscretized values of hypertension

Systolic Blood Pressure (mm/hg)		Hypertension	
		Men (%)	Women (%)
Normal	$\leq$ 120	43	52
High Normal	>120 – 150	41	34
Hypertension	> 150	16	14

As a result shown in Table 12, the finding found that there were relatively few participants at the higher stage of hypertension. Approximately half of the participants for each gender had blood pressure level in the normal range. After rediscretizing to the new range of blood pressure levels, in addition to the first four factors, the predictive accuracy is proved to more efficient and to have a better performance.

At initial examination of the feature selection phase, obesity is one of the relative risks but not selected by the selection models. Nevertheless, obesity was found to be relative associated with coronary heart disease. Thus, this study also explored the rediscretization task on obesity risk factors as the sixth factors to investigate the better accuracy in prediction, by adding to five factors above. The percentages of each interval of attribute obesity can be summarized in the Table 13.

As a result, presented in Table 13, most of participants have the normal or a little bit overweight. In the medical value, people would be detected to be critical overweight or have the obesity when they must have the value of obesity of more than 25. In our

study, you can be detected to be risky to have a chance of coronary heart disease occurrence when you have a value of BMI ranged from 23 to 25. Therefore, if people take the discretized value of obesity as consideration, they can have a primary prevention for avoiding overweight, which might finally lead to coronary heart disease.

Table 13: The ratio of Participants with rediscretized values of Obesity

Discretized Values	Participants (%)
19-22	22
23-25	36
26-30	26
>30	16

Referring to the phase of feature selection, triglyceride was explored to be the critical risk factors for CHD, because it was appeared in every attribute selection methods. Thus, it indicates that triglyceride is the significant major risk factors for Thai people. Triglycerides, usually considered as having levels higher than 200 mg/dl., are a type of fat stored in human's blood, which are derived from consuming fat and high calories food. In this research, triglyceride is one of the major risk factors to be rediscretized the medical values to test the better classification performance due to difference of individuals on the basis of both physical and behavioral issues.

Normally, triglyceride is considered adjusting with HDL, and LDL. Accordingly, this study considered to examine the better accuracy by rediscretizing the attribute of triglyceride, in addition to rediscretized five factors conducted earlier. Triglyceride has been tried rediscretizing from 2 to 10 bins similar to other discretized factors, and has been chosen a 4-bin as the best accuracy for our dataset.

Table 14: Characteristics of Participants with rediscretized values of triglyceride

Triglyceride (mg/dl)		Participants (%)
Normal	0 – 130	36
Normal to borderline High	131 – 179	26
Borderline High to High	180 – 259	25
High	> 259	13

According to Table 11, the rediscretized values of triglyceride were categorized into 4 bins, due to the best accuracy ranged from 2 to 10 bins. The new range of triglyceride values was divided starting from 0 – 130 mg/dl, 130 – 179 mg/dl, 180 – 259 mg/dl, and more than 259 mg/dl. Compared with the predefined clinical value, which was categorized into only 2 classes with the criteria of 0 -200 and more than 200 mg/dl., the new rediscretized values can provide more details. Cutoffs for discretized values of triglyceride provided some characteristics of participants, was shown in Table 14.

As seen in the Table 14, the people who has triglyceride in the range of 0 – 130 mg/dl is accounted for 36%, in the range of 131 – 179 mg/dl is accounted for 26%, in the range of 180 – 259 mg/dl is accounted for 25%, and in the range of more than 259 mg/dl is accounted for 13%. The finding found that most of participants in this study ware categorized in the normal to borderline high level. Additionally, the experimental result of this finding provided some significant information for people who were in the normal to borderline high level. In this level, there are 5225 participants involved. Within 5,225 people, there are 311 of them that has the heart disease, and 12 of them has the triglyceride less than 150 mg/dl. Furthermore, in the level of borderline high to high triglyceride value, 4,950 people were categorized in this level and 543 of them had the heart disease. With this number, there were 114 participants that have triglyceride value less than 200 mg/dl. Therefore, the rediscretization in this study can help as an early warning for people who look healthy but actually, they can have a chance to get the CHD, even they have the triglyceride value less than 200 mg/dl or even less than 150 mg/dl. They can have the primary prevention for being aware of eating food, exercise, and other positive behavior that will inverse with incurring a chance of CHD.

Consequently, rediscretization on the seventh major risk factor which is triglyceride is proved to produce the better predictive accuracy. The performance evaluation of comparative 7 rediscretized major risk factors conducted on the experiment is shown in Table 15.

Table 15: Comparison of the accuracy of different algorithms between with feature selection, without feature selection and Discretization, and with discretization and feature selection for 7 major risk factors

		Accuracy (%)			
		NB	J48	MLP	SVM
no discretization no feature selection		83.18	81.67	80.67	85.55
no discretization but feature selection with RAE		88.82	87.55	83.59	90.92
discretization + RAE	4 Factors	90.90	88.82	87.59	91.92
	5 Factors	91.33	89.02	90.55	92.58
	6 Factors	92.05	90.67	91.89	93.10
	7 Factors	93.47	92.18	92.57	**95.50**

According to Table 15, the full dataset of 20,000 instances are employed for rediscretization. As the empirical studies in the earlier experiment proved that RAE with Bayes is the best hybrid model for selecting the relevant risk factors, rediscretization task in this phase is conducted and classified applying RAE with Bayes, based on accuracy measurement. Table 15 demonstrates the accuracy comparison of 4 different classifiers, which are Naïve Bayes, J48, MLP, and SVM, between without feature selection and without discretization, without discretization and feature selection by RAE, and with feature selection by RAE and with discretization for 4, 5, 6, and 7 risk factors respectively.

From the result shown in Table 15, the comparison of 3 experimental setting is shown. The predictive performance of CHD dataset for Thai people classified by 4 classifiers is presented without feature selection and without discretization in the first row. The second row presents critical risk factors associated with CHD are pioneered by feature selection but not fed into discretization method, which are classified by 4 classifiers. The last section demonstrates predictive accuracy of rediscretization method with RAE feature selection approach for

4, 5, 6, and 7 risk factors respectively. Rediscretization is set up on medical values, which is divided in 4 rounds. For the first round, we rediscretized on four major risk factors as age, Fasting Blood Sugar, Obesity and Triglyceride. In the second round, hypertension is added as the fifth major risk factors for rediscretization with the first four factors conducted in the first round. For the third round, LDL is the sixth factors included for rediscretization with the five factors performed in the second round. Finally, HDL is the last factors added on for rediscretization including the six factors performed earlier. As seen in all of the performance measurement, the results can be summarized that the proposed method explored and designed applying rediscretization produces the better predictive performance than classifying without rediscretization. Accordingly, the finding can be implied that all classifiers is achieved the improved accuracy when performing rediscretization. Additionally, rediscretization approach with RAE feature selection on 7 risk factors for Thai heart disease dataset classified by SVM produces the highest accuracy of 95.50%, when compared with predefined clinical value conducted in the experimental setup with no feature selection and with no rediscretization as 85.55%, and 90.92% as with the process of hybrid feature selection model but no rediscretization applied. Thus, the finding can infer that the more numbers of risk factors rediscretized, the more predictive accuracy are produced.

Finally, according to the result of rediscretization on clinical values shown above, we describe the characteristic of data of each rediscretized major risk factors in context. Thus, the example of data which are rediscretized from clinical values, categorized by men and women are demonstrated in Table 16.

Table 16: Comparison of Data Characteristics of Rediscretized values from major risk factors categorized by gender

	No CHD (n = 18,795)		With CHD (n = 1,205)	
	Male	Female	Male	Female
	10,436	8,359	804	401
HDL (mg/dl)				
>= 60	2,738	2,407	228	110
>=50 - 59	3,167	2,608	241	124
>=40 - 49	2,627	2,163	206	103
<40	1,724	1,361	129	64
LDL (mg/dl)				
<=100	1,932	1,536	146	73
>100 - 130	2,582	2,039	245	118
>130 - 160	2,987	2,515	230	115
>160	2,935	2,269	226	113
Obesity (BMI)				
19 - 22	2,288	1,848	176	88
23 - 25	3,744	3,024	288	144
26 - 30	2,704	2,184	208	104
>30	1,664	1,344	128	64
Triglyceride (mg/dl)				
0 - 130	3,644	2,293	263	119
>130 – 179	2,704	2,284	192	100
>179 – 259	2,536	2,315	185	93
>259	1,552	1,647	164	89

5. Conclusion and Discussion

The findings of this research can be implied that Education is the significant SES indicator that can incur CHD associated with the major risk factors. Support vector machine (SVM) is the best classifier for our dataset due to the best accuracy in every feature selection methods, and for the whole study. The proposed hybrid models of feature selection were explored to produce the better accuracy for CHD prediction on Thai people. Additionally, the objective of the research also concentrates on identifying SES factors and individual biological factors strongly associated with CHD by discovering the critical risk factors from both types which are significantly predictable for raising CHD. SES factors as education has been found as a significant indicator for raising CHD, after adjusting with major risk factors. Furthermore, SES factors as income is also revealed to be a significant predictor that have the great impact for causing CHD for Thai citizens, as well as the biological factors as gender, age, cholesterol, smoking, and diabetes.

Finally, this study proposed rediscretization approaches, which are confirmed to enhance the predictive accuracy of classifiers on CHD diagnosis. Equal Frequency Discretization with attribute selected by RAE and Bayes is proved to produce the accuracy improvements, as higher than the traditional classifiers conducted without feature selection and without rediscretization. SVM gives the highest predictive performance of 95.50% accuracy. After employing the rediscretiztion method in order to convert the predefined clinical values to the new boundaries of values with different number of intervals, we found that age, triglyceride, obesity, LDL, HDL, and hypertension are appeared to be significantly correlated with CHD. In addition, the finding can be implied that the proposed rediscretization method is proved to achieve the higher accuracy of predictive model, compared with the predefined values from medical field. Equal Frequency Discretization pioneered with attribute selection by RAE combined with Bayes generates the higher accuracy than the classification without rediscretization and with no feature selection, which indicates that the novelty of proposed rediscretization model is obviously superior to the predefined medical values. Nevertheless, in general, each individual country has its own individual characteristics, different physical conditions, lifestyles and environment. Thus, the similar parameter values are able to be measured and applied in alternative approaches for best suitable prediction incorporating with risk factors to derive the best performance measurement in predictive model.

References

[1] S. Dedkhar, “Risk factors of cardiovascular disease in rural Thai women,” The University of Arizona, https://repository.arizona.edu/handle/10150/195629, 2006.

[2] P. Vathesatogkit, P. Sritara, M. Kimman, B. Hengprasit, T. E-Shyong, HL. Wee, M. Woodward, "Associations of lifestyle factors, disease history and awareness with health-related quality of life in a Thai Population," pone, **7**(11), 2012, https:/doi.org/10.1371/0449921.

[3] G.M. Egeland, A. Tverdal, R.M. Selmer, H.E. Meyer, "Socioeconomic status and coronary heart disease risk factors and mortality: Married residents, three countries, Norway," The Norwegian Journal of Epidemiology, **13**(1), 155 – 162, 2003.

[4] R.V. Luepker, D. Rosamond, R. Murphy, J.M. Sprafka, A.R. Folsom, P.G. McGovern, H. Blackburn, "Socioeconomic Status and Coronary Heart Disease Risk Factor Trends : The Minnesota Heart Survey," 2172-2179, 2015, https:// doi:10.1161/01.CIR.88.5.2172.

[5] J.W. Lynch, G.A. Kaplan, R.D. Cohen, J. Tuomilehto, J.T. Salonen, "Do Cardiovascular Risk Factors Explain the Relation between Socioeconomic Status, Risk of All-Cause Mortality, Cardiovascular Mortality, and Acute Myocardial Infarction?," in 1996 CRPIT 9th proceeding of the Australian Data Mining Conference in Research and Practice in Information Technology, **144**(10), 934 – 942, 1996.

[6] S. Saravanakumar, S. Rinesh, "Effective Heart Disease Prediction using Frequent Feature Selection Method," International Journal of Innovative Research in Computer and Communication Engineering, **2**(1), 2767 – 2774, 2014.

[7] M. Shouman, T. Turner, R. Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients," in 9th proceedings of Australian Data Mining Conference in Research and Practice in Information Technology (CRPIT), **121**, 23-29, 2011.

[8] M. Sultana, A. Haider, "Heart Disease Prediction using WEKA tool and 10-Fold cross-validation," The Institute of Electrical and Electronics Engineers, March 2017.

[9] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques," International Journal of Computer Science and Engineering, **6**(4), 21-25, 2018.

[10] T.F. Mubaroq, E. Sugiharti, I. Akhlis, "Application of Discretization and Information Gain on Naïve Bayes to Diagnose Heart Disease," Journal of Advances in Information Systems and Technology, **1**(1), 75-82, 2019, https://doi: 10.5121/csit.2013.3305.

[11] M.M. Kirmani, S.I. Ansarullah, "Prediction of Heart Disease using Decision Tree a Data Mining Technique," International Journal of Computer Science and Network (IJCSN), **5**(6), 885-892, 2016.

[12] C. Partanapat, C. Jandaeng, C. Jaruskulchai, A Hybrid Model for Coronary Heart Disease Prediction in Thai Population, Ph.D. Thesis, Walailak University, 2019.

[13] A. Hall Mark, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.

[14] K. Mani, P. Kalpana, "An Efficient Feature Selection based on Bayes Theorem, Self Information and Sequential Forward Selection," International Journal of Information Engineering and Electronic Business, **6**, 46-54, 2016, https://doi.org/10.5815/ijieeb.2016.06.06.

[15] S. Ravichandran, V.B. Srinivasan, C. Ramasamy, "Measuring Accuracy of Classification Algorithms for Chi-Square Attribute Evaluator in MCDR," in 28th Advances in Engineering and Technology Convergence, Bangkok, Thailand, 2019.

[16] S.L. Pundir, S. Amitra, "Feature Selection using Random Forest in intrusion detection system," International Journal of Advances in Engineering & Technology, **6**(3), 1319-1324, 2013.

[17] A. Hall Mark, A. Smith Lloyd, "Practical feature subset selection for machine learning," Springer, 1998.

[18] C. Arunkumar, M.P. Souraj, S. Ramakrishnan, "A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets," in 7th International Conference on Advances in Computing & Communications (ICACC), 209-217, 2017, https:// doi.org/10.1016/j.procs.2017.09.127.

[19] R. Kohavi, G.H. John, "Wrapper Subset Evaluation," Artificial Intelligence, 273-324, 1997.

[20] B. K. Bhardwaj, S. Pal, "Data Mining: A prediction for performance improvement using classification," International Journal of Computer Science and Information Security (IJCSIS), **9**(4), 136-140, 2011.

[21] D. N. Bhargava, G. Sharma, D. R. Bhargava, M. Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining," International Journal of Advanced Research in computer Science and Software Engineering, **3**(6), 1114-1119, 2013.

[22] S. Bhatia, P. Prakash, G.N. Pillai, "SVM based decision support system for heart disease classification with inter-coded genetic algorithm to select critical features," in 2008 Proceedings of the World Congress on Engineering and Computer Science (WCECS), San Francisco, USA, 2008. http://www.iaeng.org/publication/WCECS2008.

[23] E. Ferdousy, M. Islam, M. Matin, "Combination of Naïve Bayes Classifier and K-Nearest Neighbor in the Classification Based Predictive Models," Computer and Information Science, **6**(3), 48-56, 2013, https://doi:10.5539/cis.v6n3p48.2013.

[24] B. Abdullah, I. Abd-Alghafar, G. I. Salama, A. Abd-Alhafez, "Performance evaluation of a genetic algorithm based approach to network intrusion detection system," in 13th international conference on aerospace sciences and aviation technology, Military Technical College, Kobry Elkobbah, Cairo, Egypt, 1-17, 2009. https:// doi: 10.21608/asat.2009.23490.

[25] J. Anil, Z. Douglas, "Feature selection: Evaluation, application, and small sample performance," in 1997 IEEE transactions on pattern analysis and machine intelligence, **19** (2), 153-158, 1997.

[26] M.A. Jabbar, B.L. Deekshatulu, P. Chandra, "Heart disease prediction using lazy associative classification," in 2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 40-46, 2013.

[27] H. Liu, R. Setiono, "Feature selection via Discretization," in 1997 IEEE Transactions on Knowledge and Data Engineering, **9**(4), 642-645, 1997. https://doi.org/10.1109/69.617056.

[28] A. Rajalakshmi, R. Vinodhini, K. Fathima Bibi, "Data Discretization Technique Using WEKA Tool" International Journal of Computer Science & Engineering Technology (IJCSET), **6**(8), 293-298, 2016.

[29] P.W. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, W.B. Kannel, "Prediction of Coronary Heart Disease Using Risk Factor Categories," American Heart Association, **97**(18), 1837-1847, 1998, doi.org/10.1161/01.CIR.97.18.1837.

[30] R.S. Vasan, M.G. Larson, E.P. Leip, W.B. Kannel, D. Levy, "Assessment of frequency of progression to hypertension in non-hypertensive participants in the Framingham Heart Study: a cohort study," Lancet, **358**, 1682–1686, 2001, doi: 10.1016/S0140-6736(01)06710-1.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 423-438 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Simulation Based Proactive Approach for Smart Capacity Estimation in the Context of Dynamic Positions and Events

Naeem Ahmed Haq Nawaz[*,1], Hamid Raza Malik[1], Ahmed Jaber Alshaor[1], Kamran Abid[2]

[1]*Department of Computer Science, Umm Al-Qura University, Makkah Al-Mukarmah, 24381, Saudi Arabia*

[2]*Department of Electrical Engineering, University of the Punjab, Lahore, 54590, Pakistan*

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 31 October, 2020
Online: 20 November, 2020

Keywords:
Smart
Capacity
Estimation
WSN
IoT
Sensor topologies
Position
Proactive approach
Cloud

ABSTRACT

Technologies are growing with the passage of time and providing solutions for the existing problems. One of such problems is to manage the crowd according to available capacity. Especially when the crowd density is dynamic because of dynamic position of the persons (Pilgrims). Not only dynamic position of the pilgrims makes crowd capacity dynamic but also special events increase and decrease in number of pilgrims that affect the capacity in a specific place. Therefore, it is not an easy job to estimate the available capacity according to the dynamic position and event. To overcome such problems of dynamic position and event-based crowd management different techniques and approaches are adopted. To solve the above-mentioned problem, this paper proposes a proactive approach to estimate the space occupy by the pilgrims in different positions in a zone or level. Maximum capacity in each zone and level is define on the basis of each position. On the basis of maximum and occupy capacities, available (remaining) capacity has been estimated in a zone and level according to the events such as Pilgrimage, Ramadan, Jummah etc. The occupied and available capacities in a zone, level or in the whole building can be estimated with the help of technologies such as Wireless Sensor Network (WSN), cloud computing, Internet of Things (IoT) and sensor topologies according to position and event. Simulation shows different scenarios according to the zones and different levels of building to prove the proactive approach. According to the results, it is concluded that zones, levels and point of entrances should be allocated to the pilgrims to avoid the congestion problems. Further for massive crowd and large area multi sink solution is better than single sink solution to estimate available capacity.

1. Introduction

Capacity estimation plays a vibrant role to maintain equilibrium among the maximum, occupied, and remaining capacities so that number of the persons can be allowed to enter in a zone, level or in a building [1]. Number of the persons permitted to enter an area directly dependent on the remaining capacity [2]. Although image-based method does not depend on the crowd to carry any type of device and counting can be performed without handling of device by the crowd. But the crowd counting done on the basis of images, does not provide result in case of dark or fade environment [3]. To provide the population awareness globally, crowd analysis can be done on the basis of video imagery to calculate and estimate the density of the crowd. It is considered an important tool of analysis for crowd density estimation at public places. Other methods also involve counting by performing the detection method or by using the clustering method for crowd estimation. Due to the more crowded environment one of the recent ways is regression. It has the capability to handle large crowded area [4]. The proposed proactive approach for capacity estimation is to calculate the number of the persons entering and exiting from the certain area. It is estimated according to the maximum capacity on the occasion of different positions and events. The sensor devices have the capability to sense and collect data at a given zone, level, building or an area [5]. In a specific scenario, data may include person ID (pilgrim ID), device ID, zone and level number, date and time for entry and exit etc. The sensed data then can be forwarded to the cluster head, then to a sink or

[*] Corresponding Author: Naeem A Nawaz, Ummal Qura University, nanawaz@uqu.edu.sa

server at each level and finally to central cloud storage location in order to calculate the remaining capacity of the areas. For different statistics, cloud computing can play a vibrant role to use the stored data. With the help of IoT data can be forwarded to different stakeholder for timely information to manage the crowd according to the position and events.

To overwhelm the problem of estimation of the capacity in different position and event, proposed proactive approach is one of the efficient solutions to provide advantages including:

- Counting number of pilgrims without human intervention
- Set the maximum capacity for different positions (Prayer, Walking, Sitting, Standing)
- Capacity estimation for different zone, Mataf area and on each level
- Capacity estimation according to Extreme crowded events such as, Jummah prayer, Ramadan, Eidain, Pilgrimage
- To restrict the pilgrims if they already performed Umrah so that remaining pilgrims can get opportunity
- Lost pilgrims can be traced within the coverage area

One of the central pillars of Islam is also Pilgrimage. Approximately 3.5 million pilgrims assemble to perform pilgrimage in the holy city of Makkah [6]. In the month of Ramadan millions of people gather to perform Umrah and other prayers [7]. Pilgrims perform different supplications at different places and time to achieve the Pilgrimage or Umrah requirements. To perform the Umrah during Ramadan people persist around the Masjid Al-Haram. The estimation of available capacity inside the Masjid Al-Haram, may be calculated by number of pilgrims in a specific zone, or level minus maximum threshold of the capacity. It has been observed that it is problematic to estimate capacity for a massive dynamic crowd, the number of pilgrims entered in Masjid Al-Haram and what is the remaining capacity inside? When it exceeds to maximum capacity? In this paper, we try to solve the above-mentioned problem by exploiting WSN. We proposed a proactive approach to process the data for different zones and levels to estimate available capacity. As I myself perform the Jummah Prayer many times, The Holy Masjid Al-Haram was so much overcrowded that it became difficult to perform Sajood and Tashahud. It is because of wrong capacity estimation of zone or level.

This paper is an extension of our previous work accepted in The International Arab Conference on Information Technology ACIT'2019 [1]. The extension includes capacity estimation of dynamic position, event, study of sensor topologies and proactive approaches.

1.1 Paper Layout

In section 2, we briefly discussed the technologies used in our proposed proactive approach that includes the technologies such as cloud computing, IoT, clustering and WSN. In Section 3, there are briefly given the approaches or strategies at present used for capacity estimation. Section 4 depicts the proposed proactive approaches according to zone, level and for the whole building. Section 5 elaborates the use case, section 6 offers comparison between existing and proposed capacity estimation, section 7 defines the algorithms, section 8 simulation and results, section 9 gives discussion and section 10 conclusion and future directions.

2. Technologies for Proposed Framework and Proactive Approach

2.1. Cloud Computing

Cloud computing is one of the most important technologies for proposed proactive approach. In cloud computing, cloud resources are located at different physical places. But via internet cloud computing makes efficient cloud services available globally to a large number of consumers located on many geographical locations [8], [9]. In a cloud computing environment, wearable devices are suggested for supportive communication between wearables and cloud server. Wearable device can be accessed remotely via the Internet [10]. Moreover, its everything-as-service-model (XaaS) impacts its clients and creators with many benefits [11]-[13]. Cloud computing is cost effective for the customers and uses the strategy "pay as you go" [14][15]. To process and access information in a more accurate way, cloud based software architecture has been proposed [16], [17]. Sensors, actuators and many other embedded devices are considered as a data generation source that provide help to enable IoT edge computing to process data [18]. Cloud can store the data according to the position, event and zones for future planning to control the massive crowd.

2.2. Internet of Things (IoT)

IoT lets people/anybody and things/devices connect anyone, anything, anytime, anywhere, for any service. Although it is flexible, complex and dynamic network infrastructure as given in [19]. In many fields such as healthcare, smart homes, smart cities, retail, traffic, security, and agriculture, IoT offers multiple solutions [20]. With the help of an IoT we can monitor the elements such as wind, crops, soil, atmosphere, lights and water over a big area [21]**.** For the growth of the IoT, there is a need of plentiful resources for the development of the intelligent devices and the WSNs. The core building blocks are the same for network communication in between devices to build an Internet of the Things setup [22]. With up to date development and advancement in the field of information and communication technologies, it also brought positive changes in the advancement and development of IoT. Through low-cost electronic devices and communication IoT paradigm improves human collaboration in the physical world [23], [24]. In the same way, it is also increasing the importance of WSNs to integrate and make connections with the Internet to develop different commercial and industrial applications [25]. In order to achieve different objectives, updated information is required efficiently during pilgrimage and Umrah. To achieve the above goal, IoT can play a vital role to get updated information as needed at a particular zone or level. Information received from IoT may be used by the management such as security management, and resource management (water, chairs, Quranic scripts etc…). So that necessary actions can be performed according to updated information on time in proactive way.

2.3. Clustering and WSN

There are some limitations in WSN such as energy limitation, coverage area, network lifetime, and bandwidth. These limitations

can be overcome with the help of some clustering schemes to get efficient solution [26]. Amongst the other factors, clustering is one of the most important factors. Clustering can also play an important role to control the performance of a network [27]. Wearable devices provide help to sense and disseminate information, such as physical activities, blood pressure, blood sugar level, heart rate) [28]. To monitor diverse environmental situations such as humidity, temperature, sunlight, and pressure, a smart sensing platform based on NodeMCU-ESP8266 microcontroller board has developed and given in [29]. With the implantation of WSN data can be sensed, stored, disseminated and processed to estimate the capacity [30], [31]. As an industrial IoT is an important infrastructure, WSNs play a crucial role in it. In IoT applications, efficiency balanced verification scheme for WSNs becomes a big contest, due to the resource limitation feature of sensor [32].

A proactive approach and framework can play an important role to enable reliable and an efficient management for WSN. One of the examples is Web Services based on clouds and lightweight cross-layer design [25].

2.4. Sensor Topologies

The topologies can differ in complication from a single node connected to a collector node to fully meshed network covering a large area. Flat or hierarchical architecture are used to design sensor topologies. In flat (peer-to-peer) architecture, each node either can be a sink or sensor node that has the same communication and computational abilities in the network. In hierarchical architecture, the nodes link with their particular cluster head in close proximity. Normally, cluster heads own more storage capacity and processing power as compared to any ordinary sensor node. Commonly line, star or bus topologies are used by the sensor networks that are physically wired together [33].

In Figure 1, point-to-point, Bus and linear topologies are given. A hardwired connection between two points is called a permanent point-to-point topology. A switched connection is a type of point-to-point connection in which control can be transferred between end nodes.

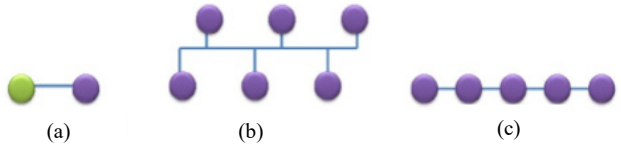


Figure 1: (a) Point-to-point, (b) Bus, and (c) Linear Network Topologies

In Bus topology, nodes are associated with a communication bus for communication. The advantages of the Bus topology are that it is simple and easy to install. But disadvantage is that the failure of central bus is the failure of the whole network, which is called a single point of failure. In Linear Topology, there is a two-way connection between the nodes. Only two terminating nodes has a single connection to their neighboring nodes at the end of the network, but all other nodes have link or connection to left and right nodes. In case of failure of a node, there is a disconnection from the network of any node connected to that node.

In Figure 2, ring, star and tree topologies are given. Network set up in a circular fashion is called ring topology. Each node in this configuration is connected to precisely two other nodes. Direction of flow of data is from the source to each node in clockwise or anticlockwise until it reaches to the intended recipient. In order to overcome this issue, a second communication ring is added by many networks that can transfer data in the opposite direction. In star topology, each node or device is connected to a single central device called hub or switch. Star topology has advantages such as easy to plan, implement, and add or remove the nodes. An intelligent node is needed as all data traffic flows through the central node. Entire network fails as a result of the failure of this node. In Tree topology, single “root node” is at the top of the hierarchy and other nodes are connected to the lower level in hierarchy in a tree manner structure. Many levels of nodes can be contained by a tree topology. This topology is extendable, and identification of the fault is easy and can be fixed easily because of simple structure. With the growth of the tree networks, it becomes more difficult to manage it.

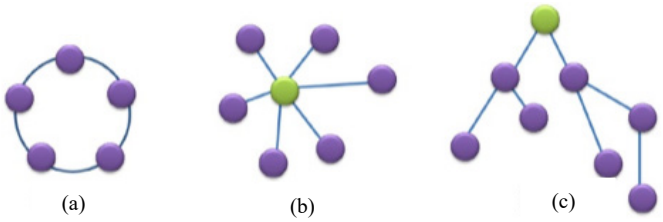


Figure 2: (a) Ring, (b) Star, and (c) Tree Network Topologies

In Figure 3, partially and fully mesh topologies are given. There are two arrangements of mesh topology: partially mesh and fully mesh. In partially mesh, some nodes are connected to more than one other node is called a partially connected mesh, while in fully mesh, every node is connected to every other node.

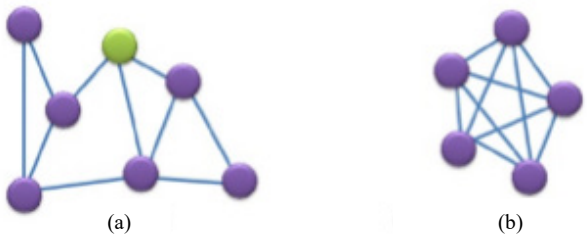


Figure 3: (a) Partially Connected and (b) Fully Connected Mesh Network Topologies

Cluster Tree topology consists of a single cluster or a multi-cluster network. Only one cluster-head (CH) is contained by a single cluster network. When all the nodes are connected to the cluster-head with single hop, the network topology becomes a star topology. More than one cluster-heads are contained by a multi-cluster network. An upper level sub-network is formed by all the cluster-heads. Nodes in different clusters communicate among themselves via their cluster heads but cannot directly talk to each other. Figure 4 illustrates the cluster tree topology, which has hierarchy architecture with the cluster-head network at the upper level and the clusters nodes at the bottom level [34]. The sink behaves as an entrance to the sensor nodes, gathering and updating the information. The sink plays a role as an interface among WSN, the external network and Internet, where the data are disseminated and processed. Sink can be single or multi sink depending upon the

requirement of the application. The goal of the single sink formation is to minimize the forwarding time and route information in the direction of a unique sink. Multi sinks are utilized for applications where tasks are distributed to different sinks. Multi sinks are used where there are high density crowd, large coverage area, redundancy, high latency, and possible high energy consumption [35]. In our scenario cluster tree or hierarchical cluster structure is suitable to collect the data from zones and levels during different positions and events.

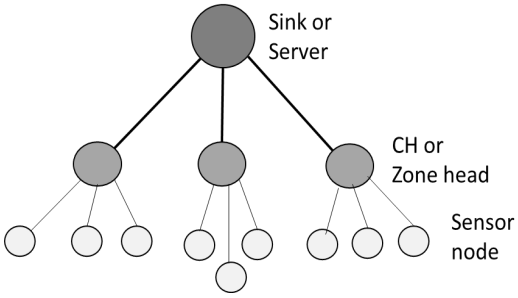


Figure 4: sensor cluster tree

During the events of Pilgrimage, Ramadan and Umrah the Holy city of Makkah (KSA) is crowded. Especially, to perform Umrah during the Ramadan all activities are around the Masjid Al-Haram, therefore it is need of the hour to organize, control and manage the crowded area smartly. The above objective can be achieved by proactive approach in smart way with the support of WSN, IoT, cloud computing and the management phases.

3. Present Capacity-Estimation Approaches

At present, capacity estimation is being done by management by taking view from the cameras or by the observation inside the Masjid Al-Haram. By cameras they discover the empty zone or level and update to the security officers to transfer the crowd to the empty or less crowded zone or level. They use the wireless system to give instructions to the security officers. By observation, the security officers check the capacity is available or not in an area where they are assigned. But they do not have auto capacity estimation approach so that they can measure how many pilgrims can be allowed in an empty space. They have just rough estimation of the available capacity. If the Zone or a level is reached to maximum capacity, the remaining crowd is diverted to the less crowded zone or level. If the Masjid Al-Haram is full then gates are covered with barrier and security officers do not permit pilgrims to get in, but they can make exit. No entry sign is turned on outside the gates. After that zones outside the Masjid Al-Haram are occupied. Sometimes zones and levels are overcrowded, and pilgrims face problem to perform the prayers or Tawaf. But currently, they do not have the smart estimation of the available capacity in a zone, level or in the whole Masjid Al-Haram. If there is smart capacity estimation with proactive approach, then pilgrims can be allowed according to the capacity available. Otherwise security stops the pilgrims at the gate and move to specified waiting area so that they do not overcrowd Masjid Al-Haram and do not cause any casualty.

4. Proposed Smart Capacity Estimation Framework and Proactive Approach

The Figure 5 represents the proposed smart capacity estimation proactive approach and framework. It provides help to estimate the capacity for zone and level to accurate the management. In order to ensure the smart capacity estimation in specific zone or level, the architecture uses WSN so that available capacity can be estimated for management automatically. The architecture provide help to understand the flow of data that is required to estimate the available capacity. It provides help to pilgrims to find out a specific zone or level where capacity is available. Figure 5 shows the overall picture of proposed framework and proactive approach for smart capacity estimation with the help of WSN, cloud storage and interaction among users by IoT for dynamic crowd management. It includes different phases: First phase involves smart capacity estimation according to time and preferable activity by pilgrims in a zone. Second phase involves smart capacity estimation according to time and their preferable activity by pilgrims at a level. Third phase involves smart capacity estimation according to time and their preferable activity by pilgrims at different levels, in order to calculate overall capacity occupied and available capacity in all zones or levels.

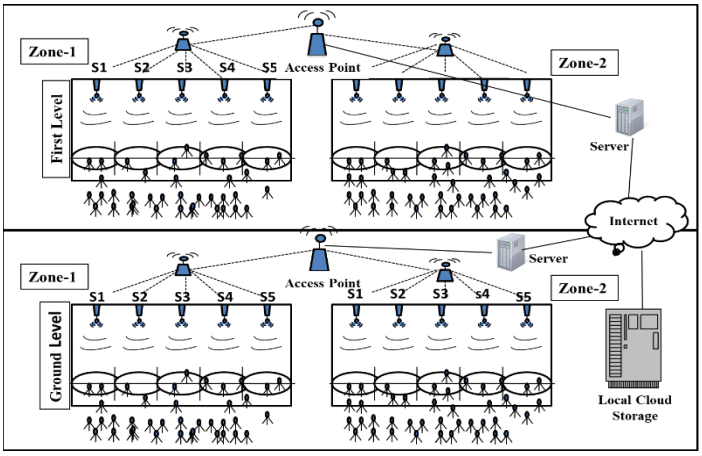


Figure 5: Proposed smart capacity estimation framework and proactive approach [1]

In the first phase, the calculation for the maximum capacity, capacity occupied by pilgrims and available capacity in the zone is done. In the second phase, head nodes or access point nodes fixed on the different zones at different levels will collect the data and forwarded to the server of different levels as update. In the third phase, data from each level from servers will be forwarded to the local cloud storage for further processing and management tasks. Detail of each proactive approach is given in subsection 4.1.

4.1. Proactive Approaches at Zone and Different Levels

Proactive approaches use to estimate the maximum, occupied and available (remaining) capacity with respect to zone, level and in a building.

The proactive approach for smart capacity estimation of zone is given as:

1- The Figure 6 depicts the steps to estimate the capacity based on a zone.

- Sensor nodes are fixed on each entry and exit point for each zone that can collect data about number of the pilgrims getting in and out of the zone.

- This information is updated to the zone access point or head node and from zone access point, it is forwarded to the level's server.
- The capacity occupied by pilgrims inside the zone can be calculated by number of pilgrims has entered minus number of the pilgrims exit.

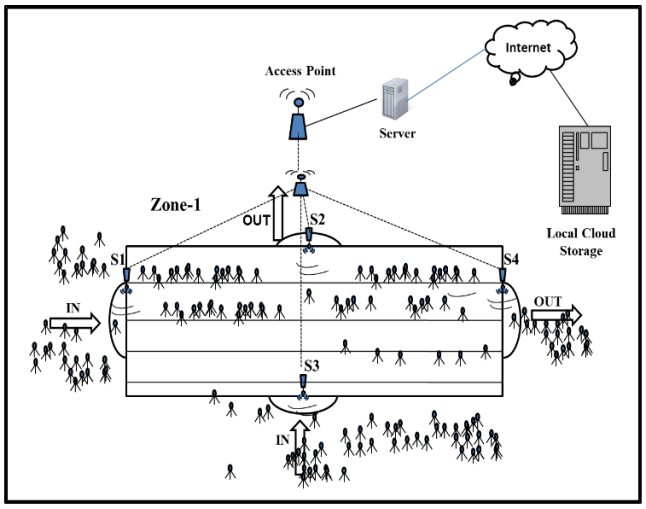


Figure 6: Smart capacity estimation of zone inside building [1]

Capacity occupied by pilgrims inside zone = number of pilgrims entered - number of the pilgrims exit

Z.IN=Entered to the zone
Z.OUT= Exit from the zone

$$Zone\ Capacity\ (Occupied) = \sum_{i=1}^{n}(Z.IN)i - \sum_{j=1}^{k}(Z.OUT)j \quad (1)$$

- The available capacity of the zone can be calculated by the maximum capacity of the zone minus total number of pilgrims inside.

The remaining capacity = maximum capacity of zone - (total number of pilgrims inside the zone).

Maximum Zone Capacity = Max (Z.C)

$$Zone\ Capacity\ (Available) = Max(Z.C) - \left(\sum_{i=1}^{n}(Z.IN)i - \sum_{j=1}^{k}(Z.OUT)j\right) \quad (2)$$

2- Proactive approach to estimate capacity at Ground Level is depicted in Figure 5.

- On each zone, sensor nodes will collect the data for pilgrims getting in or out. Afterward, it adds up or subtracts the number of the pilgrims got in and out from the head node or access point.
- The number of pilgrims getting in or out from all zones will be added up or subtracted by head nodes and then updated by the ground level server.
- If a pilgrim moves to the first level from ground level, then it will be subtracted from the ground level server but added into the first level server and vice versa.

L-IN=Entered to the level
L-OUT= Exit from the level

$$Level\ Capacity\ (Occupied) = \sum_{i=1}^{n}(L.IN)i - \sum_{j=1}^{k}(L.OUT)j \quad (3)$$

- From head node data can be forwarded to server to estimate the available capacity for ground level. If the ground level reaches to maximum capacity, then the ground level mentions no entry. Because ground level is full, it will be disturbance for the pilgrims if the pilgrims from outside entered into ground level and move to the first level. Therefore, security officers move the crowd to the first level from outside gate without entering into ground level and vice versa.

Maximum Level Capacity = Max(L.C)

$$Level\ Capacity(Available) = Max(L.C) - \left(\sum_{i=1}^{n}(L.IN)i - \sum_{j=1}^{k}(L.OUT)j\right) \quad (4)$$

3- Proactive approach to estimate capacity at First or at different levels is depicted in Figure 5:

- If the pilgrims move directly into the first level from outside gates, then there will be an addition to the ground level server and vice versa.
- If the pilgrims move into the First level via ground level, then it will be subtracted from the ground level but added into the first level and vice versa.
- From head node or access point data can be updated to the server to estimate the available capacity for first level. If the first level reaches to the maximum capacity, then at first level no entry appear at entrance. A message of maximum capacity is generated to the security officers by using IoT. The security officers move the crowd to second level from outside gate without entering into ground level or first level and vice versa.
- To avoid disturbance, pilgrims are allowed first to fill up top level then second and first level respectively. In this way first and second level pilgrims will not be disturbed.
- The data from each level forwarded to the local cloud storage to calculate the available capacity for the overall building. The data can be analyzed for different positions and events such as Ramadan, Pilgrimage, Jummah etc.

At present, most of prevailing wireless sensor devices are restricted because of memory, processing speed, efficiency and capabilities of communication. The above reasons are because of technology and economic limitations [36]. Low-power, low-cost, a multifunctional sensor device has been developed and designed by various industries for different applications, with high attention. Amongst the sensor devices one of the big issues in WSN is to classify and produce an organizational structure [37]. On behalf of clustering topologies, the processing of sensor devices can be improved which in turn improve the processing time. As a result, performance of capacity estimation will increase [38]. The collection of data will provide help to update information at zone, level, building and at crowded areas during Pilgrimage, Ramadan and Umrah for management.

The cloud storage, IoT and telecommunication play an important role in order to collect, assignment and process data locally and globally for smart capacity estimation [39]. To get results for different applications and different analysis, enormous amount of data can be sensed, stored and processed by the deployment of large number of sensor devices, cloud storage and IoT connections [40]. Detail of sensor topologies is given in section (2.4).

5. Use-Case for Smart Capacity Estimation

Data related to the pilgrims have been collected long time before by management and companies (travel agents) for pilgrimage, Umrah and visit. On the completion of e-visa process, management and companies receive information about pilgrim's residence in different sector or region. The information includes route of travel, reservation (city, hotel) and number of days to stay in different places. Pilgrims can be allocated the zones and levels according to their point of entrance according from their sector or region. As in Masjid Al-Haram, pilgrims have different positions therefore; capacities vary according to the different positions such as standing, sitting, walking and prayer positions. Capacity also depends on the events. Hence, it is important to measure the capacities according to each position and event.

5.1. Positions and Capacity

Capacity of the area is also depending on the position of the pilgrims such as standing, sitting, walking (performing Tawaf and Saee) and prayer position as given in Figure 7 and 8.In case of standing position space required for a pilgrim is less than as compared to the other position therefore, in standing position the available capacity will be maximum but in case of sitting, the pilgrim require more space than the pilgrim standing, therefore, in sitting position maximum available capacity will be reduced. In case of walking, the pilgrim required more space than sitting and standing position therefore, the maximum available capacity will be reduced as compared to sitting and standing maximum capacities.

In case of prayer position, the pilgrim needs more space than the standing, sitting, and walking position as the prayer position are dynamic such as standing (Qeyyam), sitting (Tashahud) and Sajood. Roughly, if we consider the standing position as the standard measurement then space ratio between standing and sitting position is measured as [5:9]. In case of walking position, the space ratio in between standing, and walking position is measured as [10:13]. In case of prayer position, the space ratio in between standing, and prayer position is as [1:3]. By calculating the capacities according to the crowd position, the crowd maximum capacity can be set according to standing, sitting, walking and performing prayer. At the time of prayer, the capacity will be minimum as compared to standing, sitting and walking position but in case of standing the capacity will be maximum as compared to sitting, walking and prayer position.

We can divide the sector for standing, sitting, walking and prayer. If crowd has different positions such as standing, sitting, walking and prayer within one sector or zone then we can calculate the average capacity. In case of The Holy Masjid Al-Haram, pilgrims want to see Baitullah (Kaaba), they can be managed on first, second and on top level near the balcony wall so that they can see the Kaaba in standing position. Sitting position is required when pilgrims want to recite the Holy Quran or Dhikr of Allah (Mention of Allah). They can be managed behind the Mataf area. Walking position is required in the Mataf or circular area on first, second and at top level. Prayer position can be at any sector but before prayer we need to calculate the capacity available for prayer and crowd inside the Masjid Al-Haram.

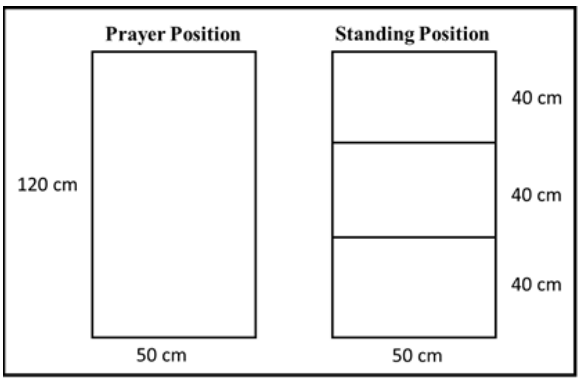


Figure 7: Capacity for prayer and standing positions

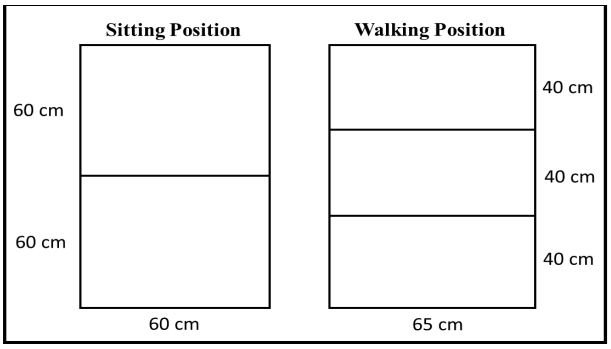


Figure 8: Capacity for sitting and walking positions

As we roughly calculated the space taken by the different position are given in Figures 7 and 8:

Standing = 40 x 50 = 2000 Sq CM

Sitting = 60 x 60 = 3600 Sq CM

Walking = 40 x 65 = 2600 Sq CM

Prayer = 120 x 50 = 6000 Sq CM

5.2. Stop and Wait

Calculate the average number of pilgrims entered in an area (Zone or sector) in unit time for example in one minute or in an hour. Average will provide help to calculate after how much time it is expected that the sector or zone will reach to its maximum limit. In this way we can control the crowd before it exceeds to its maximum limit. Usually it is observed that some time crowd exceeds the maximum limit and at prayer or Tawaf time pilgrims face a big problem which can lead to a stampede.

5.3. Move and Enter in the Zone

Rate of in and out to manage after maximum capacity, when zone and level reaches to its maximum capacity. Many pilgrims stay outside waiting for space availability. The space available for waiting pilgrims can be calculated by maximum capacity minus number of the pilgrims making the exit (going out). The space availability can be in a sector or zone or in whole area.

Table 1: Comparison between existing and proposed capacity estimation

S.No.	Existing Capacity Estimation	Proposed Capacity Estimation
i.	Initial capacity estimation is roughly calculated on the basis of previous experience and best guesses.	Initial capacity estimation will be calculated on the basis of data collected from central storage system.
ii.	Each zone capacity is calculated by the security officers or by computerized camera's system.	For each zone, automatically real time zone capacity will be estimated by zone sensors and sensor head.
iii.	Capacity available in a zone is controlled manually not automatically by security officers with the help of wireless system.	Capacity available in a zone will be calculated and automatically displayed. Zone security officer will manage the zone according to the available capacity.
iv.	There is no assurance about the exact maximum capacity estimation. Wrong estimation of maximum pilgrims' capacity can overcrowd the zone or level.	Maximum capacity for ach zone or level is already set therefore; maximum capacity alert will be generated automatically for no entry without any delay.
v.	According to dynamic positions and events there is no proper capacity records for zone and level.	Security officers can get records from local cloud storage system for future planning for each zone and level according to positions and events.
vi.	Real time capacity available cannot be calculated therefore capacity available can only be roughly estimated.	Real time capacity available will provide help to allow exact number of pilgrims to enter into a specific zone or level.
vii.	Difficult to estimate capacity available during different positions, events during prayer time and Tawaf time as each activity has different capacities that is number of pilgrims.	Capacity available during different positions, events during prayer time and Tawaf time is automatically estimated because prayer time and Tawaf time maximum capacity threshold can be set automatically according to activity.
viii.	There is no automatic way to trace out the lost pilgrim because existing system only support to manual capacity estimation and for security purpose.	Pilgrim lost can be traced out on the bases of the sensor device location with help of IoT.
ix.	CCTV cameras implementation provide help to observe and control the act of misconduct but there is no automatic way to identify the pilgrim (mis-conductor).	If the act of misconduct is done, then pilgrim (mis-conductor) will be identified on the bases of information stored on sensor device and on the cloud storage with the help of IoT.
x.	There is no such application to provide schedule, when the capacity is available for a specific building, sector and region.	Application can be developed on the basis of IoT to provide the schedule for visit to Masjid Al-Haram according to building, sector and region.

5.4. Evacuation or Exit from the Zone

In any case if the capacity exceeds to maximum limit in a zone or level, it can be managed in another zone or level. In worst case the pilgrims have to evacuate to the waiting place and put into high priority to get in when space will be available.

6. Comparison between Existing and Proposed Capacity Estimation

In the section 3, capacity estimation about existing system is described. It consists of manual capacity estimation, wireless system and computerized camera's systems. The proposed capacity estimation is discussed in the section 4. The comparison and preliminary analysis between existing and proposed smart capacity estimation systems are given in Table 1.

7. Algorithm for Proposed Capacity Estimation

We have taken the simple scenario to design the algorithms at different zones, different levels and in the whole building.

7.1. Algorithm for Zone Capacity Estimation

In zone capacity estimation, we simply check the available capacity in zone. If maximum zone capacity is greater than the number of pilgrims has entered in a zone minus number of pilgrims have exited from the zone. If the condition is true, then display the GREEN sign and also display the available capacity for pilgrims in waiting position. Otherwise display RED sign that zone is full and there is no entry.

Algorithm 1: Zone capacity estimation

```
Step-1   Begin
Step-2   GET value for Max(Z.C)
Step-3   If (Max(Z.C) > (Z.IN)-(Z.OUT))
Step-4      Display LED Turn(Green)
Step-5      Display Zone Available Capacity
Step-6   Else
Step-7      Display LED (RED)
Step-8   Endif
Step-9   End
```

7.2. Algorithm for different Level Capacity Estimation

In different levels capacity estimation, we simply check the available capacity in a level. If maximum level capacity is greater than the number of pilgrims has entered in a level minus number of pilgrims have exited from a level. If the condition is true, then display the GREEN sign for available capacity. Also display the zone number and level number where capacity is available for pilgrims in waiting position. Otherwise display RED sign that level is full and there is no entry.

Algorithm 2: Different level capacity estimation

```
Step-1   Begin
Step-2   GET value for Max(L.C)
Step-3   If (Max(L.C) > (L.IN)-(L.OUT))
Step-4      Display LED Turn(Green)
Step-5      Display Zone Number & Available Capacity
Step-6   Else
Step-7      Display LED (RED)
Step-8   Endif
Step-9   End
```

Cluster verification algorithm [31], represents the scenario of verification of cluster members (CMs) in cluster form. It discusses the number of verified and unverified CMs in a cluster. But in capacity estimation algorithm, it verifies and count the number of the pilgrims getting in or out from zone, level or in a building. All CMs remain within the cluster head transmission range for verification. But in capacity estimation algorithm, the pilgrims in the range of entry and exit sensors are verify when they make entry or exit from zone or level. CMs can enter to different clusters but in capacity estimation, it is not allowed to enter other than allocated zone or level. Zone and level are allocated so that pilgrims enter from the gate that is near to their zone and level. In this way crowd will be distributed in different zones, as a result a single path and zone will not be crowded. In the cluster verification, the number of the CM verify are listed out but in capacity estimation, it tells the available capacity for a zone and as well level and building. Cluster verification represents the number of CMs supported by a cluster head but in capacity estimation, the number of the pilgrims can be managed according to the available space in a zone or level or building. In cluster verification more than one CM can be verified. But in capacity estimation, the verification and counting of the pilgrims is done at entrance and exit one by one. In cluster verification, cluster head generate the signals for verification and in response CMs generate signals for their identification. But in capacity estimation, pilgrims (sensor devices) generate signals (that is an event driven approach) for verification and in response zone sensors verify and count the number of pilgrims getting in or out from zone.

8. Simulation and Results

Contiki OS is a lightweight open source operating system for sensor network. It was designed by Adam Dunkels, at the Swedish Institute of Computer Sciences. C programming language is used in both Contiki OS and its programs. Contiki is flexibly portable OS and it has been ported to different platforms. Texas Instruments MSP-430 in addition to Atmel AT mega series of microcontrollers are the most often used platforms. Contiki employs event-driven programming model to deal with concurrence. It is Contiki's main advantage that all processes share one stack, allowing savings of memory. In order to utilize this model Proto threads. Unconditional blocking wait is provided by Proto threads to save the state when it blocks. It jumps back to the next instruction on resumption of Proto thread. It also employs Rime stack that is a stack for sensor networks for lightweight communication. The main purpose of Rime stack is to support code reusing and simplify the implementation of sensor networks. Contiki is used in several systems, such as electric meters, alarm generating systems, remote control of house, radiation detection and streetlights control. Cooja is Java based Contiki network simulator with a graphical user interface (GUI). It is a trial system given by Contiki, and additionally empowers with original instrumentation before running it on the real target hardware. It integrates to simulate with the external tools to provide additional features to the large and small networks. Motes can be emulated, which is faster and allows simulation of larger networks, or at the hardware level. Two emulator software packages are contained by this tool: Avrora and MSPSim. Cooja can emulate multiple platforms like: TelosB/SkyMote, MicaZ mote, Zolertia Z1 mote, ESB, Wismote [41].

The Figure 9 shows full screen view of Contiki/Cooja simulator. Network simulator "Contiki/Cooja" is separated into small windows such as: Network Window, Notes Window, Simulation Control Window and Mote Output Window.

Network Window is employed to draw the WSN in random manner, linear, ellipse or manual positions. Notes Window is employed to write down the information or detail about the simulations and detail of its results. In order to start, stop, pause or reload the simulations, simulation control Window is used. The results for data packets have transmitted and received is provided by Mote Output Window. For our scenario, we use the Tmote Sky platform. Tmote Sky is one of the high data rate platforms work with low power consumption. It works well for many months to years without replacing the battery. Therefore, it is one of the most demanding wireless sensors. It is notable that options are best used in a selection manner, as the window can become disordered with too many selected options. The number for each mote is displayed by Mote IDs. Mote Output window can be used for analyzing and filtering results. On the simulated motes, LEDs are beneficial in order to observe the LED lights.

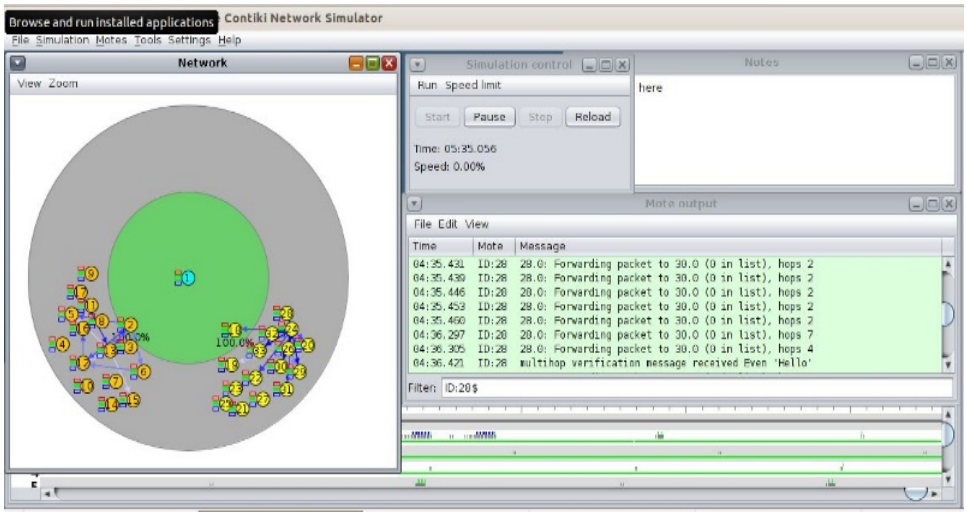


Figure 9: Full view of the Contiki/Cooja simulator

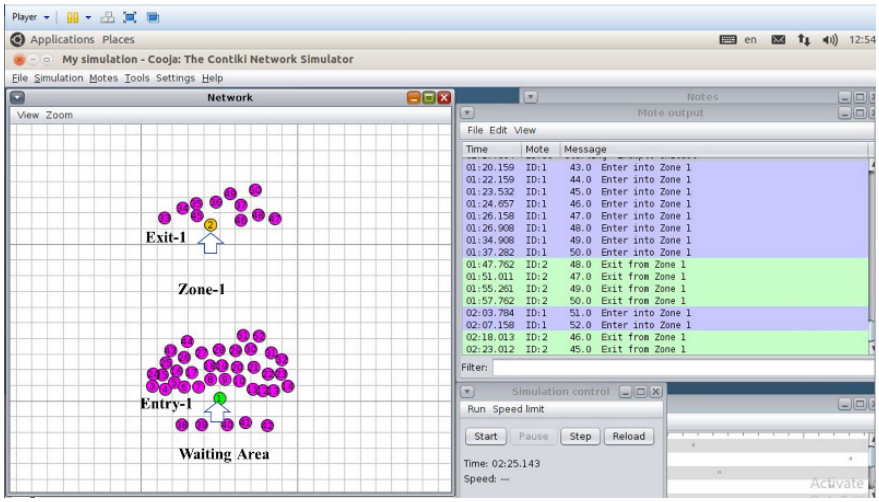


Figure 10: Single zone having maximum capacity 50

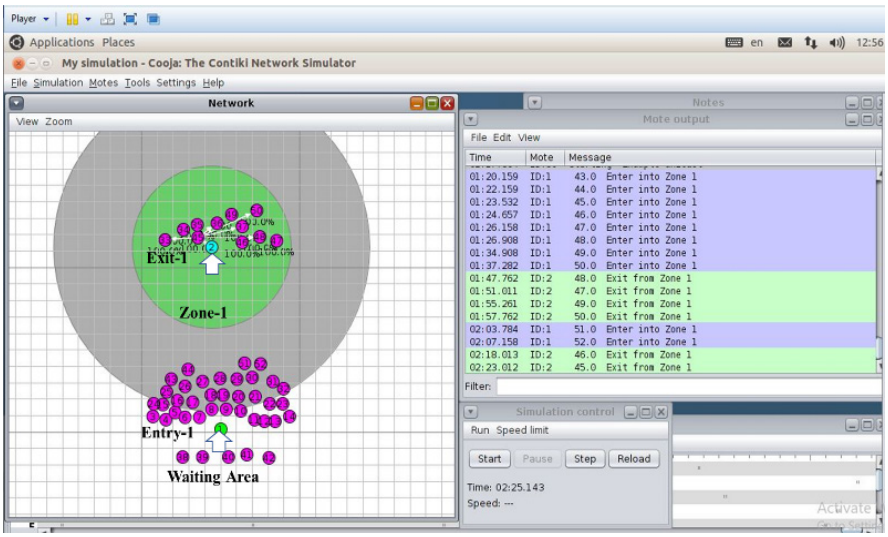


Figure 11: Single zone having maximum capacity 50 with coverage

We took the simple scenario to make understanding of the flow and function of each node as given in Figures 4 and 8. Head node function is to keep the record and status of CMs and forward record to the server for update. The status of the CM can be waiting to enter, enter and exit. Since the log file generated by the Contiki/Cooja simulation tool, we extracted the data for single and double zone in a level. For each zone we calculated the number of the pilgrims entered, exited, zone capacity occupied and available capacity for pilgrims. From this datum we can calculate the capacity for different positions and events. We also draw the table and graph for better understanding.

8.1. Single Zone Scenario

In single zone scenario, we consider one entry point and one exit point. We also set the maximum value for zone then we allow the specific number of the CMs to enter the zone and some of them allowed to exit from the zone. The simulation is shown in the Figures 10 to 18 and data is given in the Table 2. To make clarity of the figures, we consider the complete figure and then divided into its small parts (network, mote output and simulation control windows).

The simulation for maximum capacity 50 pilgrims is shown in the Figures from 10 to 14 and data is given in the Table 2.

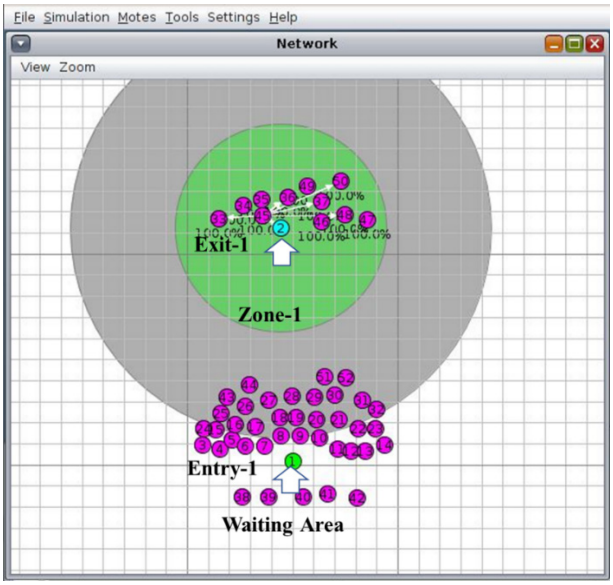


Figure 12: Network window from Figure 11

The Figures 10 and 11 represent the single zone having maximum capacity of 50 pilgrims. Figure 10 shows the capacity without coverage area while Figure 11 shows the capacity with coverage area in circular form.

The Figure 12 represents the network distribution of sensor nodes, entry point and exit point nodes in single zone having maximum capacity of 50 pilgrims. It also shows the number of nodes in waiting, occupy and exit position to calculate the available capacity.

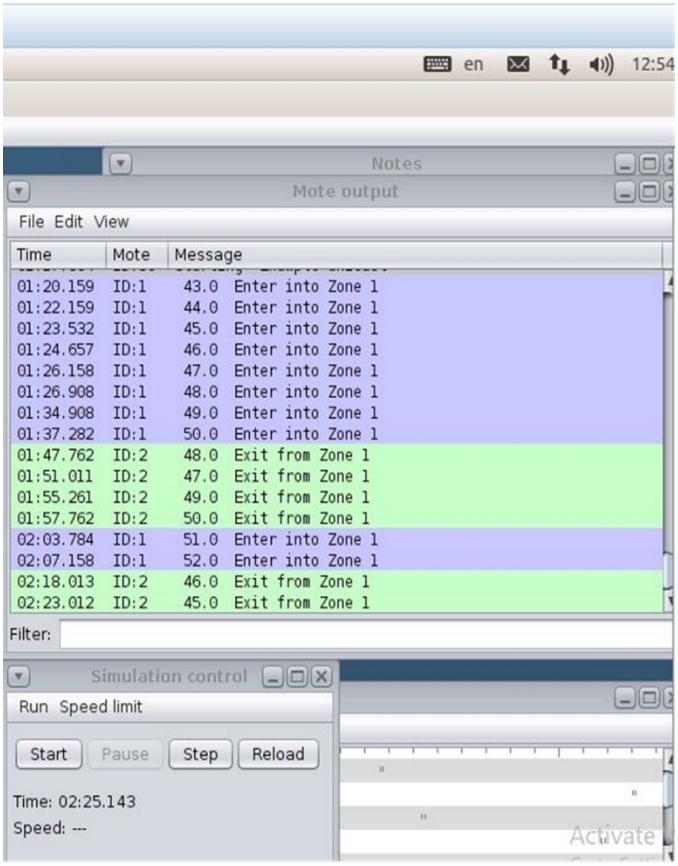


Figure 13: Mote output and simulation control windows from Figure 11

Figure 13 represents the most output with time, mote ID and its status enter or exit. The mote output is saved in the log file from where we can estimate the available capacity in a zone.

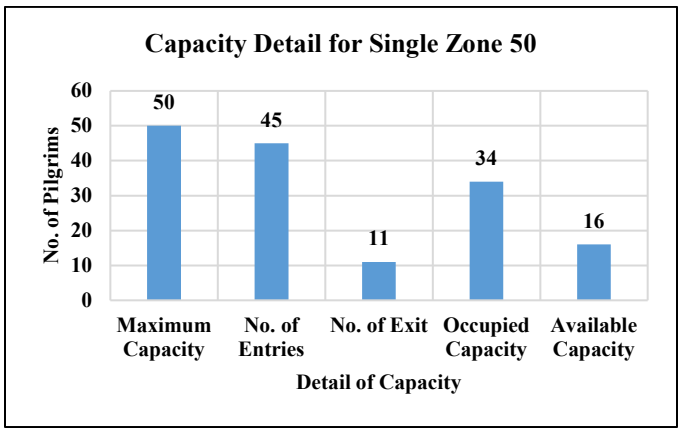


Figure 14: Capacity detail for single zone having maximum capacity 50

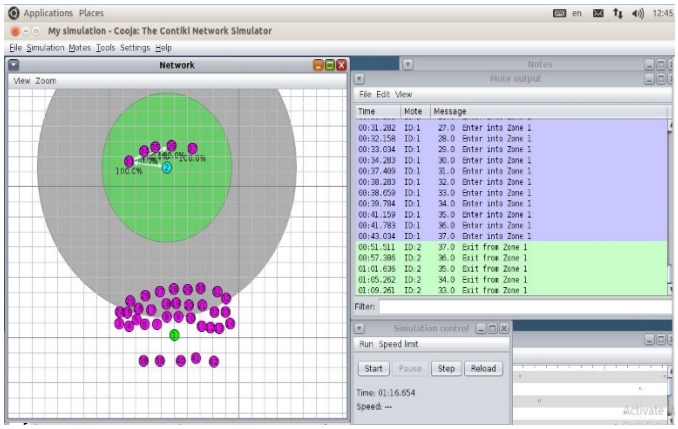


Figure 15: Single zone having maximum capacity 40 with coverage

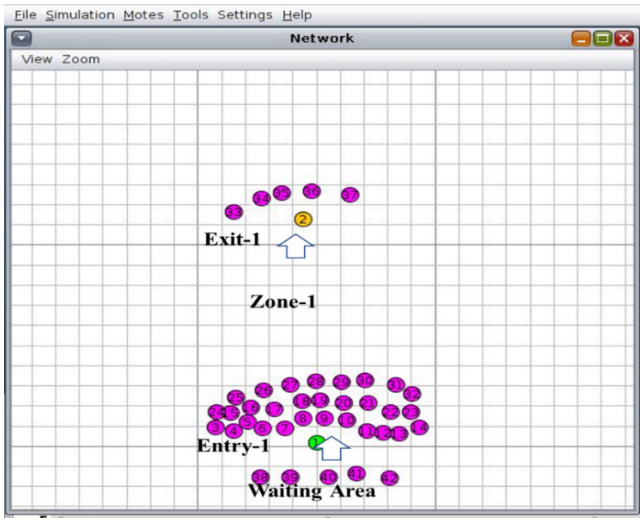


Figure 16: Network window from Figure 15

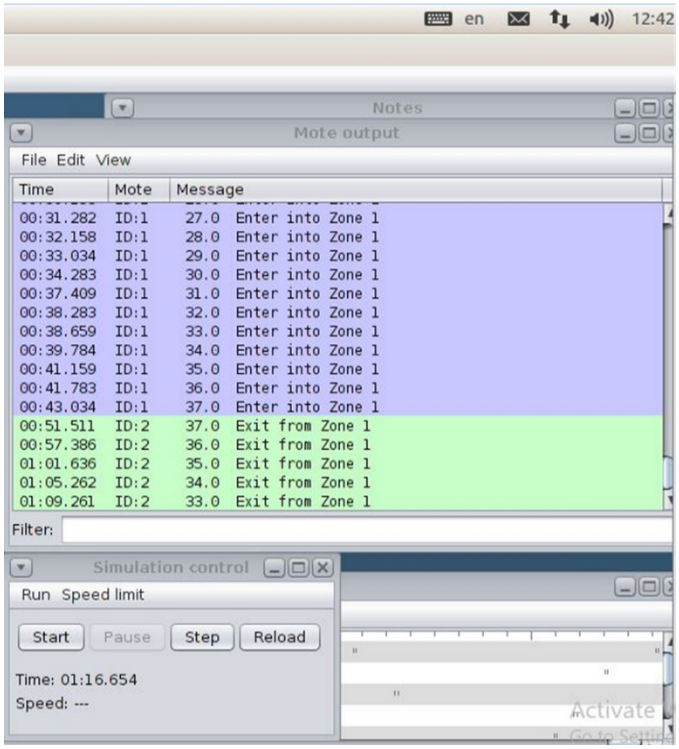


Figure 17: Mote output and simulation control windows from Figure 15

Figures from 10 to 14 considered that the maximum capacity for zone is 50 Pilgrims. Graph in Figure 14 shows that 45 of them have entered and 5 are in waiting position. 11 pilgrims have made the exit, occupied capacity in the zone is 34 and available capacity is 16.

The simulation for maximum capacity 40 is shown in the Figures from 15 to 18 and data is given in the Table 2. To make clarity of the figures, we consider the complete figure and then divided into its small parts (network, mote output and simulation control windows).

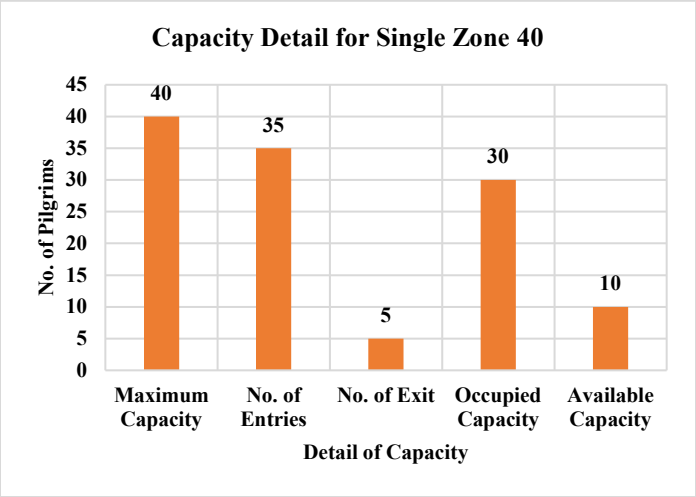


Figure 18: Capacity detail for single zone having maximum capacity 50

Figures from 15 to 18 show the maximum capacity for zone is 40 Pilgrims. Graph in Figure 18 shows that 35 out of 50 pilgrims have entered and 5 are in waiting position. 5 pilgrims have made the exit, occupied capacity in the zone is 30 and available capacity is 10.

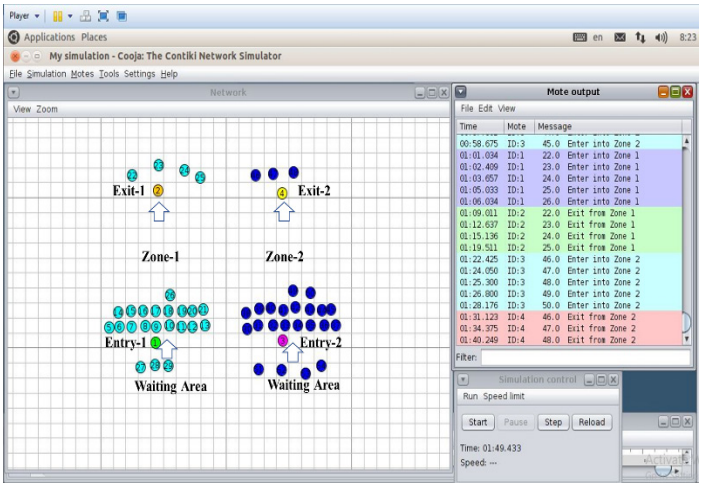


Figure 19: Each zone has maximum capacity 25 but total 50

8.2. Different Levels and Double Zone Scenarios

In different levels and double zone scenarios, we consider one entry point and one exit point in each zone and also consider that each level has double zone. We also set the maximum value for each zone and each level. We allowed the specific number of the nodes to enter in each zone and some of them allow to exit from each zone. The simulation is shown in the Figures 19-32 and data are given in the Table 2.

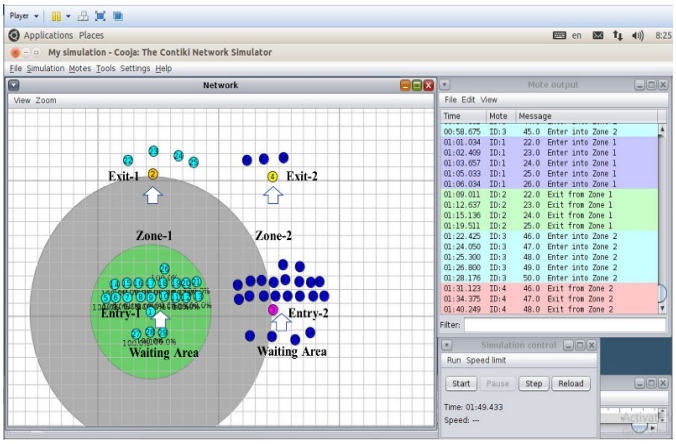


Figure 20: Each zone has maximum capacity 25 but total 50 with coverage

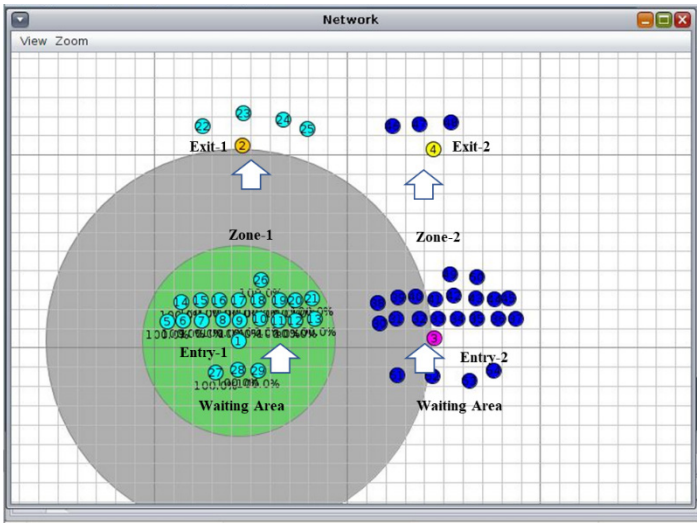


Figure 21: Network window from Figure 20

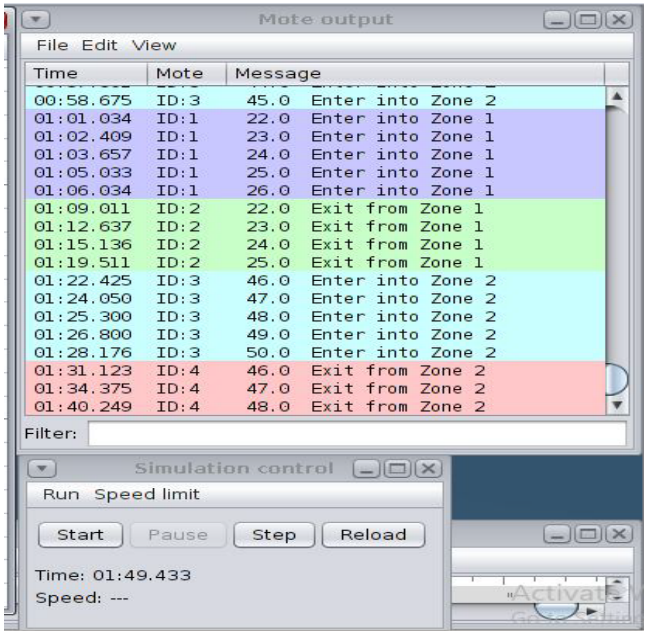


Figure 22: Mote output and simulation control windows from Figure 20

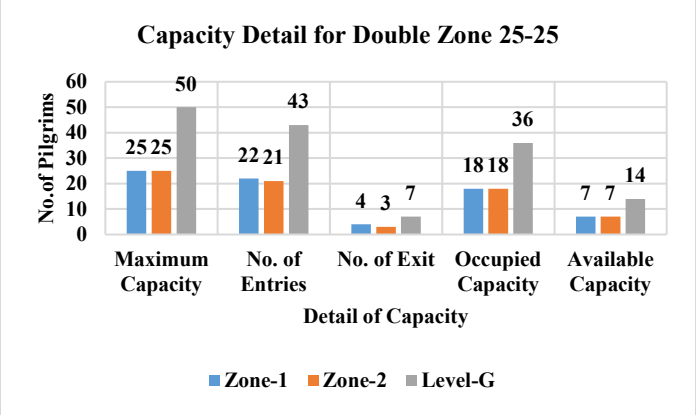


Figure 23: Capacity detail for double zones having maximum capacity 25 in each but Level-G has maximum capacity 50

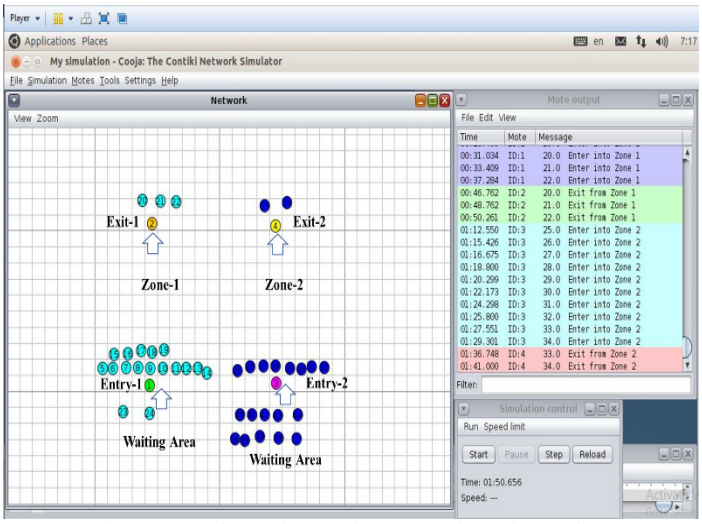


Figure 24: Each zone has maximum capacity 20 but total 40

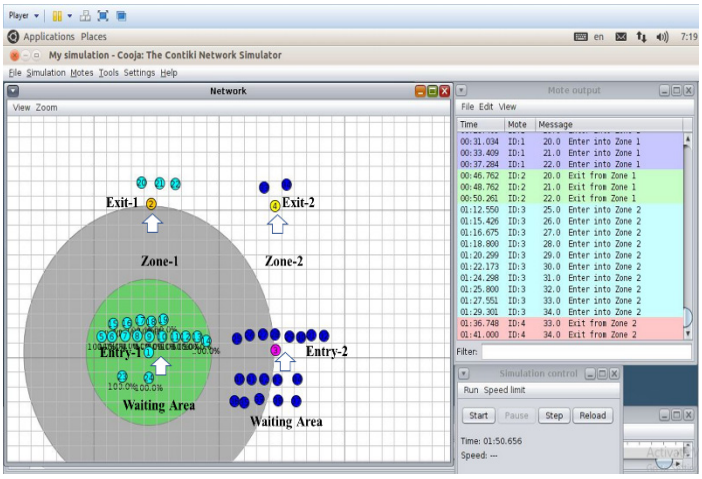


Figure 25: Each zone has maximum capacity 20 but total 40

In Figures 19 to 23, it is considered that the maximum capacity for zone-1 and zone-2 is 25. Zone-1 and zone-2 made the level-G and total capacity is 50 Pilgrims. In zone-1, 22 have entered and 3 are in waiting position. 4 pilgrims have made the exit, occupied capacity in the zone is 18 and available capacity is 7. In zone-2, 21 have entered and 4 are in waiting position. 3 pilgrims have made the exit, occupied capacity in the zone is 18 and available capacity is 7. From the log file or zone-1 and zone-2, we calculated the capacities for Level-G. Level-G has maximum capacity 50 pilgrims. 43 have entered and 7 are in waiting position. 7 pilgrims have made the exit, occupied capacity in the zone is 36 and available capacity is 14.

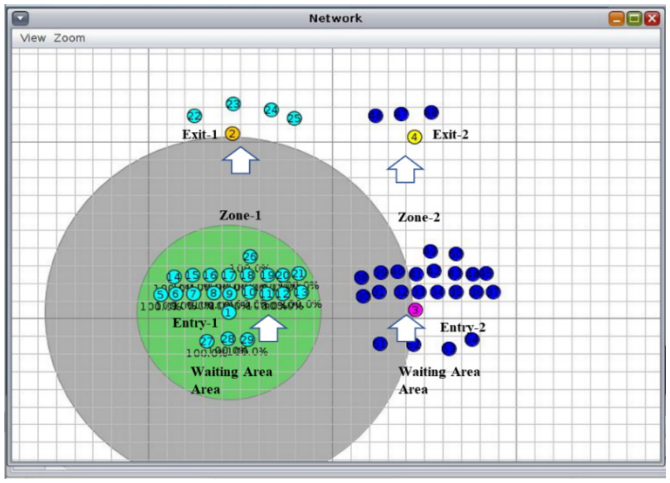


Figure 26: Network window from Figure 25

In Figures 24 to 28, it is considered that the maximum capacity for zone-1 and zone-2 is 20 pilgrims. Zone-1 and zone-2 made the level-1 and total capacity in level-1 is 50 Pilgrims. In zone-1, 18 have entered and 2 are in waiting position. 3 pilgrims have made the exit, occupied capacity in the zone is 15 and available capacity is 5. In zone-2, 10 have entered and 10 are in waiting position. 2 pilgrims have made the exit, occupied capacity in the zone is 8 and available capacity is 12. From the log file or zone-1 and zone-2, we calculated the capacities for Level-1. Level-1 has maximum capacity 40 pilgrims. 28 have entered and 12 are in waiting position. 5 pilgrims have made the exit, occupied capacity in the zone is 23 and available capacity is 17.

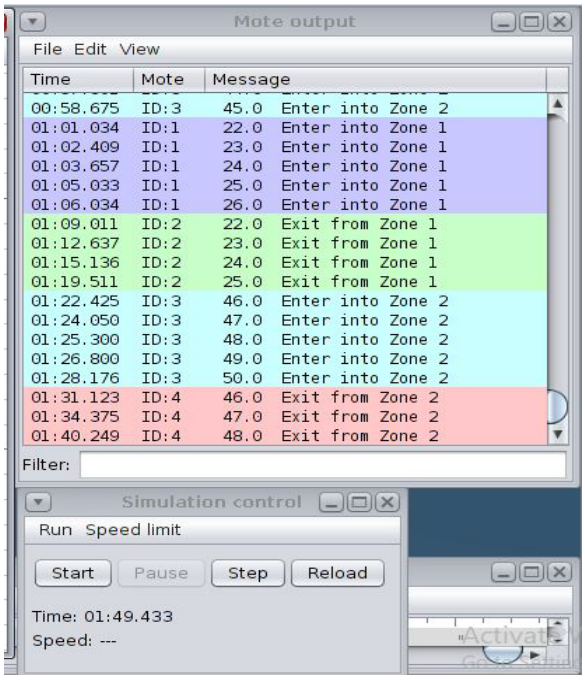


Figure 27: Mote output and simulation control windows from Figure 25

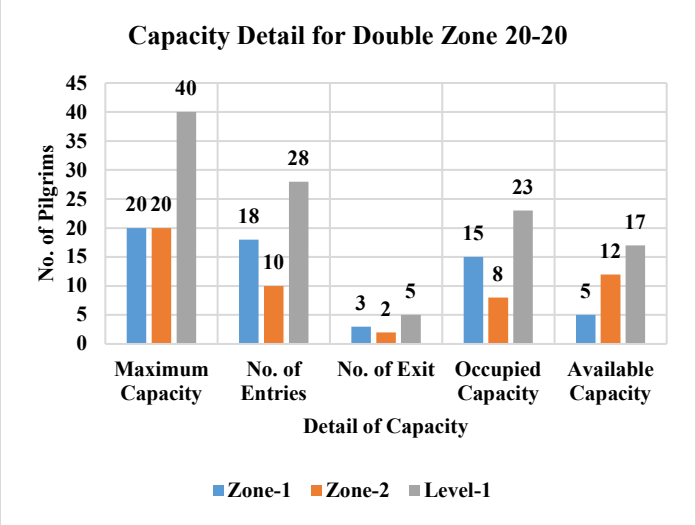


Figure 28: Capacity detail for double zones having maximum capacity 20 in each but Level-1 has maximum capacity 40

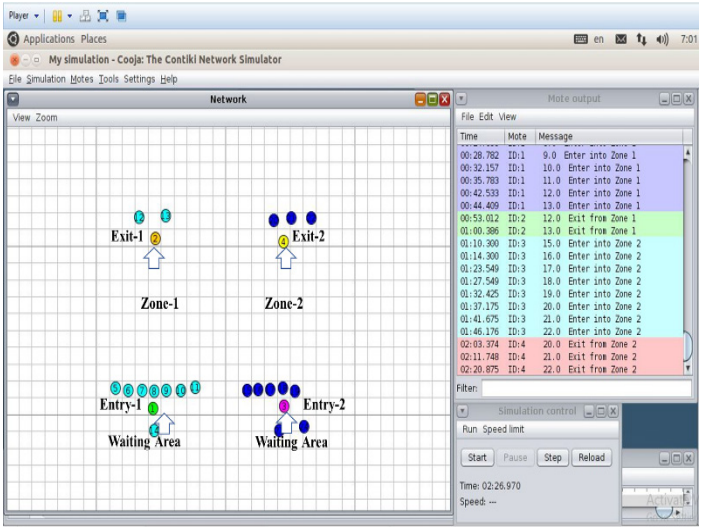


Figure 29: Each zone has maximum capacity 10 but total 20

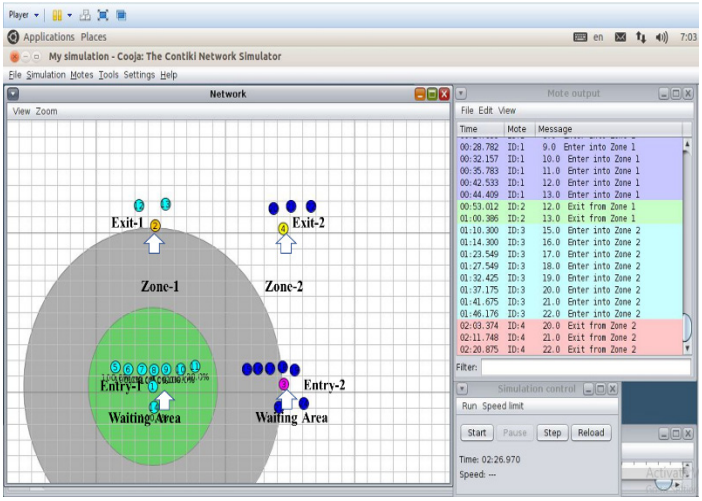


Figure 30: Each zone has maximum capacity 10 but total 20 with coverage

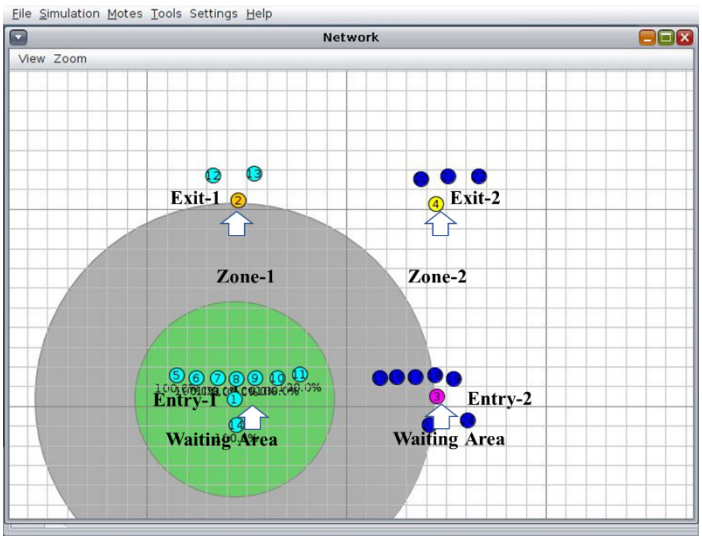


Figure 31: Network window from Figure 30

In Figures 29 to 32, it is considered that the maximum capacity for zone-1 and zone-2 is 10 pilgrims. Zone-1 and zone-2 made the level-2 and total capacity in level-2 is 20 Pilgrims. In zone-1, 9 have entered and 1is in waiting position. 2 pilgrims have made the exit, occupied capacity in the zone is 7 and available capacity is 3. In zone-2, 8 have entered and 2 are in waiting position. 3 pilgrims have made the exit, occupied capacity in the zone is 5 and available capacity is 5. From the log file or zone-1 and zone-2, we calculated the capacities for Level-2. Level-2 has maximum capacity 20 pilgrims. 17 have entered and 3 are in waiting position. 5 pilgrims have made the exit, occupied capacity in the zone is 12 and available capacity is 8.

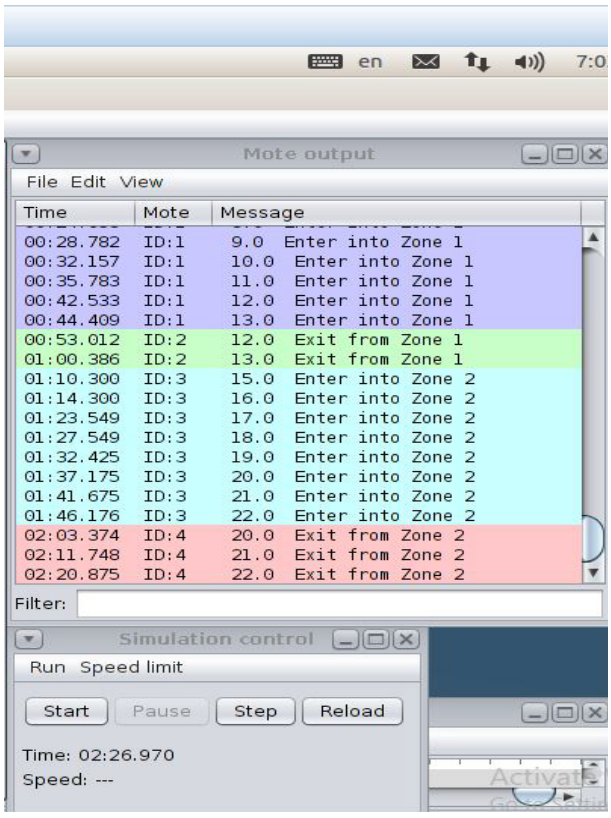


Figure 32: Mote output and simulation control windows from Figure 30

Table 2: Zones and Levels and capacities

Zone/Level	Maximum Capacity	No. of Entries	No. of Exit	Occupied Capacity	Available Capacity
Single Zone	50	45	11	34	16
	40	35	5	30	10
Level/Double Zone					
Zone-1	25	22	4	18	7
Zone-2	25	21	3	18	7
Level-G	50	43	7	36	14
Zone-1	20	18	3	15	5
Zone-2	20	10	2	8	12
Level-1	40	28	5	23	17
Zone-1	10	9	2	7	3
Zone-2	10	8	3	5	5
Level-2	20	17	5	12	8

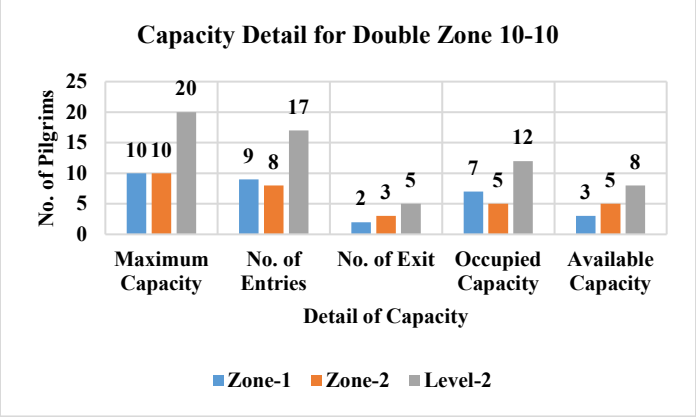


Figure 33: Capacity detail for double zones having maximum capacity 10 in each but Level-2 has maximum capacity 20

9. Discussion

The present analysis deals with capacity estimation for pilgrims at crowded place. The prime focus is on the allotment of zone and level number. This is performed according to moving status and events, single or multi sink, entry and exit point.

The concern of discussion is the allotment of zone and level number to pilgrims. According to discussion allotment of zone and level number to pilgrims should be specific (fixed) or flexible for dynamic position and events accordingly. Specific zone and level number according to dynamic positions and events is preferred because it is noticed if allotment is flexible then some zone and level number are more crowded than available capacity while others remain empty or less occupied. Furthermore, prayer position requires more area than standing and sitting position comparatively as given in section (5.1). As compared to routine days, pilgrimage, Ramadan and Jummah needs more area. The above mention reasons lead to non-uniform distribution of crowd. What is more, pilgrims having flexible zone and level move in different directions. It can result in congestions (bottleneck) that can lead to stampede. But in particular case, pilgrims are allotted the specific tracks to approach their zone and level number, that is why, chances of bottleneck and stampede are lessened or minimized. In flexible case, pilgrims are unable to move flexibly from overcrowded to less crowded zone or level (where capacity is available). This is because they have right to enter any zone and level number. But in particular case, pilgrims are allotted by the zone and level number. They can only move into allotted zone and level number. As allotment is performed considering maximum capacity of zone or level (a proactive approach), the zone or level does not get overcrowded. Place can be occupied within the zone on the basis of first come and first serve.

Entry point and exit point is the second point of discussion. In case of flexible, the pilgrims can enter and exit from any one of the points. Again, some entry points become more crowded than the other and cause bottleneck and there are chances of stampede. Moreover, pilgrims' attempt of entry and making exit from same location also causes bottleneck and stampede. But in specific case, each zone and level have particular point of entry and exit. In this way, chances of problem of bottleneck and stampeded can be overcome. Furthermore, pilgrims can conveniently access the transport of their residential area.

Third point of discussion is that there should be a single or multiple sink for WSN system. Massive crowd, scalability, latency, reliability, and network lifetime are the numerous challenges that may be created by the design of WSNs. The question whether sink should be single or multi sink is dependent upon its application. In the case of single sink and massive crowd, single sink has to control all traffic. The streaming of data traffic will be in one direction which causes congestion. There are probabilities that single sink is ineffective to control massive amount of data which causes the drop of data. This drop of data will cause the high latency rate. However, there are separate solutions to control the massive crowd and data in case the multi sinks are used. In the case of multi sink, data flow will be in multiple directions rather in one direction. Probability of congestion can be minimized in this way. In order to avoid the congestion and to distribute the task among sinks different type of data traffic or task can be allotted to different sinks. If the data is dropped on one path, it can be received on the alternative path. In this way latency rate can be minimized. When

data is received from different paths and sinks, it will provide the data redundancy. Single sink solution is not appropriate in case of long-distance communication because transferring data to long distance with the support of sensor nodes means consumption of energy to a lot degree. This consumption of energy will lead the sensor node die early and hence it will shorten the network life span. But in case of multi sink, the sensor nodes will get the shortest path as multi sink will divide the path and will consume less energy by increasing the network life span. Increasing of lifetime of network and receiving of data from multiple courses increase the reliability of the network. In our case, as data is massive, crowd is large, dynamic, and on large area therefore WSN required the multi sinks solution.

10. Conclusion and Future Directions

This paper provides an overview of a dynamic positions and events, proactive approach with framework and its implications towards smart capacity estimation using WSN and cloud computing in the IoT paradigm.

We discussed the framework and proactive approach in the context of dynamic position and events for smart capacity estimation. We examined how smart capacity estimation automate and minimizes time to manage available capacity in a zone or level by using framework and proactive. Explanation is done by use-case of masjid Al-Haram for smart capacity estimation for crowd management. We also compared the existing capacity estimation and smart capacity estimation systems. We also defined the algorithms with respect to Zone and Level. We also ran different simulation scenarios and also discussed the results. In future we will calculate the capacity according to the different position. The dynamic capacity can be calculated on the basis of time frame and capacity can be maximized and minimized according to time frame. In future algorithms can be defined how to maintain social distancing, how to calculate the capacity with social distancing and find the empty space with direction? Algorithms devised how to define the exit plan to move out without facing the congested path. Smart capacity estimation with Sensor and without sensor. Without sensor, capacity of pilgrims can be counted by putting the manual counter. If proposed proactive approach and framework for smart capacity estimation is commercially applied, it can provide help in different application developments, such as: 1. Capacity available in a zone or at a level can be calculated precisely. 2. To estimate the occupied and remaining capacity in a building or in an area. 3. Alert generation before the zone or level reached to its maximum threshold. 4. Positions and Events based capacity estimation. 5. Capacity estimation on the basis of social distancing. 6. Calculation of the social distance. 7. Drinking water estimation to supply the water in different zones and levels. 8. On the basis of wheelchairs' requests, wheelchairs supply plan according to demand in different zones and levels. 9. Estimation of the Holy Quran scripts according to different languages. 10. If water, chairs or the Holy Quran scripts are not enough in a zone or level then generate the alert to workers for supply. 11. Sharing of zone capacity data with different zone security officers to manage the crowd capacity. 12. In case of emergency, provide path for evacuation from affected zone or level to outside.

References

[1] N.A. Nawaz, H.R. Malik, A.J. Alshaor, "A framework for smart capacity estimation at crowded area using WSN," in Proceedings - 2019 International Arab Conference on Information Technology, ACIT 2019, 159–164, 2019, doi:10.1109/ACIT47987.2019.8991037.

[2] H. Paloheimo, M. Lettenmeier, H. Waris, "Transport reduction by crowdsourced deliveries – a library case in Finland," Journal of Cleaner Production, **132**(2016), 240–251, 2016, doi:10.1016/j.jclepro.2015.04.103.

[3] W. Xi, J. Zhao, X.Y. Li, K. Zhao, S. Tang, X. Liu, Z. Jiang, "Electronic frog eye: Counting crowd using WiFi," in Proceedings - IEEE INFOCOM, 361–369, 2014, doi:10.1109/INFOCOM.2014.6847958.

[4] C.C. Loy, K. Chen, S. Gong, T. Xiang, Crowd Counting and Profiling: Methodology and Evaluation, 347–382, 2013, doi:10.1007/978-1-4614-8483-7_14.

[5] B. Guo, Z. Wang, Z. Yu, Y. Wang, N.Y. Yen, R. Huang, X. Zhou, "Mobile Crowd Sensing and Computing: The review of an emerging human-powered sensing paradigm," ACM Computing Surveys, **48**(1), 1–31, 2015, doi:10.1145/2794400.

[6] S. Hashish, M. Ahmed, "Efficient wireless sensor network rings overlay for crowd management in Arafat area of Makkah," in 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, SPICES 2015, 1–6, 2015, doi:10.1109/SPICES.2015.7091437.

[7] H. Elazhary, "Cloud-based Context-aware Mobile Applications and Framework for Hajj and Umrah Management," International Journal of Computer Trends and Technology, **47**(2), 106–115, 2017, doi:10.14445/22312803/ijctt-v47p115.

[8] F. Tao, J. Panneerselvam, T. Holding, L. Liu, "A cloud-based sustainable business model for effective ICT provision in higher education," in Proceedings - 9th IEEE International Symposium on Service-Oriented System Engineering, IEEE SOSE 2015, 222–228, 2015, doi:10.1109/SOSE.2015.9.

[9] T. Gamage, D. Anderson, D. Bakken, K. Birman, A. Bose, C. Hauser, K. Maheshwari, R. Van Renesse, Mission-critical cloud computing for critical infrastructures, 1–16, 2017, doi:10.1201/b16908.

[10] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C.H. Youn, "Wearable 2.0: Enabling Human-Cloud Integration in Next Generation Healthcare Systems," IEEE Communications Magazine, **55**(1), 54–61, 2017, doi:10.1109/MCOM.2017.1600410CM.

[11] K. Nowicka, "Competitiveness of Polish enterprises in relation to the potential of cloud computing," Kwartalnik Nauk o Przedsiębiorstwie, **42**(1), 85–99, 2017, doi:10.5604/01.3001.0010.0158.

[12] N.A. Nawaz, A. Waqas, Z.M. Yusof, A. Shah, "A framework for smart estimation of demand-supply for crowdsource management using WSN," in ACM International Conference Proceeding Series, 1–5, 2017, doi:10.1145/3018896.3025140.

[13] Y. Mao, C. You, J. Zhang, K. Huang, K.B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys and Tutorials, **19**(4), 2322–2358, 2017, doi:10.1109/COMST.2017.2745201.

[14] S.H.H. Madni, M.S.A. Latiff, Y. Coulibaly, S.M. Abdulhamid, "Resource scheduling for infrastructure as a service (IaaS) in cloud computing: Challenges and opportunities," Journal of Network and Computer Applications, **68**(2016), 173–200, 2016, doi:10.1016/j.jnca.2016.04.016.

[15] S. Goyal, "Public vs Private vs Hybrid vs Community - Cloud Computing: A Critical Review," International Journal of Computer Network and Information Security, **6**(3), 20–29, 2014, doi:10.5815/ijcnis.2014.03.03.

[16] C. Stergiou, K.E. Psannis, B.G. Kim, B. Gupta, "Secure integration of IoT and Cloud Computing," Future Generation Computer Systems, **78**(2018), 964–975, 2018, doi:10.1016/j.future.2016.11.031.

[17] N. Pavón-Pulido, J.A. López-Riquelme, R. Torres, R. Morais, J.A. Pastor, "New trends in precision agriculture: a novel cloud-based system for enabling data storage and agricultural task planning and automation," Precision Agriculture, **18**(16), 1038–1068, 2017, doi:10.1007/s11119-017-9532-7.

[18] M.A. Zamora-Izquierdo, J. Santa, J.A. Martínez, V. Martínez, A.F. Skarmeta, "Smart farming IoT platform based on edge and cloud computing," Biosystems Engineering, **177**(2019), 4–17, 2019, doi:10.1016/j.biosystemseng.2018.10.014.

[19] E.A. Shammar, A.T. Zahary, "The Internet of Things (IoT): a survey of techniques, operating systems, and trends," Library Hi Tech, **38**(1), 5–66, 2019, doi:10.1108/LHT-12-2018-0200.

[20] J. Qiu, Z. Tian, C. Du, Q. Zuo, S. Su, B. Fang, "A survey on access control in the age of internet of things," IEEE Internet of Things Journal, **7**(6), 4682–4696, 2020, doi:10.1109/JIOT.2020.2969326.

[21] R. Zheng, T. Zhang, Z. Liu, H. Wang, "An EIoT system designed for ecological and environmental management of the Xianghe Segment of China's Grand Canal," International Journal of Sustainable Development and World Ecology, **23**(4), 372–380, 2016, doi:10.1080/13504509.2015.1124470.

[22] X. Li, D. Li, J. Wan, A. V. Vasilakos, C.F. Lai, S. Wang, "A review of industrial wireless networks in the context of Industry 4.0," Wireless Networks, **23**(1), 23–41, 2017, doi:10.1007/s11276-015-1133-7.

[23] M. Torres-Ruiz, J.H. Juárez-Hipólito, M.D. Lytras, M. Moreno-Ibarra, "Environmental noise sensing approach based on volunteered geographic information and spatio-temporal analysis with machine learning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 95–110, 2016, doi:10.1007/978-3-319-42089-9_7.

[24] J.H.J. Hipólito, M.A.M. Ibarra, M. Torres-Ruiz, G. Guzmán, R. Quintero, Innovation on user-generated content for environmental noise monitoring and analysis in the context of smart cities, 490–519, 2018, doi:10.4018/978-1-5225-7033-2.ch021.

[25] A. Kumar, M. Zhao, K.J. Wong, Y.L. Guan, P.H.J. Chong, "A comprehensive study of IoT and WSN MAC protocols: Research issues, challenges and opportunities," IEEE Access, **6**(2018), 76228–76262, 2018, doi:10.1109/ACCESS.2018.2883391.

[26] K. Zkik, G. Orhanou, S. El Hajji, "A new secure framework in WSNs using ECC: Medical application," in Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017, 1–7, 2018, doi:10.1109/ICEngTechnol.2017.8308144.

[27] T. hoon Kim, C. Ramos, S. Mohammed, Smart City and IoT, Future Generation Computer Systems, 159–162, 2017, doi:10.1016/j.future.2017.03.034.

[28] F. Xiao, Q. Miao, X. Xie, L. Sun, R. Wang, "Indoor Anti-Collision Alarm System Based on Wearable Internet of Things for Smart Healthcare," IEEE Communications Magazine, **56**(4), 53–59, 2018, doi:10.1109/MCOM.2018.1700706.

[29] M. Al-Kuwari, A. Ramadan, Y. Ismael, L. Al-Sughair, A. Gastli, M. Benammar, "Smart-home automation using IoT-based sensing and monitoring platform," in Proceedings - 2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering, CPE-POWERENG 2018, 1–6, 2018, doi:10.1109/CPE.2018.8372548.

[30] W. He, G. Yan, L. Da Xu, "Developing vehicular data cloud services in the IoT environment," IEEE Transactions on Industrial Informatics, **10**(2), 1587–1595, 2014, doi:10.1109/TII.2014.2299233.

[31] N.A. Nawaz, A. Waqas, Z.M. Yusof, A.W. Mahesar, A. Shah, "WSN based sensing model for smart crowd movement with identification: An extended study," Journal of Theoretical and Applied Information Technology, **95**(5), 975–988, 2017.

[32] X. Li, J. Niu, S. Kumari, F. Wu, A.K. Sangaiah, K.K.R. Choo, "A three-factor anonymous authentication scheme for wireless sensor networks in internet of things environments," Journal of Network and Computer Applications, **103**(2018), 194–204, 2018, doi:10.1016/j.jnca.2017.07.001.

[33] M.J. McGrath, C.N. Scanaill, M.J. McGrath, C.N. Scanaill, Sensor Network Topologies and Design Considerations, 79–95, 2013, doi:10.1007/978-1-4302-6014-1_4.

[34] D.L. Guidoni, F.S.H. Souza, J. Ueyama, L.A. Villas, "RouT: A routing protocol based on topologies for heterogeneous wireless sensor networks," IEEE Latin America Transactions, **12**(4), 812–817, 2014, doi:10.1109/TLA.2014.6868887.

[35] H.R. Faragardi, H. Fotouhi, T. Nolte, R. Rahmani, "A Cost Efficient Design of a Multi-sink Multi-controller WSN in a Smart Factory," in Proceedings - 2017 IEEE 19th Intl Conference on High Performance Computing and Communications, HPCC 2017, 594–602, 2018, doi:10.1109/HPCC-SmartCity-DSS.2017.77.

[36] E. Fadel, V.C. Gungor, L. Nassef, N. Akkari, M.G. Abbas Malik, S. Almasri, I.F. Akyildiz, "A survey on wireless sensor networks for smart grid," Computer Communications, **71**(2015), 22–33, 2015, doi:10.1016/j.comcom.2015.09.006.

[37] H. Toral-Cruz, F. Hidoussi, D.E. Boubiche, R. Barbosa, M. Voznak, K.I. Lakhtaria, A survey on wireless sensor networks, 171–210, 2015, doi:10.4018/978-1-4666-8687-8.ch006.

[38] R. Zhu, Y. Qin, J. Wang, "Energy-aware distributed intelligent data gathering algorithm in wireless sensor networks," International Journal of Distributed Sensor Networks, **7**(1), 235724, 2011, doi:10.1155/2011/235724.

[39] A. Waheed Khan, A.H. Abdullah, M.H. Anisi, J. Iqbal Bangash, "A comprehensive study of data collection schemes using mobile sinks in wireless sensor networks," Sensors (Switzerland), **4**(2), 2510–2548, 2014, doi:10.3390/s140202510.

[40] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," Information Systems, **47**(2015), 98–115, 2015, doi:10.1016/j.is.2014.07.006.

[41] E.N. Sharad, I.K. Aulakh, "Evaluation and implementation of cluster head selection in WSN using Contiki/Cooja simulator," Journal of Statistics and Management Systems, **23**(2), 407–418, 2020, doi:10.1080/09720510.2020.1736324.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 439-446 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Effective Learning of Tax Regulations using Different Chatbot Techniques

Rafael Mellado-Silva*, Antonio Faúndez-Ugalde, María Blanco-Lobos

School of Commerce, Faculty of Economic and Administrative Sciences, Pontificia Universidad Católica de Valparaíso, Valparaíso, 2540000, Chile

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 01 November, 2020
Online: 20 November, 2020

Keywords:
Chatbot
Tax Learning
Automation
Decision tree

ABSTRACT

Teaching tax-related regulations have always been a challenge due to the inclusion of external variables that hinder the learning process, such as the high complexity of tax systems and legislation variability. Universities have sought different alternatives to support the teaching process both outside and inside the classroom, ranging from recreational activities to active learning. This article will show the experience resulting from using a chatbot to support learning in accounting students for the teaching of tax regulations related to the Chilean tax system, comparing two types of tools, on the one hand, a free conversation chatbot using natural language processing versus a rule-based chatbot driven by a decision tree. The experimentation process was carried out with 50 higher education students, divided into an experimental group and a control group, in two different courses. The results obtained demonstrated in both cases greater effectiveness of the use of the chatbot in learning the tax matter, both in the free conversation chatbot where the experimental group obtained a 15.7% improvement versus the control group that obtained a 1.05% improvement, as in the chatbot that applied decision tree where the experimental group obtained a 32% improvement versus the control group with 5.2%. Considering the complexity of the content in tax matters and the little innovation in the existing teaching subjects in the area and that the students improve learning using both chatbot tools compared to other learning techniques, is considered a relevant contribution to teaching innovation.

1. Introduction

This article is the continuation of a lengthy investigation developed in courses on the subject of taxation [1], which seeks to generate support for learning the Chilean tax system through the use of a chatbot with an automated conversational system technique; these technological systems aim to establish new mechanisms that have a direct impact on student learning, especially on issues such as the tax regime of a country, an area characterized by the complexity in the application of tax regulations.

Significant studies have taken place about the behavior and learning of students [2], [3], where motivation has been one of the key factors due to its direct impact on the perception and predisposition that students have towards the teaching activities delivered by teachers [4]. In this sense, the lack of motivation is one of the most critical causes when studying the failure and dropout of educational systems [5]. Socioeconomic and health factors also influence learning processes [6], [7], as well as the design problems of the different learning resources created by teachers [8], [9]. However, what research mostly agrees with is that this is a cross-cutting problem in society [3], [10], [11].

One technique that has become a common factor is the use of technology and software to support the teaching process and enhance the students learning [12], [13]. The integration of technologies in the classroom is a constant challenge that teachers and educational communities must face, but it is necessary based on the global demands on students' development of skills [14]. Based on this, there are numerous innovations, but the most outstanding and most attractive to students are those related to digital credentials, virtual assistants, and blockchain [15], [16].

This work puts into application in educational innovation, the use of a chatbot for learning theory, and application of Chilean tax regulations. The learning processes of tax regulations are always a

*Corresponding Author: Rafael Mellado-Silva, Pontificia Universidad Católica de Valparaíso, +56 32 2273340, Email: rafael.mellado@pucv.cl

great challenge for education and tax entities [17], [18]. In the case of Chile, the last ten years have seen critical structural changes in the tax system related to income tax, the reforms of which have been analyzed and studied [19], both due to the complexity that the tax system has acquired well as the particular characteristics defined for corporate taxes, which have direct effects on the economies of the countries [20]. Because of the above, different investigations have revealed the importance of teaching correctly and innovatively the subject of taxes [21]. With this, it is that innovative teaching mechanisms must be sought to produce effective learning in students.

The research objective is to compare the effectiveness in the learning of tax regulations in students of accounting careers through the use of two chatbot techniques, that is, one, through an automated conversational system and, two, based on a tree decision. This article presents the result of applying chatbots linked to the study of the tax system regulations, which cannot be exempt from applying, integrating, and using information technologies in their learning process [22].

Considering the complexity of the tax systems and the constant legislative variability produce an effect of obsolescence in the subjects, affecting the students' rejection towards these subjects and, consequently, very high rates of failure and dropout [23]. To address this problem, there were two chatbots with knowledge of tax matters. The first with a free communication with the students as applied in [1] and the second using a decision tree. The study was carried out in an N=50, dividing the students into an experimental and control group, in the accounting career in a Chilean higher education institution. The results obtained are promising, where, in the case of the free communication chatbot there is a difference of 14.65% in favor of the experimental group and in the case of the decision tree chatbot there was a difference of 26.8% in the results in favor of the experimental group.

For this article, we will consider a chatbot as a software tool that allows users to have a conversation between a human being and an artificial entity [24]. One of these technology precursors is Joseph Weizenbaum, with Eliza [25] applied to psychoanalysis with promising results [26]. As the winners of the Loebner [27] contest point out, chatbots have evolved from simple pattern matching systems to increasingly complicated patterns of computer interaction and reasoning, being applied in education [28], entertainment [29], health [30], among others. The rise of chatbots has occurred in customer service, such as Hennes & Mauritz AB, which seeks to replicate a human conversation to facilitate information collection by a customer [31]. The purpose of chatbots is to automate the process of communication and support for people [32].

At an educational level, there have been various applications to chatbots with mixed results, but most of them agree that an improvement in learning and student satisfaction is evident [33], [34], generating greater interest in the subjects test [35]. For this reason, three main factors stand out to determine their participation, the performance expectation, the effort expectation, and the habit [36], which indicates that the commitment is linked to the benefits versus the effort to which the students are subjected by supplementing your learning with a chatbot system. This is why it is critical to innovate in non-traditional areas and that the effect has a direct application in business, accounting, and auditing matters.

The research questions posed are in line with determining the effectiveness of the two tools applied and their impact on effective learning:

- Does the use of the two chatbot techniques positively contribute to learning tax regulations?
- Is learning through a chatbot more effective than through the use of gamified activities and e-learning videos?

Based on the above, we established as a research hypothesis:

Incorporating a chatbot, independent of the applied technique, for the learning process of tax regulations in higher education provides better results due to the students' academic performance.

Due to the above, a null hypothesis is also established:

The inclusion of technological tools, particularly chatbots, does not improve effective learning on tax matters compared to traditional methodologies or e-learning.

Concerning the structure of this paper, the theoretical framework that will deliver the definitions and general aspects that must be taken into account is presented below; later, other experiences applied in the matter will be seen, and then proceed with the experimental design and case study for the two techniques to be compared. This article ends with the results obtained and their corresponding discussion.

2. Related Works

Australian work by [46] shows that a significant number of accounting schools resist the adoption of new technologies in their training processes. The reasons are the lack of interest in implementing technologies with an extra workload, the little support, and the lack of resources, but the most relevant is the lack of educators' time to learn and develop skills and competencies to adopt new technologies in education systems efficiently. The work recommends considering a new model that reflects the current innovative technological nature to provide education, anywhere, anytime, and for anyone.

As far as tax learning is concerned, the work of [47] mentions how tax education faces significant challenges in the 21st century and delivers results to change current educational paradigms, prompting a more effective teaching practice based on conceptual knowledge over technical ability. In [48] it is explained how the subject of taxes is elementary to complement other subjects within accounting programs, and it seeks to find effective teaching methods using computer software, which, although it presents good results, its use is not every day. In [49] the work to support the understanding of tax systems and the effect on the systemic change that arises when different organizations are willing to learn from each other is presented. For this, real case studies are used in the companies Oxfam and Unilever. In Spain, there is the work of [50], which analyzes part of the challenges that exist for teaching tax law, considering students who are studying in a highly multidisciplinary degree; and in [51] whose economic concept is the distribution of the tax burden between buyers and sellers, it is

suggested that collaboration between teachers is more effective than the lesson study model. The work of [52] demonstrates how tax learning is necessary even in non-accounting programs, relating the interest of professionals in learning tax matters and the needs of companies.

As mentioned in [53], chatbots as learning tools are still in their infancy. The study carried out in [54] suggests that although initially arousing considerable interest in behavior, some tasks may not be interesting enough to impact subsequent interest in the broader domain of the study; for this purpose, in [55] a four-phase model of lasting interest development. In [56], it is stated that digital tools can provoke feelings of isolation and detachment caused by the lack of interaction with teachers, which is approached through the design of a hybrid-chatbot. Finally, it must be guaranteed that a chatbot's novelty does not define the end-user experience, although it is undoubtedly part of the initial interactions with these types of tools [54].

3. Solution Design

The solution was thought of in accounting students from a Chilean university, where the learning approach is related to tax regulations. In this context, it is essential to point out the Chilean tax system's difficulty due to its complexity, variability, and a very extensive law [57].

The proposed solution's objective was the implementation of two chatbots with pedagogical resources, which, using an automated conversational system and decision trees, could be applied to the context of tax regulations at the university level.

To respond to the proposed objective, two chatbots were designed, one through an automated conversational system and the other based on a decision tree, both focused on knowledge of tax regulations under the Chilean legal context. Also, the different tests and activities were designed to demonstrate the interaction and verify the effectiveness of the tools in the learning process.

The chatbot architecture is fundamental to explain its operation, and, as is well known, there are several architectures [37], [38]. In general, there is a standard operation that can be divided into three stages, as seen in Figure 1, where the user interacts with the voice manager through a device, where the analysis and generation of the query are carried out under the work context, considering that the bot must generate a response without considering the sense of intention. Finally, the generation of messages includes the planning of words, sentences, and coherence to deliver. With these aspects defined, we can establish that the user writes a question in his application, which must be answered by the chatbot and with greater complexity and volume of knowledge, then a more significant number of interactions between a bot and user [40], [41].

Decision trees are highly complex data structures sharing binary trees [42], [43]. This type of tree is characterized by storing its different nodes in a hierarchical and orderly manner under a defined criterion based on the problem addressed. Compared to linked lists, queues, or stacks, in this case, the data is not stored linearly [44].

In this case, the constraints to be applied to the decision tree have characteristics of an N-ary tree with single essential control used in binary search trees. The walkthroughs and handling of the decision tree structure are handed over to the natural language processing engine. The algorithm used by the classification and interpretation engine conforms to a conditional inference tree; consider that the Recast.ai engine handles the classification of intentions automatically, so no changes were made to it in this research.

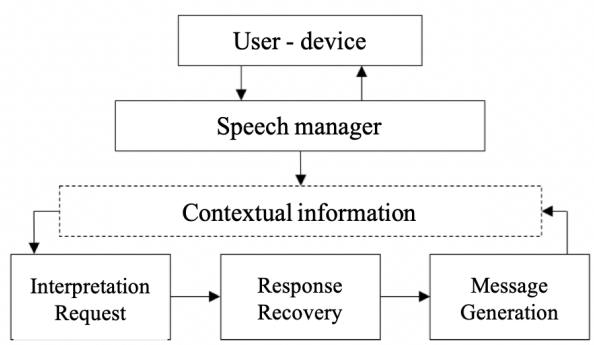


Figure 1: General chatbot architecture [39]

In this case, we start with a single node and then branch out into possible results based on the intent analysis of the natural language processing engine [45]. The decision node will represent the intention in which the user's interaction will be classified, giving rise to the existence of a terminal node that will show the result of the intention. The first node will correspond to the concept of tax and the last to specific answers.

4. Method

4.1. Pedagogical context

The learning context of this research's tax regulations focuses on the career of auditor accountant (public accountant) of a Chilean university. Thus, teaching and learning the rules that regulate tax regulations is a constant challenge, especially to achieve effective learning. The different indicators of the courses approval do not generate significant delays in the students' curricular progress. The professional profile for graduating the auditor accountant career (public accountant) considers the development of specific competencies that are directly related to the work performed:

- Generic fundamental training competencies: proactivity and responsibility; carries out its work with professional ethics.
- Specific disciplinary competencies: understand the conceptual framework of the current internal fiscal tax legislation.
- Specific professional competencies: apply the conceptual framework of the current internal fiscal tax legislation; apply the general, and special regulations of the tax legislation of the taxes levied on business income.

4.2. Learning Objectives

The Taxation courses of the auditor accountant career (public accountant) have a content structure that contributes to the learning results; in this way, the contents associated with this experiment are: (1) Understanding of the examination regulations, their practical application in the taxpayer's life cycle; (2) Understanding of the corporate tax structure.

For the technique through an automated conversational system, the learning outcomes associated with the experiment and the course are [1]:

- OL1: Recognizes basic concepts related to corporation tax to apply them to the company.
- OL2: Identify basic concepts of the different corporate taxes to compose tax returns.
- OL3: Applies basic concepts of the tax object to calculate taxes.

For the decision tree-based technique, the learning outcomes associated with the experiment and the course are:

- OL4: Defines the types of tax examination to identify the procedure applied in each of them.
- OL5: It includes the indirect control rules for their correct application in the taxpayers' life cycle.
- OL6: It includes direct inspection rules to identify the rights that assist taxpayers.

4.3. Methodology

For methodological purposes, we worked with the Taxation 1 and Taxation 3 courses of the auditor-accountant career, where the main subject of these courses is the regulations on audit and corporate taxes, respectively. The experimentation process for both chatbots is considered the same methodological steps are shown in the BPMN diagram in Figure 2.

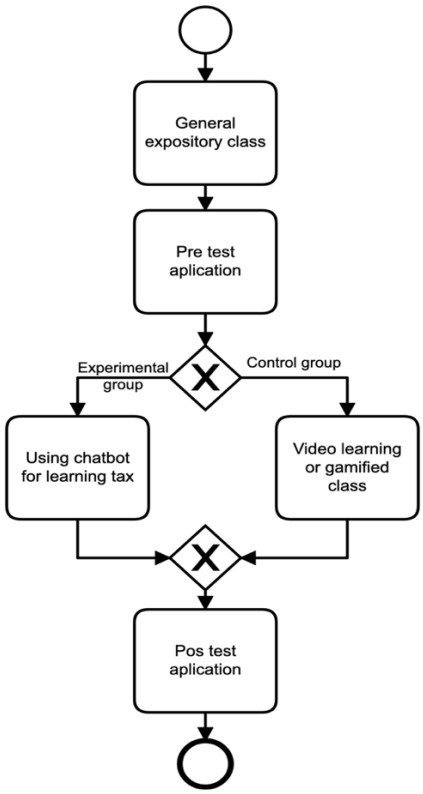


Figure 2: Stages of the methodology

First, a class of theoretical contents is held in both courses with all students without differentiating or separating individuals. This class is face-to-face and in a classroom with a projection. After that, a pre-test is applied to all students based on the contents taught in the face-to-face class of the Tax 1 and Tax 3 courses. Both courses are then divided into two groups, which we will call the experimental group and the control group. In the case of the Taxation 1 course, the use of videos through the *Moodle* platform was applied to the control group, and in the case of the experimental group, the use of the chatbot based on the decision tree technique was applied. In the case of the Taxation 3 course, the control group was given playful activities in the classroom led by a teacher, and in the case of the experimental group, the use of the chatbot with the technique was applied through an automated conversational system. A post-test is applied to both courses to measure the difference generated in the learning process. In this way, the results obtained in the tests by each of the groups of both courses will be compared, and the effectiveness of using the chatbot with the two techniques will be shown. Finally, it should be noted that, in the Taxation course, 1 of the of 16 students, eight remained in the experimental group and 8 in the control group. In the case of the Taxation course, 3 of 34 students, 18 remained in the experimental group, and 16 in the control group.

As mentioned above, two tests were applied to each course, one that was developed before dividing the group and another afterward to measure effective learning. An example of interaction in the evaluations is presented in Figure 3, using Google Forms.

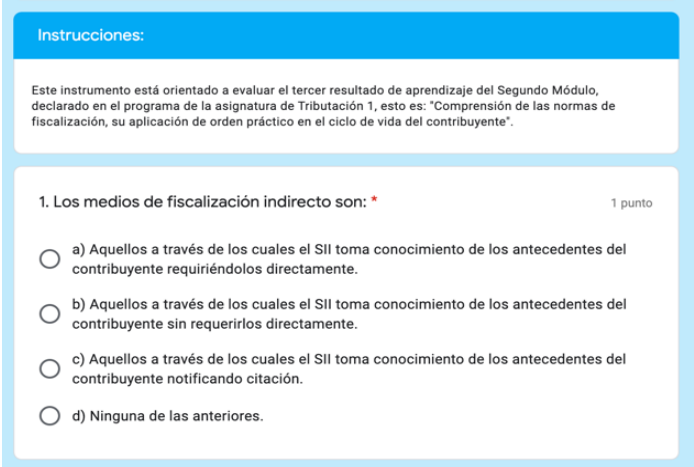


Figure 3: Example evaluation rendered by the student

4.4. Mechanics

In the Tax 1 course, the students belonging to the control group saw explanatory videos with the contents. For the experimental group, a guide to questions to be solved was delivered with the bot's support. The chatbot used was called Tribuchat, and it was linked to a Telegram chat, with which the students who made use of their mobile devices interacted under the guidance of the chatbot. Figure 4 shows an example of interaction with the bot.

In the case of the Tax 3 course, the students belonging to the control group did playful activities that included the themes addressed in the planning. The teacher-guided these activities. For the experimental group, a guide to questions to be solved was delivered with the bot's support. The chatbot was called Tribubot, and it was linked to a Facebook Messenger chat, with which the students, using their mobile devices, freely interacted with the chatbot. Figure 5 shows an example of interaction with the bot.

It is important to note that feeding both bots was carried out, in which the teachers in charge of the subjects generated the expert knowledge base on the subject and thinking about the interaction with the students, which was loaded into the tool SAP Recast.ai. Finally, interaction actions were included to measure a formative nature's learning through questions asked by the same bot to the students.

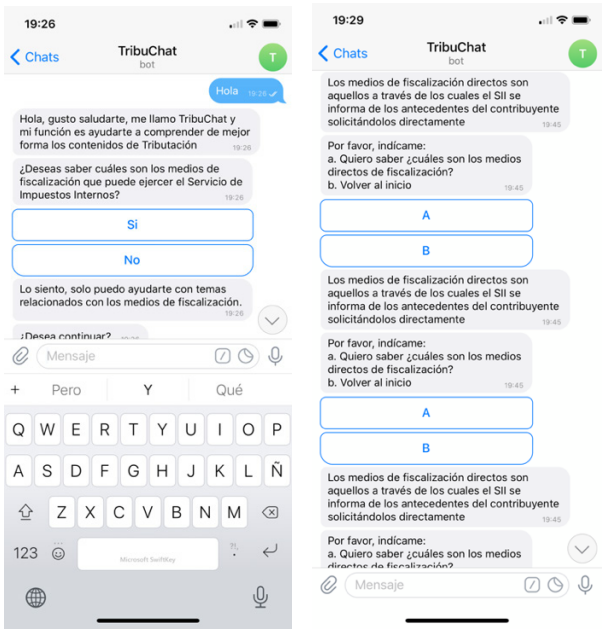


Figure 4: Chatbot group 1 interaction example

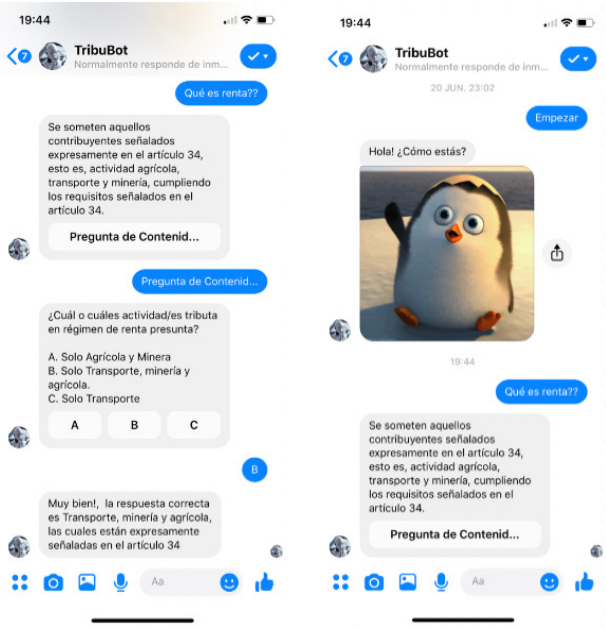


Figure 5: Chatbot group 2 interaction example

5. Results

The display of results obtained will be classified in the general results, later categorizing by gender, and finally according to performance. Table 1 shows the results obtained in the tests applied to the Taxation 1 and 3 courses, and the averages of correct, incorrect, omitted answers, and the percentage of correct answers that were.

Table 1: General Results

Course	Group	Clasification	Pretest	Postest
Tax 1 course	Control	Correct	7	7,125
		Wrong	5	4,875
		% correct	58,33%	59,38%
	Experimental	Correct	6,125	8
		Wrong	5,875	4
		% correct	51,0%	66,7%
Tax 3 course	Control	Correct	6,88	7,5
		Wrong	4,01	4
		% correct	57,3%	62,5%
	Experimental	Correct	6,56	10,39
		Wrong	4	1,87
		% correct	54,6%	86,6%

In Table 2, we can see the characterization and results in the test of the groups of Taxation 1 and 3 courses based on composition according to gender.

Table 2: Characterization by gender

Course	Tax 1 course		Tax 1 course	
Group	Control	Exp.	Control	Exp.
N	8	8	16	18
Male	25%	37,5%	43,75%	50%
% Improvement	0%	27,8%	5,7%	60,32%
Female	75%	62,5%	56,25%	50%
% Improvement	2,3%	32,3%	13,4%	58,21%

Table 3 shows groups' behavior according to the average of the Tax 1 and 3 course; for this, those below the average, equal and above, are grouped in their categories of the experimental and control groups.

Table 3: Behavior of groups according to the average

	Test	Group	Ref. $\bar{x}$	%<$\bar{x}$	%=$\bar{x}$	%>$\bar{x}$
Tax 1	Pretest	Control	7	12,5%	75%	12,5%
		Exp.	6,125	25%	62,5%	12,5%
	Postest	Control	7	25%	37,5%	37,5%
		Exp.	6,125	0%	25%	75%
Tax 3	Pretest	Control	6,88	6,25%	68,75%	125%
		Exp.	6,56	5,56%	77,78%	16,66%
	Postest	Control	6,88	0%	81,25%	18,75%
		Exp.	6,56	0%	0%	100%

In addition, to statistically validate the results, some tests were applied. Because the new experiment points directly to tax course 1, then as the results were analyzed, we begin with Table 4, where the samples' normality is appreciated through a Shapiro-Wilks Test. As can be seen, for the sample of the experimental group in the post-test, normality was rejected with a high level of significance, so that we will use non-parametric tests for the comparison of means.

Table 4: Results of Shapiro-Wilks test.

Group	Test	P-value
Control	Pretest	0,6877
	Postest	0,2993
Experimental	Pretest	0,1842
	Postest	**0,02066**

Table 5 shows the result of applying the Wilcoxon test for paired data. In this case, the difference in means for the control group between each test is not statistically significant; on the contrary, in the experimental group, it is observed that the mean of the post-test is higher with a high level of statistical significance (0.0001526 < <0.05).

Table 5: Wilcoxon test for paired data

Samples/Group		P-value
Control Pretest	Control Postest	0,08052
Experimental Pretest	Experimental Postest	**0,0001526**

In this first investigation, the data concludes that the research hypothesis is accepted, giving a difference between groups of p=9,722e-06. Specifically, the experimental group results were statistically much higher than those shown by the control group; therefore, the chatbot was more effective than the other method used.

Regarding the comparison of the two tools, only the sample data of each experimental group were taken. The research hypothesis is addressed: "The students' performance is different given the applied chatbot". To do this, first, a difference in proportion test was applied, where this test analyzes whether the percentage or proportion of a sample is statistically significance concerning another. Table 6 represents the data of the experimental group students who obtained a performance higher than the global average (control + experimental) in each post-test and the total sample of students of the experimental group.

Table 6: Proportion difference test by course

Students	Tax 1 course	Tax 3 course
Above average	7	15
Totals	8	18

Furthermore, with a p-value = 1, it can be concluded that this proportion is the same in both groups, and we cannot find a significance difference.

Finally, the result changes since the hypothesis focus only on a one-sided inequality and the test; therefore, a significance mean difference can be accepted and affirmed that the first chatbot students had a better performance (p = 0.04418). Also, it is noticeable that course 1 was much higher than those of course 2, but there was also a larger sample in that case (a little more than twice as many students), so that although the inequality between chatbots could be maintained. With the above, we have ruled out the null hypothesis.

6. Discussions

Due to what is stated in Table 1, it is evident that the control group of the Tax 1 course presents an advantage of 12.56% over the experimental group in the pretest, which was reversed in the posttest, leaving a 10.97% in favor of the experimental group; which leads us to deduce in the first instance that if the performance of the students in the experimental group is compared with the control group, the chatbot tool produced an evident improvement in learning to the detriment of the control group that took explanatory videos. This justifies the first research question about contributing positively to learning the contents related to the Chilean tax system, which is characterized mainly by the complexity of its normative structure.

The above also allows supporting the second research question on the effectiveness in the use of the chatbot. Indeed, the differences between incorrect answers and errors made are evident, where the control group presents discrete performance compared to the experimental group.

In the case of Table 1 related to the Tax 3 course, it is evidenced that the control group also presented an advantage of 4.94% over the experimental group in the pretest, reversing the situation in the posttest by 27, 82% in favor of the experimental group. The evidence is also significance when comparing the wrong questions that did not decrease.

If the behavior by gender in Table 2 is studied, it is essential to note that in the Tax 1 course, the men belonging to the control group did not show improvement (0%) versus the men in the experimental group (27.8%). In women's case, the improvement in those belonging to the control group is only 2.3% versus 32.3% in the experimental group. In the case of the Tax 3 course, the result in Table 2 is not so different compared to the Tax 1 course. Indeed, it can be found that men who make use of the chatbot show more than a 60% improvement in its results; on the other hand, those who do not use the chatbot improve by 5.7%; in the case of women, the difference is also greater, because those women who use bot the percentage of improvement is greater than 58% but those who do not use it improve 13.4%.

If the behavior is analyzed according to the means in the Tax 1 course, as shown in Table 3, the control group goes from having 75% in the average to 37.5%, leaving 37.5% above the average versus the initial 12.5%. In the case of the experimental group of 62.5% who were in the mean, only 25% remained, and those who were above the mean grew from 12.5% to 75% of the group, which together with finally remaining 0 % below the mean, then the effect of the tool applied in the experimental group is greater than that of the control group since many students being below the mean or within it rose to be at or above it. The above is a critical case to highlight and take into account for later statistical analysis. Finally, as shown in Table 3 about the Taxation 3 course, initially, 5.56% of the students who use chatbot were below the average; after using the chatbot, that number decreases to 0%, which shows a positive impact of the tool. It is not minor to identify that the control group also improves the results after applying the gamified activities.

7. Conclusions

It is already known that the teaching of tax matters is a constant challenge that expert teachers in the matter must face daily, especially considering the complexity of the contents and the repeated modifications of the legislation in the last ten years.

Throughout this paper, we have been able to see different experiences that try to improve effective learning, comparing two chatbot techniques, that is, through an automated conversational

system and another based on a decision tree, showing that independent of the technique of the chatbot used has better results than guided learning through videos or gamification techniques. The results showed that, in all cases, students who use a chatbot, regardless of the technique used, perform better than other tools, justifying their effectiveness according to what was raised in the research questions. The results show that the performance of the students of the experimental group produced an evident improvement in the learning of the contents related to the Chilean tax system, which is characterized by its complexity, derived mainly from the withdrawn legislative reforms in the last ten years.

As future work, the perspective should be taken to analyze and quantify the impact of the use of this type of tool and the cognitive load that it generates in students. Besides, it would be essential to study the change suffered by the levels of frustration of students compared to the different techniques that exist in novel and impact studies, directly related to emotionality and the application of algorithms. In this sense, an artificial emotion can be implemented as a synthetic abstraction derived from the observation of human emotions, which allows improving the effectiveness of the decision made [58]. It is also relevant to test other techniques and tools such as ROL-type games [59], gamification techniques [60], or the application of multilingual verbal communication with a focus on human-robot interaction.

Finally, it should be noted that through this research, it is shown that a chatbot improves the results in learning subjects related to tax regulations and, therefore, it is necessary to continue with studies that cover other aspects of effective teaching and learning.

Acknowledgment

Tomás Vidal Maulén, Xavier Diaz and Roberto Rojas: PUCV Research Assistants. Rafael Mellado-Silva is a beneficiary of the INF-PUCV doctoral scholarship.

References

[1] R. Mellado-Silva, M. Blanco, A. Faundez, H. De La Fuente, "Support to the learning of the Chilean tax system using artificial intelligence through a chatbot," in 2019 38th International Conference of the Chilean Computer Science Society (SCCC), 1–8, 2019, doi:10.1109/SCCC49216.2019.8966410.

[2] R. Goodman, "Problem-based learning: merging of economics and mathematics", Journal of Economics and Finance, **34**(4), 477-483, 2010, https://doi.org/10.1007/s12197-010-9154-7

[3] M. Prince, R. Felder, "Inductive teaching and learning methods: Definitions, comparisons, and research bases", Journal of Engineering Educaction, **95**(2), 123-138, 2006, https://doi.org/10.1002/j.2168-9830.2006.tb00884.x

[4] D. Schunk, B Zimmerman, "Motivation and Self-Regulated Learning: Theory, Research and Applications", Routledge, **1**, 38-56, 2012, https://books.google.cl/books?id=8VCLdg2DmoQC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

[5] C. Cofer, M. Appley, "Motivation: Theory and research", John Wiley, 1, 958, 1964, https://books.google.cl/books/about/Motivation.html?id=O7x_KgAACAAJ&redir_esc=y

[6] Guerrero, E. Hishinuma, N. Andrade, S. Nishimura, V. Cunanan, "Correlations among Socioeconomics and Family Factors and Academic, Behavioral, and Emotional Difficulties inn Filipino Adolescents in Hawai", International Journal of Social Psychiatry, **52**(4), 343-359,2006, doi:10.1177/0020764006065146

[7] P. Lantz, J. House, J. Lepkowski, D. Williams, R. Mero, J. Chen, "Socioeconomics Factors Health Behaviors, and Mortality: Results From a Nationally Representative Prospective Study of US Adults", JAMA, **279**(21), 1703-1708, 1998, doi:10.1001/jama.279.21.1703

[8] V. Aleyen, E. Stahl, S. Schworm, F. Fischer, R. Wallace, "Help Seeking and Help Design in Interactive Learning Environments", Review of Educational Research, **73**(3), 277-320, 2003, https://doi.org/10.3102/00346543073003277

[9] D. Gjibels, F. Dochy, P. Van den Bossche, M. Segers, "Effects of Problem-Based Learning: A Meta-Analysis From the Angle of Assessment", Review of Educational Research, **75**(1), 27-61, 2005, https://doi.org/10.3102/00346543075001027

[10] M. Chi, M. Bassok, M. Lewis, P. Reimann, R. Glaser, "Self-Explanations – How Students Study and Use Examples in Learning to Solve Problems", Cognitive Science, **13**(2), 145-182, 1989, https://doi.org/10.1016/0364-0213(89)90002-5

[11] C. Hmelo-Silver, "Problem-Based Learning: What and How Do Students Learn?", Educational Psychology Review, **16**(3), 235-266, 2004, https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

[12] Y. Ye, K. Kishida, "Toward an understanding of the motivation of open source software developers", in 25th International Conference on Software Engineering, 419-429, 2003, doi:10.1109/ICSE.2003.1201220.

[13] W. Nuankaew, K. Phanniphong, S. Bussaman, D. Teeraputon, P. Nuankaew, "Mentoring Model in an Active Learning Culture for Undergraduate Projects, Astes Journal, **5**(4), 805-815, 2020, http://www.astesj.com/publications/ASTESJ_050495.pdf

[14] W. Lambrechts, I. Mulà, K. Ceulemans, I. Molderez, V. Gaeremynck, "The integration of competences for sustainable development in higher education: an analysis of bachelor programs in management", Journal of Cleaner Production, **48**, 65-73, 2013, doi:10.1016/j.jclepro.2011.12.034.

[15] R. Sharif, "The relations between acculturation and creativity and innovation in higher education: A systematic literature review", Educational Research Review, 28, 1-26, 2019, doi:10.1016/j.edurev.2019.100287.

[16] L. Carvalho, P. Goodyear, "Design, learning networks and service innovation", Design Studies, **55**, 27-53, 2018, doi:10.1016/j.destud.2017.09.003.

[17] E. Engelbrecht, "Adapting to changing expectations: Post-graduate students experience of an e-learning tax program", Computers & Education, **45**(2), 217.229. 2005, doi:10.1016/j.compedu.2004.08.001.

[18] Faúndez-Ugalde, R. Mellado-Silva, E. Aldunate-Lizana, "Use of artificial intelligence by tax administrations: An analysis regarding taxpayers rights in Latin American countries", Computer Law & Security Review, 38, 1-13, 2020. 2020, doi:10.1016/j.clsr.2020.105441.

[19] T. Fairfield, "The political economy of progressive tax reform in Chile", Revista de Economia Institucional, **17**(32), 129-156, 2015, https://ssrn.com/abstract=2625309

[20] T. Derya-Baskan, E. Balikcioglu, "The Effects of Firm Components of Open Joint Manufacturing Companies on Corporate Taxes", Sosyoekonomi, **26**(37), 219-226, 2018, doi: 10.17233/sosyoekonomi.2018.03.12

[21] M. Crandall-Hollick, "Higher Education Tax Benefits: Brief Overview and Budgetary Effects, Congressional Research Service **1**, 1-17, 2018, https://files.eric.ed.gov/fulltext/ED593609.pdf

[22] L. Lee, J. Hung, "Effects of blended e-Learning: a case study in higher education tax learning setting", Human-Centric Computing and Information Sciences, **5**(1), 13, 2015, doi:10.1186/s13673-015-0024-3.

[23] J. Walmsley, "Talking to Top people: Some Issues Relating to the Citizenship of People with Learning Difficulties", Disability, Handicap & Society, **6**(3), 219-231, 1991, doi:10.1080/02674649166780261.

[24] Sheehan, H. Jin, U. Gottlieb, "Customer service chatbots: Anthropomorphism and adoption", Journal of Business Research, **115**, 14-24, 2020, doi:10.1016/j.jbusres.2020.04.030.

[25] J. Weizenbraum, "ELIZA—a computer program for the study of natural language communication between man and machine", Communications of the ACM, 9(1), 36-45, 1966, https://dl.acm.org/doi/10.1145/365153.365168

[26] N. Block, "Psychologism and Beehaviorism", The Philosophical Review, **90**(1), 5-43, 1981 https://www.researchgate.net/publication/265578700_Psychologism_and_Behaviorism

[27] L. Bradesko, D. Mladenic, "A Survey of Chabot Systems through a Loebner Prize Competition", in Slovenian Language Technologies Society 8th Conference of Language Technologies, 2012, https://www.researchgate.net/publication/235664166_A_Survey_of_Chatbot_Systems_through_a_Loebner_Prize_Competition

[28] J. Jia, "CSIEC: A computer assisted English Learning chatbot based on textual knowledge and reasoning", Knowledge-Based Systems, **22**(4), 249-255, 2009, https://doi.org/10.1016/j.knosys.2008.09.001

[29] Lee, S. Jung, S. Kim, G. Lee, "Example-based dialog modeling for practical

multi-domain dialog system", Speech Communication, **51**(5), 466-484, 2009, https://doi.org/10.1016/j.specom.2009.01.008
[30] R. Crutzen, G. Peters, S. Portugal, E. Fisser, J. Grolleman, "An Artificially Intelligent Chat Agent That Answer Adolescents Questions Related to Sex, Drugs and Alcohol: An Exploratory Study", Journal of Adolescent Health, **48**(5), 514-519, 2011, https://doi.org/10.1016/j.jadohealth.2010.09.002
[31] P. Brandttzaeg, A. Følstad, "Why People Use Chatbots", Internationnal Conference on Internet Science, 377-392, 2017, doi:10.1007/978-3-319-70284-1_30.
[32] H. Shah, K. Warwick, J. Vallverdú, D. Wu, "Can machines talk? Comparison of Eliza with modern dialogue systems", Computers in Human Behaviors, **58**, 278-295, 2016, doi:10.1016/j.chb.2016.01.004.
[33] O. Simpson, "Supporting Students in Online, Open and Distance Learning", 1, 348-362, Routledge, 2018, https://doi.org/10.4324/9780203417003
[34] J. Lee, "Online support service quality, online learning acceptance, and student satisfaction", The Internet and Higher Education, **13**(4), 277-283, 2010, https://doi.org/10.1016/j.iheduc.2010.08.002
[35] L. Fryer, K. Nakao, A. Thompson, "Chatbot learning partners: Connecting learning experiences, interest and competence", Computers in Human Behavior, 93, 279-289, 2019, https://doi.org/10.1016/j.chb.2018.12.023
[36] F. Almabri, D. Bell, M. Merhi, "Understanding Student Acceptance and Use of Chatbots in the United Kingdom Universities: A Structural Equiation Modelling Approach", in 2020 6th International Conference on Information Management (ICIM), IEEE, London, United Kingdom, 284-288, 2020, doi:10.1109/ICIM49319.2020.244712.
[37] S. Ghose, J. Barua, "Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor", in 2013 International Conference eon Informatics, Electronics and Vision (ICIEV), 1-5, 2013, doi: 10.1109/ICIEV.2013.6572650
[38] M. Yan, P. Castro, P. Cheng, V. Ishakian, "Building a Chatbot with Serverless Computinng", in Proceedings of the 1st International Workshop on Mashups of Things and APIs, **5**, 1-5, 2016, doi: 10.1145/3007203.3007217
[39] Braun, A. Hernandez, F. Matthes, M. Langen, "Evaluating Natural Language Understanding Services for Conversational Question Answering Systems", in Proceedings of the 18th Annual SIGdial Meeting of Discourse and Dialogue, 174-185, 2017, doi: 10.18653/v1/W17-5522
[40] M. Dahiya, "A Tool of Conversation: Chatbot", International Journal of Computer Sciences and Engineering, **5**, 158-161, 2017, https://www.researchgate.net/publication/321864990_A_Tool_of_Conversation_Chatbot
[41] S. Abdul-Kader, J. Woods, "Survey on Chatbot Design Techniques in Speech Conversation Systems", International Journal of Advanced Computer Science and Applications, **6**(7), 2015, http://dx.doi.org/10.14569/IJACSA.2015.060712
[42] C. Wang, A. Wang, J. Xu, Q. Wang, F. Zhou, "Outsourced privacy-preserving decision tree classification service over encrypted data", Journal of Information Security and Applications, **53**, 788-813, https://doi.org/10.1016/j.jisa.2020.102517
[43] Wang, Q. Wang, F. Nie, Z. Li, W. Yu, F. Ren, "A linear multivariate binary decision tree classifier based on K-means splitting", Pattern Recognition, 107, 423-437, 2020, doi:10.1016/j.patcog.2020.107521.
[44] M. Van den Boomen, M. Spaan, R. Schoenmaker, A. Wolfert, "Untangling decision tree and real options analyses: a public infrastructure case study dealing with political decisions, structural integrity and price uncertainty", Construction Management and Economics, **37**(1), 24-43, 2019, doi:10.1080/01446193.2018.1486510.
[45] S. Garg, "An evaluation of investor acceptability for physical gold using classification (Decision Tree)", Materials Today: Proceedings, FALTA VOLUMEN, FALTAN PAGINAS, 2020, doi:10.1016/j.matpr.2020.06.177.
[46] K. Watty, J. McKay, L. Ngo, "Innovators or inhibitors? Accounting faculty resistance to new educational technologies in higher education", Journal of Accounting Education, **36**, 1-15, 2016, doi:10.1016/j.jaccedu.2016.03.003.
[47] N. Ahamad, M. Lai, "Undergraduate tax education in the new millennium: Technical vs. conceptual", in 2010 International Conference on Science and Social Research, 602-606, 2010, doi:10.1109/CSSR.2010.5773852.
[48] J. Craner, A. Lymer, "Tax education in the UK: a survey of tax courses in undergraduate accounting deegrees", Accounting Education, **8**(2), 127-156, 1999, doi:10.1080/096392899330973.
[49] P. Senge, M. Dow, G. Neath, "Learning together: new partnerships for new times", Corporate Governance: The international journal of business in society, **6**(4), 420-430, 2006, doi: 10.1108/14720700610689531
[50] P. Marín, "Enseñanza y aprendizaje del derecho tributario, y ejercicio profesional, en ADE", Revista de Educación y Derecho, **12**, 129-150, 2015,
[51] M. Pang, F. Marton, "Beyond 'lesson study': Comparing two ways of facilitating the grasp of some economic concepts", Instructional Science, **31**(3), 175-194, 2003, https://doi.org/10.1023/A:1023280619632
[52] M. Lai, Y. Zalilawati, M. Amran, K. Choong, "Quest for Tax Education in Non-Accounting Curriculum: A Malaysian Study", Asian Social Science, **9**(2), 154, 2013, doi:10.5539/ass.v9n2p154.
[53] D. Gonda, J. Luo, Y. Wong, C. Lei, "Evaluation of Deeveloping Educational Chatbots Based on the Seven Principles for Good Teaching", in 2018 IEEE Internationnal Conference on Teachingn, Assessment, and Learning for Engineering (TALE), 446-453, 2018, doi:10.1109/TALE.2018.8615175.
[54] L. Fryer, M. Ainley, A. Thompson, A. Gibson, Z. Sherlock, "Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners", Computers in Human Behavarior, **75**, 461-468, 2017, https://doi.org/10.1016/j.chb.2017.05.045
[55] S. Hidi, K. Renninger, "The Four-Phase Model of Interest Development", Educational Psychologist, **41**(2), 111-127, 2006, https://doi.org/10.1207/s15326985ep4102_4
[56] Wu, C. Lin, Y. Ou, C. Liu, W. Wang, C. Chao, "Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot", IEEE Access, 8, 77788-77801, 2020, doi:10.1109/ACCESS.2020.2988252.
[57] G. Campos, "Complejidad del Sistema Tributario", PriceWaterhouseCoopers, 2017. https://www.pwc.com/cl/es/prensa/columnas-de-opinion/complejidad-del-sistema-tributario.html
[58] D. Cabrera, C. Cubillos, A. Cubillos, E. Urra, R. Mellado, "Affective Algorithm for Controlling Emotional Fluctuation of Artificial Investors inn Stock Markets", IEEE Access, **6**, 7610-7624, 2018, doi:10.1109/ACCESS.2018.2802781.
[59] R. Mellado, B. Melgarejo, C. Velasquez, C. Cubillos, S. Roncagliolo, N. Gonzalez, "ROLE video game tool for teaching myths and legends to school basic students", in 2018 37th International Conference of the Chilean Computer Science Society, 1-8, 2018, doi:10.1109/SCCC.2018.8705258.
[60] R. Mellado, M. Blanco, C. Cubillos, H. Fuente, A. Faúndez, "Learning of Information Technologies in Administration Science through gamification techniques" in 2019 IEEE Chilean Conference on Electrical, Electronics, Engineering, Information and Communication Technologies, 1-7, 2019, doi:10.1109/CHILECON47746.2019.8988017.

ASTES

Advances in Science, Technology and Engineering Systems Journal, Vol. 5, No. 5, 444-451 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Newton-Raphson Algorithm as a Power Utility Tool for Network Stability

Lambe Mutalub Adesina[1,*], Ademola Abdulkareem[2], James Katende[3], Olaosebikan Fakolujo[4]

[1]*Kwara State University Malete, Dept. of Electrical and Computer Engineering, Faculty of Eng'g and Technology, 241104, Nigeria*

[2]*Covenant University Sango Ota, Dept. of Electrical and Information Engineering, College of Engineering, 112233 Nigeria*

[3]*University of Namibia, Department of Electrical and Computer Engineering, Faculty of Technology, 13301 Namibia*

[4]*University of Ibadan, Dept. of Electrical and Electronics Engineering, Faculty of Technology, 200284 Nigeria*

ARTICLE INFO

Article history:
Received: 28 July, 2020
Accepted: 28 August, 2020
Online: 21 September, 2020

Keywords:
National power reform
Network reliability
Newton – Raphson Software
Operational parameters
Revenue generation
System collapse
Utility feeder network

ABSTRACT

Nigerian power utility companies particularly the distribution and generation aspects were recently in the process of national power reform converted from public to private service by privatization. Prior to these development, power utility companies' performance is low due to poor operational style that leads to inadequate revenue generation. Thus, the task before the privatized companies includes autonomy, high reliability operation and brake-even management. To achieve these goal, frequent outages and system collapses must be minimized. One of the methods of achieving this is using power flow to improve the reliability of power system which will subsequently improve other lacking factors. A developed software for Newton-Raphson power flow was tested with a known solution network and the results obtained are accurate and reliable. Therefore, this paper presents an application of this software on real-time transmission network. Nigerian 330kV transmission grid is considered as case study. The power flow analysis of this grid was carried out and the network operational parameters were obtained. These results are stated and carefully analyzed. In practice, power utility distribution network of medium voltage of 11kV feeder was also tested with this NR-Software in ascertaining network reliability and in the course of adding public transformers to utility feeder network.

1. Introduction

This paper is an extension of work originally presented in IEEE Nigeria Computer Conference, where a computer software package was developed using Newton-Raphson power flow algorithm [3]. The package was tested with a known solution network and the obtained results are accurate and reliable[3, 8]. Nigerian population is exponentially increasing with the attendant increase in demand for electric power supply proportionally. This becomes a serious task for the power utility companies in the country to meet the huge energy or power demand of the consumers continuously and safely. Also, power utility companies need to minimize the cost of running the network whether distribution or transmission for smooth, economical and profitable operation of the system. Adding equipment such as generators, lines, buses etc makes the network more prone to faults [3, 4, 8]. Thus, power system network could be interpreted as an interconnection of generators, buses, transformers and lines to give electricity supply to load at several points on the system[3, 4, 8, 11]. However, regardless of the number of load points on the system, the faults should be minimized and the load be made to operate at maximum efficiency with safety so as minimize losses. Therefore, power flow study is employed to study the power system network, and its behavior when it is subjected to any expansion or upgraded for any system.

Power flow studies are tools that are significantly used in system planning and design of power system. It is also used in future expansion and optimization of the existing systems in enhancing reliable performance. The main parameters obtained from typical power flow study include the magnitude and phase angle of the bus bars' voltages as well as the active and reactive power flowing between the buses [5, 9, 10]. Power flow equations when solved gives the steady state condition of every bus bar

*Corresponding Author: Lambe Mutalub Adesina, Nigeria, +2348033207586 lambe.adesina@kwasu.edu.ng

voltage of the network. The problem associated with these power flow equations is their non-linearity and difficulty in obtaining solution by mathematical calculation. These non-linear equations are normally solved by iteration techniques [5, 11, 12]. Therefore, power flow solution problems are an iterative process which involves assigning assumed values to the unknown bus bar voltages and subsequent calculation of new voltage values for the bus bars from assumed values for the remaining bus bars with the specified active power and reactive power or voltage magnitude. New set of voltage values for the bus bars are obtained which are eventually used to determine new set of bus bar voltages by a pre-defined algorithm. This iteration process is updated until the voltage changes between the last and just preceding iteration is less than a specified tolerance value for all the bus bars [3, 4, 8, 11]. From literature, most algorithms often result to high number of iterations and hence take long time to complete the required iterations. For a large network, high number of iterations is expected depending on the type of power flow technique and algorithm used. Several iterative techniques are used in power flow studies which include Gauss, Gauss Seidel, Newton Raphson (NR), Fast Decoupled, etc [3, 8]. Many research papers on related topic had been written and published using the above mentioned techniques but their iteration level is high resulting to long period of convergence [3, 8, 9, 12]. Thus, a software package in this regards is predicted to improve the convergence period.

2. Newton – Raphson Algorithm

The technique illustrates a vector $x \in IR^N$ such that,

$$F(x) = 0 \tag{1}$$

where,

F = Set of 'N' Nonlinear equations.

In Newton-Raphson (NR) Techniques, the vector of state variable is determined by performing an expansion of equation (1) by Taylor's series, assuming an initial estimate X (0). Neglecting error in higher order terms of Taylor series; solution for x is obtained using equation (2),

$$x^{(k)} = x^{(k-1)} - J^{-1} F(x^{(k-1)}) \tag{2}$$

where,

k = iteration,

J = A square matrix of same dimension as x and F, and its entries are partial derivatives defined as[3,4,6],

$$J_{pq} = \partial F_p / \partial x_q \,|_{x^{(k-1)}} \tag{3}$$

where,

p, q = Buses.

J_{pq} = Jacobian matrix of equation (1).

When this mathematical expression is related to power flow solution, then power mismatch is forced to zero via the derivatives of mismatch of the power flow equations. Thus, bus p equation is defined as [1,3],

$$P_p + jQ_p = E_p I_p^* \tag{4}$$

where

$$I_{P=} \sum_{K=1}^{NB} Y_{pk} \; E_k \tag{5}$$

Then,

$$P_p + jQ_p = E_p \sum_{K=1}^{NB} Y_{pk} E_k^* = |E_p|^2 Y_{pp}^* + \sum_{\substack{K=1 \\ K \neq p}}^{NB} Y_{pk}^* E_p E_k^* \tag{6}$$

For p= 1, 2, 3 …… NB, excluding the slack bus.

In starting iteration, initial voltages often assumed and used. The active power P and reactive power Q evaluated from equation (6) is deducted from the scheduled active power P_2 and reactive power Q_s at the bus. The difference is considered as errors stored. Polar co-ordinates are often used for voltage evaluations. In this approach, the voltage magnitudes and phase angles are treated separately as different variables. Consequently, bus injection equations are differentiated with respect to all variables at each bus. Therefore, the power mismatch for each of the buses are evaluated using equations (7) and (8) [2,7].

$$\Delta P_p = \sum_{k=1}^{NB} \frac{\partial P_p}{\partial \theta_k} \Delta\theta_k + \sum_{K=1}^{NB} \frac{\partial P_p}{\partial |E_{k|}} \Delta |E_k| \tag{7}$$

$$\Delta Q_p = \sum_{K=1}^{NB} \frac{\partial Q_p}{\partial \theta_k} \Delta\theta_k + \sum_{K=1}^{NB} \frac{\partial Q_p}{\partial |E_k|} \Delta |E_k| \tag{8}$$

The resulting partial differentials from equations (7) and (8) are put in a Jacobian matrix format so that these equations (7) and (8) are representable in matrix vector form as written in equation (9),

$$\begin{bmatrix} \Delta P_1 \\ \Delta P_{NB-1} \\ - - \\ \Delta Q_1 \\ \Delta Q_{NB-1} \end{bmatrix} = \left[\begin{array}{ccc|ccc} \frac{\partial P_1}{\partial \theta_1} & \cdots & \frac{\partial P_1}{\partial \theta_{NB-1}} & \frac{\partial P_1}{\partial |E_1|} & \cdots & \frac{\Delta P_1}{\Delta |E_{NB-1}|} \\ \frac{\partial P_{NB-1}}{\partial \theta_1} & \cdots & \frac{\partial P_{NB-1}}{\partial \theta_{NB-1}} & \frac{\partial P_{NB-1}}{\partial |E_1|} & \cdots & \frac{\Delta P_{NB-1}}{\partial |E_{NB-1}|} \\ - & - & - & - & - & - \\ \frac{\partial Q_1}{\partial \theta_1} & \cdots & \frac{\partial Q_1}{\partial \theta_{NB-1}} & \frac{\partial Q_1}{\partial |E_1|} & \cdots & \frac{\Delta Q_1}{\partial |E_{NB-1}|} \\ \frac{\partial Q_{NB-1}}{\partial \theta_1} & \cdots & \frac{\partial Q_{NB-1}}{\partial \theta_{NB-1}} & \frac{\partial Q_{NB-1}}{\partial |E_1|} & \cdots & \frac{\Delta Q_{NB-1}}{\partial |E_{NB-1}|} \end{array}\right] \begin{bmatrix} \Delta\theta_1 \\ \Delta\theta_{NB-1} \\ - - \\ \Delta |E_1| \\ \Delta |E_{NB-1}| \end{bmatrix} \tag{9}$$

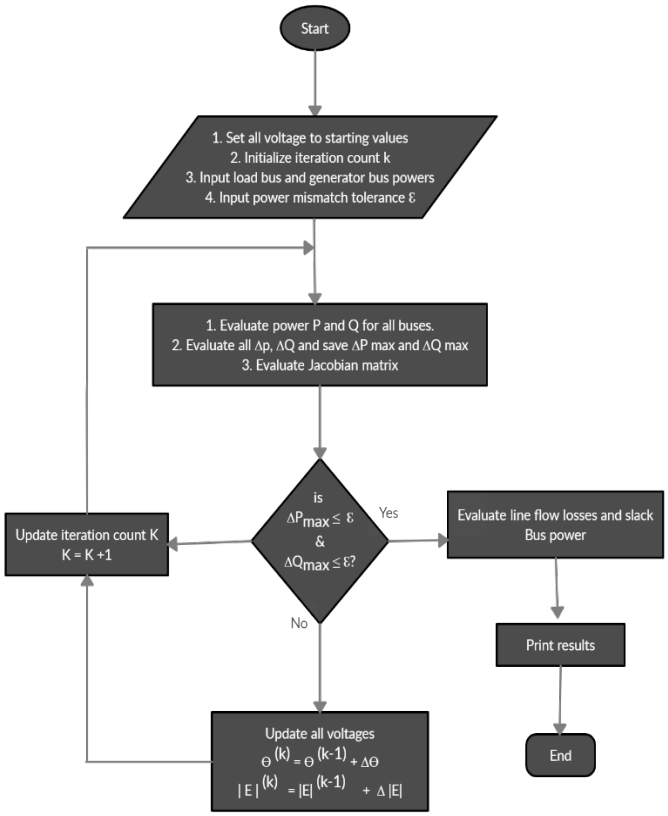
To fasting convergence, multiply voltage corrections by a constant called acceleration factor at each iteration end [3,8,12] as;

$$Q_p^{(k)}(\text{Accel}) = Q_p^{(k-1)} + \propto \Delta Q_p^{(k-1)} \tag{10}$$

$$|E_p|^{(k)}(\text{Accel}) = |E_p|^{(k-1)} + \propto \Delta |E_p|^{(k-1)} \tag{11}$$

where, $\propto$ = acceleration Factor, $0.7 < \propto > 1.4$

The common procedures of Newton-Raphson power flow solution are displayed in flowchart shown in Figure 1.



2.1 Developed NRPF Software Package

In solving power flow solution problem using the Newton-Raphson Technique, some computer programs were written to create the software package. The developed NRPF package was implemented for a small power system network using a personal computer (PC) in the paper presented at IEEE Nigeria Computer Conference [3]. The interface of package and it users are interactively designed. The arrangement of the package was done to limit the PC core memory and made possible by overlaying technique.

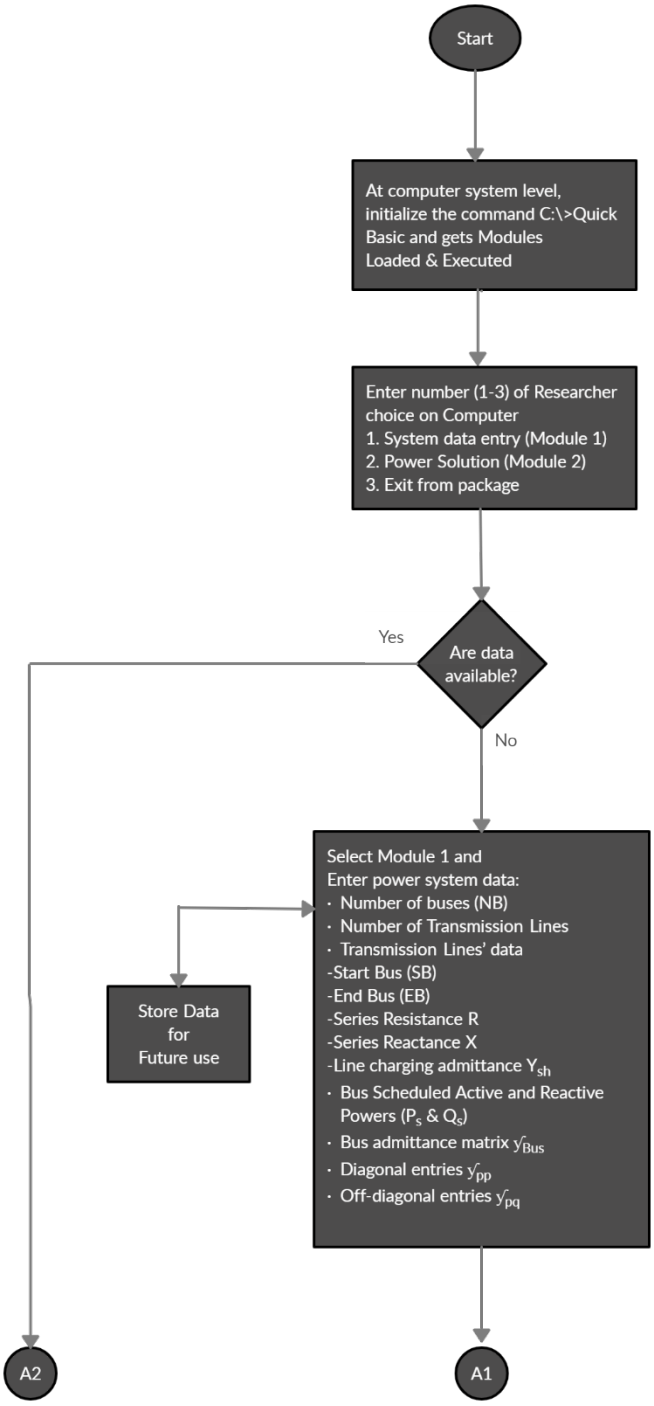
Overlaying reduced the working memory of the computer and allows the package to be put into modules. With this overlaying, the package operation is in 3 steps and named as;

1. NRPF Supervising module
2. ADESINA 1 Power system data entry and Management
3. ADESINA 2 NR power flow solution.

Only one out of the three modules are allowed to resident in core memory of computer at a time. The Quick Basic language command is used to interface the supervising module (NRPF) to each of the other two modules. Computer programs are coded in Microsoft's Quick Basic. These two modules are with their respective complete subroutine procedures.

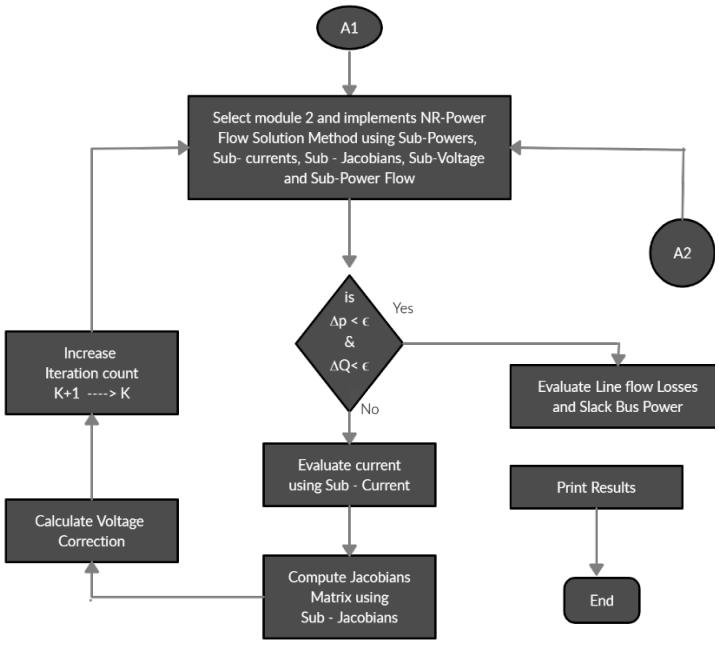
A. Supervising Module (NRPF)

NRPF is more or less act as Controller of the package structure because it allows the user to select which of the two modules is to be executed. At the starting point on computer, the command is C: \>Quick Basic NRPF. This implies loading the module into memory of computer and executes the process. Having activated the module, the computer displays the contents in block 3 of the flow chart in Figure 2



B. System Data Entry Module (ADESINA 1)

This module permits operational or network data input into computer by the user. It loads and executes the selected option on NRPF's menu. The Personal computer (PC) interactively ask for the under listed items.



1. Number of buses (NB) in the network
2. Number of transmission lines (NL)
3. Transmission line data
   - Start bus, (SB)
   - End Bus, (EB)
   - Series resistance, R
   - Series reactance, X
   - Line charging admittance, Y_{sht}
4. The bus scheduled active power P_S and reactive powers Q_S. Bus admittance matrix, Y_{bus}, are formulated. The non – availability of in – built complex arithmetic functions in Quick Basic, the real and imaginary parts of Y_{bus} are calculated separately. The admittance matrix, Y_{bus}, building algorithms are presented as follows;

The diagonal elements (Y_{PP}) of the Y_{bus} matrix is calculated by adding lines' admittances and line charging admittance.

The off-diagonal elements (Y_{pq}) are the negatives of the admittance buses p and q. Where there are no transmission lines between bus p and bus q, then the off-diagonal element is zero.

The aforementioned data above which already entered as module 1 are stored in data files for future use. This means module 1 do not needs calling since the data are already in memory and kept in data files. Exiting from module 1, the NRPF module is reloaded into memory and executed.

C. Power Flow Solution Module (ADESINA 2)

ADESINA 2 module implements the power flow solution technique. It is loaded into memory and executed provided the data files (ADESINA 1) earlier created had been inputted. When it is noticed that the data files are not in memory then computer gives error signal to show that file data is not yet in computer memory. This imply that, in running this module, the user must ascertain that system data entry module has been executed.

However, it is mentioned earlier that Quick Basic programming language have no in – built complex arithmetic library functions. Due to this lack, power flow solution would need reformulation in rectangular co-ordinates so that parameters in complex form can be effectively manipulated. The module is divided into a main programme and five under listed subroutines.

1. Sub Powers
2. Sub Currents
3. Sub Jacobian
4. Sub Voltage
5. Sub Power flows

The flow process of the program's logic is illustrated in Figure 2.

3. Description of Nigerian 330kVTransmission Grid

The Nigerian transmission grid is shown in figure 3. The grid consists of thirty-two buses including nine generator buses and thirty – four transmission lines. There are three power transformers which are usually the power regulation type. These transformers are usually the power regulation type. Each generation station, depending on its installed generation capacity has a number of generators. Some of the generators in the station are made to supply to the grid, while the remaining ones are considered as reserves. The company's installed generation capacities are shown in Table 2 [13] . Table 3 [13] shows the number of units of generators found in each station together with corresponding company's power factor specification. The generating stations have their installed generation capacity and has a number of generator units. Part of installed generators in the station give supply to the grid, while the remaining ones provides the spinning reserves. Available data shows that the total installed generation capacity is 3860MW.

Table 1: Names and Symbols of Buses

Bus No	Bus Name	Bus Symbol
1	Sapele	SPG
2	Osogbo	OS
3	Aiyede	AY
4	Ikeja West	IK
5	Akangba	AK
6	Benin	BN
7	Sapele (L)	SPL
8	Onitsha	ON
9	Enugu	EN
10	Alaoji	ALO
11	Akure	AKR
12	Benin-B	BNB
13	Delta (G)	DTG
14	Ajaokuta	AJ
15	Egbin (G)	EGBG
16	Afam (G)	AFG
17	Portharcourt	PT
18	Aladja	AL
19	Makurdi (G)	MKG
20	Ikom (G)	IKG
21	New-haven	NH

Table 2: Company's Instilled Generation Capacities [13]

Bus No.	Station Name	Installed Capacity (MW)	Type of Station
1.	Sapele	1020	Thermal
13.	Delta	820	Thermal
15	Egbin (G)	1320	Thermal
16.	Afam (G)	700	Thermal
19.	Makurdi (G)	0	Proposed
20.	Ikom	0	Proposed
Total installed Generation		3860	

Table 3: Generator Power Factor Specification for Generation Companies [13]

Substation Location	Number of Units	Power Factor (PF)
Afam I	2	0.80
Afam II	4	0.80
Afam III	4	0.80
Afam IV	6	0.80
Delta I	2	0.80
Delta II	6	0.80
Delta III	6	0.80
Sapele I	4	0.90
Sapele II	4	0.80
Egbin	6	0.80

Table 4: Bus Power and Voltage Specifications in per unit (p.u) [13].

BUS No	P_g(MW)	Q_g(MVAr)	P_{sp}(MW)	Q_{sp}(MVAr)	V(kV)
1	10.2	7.6	0	0	1.0
2	0	0	1.008	0.487	-
3	0	0	1.922	0.897	-
4	0	0	4.55	2.19	-
5	0	0	13.2	5.865	-
6	0	0	2.40	0.75	-
7	0	0	5.10	2.55	-
8	0	0	1.234	0.695	-
9	0	0	2.53	1.226	-
10	0	0	1.638	0.784	-
11	0	0	0.451	0.208	-
12	0	0	1.856	0.896	-
13	8.2	6.15	0	0	1.0
14	0	0	2.20	0.96	-
15	13.2	9.9	0	0	1.0
16	7	5.25	0	0	1
17	0	0	3.39	1.64	-
18	0	0	1.397	0.579	-
19	0	0	0	0	1
20	0	0	0	0	1
21	0	0	0.14	0.57	-

Table 5: Impedance and Shunt admittance of the Transmission Lines in per unit [13]

S/No	SB	EB	Z	YSHT
1	1	7	0 + j 0.0109	0
2	2	3	0.0041 + j 0.0349	0.437
3	2	4	0.0042 + j 0.0344	1.8
4	2	11	0.0037 + j 0.028	0.377
5	3	4	0.0049 + j 0.0416	0.521
6	4	5	0.0003 + j 0.0023	0.13
7	4	6	0.0025 + j 0.0195	4.648
8	4	15	0.0011 + j 0.0086	0.514
9	6	7	0.0005 + j 0.0035	0.832
10	6	8	0.0024 + j 0.0203	1.042
11	6	11	0.0053 + j 0.0483	0.577
12	6	12	0. + j 0.04	0
13	6	14	0.0035 + j 0.028	1.49
14	7	18	0.0011 + j 0.0095	0.478
15	8	9	0.0148 + j 0.01826	0
16	8	10	0.0023 + j 0.0199	1.048
17	8	21	0.0034 + j 0.0292	0.365
18	9	19	0.0018 + j 0.0139	0.208
19	9	20	0.0015 + j 0.0123	0.616
20	10	16	0.0005 + j 0.0035	0.208
21	12	13	0.0054 + j 0.0173	0.612
22	13	18	0.0011 + j 0.0088	0.135
23	16	17	0.0007 + j 0.0049	0.25

Table 6: Bus Power and Voltage Specifications in per unit

BUS NO	Pg (MW)	Qg (MVAr)	Ps(MW)	Qs(MVAr)	V(KV)
1	10.2	7.6	0	0	1.0
2	0	0	1.008	0.487	-
3	0	0	1.922	0.897	-
4	0	0	4.55	2.19	-
5	0	0	13.2	5.865	-
6	0	0	2.4	0.75	-
7	0	0	5.1	2.55	-
8	0	0	1.243	0.695	-
9	0	0	2.53	1.226	-
10	0	0	1.638	0.784	-
11	0	0	0.451	0.208	-
12	0	0	1.856	0.896	-
13	8.2	6.15	0	0	1.0
14	0	0	2.2	0.96	-
15	13.2	9.9	0	0	1.0
16	7.0	5.25	0	0	1.0
17	0	0	3.39	1.64	-
18	0	0	1.397	0.579	-
19	0	0	0	0	1.0
20	0	0	0	0	1.0
21	0	0	1.140	0.570	-

3.1. Software Application to Sub-Nigerian Transmission Grid.

In this section, the developed software is applied to a Sub-Nigerian Transmission Grid which covers the southern region of the country. This grid network is shown in Figure 3. Per unit impedance and Shunt admittance of the transmission lines in the network diagram of this figure 3 are shown in Table 5. Other parameters required in data input into computer includes;

$$Number\ of\ Buses\ =\ 21$$
$$Number\ of\ Lines\ =\ 23$$
$$Slack\ Bus\ Number\ =\ 1$$
$$Slack\ Bus\ Voltage\ =\ 1.0\ +\ j\ 0$$
$$Power\ mismatch\ tolerance\ \epsilon\ =\ 0.001\ per\ unit.$$

The bus scheduled powers for all non-generator buses are already shown in Table 3 above.

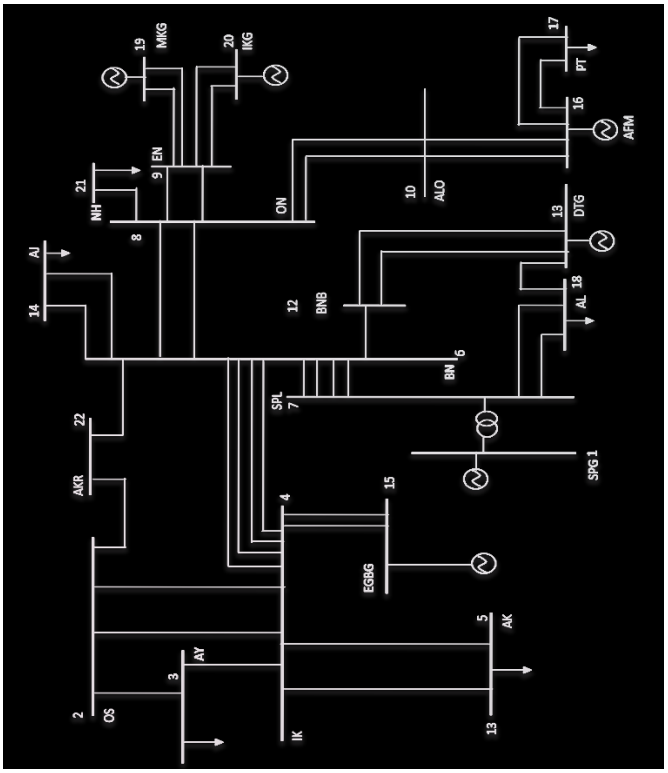


Figure 3: Nigerian 330kV Transmission[3,13]

ADESINA 1 module input data and store in computer memory. These data include; Number of bus bars, Number of transmission lines, Send-end bus bars number, Receiving-end busbar number, Line impedance, Line Susceptance, Bus scheduled active power P_{sp} and reactive power Q_{sp}. It is followed by the formulation of bus admittance matrix. ADESINA 2 module executes the bus voltage calculations which is initialized by slack bus bar assumed voltage, V = 1.00 + j 0 p.u. The system solution converged in third iteration with mismatched power tolerance, ϵ = 0.001 per unit. The power flow solution of this case study is shown in Table 7 to Table 9.

Table 7: Calculated Bus Voltages of the Network in per unit

Bus	Voltage (per unit)
1	1.0+ j 0
2	1.057 + j 0.062
3	1.051 + j .055
4	1.043 + j .045
5	1.043 + j .045
6	1.027 + j .025
7	1.021 + j .019
8	1.031 + j .029
9	1.047 + j .048
10	1.032 + j .029
11	1.046 + j .049
12	1.024 + j .022
13	1.0 + j 0
14	1.027 + j .026
15	1.0 + j 0
16	1.0 + j 0
17	1.032 + j .029
18	1.022 + j .019
19	1.047 + j .048
20	1.047 + j .048
21	1.031 + j .03

Table 8: Evaluated Bus Voltage Magnitudes, Phase Angles, Active and Reactive Powers

Voltage (per unit)			Power (per unit)	
Bus No	Magnitude	Phase Angle (degree)	Active (MW)	Reactive (MVAr)
1	1	0	0	0
2	1.059	3.357	0	2.203
3	1.052	2.996	0	0.479
4	1.044	2.471	0	3.806
5	1.044	2.471	0	0.065
6	1.027	1.395	0	4.294
7	1.021	1.066	0	0.656
8	1.031	1.611	0	1.228
9	1.048	2.625	0	0.724
10	1.032	1.61	0.001	0.628
11	1.047	2.682	0	0.477
12	1.024	1.231	0	0.306
13	1	0	0	0
14	1.027	1.45	0	0.745
15	1	0	0	0
16	1	0	0.001	0
17	1.032	1.61	0	0.125
18	1.022	1.065	0	0.306
19	1.048	2.625	0	0.416
20	1.048	2.625	0	0.308
21	1.031	1.667	0	0.182

Table 9: Calculated Power Flows between the Buses

Buses		Power Flow (per unit)	
From	To	Active	Reactive
1	7	1.746	-1.93
7	1	-1.746	2.004
2	3	-0.221	0.495
3	2	0.221	0.157
2	4	-0.545	0.721
4	2	0.544	-0.063
2	11	-0.516	0.704
11	2	0.514	-0.047
3	4	-0.264	0.744
4	3	0.263	0.37
4	5	0	0.518
5	4	0	0.518
4	6	-1.136	1.278
6	4	1.132	-0.226
4	15	-5.834	5.222
15	4	5.778	-3.786
6	7	-1.936	2.477
7	6	1.933	-0.568
6	8	0.219	0.768
8	6	-0.219	1.14

6	11	0.539	0.592
11	6	-0.541	1.364
6	12	-0.075	1.029
12	6	0.075	0.866
6	14	0.036	0.953
14	6	-0.036	0.946
7	18	0.01	0.09
18	7	-0.01	0.303
8	9	0.112	0.193
9	8	-0.112	0.374
8	10	0.004	0.226
10	8	-0.004	0.329
8	21	0.035	0.28
21	8	-0.035	0.274
9	19	0	0.071
19	9	0	0.071
9	20	0	0.071
20	9	0	0.071
10	16	-9.375	10.947
16	10	9.307	-5.622
12	13	-1.571	0.958
13	12	1.553	-0.902
13	18	2.434	-1.451
18	13	-2.445	3.069
16	17	6.714	-5.047
17	16	-6.768	6.503

4. Discussion of Results

Sapele generating station (Bus 1) of the power system network of figure 3 was chosen as the Slack bus for this analysis. This is because it is one of the sources of supply to the grid. The voltage of generator buses is constrained to 1.0 < 0 per unit.

In Table 8, the magnitude of the bus voltages varies from 1.021 to 1.059 per unit, apart from generator buses that are having 1.00 per unit as constraint value. The bus bar voltages of 0.95 and 1.05 per unit are recognized as the minimum and maximum voltage values for a standard power system grid [1, 2]. Therefore, buses that are having voltage magnitude whose values falls above 1.05 per unit are regarded as overloaded buses[3,8]. From Table 8, buses 2 (Osogbo) and 3 (Aiyede) falls in to this category of overloaded buses. Apart from these 2 bus bars, the calculated voltages of other bus bars in the grid are within the allowable range and therefore such bus bars are not overloaded. Overloaded busbars are often experienced due to high electricity demand as well as long distance between the busbars and the supply sources (generating stations). However, it is explicit in the study that two generation sources are not presently existing in the Nigerian power utility 330kV grid, but they are proposed for future expansion. These generation sources are Makurdi (MKG) and Ikom (IKG).

Table 9 shows the network flow pattern of the power flow solution results. It could be seen that these results justified a steady state condition of the grid because the transmission lines are not overloaded. Giving a clear picture, Table 9 illustrates the flow of active and reactive power from bus bar 2 (Osogbo) to bus bar 3 (Aiyede) measured at Aiyede are - 0.221MW and 0.495MVAr per unit respectively. But measuring the returns at Osogbo, the active and reactive powers are 0.221MW and 0.157MVAr per unit respectively which shows that the active power loss on the line is negligible. While lesser reactive power flow is measured at Osogbo, signifying losses.

Also, from Table 9, the Total Power Loss is obtained to be - 0.220 + j27.054 per unit while Table 2 shows that the total generation installed capacity is 3860MW. However, it's rare to have generation companies operating at an installed generation capacity. Therefore, it is assumed in this study that 75% of installed generation capacity is the generation level of the grid network for Nigeria. Consequently, 2895MW being 75% of installed generation capacity and of course the country generation level. From the above, total active power loss is 0.22 per unit which implies 22MW energy loss. Base on this value and generation capacity, the percentage energy loss in this network becomes 0.76%. This mean that energy lost in the network is very little and negligible. The results in table 9 indicates that the slack bus power S_1 is 1.746 + j2.004 per unit at Sapele Generating station while the power S_2 at bus just immediately after the generation bus termed Sapele load bus is 1.934 + j0.658 per unit. It is observed that S_2 is greater than S_1 because power S_1 has already been stepped up due to the applied step-up transformer that jacked the generated voltage up for transmission purposes.

In addition, the power utility data shown in table 2 shows that the installed generation capacities (active power) in Megawatts for bus bar 1 (the slack bus) is 1020MW (or10.20 per unit). While, the power flow solution results shows that the slack bus power (active) is 1.746 per unit, which of course is lower than the installed generation capacity of 10.20 per unit. This, however, confirms that the use of bus 1 as the slack bus is realistic. The power flow solution converged at eight iterations and the results obtained are quite reasonable, interpretable and reliable. This makes the software package better compared to previous research works where power flow solution either converged at high iteration level or the system did not converge at all leading to no solution.

It was observed that the following transmission lines are either having high reactive power than the active power or both powers are having almost the same values. The affected lines include; line 1-7, 2-4, 2-11, 4-6, 6-7, 10-16, 13-18 and 16-17. This condition is not tolerated in power system as it causes system instability. Therefore, installation of Capacitor bank or other means of absorbing reactive power on those lines is highly recommended to power utility management.

5. Conclusion

This paper presents an application of the developed NR software package to the power flow analysis of Nigerian 330 kV regional power utility grid. The grid consists of 21 buses and 23 transmission lines. Often, a bus bar, preferably a generator bus bar is used as slack bus. Consequently, bus bar 1 (Sapele bus bar) was chosen as the slack bus and its voltage is 1.0 < 0 per unit for this analysis. The voltage of the voltage controlled bus bars (or generator) were also constrained to 1.0 < 0 per unit. The results of the analysis converged at the third iteration. This early convergence of the NR solutions satisfied the theoretical reports of Newton – Raphson iterative algorithm. In addition, early

convergence may not also be unconnected with the use of real life power system grid having reliable operational parameters or data. The results of the power flow analysis for this Nigerian 330kV regional transmission grid are presented and discussed. These results are accurate and reliable. The transmission lines with high reactive power will deserves an installation of equipment that could reduce or control the reactive power on the lines. On the same note, the bus bars with voltage values not fall within the specified range would require the special attention of the power utility company; in this case, Transmission Company of Nigeria (TCN). Furthermore, the package can also be applied to determine the stability of medium voltage distribution feeders.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors sincerely appreciate the Transmission Company of Nigeria (TCN) for providing the data used in this research work.

References

[1] A. Abdulkareem, Power Flow Analysis of Abule-Egba 33-kV Distribution Grid System with real network Simulations., IOSR J. Electr. Electron. Eng. **9,** 67–80, 2014.
[2] I.A. Adejumobi, G.A. Adepoju, K.A. Hamzat, Iterative Techniques for Load Flow Study: A Comparative Study for Nigerian 330kv Grid System as a Case Study, Int. J. Eng. Adv. Technol. **3,** 153–158, 2013.
[3] L.M. Adesina, J. Katende, G.A. Ajenikoko, Development and Testing of Q-Basic Computer Software for Newton-Raphson Power Flow Studies, in: 2019 2nd Int. Conf. IEEE Niger. Comput. Chapter, 1–6, 2019.
[4] R. Alqadi, M. Khammash, An Efficient Parallel Gauss-Seidel Algorithm for the Solution of Load Flow Problems, Int. Arab J. Inf. Technol. **4**, 148–152, 2007.
[5] J. Chakavorty, M. Gupta, A New Method of LoadB_ Flow Solution of Radial Distribution Networks, in: 2011.
[6] H.-C. Kim, N. Samaan, D.-G. Shin, B.-H. Ko, G.-S. Jang, J.-M. Cha, A New Concept of Power Flow Analysis, J. Electr. Eng. Technol. 2, 2007.
[7] M.A. Kusekwa, Load Flow Solution of the Tanzanian Power Network Using Newton-Raphson Method and MATLAB Software, Int. J. Energy Power Eng. **3,** 277, 2014.
[8] L.M. Adesina, O. Ogunbiyi, G. Ajenikoko, Newton-Raphson Algorithm for Power Flow Solution and Application, **9**, 1–17, 2020.
[9] V.K. Shukla, A. Bhadoria, Understanding Load Flow Studies by using PSAT, in: 2013.
[10] S. Singh, Optimal Power Flow using Genetic Algorithm andParticle Swarm Optimization, IOSR J. Eng. **2**, 46–49, 2012.
[11] S. Singh, Load Flow Analysis on Statcom Incorporated Interconnected Power System Networks Using Newtonraphson Method, IOSR J. Electr. Electron. Eng. **9,** 61–68, 2014.
[12] M. Syai'm, A. Soeprijanto, Improved Algorithm of Newton Raphson Power Flow using GCC limit based on Neural Network, Int. J. Electr. Comput. Sci. **12,** 7–12, 2012.
[13] Transmission Company of Nigeria, Line parameters, Bus specified power and voltages, Annu. Mag. Natl. Grid. 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 447-454 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Electronic Warfare Methods Combatting UAVs

Miroslav Kratky[*,1], Vaclav Minarik[1], Michal Sustr[2], Jan Ivan[2]

[1]*University of Defence, Department of Air Defence, 662 10, Czech Republic*

[2]*University of Defence, Department of Fire Support, 662 10, Czech Republic*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 15 October, 2020
Online: 20 November, 2020

Keywords:
Unmanned Aerial System – UAS
Unmanned Aerial Vehicle – UAV
Air Defence
Electronic War-fare
Cybernetics
Neural Network

ABSTRACT

This paper describes methods of eliminating Unmanned Aerial Vehicles (UAV) non-destructively, using Electronic Warfare Methods. The aim is to introduce certain methods of UAV detection and elimination in a complex environment and terrain, e.g., in an urban and battlefield environment, that will result in finding the control device position and the UAV itself. Neural networks, cyber penetration elements, and wireless network scanning programs are all used to address this issue. The output of this article is a new concept of a comprehensive solution, which can be implemented into the existing complex system of electronic defence against UAVs, e.g., within the allied base. Conclusions will be also used to further improve the above-mentioned topics at the authors' workplace, within the frame of long-term projects and specifically as a part of solutions applicable to the force protection of combat support units, namely field artillery, which is described here in detail.

1. Introduction

This paper is an extension of the work originally presented in proceedings of the "2019 International Conference on Military Technologies" (ICMT), Brno, Czech Republic [1]. The original material was enriched by the proposed concept applied in the urban area. The whole section proposing a possible concept applicable to artillery units was added.

A very dynamic development of Unmanned Aerial Systems (UASs) is becoming highly visible as they are used in all areas of human activities [2].

With the gradual widening of the UASs complexity and complementing other features, there is a demand for more complex security. The reason is that both military forces and terrorists have increasingly used Unmanned Aircraft Systems to plan, prepare, and execute attacks on ally and partner's forces and on "soft targets" in the civilian sector. Preventing, protecting, and recovering from such attacks require a cross-governmental approach, bridging the different efforts that Allies, partners, NATO, other international organizations, industry, and academia are making on this topic, both in the military and in the civilian domain. The threat description, probable scenarios, and protection models concepts are well described now by several trustworthy documents, e.g., in [3].

In another word, simultaneously with UASs development, we have been witnessing the counter-UAS (C-UAS) systems rapid improvement as well. Moreover, the C-UAS systems development concerns not only the direct defence against flying apparatuses themselves, but this also applies to the whole loop of the air defence system: detection, command-control system and elimination itself. The whole engagement process is now a highly complex, sophisticated, and, from a scientific point of view, multidisciplinary one.

Thus, after transforming from simple radio-controlled machines into sophisticated, "smart", digitally-controlled UAVs, an opportunity was developed to combat the whole UAS not only with standard methods, but – apart from another ones - also with cybernetic methods. Contemporary modern UAV can be seen as a small computer or a mobile phone with the ability to fly. This connects the current IT security issues and air defence issues together with the UAVs.

During the next development, it will be necessary to take into account the possibility of anti-aircraft defence congestion caused by the high availability of micro UAVs at a very low affordable price. Therefore, the developed means should be able to operate a large number of UAVs at the same time and at a minimum cost.

[*]Corresponding Author: Miroslav Kratky, Email: miroslav.kratky@unob.cz

2. Methods of Cyberattack Applied to UAVs

In today's digital age, the growth of cyberattacks can be seen not only on personal computers, servers, and mobile phones, but also with the introduction of "internet of things", on any device that is able to connect and communicate over the network. This opens the possibility of cyberattacks on most types of UAVs too. Cybernetic methods of attack, like conventional methods, can be divided into two basic groups according to the expected effect. These methods are non-destructive and destructive.

2.1. Non-destructive Methods of Cyberattack

Non-destructive cybernetic methods of combat are understood as methods in which there is no direct destruction of any particular component of the affected system. This group includes the most contemporary cyberattacks. We can further divide these attacks into several subareas (in detail: [4]). Graphical representation of non-destructive cyberattack methods on UASs is depicted in Fig. 1.

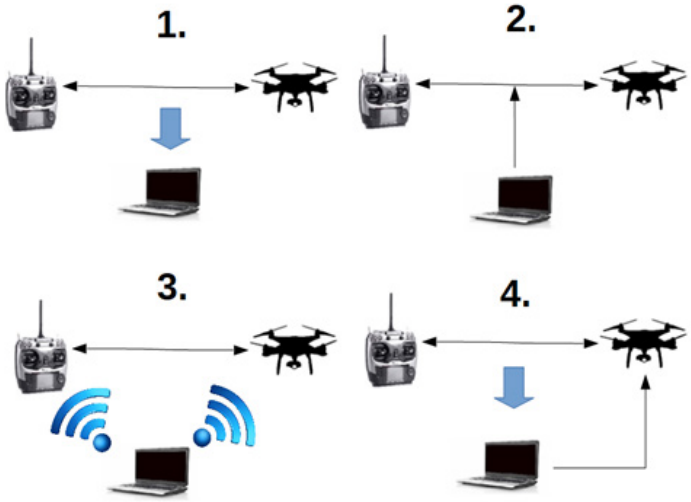


Figure 1: Graphical representation of non-destructive cyberattack methods applied to UAV´s

1) Leakage of Information

This type of attack results in disclosure or leakage of protected information. The advantage is its detection difficulty and in most cases the speed of attack. In UAVs, this attack is about getting the downlink channel information providing the UAV's mission or about getting the data to find the password for Wi-Fi communication.

2) Disturbance of Integrity

This subgroup includes attacks where the UAV's data are destroyed, damaged, or changed. This type of attack is very well detectable, but in most cases until after its accomplishment.

3) Denial of Service

Denial of Service (DOS) attacks make it impossible to use a particular service or system of UAS. The C-UAS defender will focus on a particular access channel or a specific service and will disable real-time activity by systematically sending requests (see DoS attack below). This attack is very visible and is usually suppressed quite quickly.

4) Unlawful Use of Information

These attacks focus on using the obtained information to access non-public parts of the system or to use certain services without authorization. In UAVs, for example, the acquired password can be used to decrypt the intercepted communication or to take over its control.

2.2. Destructive Methods of Cyberattack

The destructive C-UAS cybernetic methods of combat have a direct impact on the part of the attacked system that is physically irreversibly damaged as a result of this activity. These methods mainly use the vulnerability in the lower layers of the OSI (Open Systems Inter-connection) model and focus primarily on individual hardware components that are used in multiple systems. The operation of these methods is primarily based on a mechanical damage. For example: the cybernetic attack induces a collision of mechanically moving components or the cybernetic attack forces a battery or some other component to overheat and thus damage themselves.

3. Signal Detection of the UAV

When combatting mini and micro UAVs, one of the biggest problems is the detection and identification of the UAS itself, especially in an urban densely built area. Using radar or other detection methods with optical equipment (in the visible or IR optical band) is often considerably complicated by many fixed obstacles [5]. Methods using specific acoustic characteristics of the UAVs are also limited due to the interfering ambient noise [6].

The best choice in such environments is, therefore, the detection and localization of signals transmitted by the UAV itself or its control station.

These signals can be relatively well detected and identified due to their known specific transmission frequencies and known encoding.

However, this detection becomes considerably harder if the UAV is managed only by Wi-Fi in an electronically complex environment. Contemporary modern cities are full of devices that use the same Wi-Fi standards, and that hides the UAV' control signal among the other Wi-Fi networks within its range [7].

3.1. Localization Position of the remote control station or UAV

An UAV recognition method can be used, utilizing the MAC address to locate the position of the control station or the Wi-Fi standard to locate the position of the connected UAV itself. Each producer has a certain range of initial MAC address characters, which makes it uniquely identifiable. However, a problem occurs when the Wi-Fi module is modified or if the UAV MAC address is changed by software. Both of these methods are realizable by at least an average capable IT man.

3.2. Analysis of Data Flow Using Neural Network

One of the progressively evolving technologies is the technology of neural networks (NN). The main domain of these networks is their relatively rapid analysis of a large volume of data and the data subsequent evaluation. Neural networks, unlike algorithmic solutions, do not use serial computations. The task is solved simultaneously with several layers of neurons that interact

with each other. Neural network inputs can also be parameters of the UAS Wi-Fi traffic, such as the number of frames, their size, and generally the data flow over time.

Thanks to NN advantages, especially in the field of data processing, i.e., their ability to take in a lot of inputs, process them to infer hidden as well as complex, nonlinear relationships, NNs are playing a big role in signal characteristic recognition. Based on these features, the NN should be able to recognize which device it is both the UAV itself and the control station as well. Thus, the next part is devoted the NN application to the UAS detection and identification.

4. The UAS Signal Detection Using Neural Network

Neural networks can be used in the process of data mining. This is currently an increasingly inflected term. In general, it can be understood as a process aimed at discovering dependencies or finding the required information in a large volume of usually experimentally obtained data. The output is a certain knowledge that can be used in solving a decision problem, predicting values for other new data, or simply understanding a certain phenomenon or context. [8]

In principle, three methods can be used to identify UASs using 802.11 standards. UAS can be identified by the SSID, which is the name of the access point. However, this can be changed very easily. Furthermore, we can use the identification according to the MAC (Media Access Control) address, because in the MAC address the first six characters identify the manufacturer. This method is already suitable for use, but the MAC address can also be changed. The third method is based on the analysis of the data flow and the creation of the so-called fingerprint. For this purpose, a neural network can be successfully used, which has the task of analysing the data flow to decide whether it is intercepted traffic originating from UAS communication or not.

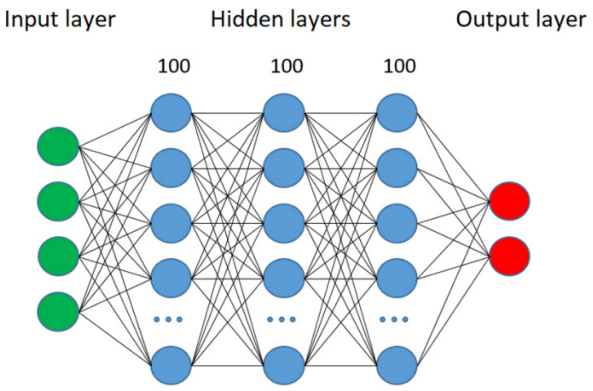


Figure 2: Topology of the created neural network

Input data is a key element in training neural networks. The ability of the network to learn and evaluate, or rather to evaluate with sufficient accuracy, depends on the appropriate selection of individual parameters and the input data set.

The selection of input parameters was made on the basis of knowledge obtained by analysis of data flows originating from the tested UAS and several other applications (e.g. Skype, YouTube). Analysis of data streams revealed that one of the determining parameters is the number of unique frame sizes in a given sample. This is due to the nature of the transmitted stream, where certain activities or communications create a larger number of unique frame sizes than others. The second determining parameter is the number of frames transmitted per second. This is mainly related to the overall data flow, but it also in some way represents the disposition of the intercepted communication. The average frame size was chosen as the third parameter. It expresses whether the captured traffic contains rather smaller or larger frames. The size of the frames is affected by the disposition of the transmitted data. The last parameter was the ratio of the number of frames with a size greater than 100 B and less than 100 B. This value was chosen mainly based on the analysis of all data streams, where different operations have this ratio different. The topology of created neural network is depicted in the Fig. 2.

Based on these parameters, the neural network can distinguish whether it is UAS operation even when masking the SSID or changing the MAC address of the transmitted frames.

5. Possibilities for UAS Elimination Using EW Methods

5.1. Possibilities of Penetration into the UAV Communication Channel

In general, the resilience of security depends on several factors, of which the quality of the encryption used and the length of the password usually have the greatest influence. Specifically, WPA2 security uses the CCMP (Counter Cipher Mode Protocol) algorithm based on the Advanced Encryption Standard (AES) encryption algorithm. The length of the password that can be used with WPA2 is between 8-63 characters. Security resilience depends not only on the length of the password, but also on the character set used. It is also advisable to choose passwords that are not contained in dictionaries, or that do not contain word forms or diminutives [7].

Brute force Attack and Dictionary Attack by Cloud Computing

Brute force Attack and Dictionary Attack can be amplified by using Cloud Computing. In the case of a brute force attack, it is necessary to test all possible combinations in the set given by the specified parameters. Any known information about password parameters significantly speeds up the successful use of this attack. Brute force Attack will certainly find a password in the future, but it is better to use a dictionary attack first. Dictionary attack is based on creating as big as possible a dictionary of known words, which increases the chance of an earlier password being discovered. Commonly used high-performance computer sets can test even with a GPU (Graphic Processor Unit) of only 400 KHps (Kilo Hash per second) when used for WPA2 encryption. The way to amplify computing power is called Cloud Computing. With these services, it is possible under certain circumstances to break relatively long passwords in short or real time. These services can be rented from technology companies such as Google or Amazon, from 10 minutes for the duration of the rental [9].

Rainbow Table

In some cases, the "Hashing Function" format is used for encrypted communication. This function recalculates the password entered by the user and the output is "Hash", which is a transformation of a string to a given number of characters. These

hashes are precalculated within the "Rainbow Table" to make it easier to crack the password [10].

DoS Attack

Denial of Service attack is an attack which transmits recurring deauthentication frames that result in disconnection of communicating devices on 802.11 [11].

KRACK Method

The Key Reinstallation Attack (KRACK) method is one of the relatively newly discovered methods focusing on a certain vulnerability of the Wi-Fi standard. It uses a four-way authentication process, in which communication is established between the AP (Access point) and the client. Simply put, it works on the principle of delaying the response to the third message during the four-way authentication sent to the access point. As a result, the AP sends a new 3rd message with an increased "counter number". When it receives the 3rd message for the first time, the client installs the GTK (Group Transient key) and PTK (Pairwise Transient Key) keys and sends the 4th message, which is then held. Upon receiving a retransmitted 3rd message with an increased counter number, the client reinstalls the previously installed GTK and PTK keys and resends the 4th message. Subsequently, both 4th messages are left to pass to the AP. Knowledge of changing the original and later reinstalled GTK and PTK keys can then be used to decrypt the communication [12].

5.2. Options after Successful Penetration into the UAV System

From the attacker's point of view, there are several ways to penetrate the system, differing in danger and enabling different goals to be met [1].

Secret Observation

The most inconspicuous way to start a penetration is a simple observation of the video stream or telemetry data, which are received by the UAS control station. This makes it possible to identify the equipment of the UAS with specific sensor (measuring) devices or to identify the operator' area of interest.

Sending Unobtrusive Commands

The operator' control can be affected by sending confusing or conflicting commands. Such an effect on the UAS may lead to the termination of the UAS task, as the operator will think it is a technical fault.

Complete take over the UAS

For the complete takeover of the UAS, it is necessary to disconnect the original RC and prevent it from being reconnected. To do this, you need to know the password and the ability to change it in real time as well. Successful takeover largely depends on the manufacturer of the individual components or the UAS model itself.

6. The Proposed Concept of the EW C-UAS System

6.1. Basic principles and functioning of the proposed concept

The basic principles of system operation are clearly shown in the Figure 3. The whole system is divided into one main and two support sections, each of which brings a different ability, or added value. The proposed concept [6] may represent the addition of another element of complex defence against micro and mini UAS.

The concept includes devices to detect signals and record Wi-Fi data in passive mode and send deauthentication frames in active mode. Within one computer and control centre, information are collected from one or more devices.

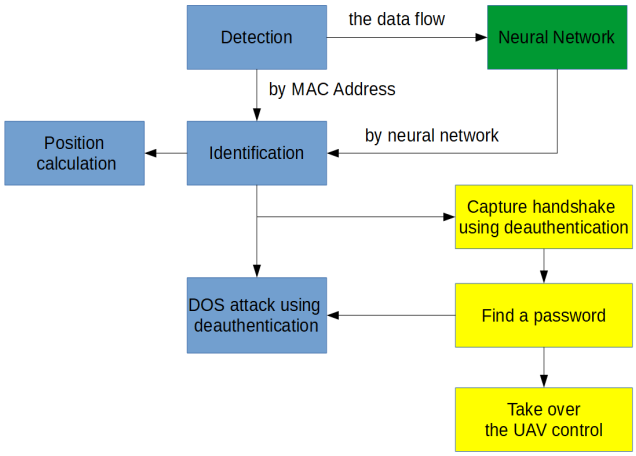


Figure 3: The proposed neural network

Detection, Identification, Position calculation and DoS Attack Using Deauthentication

The main section (shown in blue) provides the basic capabilities of the system and is the only one that can act automatically in a limited way. It contains 4 blocks providing initial signal detection (capture), device identification according to MAC address, position calculation based on information from one or more sensors and, if necessary, the DOS attack itself. Upon successful capture and identification of an unwanted UAS, we are able to send a deautentication frame to the nearest sensor to disconnect communication between the RC and the UAS or between the RC and the display device. After sending only one or a few deauthentication frames, the connection is established automatically, but if the deauthentication frames are still sent, the connection will not be established. [9]

Neural Network

The first support section (shown in green) evaluates the intercepted data stream using a neural network (explained in more detail in Section 4) and, based on the results, identifies the type of device or the type of traffic in the intercepted communication. This information is passed on to the main section for subsequent specification [6].

Possibilities to Find Password

The second support section (shown in yellow) aims to bring the ability to take control of the UAV. The process usually begins by sending a deauthentication frame and then capturing a 4-way handshake. This provides a signal pattern to perform the password retrieval process. After successfully finding the password, it is possible to take over the control, either completely or partially, or penetrate into the communication system. [6]

6.2. The proposed concept applied in the urban area

For practical use in urban areas, e.g., the device *Alfa AWUS036ACM* (see Fig. 4) can be used as a sensor and

transmitter, providing sufficient omnidirectional range, supporting 802.11 a / b / g / n / and ac standards, and operating in the 2.4 GHz and 5 GHz frequency bands, respectively. Another indisputable advantage of this device is its low price in combination with normal commercial availability [6].



Figure 4: Device *Alfa AWUS036AC*

Figure 5 shows an object with a marked critical defence point (red circle), access roads (3, 5), high-rise areas (1,2,4), and a low building (6), which poses the greatest risk due to the possibility of direct visibility to defended point

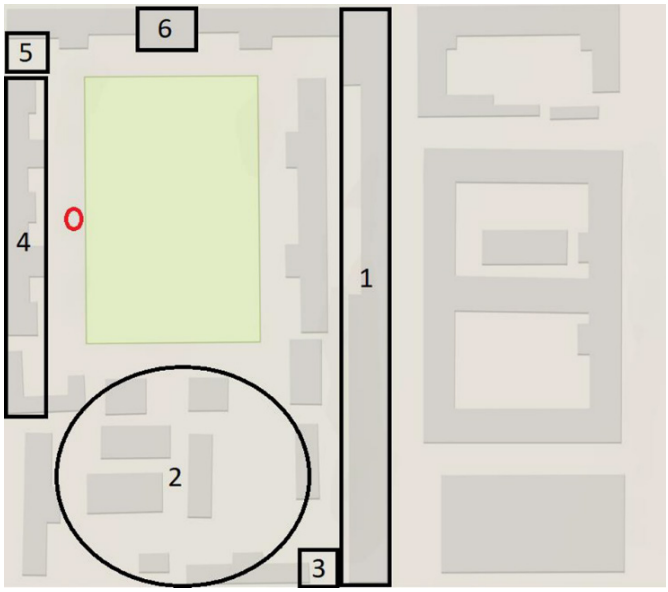


Figure 5: Defended critical point

It is advisable to place the sensors so that they cover most of the area and selected places outside the area from which the UAS control could be performed. The Alfa AWUS036AC device, which is connected to a certain small computer, e.g. *Raspberry Pi*, is taken as a sensor. All sensors would then be connected to a central point where evaluation and control would take place.

Despite of this that the quantity of sensors seems to be a little considerable, by using low-price hardware mentioned above, we are able to cover the protected area completely. Connection to the EW C-UAS system proposed in subsection 6.1 can sustain full control of this defended critical area.

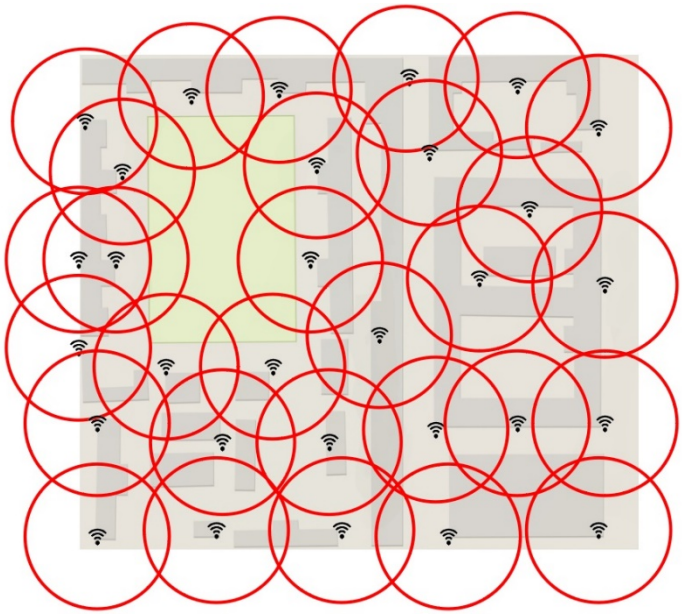


Figure 6: Defended critical point

6.3. Partial conclusion of Section 6

The previous parts described the general principles and division of cyber-attacks and the operation of the 802.11 standard of interest. To detect UAS using this standard was developed by applying the neural network and the concept in which it is embedded. However, instead of a neural network, it would be possible to use other methods of genetic programming, but the use of a neural network is suitable for this task.

Within this theoretical concept, the use of an existing HW and its deployment on the example of a defended object in a non-war area was also proposed. The use of the system in the military field is dealt with in the following section.

7. The proposed concept applied to artillery units

Based on knowledge from current conflicts, especially from the war in Ukraine and Syria, it is clear that UAVs - small, light, and cheap unmanned aerial vehicles pose a great risk to all combat units. The availability of these tools, which are available for small amounts of money, together with their capabilities makes them an ideal means of conducting aerial reconnaissance.

Findings from the conflicts in Ukraine and the Middle East clearly show that these instruments, when applied to fire support units, specifically artillery, can be used very effectively, especially for:

- reconnaissance (uncovering the battle group),
- target acquisition,
- artillery fire control,
- evaluation of the effectiveness of fires,
- perform attacks using explosives,
- perform attacks, using weapons s of mass destruction (chemical, biological, and possibly also radioactive "dirty" bombs).

In recent years, we can find countless combat situations, where small UAVs have successfully performed the tasks mentioned above. To illustrate this, several examples of the use of small UAVs are given in the table No 1.

Table 1: Examples of small UAV attacks

Year	Example
2005	Al-Qaida used Chinese made remote control model airplanes to recon Pakistan security forces prior attacks. It was also weaponized with IED.[13]
2014	Islamic state used small quadcopters for recon missions of Syrian military bases prior to ground attacks. [13]
2015	Drone attacks on military bases in Ukraine. Each drone was equipped with one grenade (ZMG-1 type). [14]
2015	Several attacks on ammunition depots at Svatovo, Ukraine. Drones were equipped with grenades. [14]
2016	Two French Special Forces soldiers were injured and an exploding ISIS drone killed two Kurdish fighters. [14]
2019	Drone attacks struck two key oil installations inside Saudi Arabia [15].
2020	Pro-Kremlin mercenaries use an unmanned aerial vehicle to drop a rocket-propelled grenade on Ukrainian emplacements. [16]

These examples clearly show the possibilities of using small UAVs and their ability to cause significant losses. The use of small commercial UAVs in combat operations can provide the enemy with key information about the positions of our own troops and their manoeuvres. In combination with explosives or weapons of mass destruction, it is possible to cause significant losses in technique and manpower. For this reason, it is necessary to be prepared for this variant and be able to defend effectively against these types of attacks.

7.1. Artillery units in the regular operations and proposed concept of application

In a peer-to-peer war, all kinds of military forces and resources are involved in combat operations. For simplicity, examples of the application of the proposed concept of protection against small UAVs are reduced only to artillery units. However, a similar approach can be used for all types of units and their specializations wherever there is a real possibility of attack by these UAVs.

The artillery battalion (battery) is conducting its operations within a given position area of artillery (PAA) within the zone of attack (defence) of the brigade (task force). Part of this PAA are also battalion (battery) Fire Direction Centres (FDC). Fig. 7 gives an example of PAA [17].

The danger emerging from the use of small UAVs is, in particular, the possibility of uncovering a combat formation and subsequent directing fire on firing units and command posts, and attacks using explosives placed on the UAV or attacks using weapons of mass destruction. Artillery units are usually deployed in a hidden position, outside the firing positions, when they do not conduct fires. Weapon systems occupy firing positions only when they are firing. Firing positions and firing points are usually in the open space. At this point, they are vulnerable and easy targets.

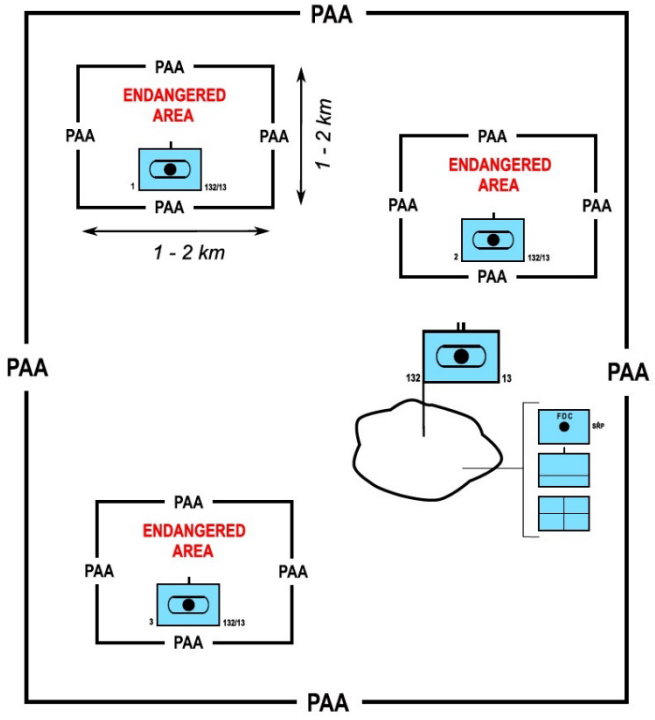


Figure 7: The position area of artillery [18]

One of the main risks the artillery is facing is the counterbattery fires. Artillery units are always priority targets and it is necessary to avoid the risk of detection, which precedes the implementation of counterbattery fire. The tactic of using small UAVs by enemy reconnaissance and diversion groups has always been a danger. In connection with the development of capabilities and possibilities of using not only military, but also small commercial UAVs, new risks are emerging and their frequency may be higher due to their massive spread and availability.

With a standard time of one fire mission and time to leave the firing position, the probability of successful counterbattery fire is low. In the case of proactive counterbattery activity, where artillery units are actively searched for and neutralized without unmasking themselves by firing, the probability of success is much greater than in the case of reactive counterbattery fire when it is reacted to firing artillery. The enemy gains an advantage by knowing the coordinates of the firing positions and starts the fire mission when the cannons are taking up firing points to conduct fire. It is this key information that can be easily obtained using small UAVs. They are able to uncover a combat formation, detect individual artillery weapon sets and already in the phase of the manoeuvre into the firing position, warn the enemy, and provide sufficient information for the preparation and execution of fire. At the same time, they can provide a detailed evaluation of battle damage assessment (BDA) after firing.

Another variant is the use of UAVs as carriers of explosive devices, which can be used to destroy targets. In the case of artillery, a suitable target are FDCs. Destruction of them will

eliminate the entire firing battery from combat for a period of time. Alternatively, it is possible to destroy individual artillery weapon sets, which cause a deficiency in fire support.

Based on these findings, the proposed concept of detection and neutralization of UAVs is a very effective tool for reducing the risk of detection by these means and the subsequent elimination of artillery units by enemy counterbattery fire, or explosive devices. From the point of view of the effective use of the proposed concept, it is necessary that the established measures reflect the tactics and procedures of artillery fire units. The biggest restrictions will be mainly area requirements. The firing battery of self-propelled cannon howitzers is deployed in the area of firing positions, usually 1 - 2 x 1 - 2 km, depending on the position conditions and the combat task [19]. This space must be covered with the ability to neutralize the UAV - this can be achieved by using a directional antenna with sufficient gain or power, which will ensure the required range. The individual antennas should be connected to an automated system, from which their focus and modes of operation will be controlled.

Firing units are most exposed to observation at the moment of performing fire missions, when they are deployed in firing positions. A suitable variant is the placement of equipment for neutralization of UAVs on vehicles of artillery reconnaissance/survey units, or on artillery weapons sets [20]. To a minimal extent, it would be sufficient to place this device on the lead gun of fire platoons, or vehicle of the platoon commander. However, this depends on the possibility of increasing the scope of the opportunity of neutralizing the enemy UAV within the proposed solution. At command points (FDCs), it would be sufficient to place the equipment on one vehicle of FDC.

7.2. Artillery units in asymmetric operations and proposed concept of application

The use of artillery in an asymmetric conflict has its specifics. Firing units are usually located on permanent or forward operational bases and provide fire support to units performing framework operations in the area of responsibility. They do not manoeuvre and are constantly at the firing points. In an asymmetrical operation, the risk of counterbattery fire is minimal according to current experience. However, commercially available UAVs can radically change this state if used correctly [21]. The danger of the UAV being used by the enemy lies mainly in the possibility of uncovering the combat formation, finding out the position of the firing point of individual artillery weapon sets, and possibly leading an attack on them using explosives placed on the UAV. In an asymmetrical environment, this way of conducting combat is one of the few ways to put the fire units located in the base area in danger. A coordinated attack can cause significant technical losses and limit the ability to provide artillery fire support in the area of operations. It is therefore appropriate to consider how to defend against such an attack and to provide protection for the artillery fire support units and other base personnel. Placing the defence EW device on the base in such a way, which achieves coverage of the base perimeter, is a suitable variant of solving the problem. In the event of a breach of the base perimeter, the operator is able to ensure the protection of base members and prevent the UAV from flying over the base within the proposed method.

7.3. Partial conclusion of Section 7

This concept of protection against small UAVs will ensure that units are not attacked or uncovered and the risk of losing artillery fire support will not increase. At the same time, the simplicity of the proposed solution will not place excessive demands on interventions in the construction of vehicles or buildings. The solution can be applied to all types of troops and its specialization, wherever the risk of using small UAVs in the combat activities of the enemy can be assumed.

8. Conclusion

UAV defence in a cyber environment combines elements of cyber security and air defence. The development of new methods for attack and defence is necessary due to the dynamics of the development of these areas, which is confirmed by current knowledge from the fighting in Nagorno-Karabakh, where unmanned aerial vehicles and artillery play a major role. One of the possible ways is the proposed solution for penetration into UAS control systems.

The methods described in this article focus on supplementing and extending existing complex solutions of UAV defence models. Their application clearly achieves an increase in capability in the implementation of defensive measures and the fight against the UAV opponent. Simplicity and ease of materials and new inexpensive technical means represent an advantage for application in practice, e.g. in the military environment, as described and suggested by the examples of artillery units.

As the main achievement, authors consider a practical demonstration of the possibility for identifying the device based on the characteristics of the data frames. For this purpose, a neural network was developed which, based on the entered parameters, can evaluate whether the intercepted traffic comes from UAV communication.

As all the UAS defence systems are usually aimed at a certain type of opponent, this concept is not universal. It focuses on complementing the existing comprehensive defence model with more options, thus creating the ability for the defence institutions to fight more effectively with highly sophisticated adversary means.

Subsequent research in this area could be aimed at expanding the input parameters of the neural network and output neurons. The neural network should then be able to identify not only the device or the type of data stream transmitted during UAV operation, but also other types of operation. The whole system could then be used for overall monitoring - for the purpose of using Wi-Fi networks in the area of interest. However, extending this system to such a level could mean a significant intrusion on the privacy of individual users. If implemented in practice, this problem would be forced by the author to solve both at the technical and legal level.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was performed as a part of the project "PROKVES" for Development of Electro – Military Departments at Faculty of Military Technology and "SPECIFIC RESEARCH" at Faculty of Military Leadership, University of Defence in Brno. They have been financed from the Institutional Support Department funded by the Ministry of Defence Czech Republic.

References

[1] V. Minařík a M. Krátký, "Cybernetics fight against the UAV". 2019 International Conference on Military Technologies (ICMT), Brno, Czech Republic: Institute of Electrical and Electronics Engineers Inc., 2019, 8870103.

[2] M. Kratky and J. Farlik "Countering UAVs – the Mover of Research in Military Technology". Defence Science Journal 2018, **68**(5), 460-466. https://doi.org/10.14429/dsj.68.12442.

[3] NATO C-UAS Working Group, "NATO Countering Class I Unmanned Aircraft Systems (C-UAS) Handbook". Version 1.8. Brussels, BEL 2020.

[4] J. Kolouch, "Kybernetické útkoy" 2019. https://csirt.cesnet.cz/_media/cs/documents/kyberneticke_utoky.pdf

[5] T. Kornelly, J. Casar, V. Stary and J. Farlik, "Mobile phone optical sensor usage for navigation tasks" 2017 International Conference on Military Technologies (ICMT), Brno, 2017, 671-675. doi: 10.1109/MILTECHS.2017.7988842.

[6] V. Minařík, "Elektronický boj v obraně proti bezpilotním prostředkům". Brno, 2019. Ph.D. exam thesis. University of Defence.

[7] V. Minařík a M. Krátký, "The Non-destructive Methods of Fight Against UAVs". 2017 International Conference on Military Technologies (ICMT), Brno, 2017, 690-694. doi: 10.1109/MILTECHS.2017.7988845.

[8] M. Lastovka, "Data mining jako moderní prostředek analýzy dat a možnosti jeho uplatnění v podmínkách AČR". Brno, 2015. Diploma thesis. University of Defence.

[9] "Ethical Hacking and Countermeasures v10". EC-council, 2019. https://iclass.eccouncil.org/.

[10] "Rainbow tables tajemství zbavené" [Online article]. 2015 [cit. 2019-01-5]. Available from: https://www.soom.cz/clanky/1165--Rainbow-tables-tajemstvi-zbavene.

[11] J. Kolouch, P. Bašta, "CYBERSECURITY". Praha: CZ.NIC, z. s. p. o, 2019. ISBN 978-80-88168-34-8.

[12] M. Vanhoef, F. Piessens, „Key Reinstallation Attacks: Forcing Nonce Reuse in WPA2" [Online article]. 2017 [cit. 2019-01-5].

[13] J. R. Bunker, Terrorist and insurgent unmanned aerial vehicles: use, potentials, and military implications. Carlisle, Pennsylvania: US Army war college press, 2005.

[14] Russian Drone With Thermite Grenade Blows Up a Billion Dollars of Ukrainian Ammo. Popularmechanic [online]. 2017, 27. 7. 2017, [cit. 2020-08-13]. Dostupné z: https://www.popularmechanics.com/military/weapons/news/a27511/russia-drone-thermite-grenade-ukraine-ammo/.

[15] Donbas: Russian militants use drone to attack Ukrainian positions. 112UA [online]. 1. 6. 2020, [cit. 2020-08-13]. Dostupné z: https://112.international/conflict-in-eastern-ukraine/donbas-russian-militants-use-drone-to-attack-ukrainian-positions-51819.html

[16] J. Ivan, K. Šilinger, L. Potužák. "Target acquisition systems suitability assessment based on joint fires observer mission criteria determination". In: Madani K., Gusikhin O. ICINCO 2018 - Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics. Portugal: SciTePress, 2018, roč. 1, 397-404. ISBN 978-989-758-321-6.

[17] B. Čermák, R. Hegyi, L. Potužák, M. Kalina, "Bojové použití dělostřelectva", Military publication Pub-35-14-01, TRAIDOC Vyškov, 2013.

[18] J. Ivan, "Situační značky a zkratky pro dělostřelectvo: (vojenská symbolika a taktické značky pro dělostřelectvo dle APP-6)". University of Defence in Brno. ISBN 978-80-7582-122-5.

[19] M. Bláha, K. Šilinger, "Application support for topographical-geodetic issues for tactical and technical control of artillery fire". In: International Journal of Circuits, Systems and Signal Processing, **12**, 48-57. ISSN 1998-4464.

[20] IVAN, Jan and Jan POTUŽÁK, Jiří ŠOTNAR. Dělostřelecká rekognoskace pro zabezpečení činnosti autonomních zbraňových systémů a základní požadavky na rekognoskační jednotky. *Vojenské rozhledy*. 2019, **28** (4), 063-077. ISSN 1210-3292 (print), 2336-2995 (on-line). Available at: www.vojenskerozhledy.cz.

[21] KARBER, Phillip A. Lessons Learned from the Russo-Ukrainian War: Personal Observations [online]. 8 July 2015. Available at: https://www.researchgate.net/publication/316122469_Karber_RUS-UKR_War_Lessons_Learned.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 450-456 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Solution Applying the Law on Road Traffic into A Set of Constraints to Establish A Motion Trajectory for Autonomous Vehicle

Quach Hai Tho[1,*], Huynh Cong Phap[2], Pham Anh Phuong[3]

[1]*University of Arts, Hue University, 530.000, Vietnam*

[2]*Vietnam - Korea University of Information and Communication Technologies, Da Nang University, 550.000, Vietnam*

[3]*Faculty of Information Technology, University of Education, Da Nang University, 550.000, Vietnam*

ARTICLE INFO

Article history:
Received: 12 May, 2020
Accepted: 06 June, 2020
Online: 18 June, 2020

Keywords:
Autonomous Vehicle
Model Predictive Control
Path Planning
Motion Planning
Intelligent transportation systems

ABSTRACT

With a model predictive control approach and to set the motion trajectory for an autonomous vehicle in situations where emergency braking cannot be performed, in this paper, we propose a solution to apply the law on road traffic into a set of constraints and thereby build an objective function to create motion trajectory for autonomous vehicle. The newly created trajectory is created to improve performance and enhance the ability to avoid obstacle but ensure an optimal global trajectory. The performance of this solution is assessed through simulation with different scenarios, from which there are applied research orientations on the problem of autonomous vehicle in practice.

1. Introduction

In recent years, many researches on autonomous vehicle problems have been carried out based on basic components such as navigation systems, environmental perception, planning and control [1,2,3]. In these components, motion planning is an important function to determine motion process of the vehicle. It provides the coming target of the vehicle by using the information received from the environment and navigation system. Therefore, the planning components must consider not only the vehicle elements but also the legal, ethical, and environmental changes through perception data received in the system to ensure reliability and safety when participating in traffic.

One of the challenges and tasks that researchers are interested in the field of autonomous vehicles is to create the optimal motion trajectory. It includes certain criteria such as creating smooth movement, creating comfort and achieving good energy efficiency, and must meet the restrictions arising during the operation of the vehicle such as the provision about the road traffic law.

The problem of determining the optimal path in complex environmental conditions has been studied with different approaches [4] such as potential field technology, searching techniques on graphs, or model-based techniques [5], in which the approach of interest in research is model-based. These methods are widely used for structured road and optimal trajectory that will be determined from a set of candidate trajectories. However, one difficulty faced by model-based motion planning solutions is how to effectively sample candidates in trajectory space. Since the model-based methods basically reach the lower end of the final target trajectory with optimal techniques, they require a sufficient amount of resources to obtain a large number of candidates to find the global optimal trajectory. Therefore, the optimal of model sample trajectories does not allow real-time application.

In order to solve the above-mentioned difficulties, in recent years, the planning methods of continuous air-space planning with the method based on optimal control and model predictive control (MPC) [6,7] has attracted research interest. The MPC approach usually uses nonlinear optimization technique, which will solve the problem with an iterative process for optimal control of the predictive horizon and that is the advantage of this method. From which the system built is capable of handling constraints imposed to ensure safety and facilitate movement planning in complex environmental conditions. This method can meet the requirements implementing real time.

In this paper, we propose a solution to set the motion trajectory for autonomous vehicle for a given period of time with constraints

Corresponding Author: Quach Hai Tho, Email: qhaitho@hueuni.edu.vn

built from the rules of road traffic law. The goal is to improve not only the calculation efficiency but also the uncertainty in the perceptual data of the proposed environment and vehicle system. The experimental simulation process in obstacle avoidance situations, overtaking other vehicles and decision to change lanes, etc. is carried out in a short distance, so that the problem is solved focusing on a prediction interval, a single MPC phase.

In the next section, we will introduce the set of constraints and the basic principles of the constraint clause, which are built from the provisions of road traffic law when participating of autonomous vehicle. Finally, we present the empirical part with the simulation process and conclude with some suggestions for further research directions for the problem of autonomous vehicle.

2. Building solutions for establishing motion trajectory

In the process of developing solutions to set the motion trajectory for autonomous vehicle, the factors need to be calculated and considered such as the dynamic system, the set of constraints in the operation of the vehicle, limitations on the operating environment such as the structure of the road system, the regulations of the law, the obstacles on the road ... The aim of the problem is to find a trajectory that is safe and feasible.

In this paper, we propose an approach to use model predictive control (MPC) to perform the motion trajectory planning, in which the expected trajectory is updated in subsequent stages of the model and build the cost function with logical constraints for the problem of generating the optimal motion trajectory.

However, in this paper, we only solve this motion planning problem in the case the autonomous vehicle is moving on a straight path and has a lane mark for overtaking. At the same time, the process of determining the trajectory is carried out in a predictive range of a predictive control phase, from which we can model roads, vehicles and other objects in the Cartesian coordinate system (x, y) overall, where the x-axis is the vertical direction and the y-axis is the horizontal direction of the road. Each vehicle and other objects are described by their position (S_x, S_y) within the boundary of the lane. The system state representations include the orientation angle θ_r, the distance from the boundary to the vehicle d_r, the reference arc length S_r and $K_r(S_r)$ are parameters of the arc curvature of the reference curve.

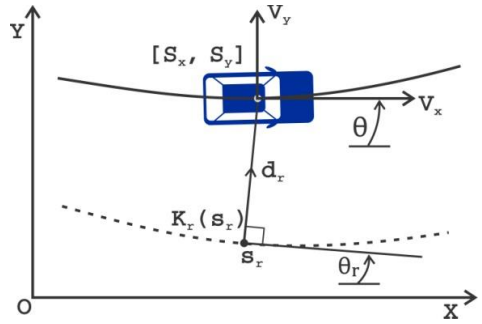


Figure 1. Vehicle model and reference curve

To facilitate the performance, in this paper we show the structure of the road as a system of roads defined by adjacent lanes of arbitrary shape and curvature. We assume and consider the i ($lane_i$) to be a path defined by the left boundary (B_{L_i}) and the right boundary (B_{R_i}). Each such path is defined as a polyline and is a combination of all lanes at a given time interval ($lane(s) = \bigcup_i lane_i$).

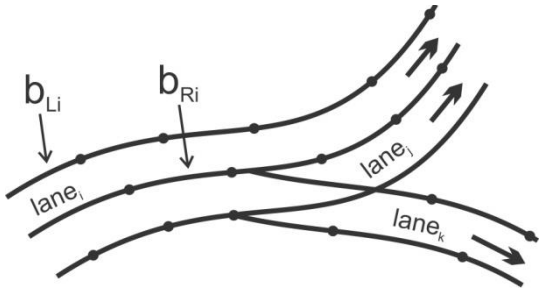


Figure 2. Road system model

The main idea when building a motion trajectory solution is to use a model predictive control (MPC) method, in which the set of constraints is constructed as logical propositions, including constraints on legal and ethical issues of traffic behavior. Based on the provisions of Vietnam traffic law, we perform logical clauses to introduce the constraint set in traffic as follows:

Rule 1: If the vehicle discovers obstacle ahead, it must slow down.

Rule 2: If it overtakes another vehicle, it must do so on the left side.

Rule 3: If it overtakes another vehicle on the right side, it must satisfy the condition that the vehicle ahead has given a signal to turn left or is turning left.

Rule 4: If it is traveling in an area where overtaking is prohibited or where the weather conditions do not ensure safe, it cannot overtake another vehicle.

Rule 5: If it changes lane or changes direction, it must give a turn signal.

Rule 6: If it changes direction, it must give way when detecting a vehicle ahead that is rudimentary vehicles and pedestrians.

Rule 7: If it moves in a residential area, the movement speed of the vehicle does not exceed 50 km/h and outside the residential area is 80 km/h.

Rule 8: If it performs changing lane and overtaking, it cannot do so continuously.

However, the issue of motion control decision depends on many situations, different states of the object types, and from which the vehicle's motion trajectory is set appropriately. Therefore, for each class the object can be mapped to a single homotopy layer of trajectory. To describe these homotopy layers, we will also build with corresponding logical propositions. For example, the logical proposition of the homotopy class is described when meeting obstacles as follow:

Rule 9: Vehicle in motion, if it overtakes other vehicles, it must overtake on the left side or on the right side.

Rule 10: Vehicle in motion, if it avoids obstacle, it will overtake to the left or to the right of the obstacle and reduce its movement speed.

As we all know, the essence of predictive control is to use an explicit model of the object to calculate optimal variables controlled through optimization methods. Specifically, to control the prediction for an object, we need to perform the following steps: Building a predictive model, defining target functions and constraint conditions, and finally solving the optimal problem. At the same time a number of predetermined conditions such as

communication between vehicles and future trajectories are determined [8, 9] during the construction of the trajectory of the motion plan.

With the idea of using the model predictive control and the first part of the proposed vehicle model with the state vector $\boldsymbol{\omega}$, control vector $\boldsymbol{u}$ as the basis for the research problem as follows:

$$\omega = \left[s_x, v_x, s_y, v_y\right]^T \quad (1)$$

$$u = \left[a_x, a_y\right]^T \quad (2)$$

where s_x, s_y represent vertical and horizontal position, v_x, v_y are velocity and a_x, a_y are vehicle acceleration along the x, y axes of the vehicle in inertial frame.

Then the dynamic model of the vehicle is represented with zero matrices of proper dimension, as follows:

$$\dot{\boldsymbol{\omega}}(t) = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & A \end{bmatrix} \boldsymbol{\omega} + \begin{bmatrix} B & \mathbf{0} \\ \mathbf{0} & B \end{bmatrix} \boldsymbol{u} \quad (3)$$

where $A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

In this problem, we make the assumption that the control vector $\boldsymbol{u}$ is a constant function at each time step τ. Therefore, the dynamical model of the vehicle is represented approximately with initial values including state vector $\boldsymbol{\omega}(k)$ and control vector $\boldsymbol{u}(k)$ in time interval $[k\tau, (k+1)\tau]$ as follows:

$$\omega(k+1) = \begin{bmatrix} A^d & 0 \\ 0 & A^d \end{bmatrix} \omega(k) + \begin{bmatrix} B^d & 0 \\ 0 & B^d \end{bmatrix} u(k) \quad (4)$$

where $A^d = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix}$, $B^d = \begin{bmatrix} \frac{1}{2}\tau^2 \\ \tau \end{bmatrix}$

In order to achieve computational efficiency with the limitations of the vehicle's dynamic system as well as the provisions about motion direction when overtaking obstacles on the road, it is necessary to consider the constraints on the state $\boldsymbol{\omega}$**(.),** the input control signal $\boldsymbol{u}$**(.)** and the motion direction of the vehicle $\boldsymbol{\theta}$**(.)** must meet the following conditions so that when the motion trajectory is built, it is feasible as follows:

$$\omega \in [\omega_{min}, \omega_{max}] \quad (5)$$

$$u \in [u_{min}, u_{max}] \quad (6)$$

$$\theta \in [\theta_{min}, \theta_{max}] \quad (7)$$

In which, the upper and lower bound values are calculated as follows:

$$\omega_{min} = \left[0,0, s_{min_y}, v_{min_y}\right]^T \quad (8)$$

$$\omega_{max} = \left[\infty, v_{max_x}, s_{max_y}, v_{max_y}\right]^T \quad (9)$$

$$u_{min} = \left[a_{min_x}, a_{min_y}\right]^T \quad (10)$$

$$u_{max} = \left[a_{max_x}, a_{max_y}\right]^T \quad (11)$$

$$\theta = \arctan\left(v_y / v_x\right) \quad (12)$$

$$v_y \in \left[v_x \tan(\theta_{min}), v_x \tan(\theta_{max})\right] \quad (13)$$

In addition, in this problem we incorporate a set of constraints that are the provisions of the road traffic law. These provisions are considered as logical clauses and will perform the conversion into a set of linear inequalities with integer variables.

The transformation is done as follows: in the literals clause is represented as an indivisible statement with a linear expression on the state variables. The literals use logical operations like conjunction $(\wedge)$, disjunction $(\vee)$, exclusive - OR $(\oplus)$, implication $(\rightarrow)$, equivalence $(\leftrightarrow)$, negation $(\bar{A})$ to represent. At the same time, these literals will be represented as a binary variable $\boldsymbol{\delta}(.)$ That only accepts 0 or 1 values, if a proposition is true then $\boldsymbol{\delta}(.)$, otherwise it is false then $\boldsymbol{\delta}(.) = 0$.

For example, to perform the constraint laws regulating about speed (Rule 1), travel speed (Rule 7) along with the structural constraints of the vehicle as follows: To ensure safety when participating in traffic then at position $\boldsymbol{x} \in [200m, 1000m]$ from the starting position, will limit the vehicle movement speed to $\boldsymbol{v} \leq 50km/h$ (Figure 3), there will be 3 literals defined $P_1 = [x(k) \geq 200]$, $P_2 = [x(k) \leq 1000]$ and $P_3 = [v_x(k) \leq 50]$, and the logical clauses expressed at any point in the segment are $\forall k \geq 0, (P_1 \wedge P_2) \Longrightarrow P_3$.

Or we represent it after assigning a binary variable $\boldsymbol{\delta}_i(k) = 1 \Rightarrow P_i$ with $i \in \{1,2,3\}$, $\forall k \geq 0$ as follows:

$$\delta_1(k) = 1 \Rightarrow x(k) \geq 200 \quad (14)$$

$$\delta_2(k) = 1 \Rightarrow x(k) \leq 1000 \quad (15)$$

$$\delta_3(k) = 1 \Rightarrow v_x(k) \leq 50 \quad (16)$$

$$-\delta_1(k) + \delta_3(k) \leq 0 \quad (17)$$

$$-\delta_2(k) + \delta_3(k) \leq 0 \quad (18)$$

$$\delta_1(k) + \delta_2(k) - \delta_3(k) \leq 1 \quad (19)$$

By the Big-M method [10], we can convert inequalities (14), (15) and (16) into the following:

$$x(k) \geq 200 - M\left(1 - \delta_1(k)\right) \quad (20)$$

$$x(k) \leq 1000 - M\left(1 - \delta_2(k)\right) \quad (21)$$

$$v_x(k) \geq 50 - M\left(1 - \delta_3(k)\right) \quad (22)$$

where M is a positive constant with great value. In this problem, we choose $M = 10^9$

Let $\boldsymbol{K} = T/\tau$ be the number of time steps in the prediction horizon with T be the time interval, $\boldsymbol{\omega} = \{\boldsymbol{\omega}(0), \ldots, \boldsymbol{\omega}(K)\}$ and $\boldsymbol{u} = \{\boldsymbol{u}(0), \ldots, \boldsymbol{u}(K)\}$ are state vectors, control trajectory vectors

in the given time horizon, $\boldsymbol{\omega}_r$ is the reference trajectory of the vehicle, which may depend on time, status, or traffic law.

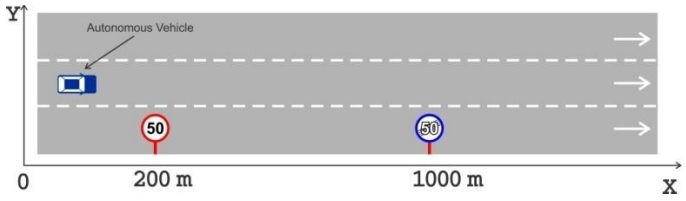


Figure 3. Speed limit ramp

To determine the objective function for this solution, we will introduce a new value vector variable $\boldsymbol{\delta}(k) = \{0,1\}^m$ and $\boldsymbol{\delta}_r$, where $\boldsymbol{\delta}(k)$ is a collection of all the binary variable results from the rebuilding the provisions of law on road traffic into linear inequalities and $\boldsymbol{\delta}_r$ is the reference trajectory for the binary variables where we can make options on some binary states.

Thus, at time $t = 0$, the optimal problem of the model predictive control with the objective function can be written as follows:

$$\min_{u,\delta} J(\omega, u, \delta) = \sum_{k=0}^{K}\left(\|\omega(k) - \omega_r(k)\|_Q^2 + \|\delta(k) - \delta_r(k)\|_R^2 + \|u(k)\|_S^2 + \|\Delta u(k)\|_W^2\right) \quad (23)$$

Conditions satisfied

$$\omega(k+1) = \begin{bmatrix} A^d & 0 \\ 0 & A^d \end{bmatrix}\omega(k) + \begin{bmatrix} B^d & 0 \\ 0 & B^d \end{bmatrix}u(k) \quad (24)$$

$$\text{with } k = [0, \dots, K-1]$$

$$\omega(k) \in [\omega_{min}, \omega_{max}] \text{ with } k = [0, \dots, K] \quad (25)$$

$$\omega_{min} = \left[0,0, s_{min_y}, v_{min_y}\right]^T \quad (26)$$

$$\omega_{max} = \left[\infty, v_{max_x}, s_{max_y}, v_{max_y}\right]^T \quad (27)$$

$$u(k) \in [u_{min}, u_{max}] \text{ with } k = [0, \dots, K] \quad (28)$$

$$u_{min} = \left[a_{min_x}, a_{min_y}\right]^T \quad (29)$$

$$u_{max} = \left[a_{max_x}, a_{max_y}\right]^T \quad (30)$$

$$\theta \in [\theta_{min}, \theta_{max}] \text{ with } \theta = \arctan(v_y/v_x) \quad (31)$$

$$v_y(k) \in [v_x(k)\tan(\theta_{min}), v_x(k)\tan(\theta_{max})] \quad (32)$$

$$\text{with } k = [0, \dots, K]$$

$$\Delta u(k) = u(k) - u(k-1) \text{ with } k = [0, \dots, K] \quad (33)$$

$$C\begin{bmatrix} \boldsymbol{\omega} \\ \boldsymbol{\omega}_r \\ \boldsymbol{\delta} \end{bmatrix} \le D \quad (34)$$

where the matrices C, D, Q, R, S and W are positive weight matrices of proper dimension, the final constraint (34) is the set of all provisions of road traffic law into linear inequality is represented in matrices.

The objective function (23) will be quadratic if the values of ω_r and δ_r are independent or linearly dependent on other variables. Therefore, solving this optimization problem will result in the global optimal trajectory being used so that it will then move to the controller at a lower level in tracking this trajectory.

Thus, given this optimal problem with the correlation and integration issues between vehicle dynamics and the provisions of the road traffic law, this solution can be used to effectively handle during the motion planning process with different situations during travelling. In order to achieve the optimal motion trajectory, it is necessary to process the cost function to a minimum and at the same time constraints including vehicle dynamics constraints and rules of participating in traffic after converting into logical constraints, it should be separated.

From the idea raised for this problem and in order to evaluate the effectiveness of the solution for the specific operating environment of the vehicle, the cost function is designed to optimize operational control with the initial condition that the vehicle velocity is constant and the horizontal deviation will change over time during the travel distance. Therefore, the expression $\|\delta(k) - \delta_r(k)\|_R^2$ in the optimal problem (23) will not be considered, so the cost function built in this study is given as follows:

$$J = \sum_{k=0}^{K}\left(q_1(v_x(k) - v_r)^2 + q_2(y(k) - y_r(k))^2 + q_3\left(v_y(k)\right)^2 + q_4(a_x(k))^2 + q_5\left(a_y(k)\right)^2 + r_1(a_x(k) - a_x(k-1))^2 + r_2\left(a_y(k) - a_y(k-1)\right)^2\right) \quad (35)$$

3. Experimental assessment

In order to ensure the reliability, objectivity and effectiveness of the solution, we have conducted simulations with different scenarios such as assessments on the roads with speed limit signs, determining the motion trajectory of the vehicle when there is an obstacle or a motion trajectory of the vehicle when overtaking in the same direction.

In the process of checking and evaluating the given solution, we conduct empirical simulation of processes in the Matlap environment. In which the calculation (using the SI measurement system) with the sampling period $\tau = 0.2s$ and the predictive horizon $T = 10s$. During the simulation, vehicles with recorded trajectory will be considered vehicles ahead and the autonomous vehicle will be behind these vehicles. The parameters used in the experimental process are given as follows:

$$\boldsymbol{\omega}_{min} = [0,0,-4,-2]^T, \boldsymbol{\omega}_{max} = [1500,80,10,5]^T, \boldsymbol{u}_{min} = [-5,-1]^T, \boldsymbol{u}_{max} = [5,\ 1]^T, \boldsymbol{\theta}_{min} = -0.5, \boldsymbol{\theta}_{max} = 0.5, q_1 = 1, q_2 = 3, q_3 = 5, q_4 = 2, q_5 = 4, r_1 = 6, r_2 = 16.$$

The location of the experimental vehicle is in the right lane with the original position at the coordinates (0, 0), the original and constant speed during the experiment of the vehicle is $v_x(0) = v_r = 60\text{km/h}$. The trajectory switches to the I/O controller with a simulation time of $0.05s$.

Scenario 1: In this scenario, the autonomous vehicle will move into a road with limited speed (Figure 4). At the beginning, the traveling speed of autonomous vehicles is stable $v_r = 60\text{km/h}$. When entering the prescribed speed range, with the input values and constraints for the control system will be able to make the vehicle's movement speed change. The control system will perform the braking operation to reduce vehicle speed to a safe limit with restrictions on road traffic rules. With such a driver assistance process, it shows the advantages of controlling the longitudinal force, horizontal force and acceleration the vehicle, as well as implementing the vehicle's movement plan to be complete when coming to the road systematic motion velocity regulation.

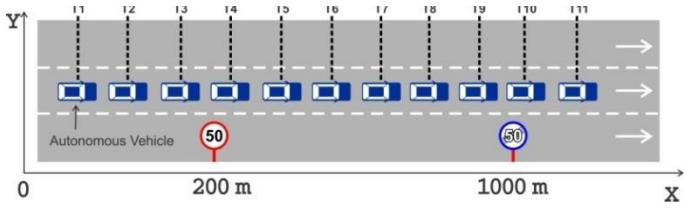


Figure 4. The vehicle moves over the speed limit segment

During the simulation process, with the parameters shown, the construction of constraints according to Rule 7 is carried out as follows:

When the vehicle moves into position T4 $\in [200m, 1000m]$, it will limit the movement speed of the vehicle is $\boldsymbol{v} \leq 50km/h$, there will be 3 literals defined $P_1 = [x(k) \geq 200]$, $P_2 = [x(k) \leq 1000]$ and $P_3 = [v_x(k) \leq 50]$, and the logical propositional represented at all times in the segment is $\forall k \geq 0, (P_1 \wedge P_2) \Longrightarrow P_3$. The constraints set is constructed to include:

$$\mathrm{x(k)} \geq 200 - \mathrm{M}\big(1 - \delta_1(\mathrm{k})\big)$$

$$\mathrm{x(k)} \leq 1000 - \mathrm{M}\big(1 - \delta_2(\mathrm{k})\big)$$

$$\mathrm{v_x(k)} \geq 50 - \mathrm{M}\big(1 - \delta_3(\mathrm{k})\big)$$

$$-\delta_1(\mathrm{k}) + \delta_3(\mathrm{k}) \leq 0$$

$$-\delta_2(\mathrm{k}) + \delta_3(\mathrm{k}) \leq 0$$

$$\delta_1(\mathrm{k}) + \delta_2(\mathrm{k}) - \delta_3(\mathrm{k}) \leq 1$$

where M is a positive constant with great value. In this simulation, we choose $M = 10^9$

The simulation results show that the movement speed of the vehicle from position T4 to position T10 according to the vehicle's planning of movement decreases below the speed of $\boldsymbol{v_r} \leq 40km/h$, so the motion controller has been effectively applied.

Scenario 2: In this scenario, the autonomous vehicle will move in the same lane with other vehicles and the autonomous vehicle will perform the right overtaking for the vehicle ahead (Figure 5). At the beginning, the autonomous vehicle and other vehicles are in one lane with the distance from autonomous vehicle to the vehicles ahead is 30m, the initial speed of the vehicles ahead is 30km/h and of autonomous vehicles is 60km/h.

At the time of T3, the vehicle in front of the vehicle moved at a constant speed, in this situation, the vehicle would either self-drive or reduce the speed of $\boldsymbol{v_r} \leq 30km/h$ or perform an overtaking operation or an accident will occur. During the simulation, we choose to overtake and to ensure that overtaking does not violate road traffic rules, autonomous vehicle must perform overtaking in accordance with Rule 2.

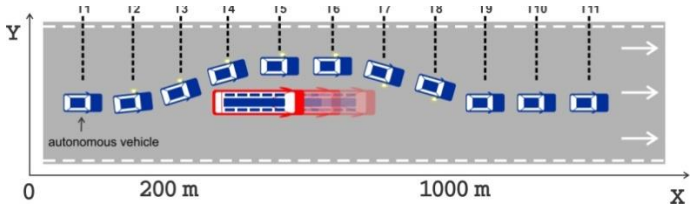


Figure 5. Autonomous vehicle passing to the right of the vehicle ahead

To set the control input values, we will describe the vehicles in front surrounded by rectangles. Each different object will have rectangles of different sizes and positions respectively. In this simulation, the position of the vehicle in front is $S_{obs} = (400,1)$ and the rectangle surrounding the vehicle in front is $L_{obs} = 12.5\text{m}$ and the width of $W_{obs} = 2.9\text{m}$. The rectangle around the vehicle in front of O_{obs} is defined in the coordinate system as $[x_o(\mathrm{k}) - L_{obs}, x_o(\mathrm{k}) + L_{obs}]$ x $[y_o(\mathrm{k}) - W_{obs}, y_o(\mathrm{k}) + W_{obs}]$ with $\forall k \geq 0$.

Thus, the constraint set is constructed to avoid collisions by overtaking on the right side as follows:

$$\delta_1(\mathrm{k}) = 1 \Rightarrow \mathrm{x(k)} \leq \mathrm{x_o(k)} - \mathrm{L_{obs}}$$

$$\delta_2(\mathrm{k}) = 1 \Rightarrow \mathrm{x(k)} \geq \mathrm{x_o(k)} + \mathrm{L_{obs}}$$

$$\delta_3(\mathrm{k}) = 1 \Rightarrow \mathrm{x(k)} \leq \mathrm{y_o(k)} - \mathrm{W_{obs}}$$

$$\delta_4(\mathrm{k}) = 1 \Rightarrow \mathrm{x(k)} \geq \mathrm{y_o(k)} + \mathrm{W_{obs}}$$

$$\delta_1(\mathrm{k}) + \delta_2(\mathrm{k}) + \delta_3(\mathrm{k}) + \delta_4(\mathrm{k}) = 1$$

or represented by the following inequalities:

$$x(\mathrm{k}) \leq (\mathrm{x_o(k)} - \mathrm{L_{obs}}) - \mathrm{M}(1 - \delta_1(\mathrm{k}))$$

$$\mathrm{x(k)} \geq (\mathrm{x_o(k)} + \mathrm{L_{obs}}) - \mathrm{M}(1 - \delta_2(\mathrm{k}))$$

$$\mathrm{x(k)} \leq (\mathrm{y_o(k)} - \mathrm{W_{obs}}) - \mathrm{M}(1 - \delta_3(\mathrm{k}))$$

$$\mathrm{x(k)} \geq (\mathrm{y_o(k)} + \mathrm{W_{obs}}) - \mathrm{M}(1 - \delta_4(\mathrm{k}))$$

$$\delta_1(\mathrm{k}) + \delta_2(\mathrm{k}) + \delta_3(\mathrm{k}) + \delta_4(\mathrm{k}) = 1$$

where M is a positive constant with great value. In this simulation, we choose $M = 10^9$

The simulation results show that the moving speed of the vehicle from position T1 to position T11 according to the movement plan of the vehicle is almost unchanged $\boldsymbol{v_r} \approx 60km/h$, but the steering angle of the vehicle begins to change from position T2 to position T5 is $\boldsymbol{\theta}(k) \approx -0.3\,rad$, from position T6 to position T8, it decreases to $\boldsymbol{\theta}(k) \approx 0.3\,rad$ and from position T9 the steering angle $\boldsymbol{\theta}(k) = 0\,rad$. The change of steering angle can be explained as follows, from position T2 the vehicle begins to move the steering angle to the right from the original movement direction to avoid collision and proceed to overtake to the right side

of the object. After overtaking the object, at the position of T6, the autonomous vehicle will move back to the left side to return e original direction of movement. From position T9, the autonomous vehicle has passed the object and adjusted the angle back to the original to return to the original motion trajectories.

Scenario 3: In this scenario, the autonomous vehicle will move through an intersection, no traffic lights, no priority roads and no priority vehicles (Figure 6). As illustrated in Figure 6, there will be the following situations: i) the autonomous vehicle crosses the intersection before other vehicles, ii) the autonomous vehicle overtakes the intersection after vehicle 01 and before vehicle 03, iii) the autonomous vehicle overtakes the intersection after vehicle 03 and before vehicle 01, iv) the autonomous vehicle crosses the intersection after all the vehicles. In these situations, because the direction of vehicle 02 does not affect the movement of the autonomous vehicle, the presence of vehicle 2 is not considered in these situations.

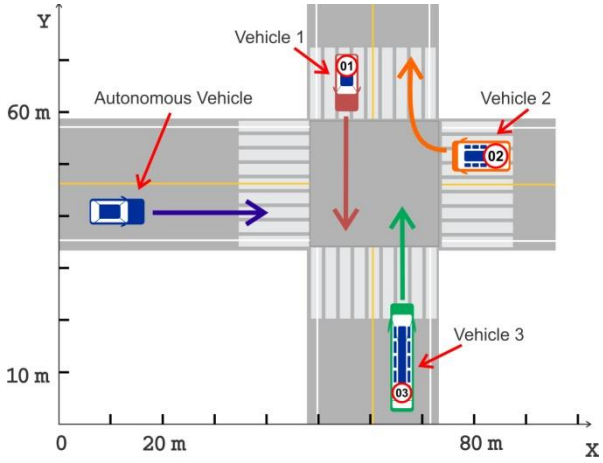


Figure 6. The vehicle moves into the intersection

To perform this simulation, we need to understand some rules when traveling through the intersection according to the provisions of Vietnam's road traffic law as follows: Is there a vehicle to the intersection?, is there a priority vehicle or not?, is there a priority path?

Vehicles are considered to enter the intersection when the front wheel has crossed the white line of the zebra crossing. In any circumstance, a vehicle that enters the intersection is given priority to go first.

For priority vehicles, the order of vehicles prioritized includes fire engines, military vehicles, police vehicles, ambulances, and other priority vehicles such as dyke protection vehicles, convoy of vehicles with police leading the way, vehicles on missions to overcome natural disasters, epidemics or vehicles on emergency duty as prescribed by law. Priority vehicles will be moved in front of other vehicles, meaning that the non-priority vehicles must give way to the priority vehicles in traffic.

For priority roads, vehicles located on non-priority roads must give way to vehicles on priority roads and should be based on the "intersection with priority roads" signs.

If there is no vehicle in the direction, the following shall be prescribed: at the intersection, priority shall be given to the vehicle that there is no vehicle to the right hand direction, at the roundabout, the priority shall belong to the vehicle that there is no vehicle to the left hand direction. Finally, for the priority turning direction, it is regulated that the vehicle that turns right goes first, then the vehicle goes straight and finally the vehicle turns left.

In this simulation scenario, the simulated vehicles are running at a speed of 50km/h, the position of the vehicles considered in the coordinates respectively is the autonomous vehicle at position (10,45), vehicle 01 at position (45.65), vehicle 02 at position (80.55) and vehicle 03 at position (65.10), the center of the intersection is at (60.55).

The problem of this simulation is performed as follows: Let the file $[L_{av}, H_{av}]$ be the part of the autonomous vehicle and $[L_{obj}, H_{obj}]$ is the path of other vehicles in the intersection area. To avoid collisions between vehicles, only one vehicle is at any time in the intersection. Therefore, if $x \in [L_{av}, H_{av}]$, $x_{obj} \notin [L_{obj}, H_{obj}]$ with $\forall t \in (t_{\min(obj)}, t_{\max(obj)})$, ie if the autonomous vehicle is already in the intersection, other vehicles cannot move into the intersection to avoid collisions at the intersection, where the values of t_{min} and t_{max} are symbols of the starting time and the last exit time of the vehicle while in the intersection .

When planning movement for an autonomous vehicle, we need to monitor whether other vehicles have entered the intersection area or not. First we will calculate the values of $t_{\min(obj)} = \mathcal{T}(L_{obj} - x_{obj}, v_{r(obj)}, a^{max}_{r(obj)})$ with $a^{max}_{r(obj)} > 0$ is assumed to be continuous increase and $t_{\max(obj)} = \mathcal{T}(H_{obj} - x_{obj}, v_{r(obj)}, a^{min}_{r(obj)})$ with $a^{min}_{r(obj)} < 0$ is assumed to be a continuous decrease of other vehicles, where $\mathcal{T}(x, v, a)$ is a time function to travel over a distance of path x, with the current velocity v and acceleration of vehicle unchanged a. In case the vehicle cannot travel the entire distance x in a finite time, this time value will be determined as $+\infty$, the values $a^{max}_{r(obj)}$ and $a^{min}_{r(obj)}$ are introduced to replace the constant acceleration value in order to determine the error limit during the evaluation of the motion of other vehicles.

The prediction horizon $k_{\min(obj)}$ and $k_{\max(obj)}$ of other means are calculated as follows:

$$k_{\min(obj)} = \left[{}^{t_{\min(obj)}}/_{\tau} \right]$$

$$k_{\max(obj)} = \left[{}^{t_{\max(obj)}}/_{\tau} \right]$$

where τ is the number of time steps in the prediction horizon.

Thus, with $\mathcal{K}$ being the prediction horizon of the autonomous vehicle, the constraint set is built to avoid collisions when the autonomous vehicle moves through intersections written with the following cases:

1. If $k_{\min(obj)} \leq k_{\max(obj)} \leq \mathcal{K}$, it means the autonomous vehicle will have to cross the intersection before another vehicle moves into the intersection area, or the autonomous vehicle must stay outside the intersection until another vehicle leaves the intersection area, then the constraint set will be:

$$\delta_1(k) = 0 \Rightarrow x^{k_{\min(obj)}} \geq H$$

$$\delta_1(k) = 1 \Rightarrow x^{k_{\max(obj)}} \leq L$$

2. If $k_{\min(obj)} \leq \mathcal{K} \leq k_{\max(obj)}$, it means the autonomous vehicle will have to go through the intersection before another vehicle moves into the intersection area, or the autonomous vehicle must be outside the intersection to the end of the prediction horizon, , then the constraint set will be

$$\delta_1(k) = 0 \Rightarrow x^{k_{\min(obj)}} \geq H$$

$$\delta_1(k) = 1 \Rightarrow x^{\mathcal{K}} \leq L$$

3. If $\mathcal{K} \leq k_{\min(obj)} \leq k_{\max(obj)}$, it means the autonomous vehicle will have to stay outside the intersection until other vehicles have moved through the intersection, no constraints are considered at this time.

The simulation results show that the speed of the autonomous vehicle decreases when preparing to enter the intersection area and then returns to the original speed, while moving through the intersection after all other vehicles. For a comparative basis, we tried to change the position of the autonomous vehicle at the new positions (45.45). Observing the process we realized that the autonomous vehicle moved through the intersection behind vehicle 1 and 2, but in front of vehicle 3 and the speed of vehicle 3 decreases when entering the intersection to make way for the autonomous vehicles to pass.

4. Conclusions

This paper proposes a solution to create the optimal trajectory with a model predictive control approach and the constraint set built from road traffic law. This approach is suitable for complex environmental conditions because these constraints can arise from various aspects of motion planning that must comply with traffic rules.

The motion trajectory of the vehicle is created from this solution aiming to improve performance and enhance the ability to avoid obstacles but still ensure optimal global trajectory. Simulation results with situations such as avoiding obstacles, passing other vehicles, moving through intersections ... have shown that this solution achieves the set goal of improving calculation efficiency and handling of uncertainty in the perceptual data of the environment. However, with the use of a second-order linear vehicle model to create motion trajectories, this solution is only advantageous when applied in the case of the autonomous vehicle moving on a straight path. For large curves, the current model built may not be accurate.

Evaluating of safety conditions, the proposed solution has been effective. This motion planning method has created an optimal trajectory that allows an autonomous vehicle to drive along the road and avoid obstacles safely. To improve the efficiency of calculations and consider the uncertainty of perceptual data and positioning on the general trajectory, the performance of this proposed solution depends on the probabilistic model of the system to create adaptive fields. Therefore, probabilistic analysis and the representation of a driving situation need to be more integrated into actual vehicle application.

In the future, in order to increase the reliability of this solution, the settings that have been experimented by simulation will be transferred to the real environment with experimental vehicles equipped with sensors, and when experimenting on reality, a number of factors to analyze the stability of the system will be added so that the behavior of traffic participants is more accurately forecasted. Extensive implementation of this driver-assisted solution for semi-autonomous or fully autonomous vehicles in vehicle control systems will be able to minimize the amount of damage and create a safe movement plan for the future.

References

[1] Bevan, Gollee and O'Reilly, Trajectory generation for road vehicle obstacle avoidance using convex optimization, *Proc IMechE Part D: J Automobile Engineering*, pp.455–473, 2010. DOI: 10.1243/09544070JAUTO1204.

[2] Jo K, Kim J, Kim D et al, Development of autonomous car – Part I: distributed system architecture and development process, *IEEE Trans Ind Electron*, pp.7131–7140, 2014. DOI: 10.1109/TIE.2014.2321342

[3] Zhang D, Li K and Wang J, Radar-based target identification and tracking on a curved road, *Proc IMechE Part D: J Automobile Engineering*, pp.39–47, 2012. DOI: 10.1177/0954407011414462

[4] A. Houenou, P. Bonnifait, V. Cherfaoui et al, "Vehicle trajectory prediction based on motion model and maneuver recognition." in *Proc of the IEEE Conference on Intelligent Robots and Systems*, pp. 4363 – 4369, 2013. DOI: 10.1109/IROS.2013.6696982

[5] Fraichard and Howard, Iterative motion planning and safety issue, *In: Eskandarian A Handbook of intelligent vehicles. London: Springer*, pp. 1433–1458, 2012. DOI: 10.1007/978-0-85729-085-4_55

[6] M. Werling, S. Kammel, J. Ziegler et al, "Optimal trajectories for time-critical street scenarios using discretized terminal manifolds." *The International Journal of Robotics Research*, vol. 31, no. 3, pp. 346–359. 2012. DOI: 10.1177/0278364911423042

[7] S. J. Anderson, S. C. Peters, T. E. Pilutti et al, "An optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios." *Transactions on International Journal of Vehicle Autonomous Systems*, vol. 8, no. 2/3/4, pp. 190 – 216, 2010. DOI: 10.1504/IJVAS.2010.035796

[8] M. Althoff and J. M. Dolan, "Online verification of automated road vehicles using reachability analysis", *IEEE Trans. on Robotics*, vol. 30, no. 4, pp. 903 – 918, 2014. DOI: 10.1109/TRO.2014.2312453

[9] M. Althoff, D. Heß, F. Gambert, "Road occupancy prediction of traffic participants." in *Proc. of IEEE Conference on Intelligent Transportation Systems*, pp. 99–105, 2013. DOI: 10.1109/ITSC.2013.6728217

[10] Richard W. Cottle, Mukund N. Thapa, *Linear and Nonlinear Optimization (2nd ed.)*, International Series in Operations Research & 15Management Science, 2017, DOI: 10.1007/978-1-4939-7055-1.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 460-468 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Bayes Classification and Entropy Discretization of Large Datasets using Multi-Resolution Data Aggregation

Safaa Alwajidi[*,1,2], Li Yang[2]

[1]*Department of Computer Science, Western Michigan University (WMU), Kalamazoo, MI 49008, USA*

[2]*Department of Mathematics and Computer Science, University of North Carolina at Pembroke (UNCP), Pembroke, NC 28372, USA*

ARTICLE INFO

Article history:
Received: 21 May, 2020
Accepted: 10 July, 2020
Online: 24 September, 2020

Keywords:
Big data analysis
Data reduction
Data aggregation
Multi-resolution
Entropy discretization

ABSTRACT

Big data analysis has important applications in many areas such as sensor networks and connected healthcare. High volume and velocity of big data bring many challenges to data analysis. One possible solution is to summarize the data and provides a manageable data structure to hold a scalable summarization of data for efficient and effective analysis. This research extends our previous work on developing an effective technique to create, organize, access, and maintain summarization of big data and develops algorithms for Bayes classification and entropy discretization of large data sets using the multi-resolution data summarization structure. Bayes classification and data discretization play essential roles in many learning algorithms such as decision tree and nearest neighbor search. The proposed method can handle streaming data efficiently and, for entropy discretization, provide sufficient information to find the optimal split variable and the optimal split value.

1 Introduction

Big data analysis offers new discovery and values in various application areas such as industrial sensor networks [1], finance [2], government information systems [3], and connected health [4]. Big data is redundant, inconsistent, and noisy. Machine learning algorithms face extreme challenges to analyze big data precisely and quickly.

Volume and velocity are the primary characteristics of big data. The massive size of data being accumulated consumes storage and leads to management shortcomings [5]. According to a IDC report [6], the volume of data may reach 40 Zettabytes (10^{21} *bytes*) by 2020. In addition, big data often come in the format of complex data streams. For example, in Wide-area Sensor Network (WSN), autonomous sensors generate a huge amount of data from the environment and transmit the data through the network to the backend servers for further processing and analysis.

Big data analysis is challenging to computing power and resources [7]. Analysis algorithms need to scan the data at least once, leading to a time complexity of $O(N)$, where N is the number of data records. Yet many algorithms use iterative computations that require multiple passes of data access. For example, Support Vector Machine (SVM) is a well-known algorithm that requires multiple scans of the dataset with time complexity $O(N^3)$. For processing of big data, accessing data point by point is a prohibitive process.

One idea to solve the problem is to summarize the data and use the summarized data for data processing. The summarization aggregates data points into a hosting structure that can be used by most mining and learning algorithms for efficient and effective analysis. The hosting structure should: (i) require one scan of the original data for its construction; (ii) hold representative statistical measures of the original data that satisfy requirements of most of the mining and learning algorithms; (iii) update the statistical measures incrementally to deal with streaming data; and (iv) organize the statistical measures at multiple resolutions. The multiresolution view of data gives us the flexibility to choose the appropriate level of summarization that fits the available resources.

In our previous work [8], we have developed an effective technique to create, organize, access, and maintain a tree summarization structure for efficient processing of large data sets. The developed tree structure is incrementally updated and provides the necessary information for many data mining and learning algorithms. This paper is a continuation of the previous work by developing data mining and learning algorithms. Namely, this research develops algorithms for Bayes classification and entropy discretization of large data sets using the multi-resolution data summarization structure.

[*]Corresponding Author: Safaa Alwajidi, +19105216517, safaa.alwajidi@uncp.edu

Bayes classification and data discretization play essential roles in many learning algorithms such as decision tree and nearest neighbor search. The proposed method can handle streaming data efficiently and, for entropy discretization, provide sufficient information to find the optimal split variable and split value for continuous variables. Data discretization for continuous variable plays a central role in learning algorithms such as decision trees.

The rest of the paper is divided into five sections. Section 2 summarizes existing data reduction techniques. Section 3 introduces the multiresolution structure and discusses its implementation, updating, and performance. Bayes classification and entropy discretization are introduced in Section 3 and Section 4 respectively. The last section concludes the paper.

2 Techniques for Data Reduction

For efficient and effective management, transmission and processing of large data sets, many techniques have been proposed. One straightforward way to reduce data size is data compression. Data compression reduces storage, I/O operations [9, 10] and network bandwidth [11], and also helps to accelerate data transmission [12]. One problem with data compression is the overhead of compression and decompression [13]. While it preserves the complete dataset, the method does not reduce the complexity of data processing, which needs to access the original data anyway.

The size of relational data can be reduced in two ways: either to reduce the number of attributes or to reduce the number of data records. One technique to reduce the number of attributes is dimensionality reduction [14]. One example algorithm for dimensionality reduction is principal component analysis (PCA) [15], which is an expensive process. Most machine learning algorithms require multiple scans of the dataset with time complexities based on the number of data records, therefore, reducing the number of data records seems to be the way to go for big data analysis.

One example technique to reduce the number of data records is sampling. Sampling is a statistical subset technique in which a small number of data points are randomly chosen to produce an unbiased representation of the dataset. A recent study [16] shows that random sampling is the technique commonly used to gain insights in big data. However, it is difficult to quickly extract a representative random sample. A critical problem is how to handle the selection bias due to issues such as class imbalance [17, 18] and nonstationary values of data [19]. A systematic method of sampling that mediates the tensions between resource constraints, data characteristics, and the learning algorithms accuracy is needed [20]. Intricate methods to subset the big data, such as instance selection [21] and inverse sampling [22], are computationally expensive [23, 24] because of the inefficient multiple preprocessing steps. Furthermore, newly added data points change data statistical measures and require re-sampling. Moreover, sampling is not appropriate in some application areas. For example, epidemiologic research involves systematic biases due to great enthusiasm in the sampling of people in EHRs (Electronic Health Records) systems and there are often biases in the way information is obtained and recorded. A large sample size cannot overcome these problems [25].

One powerful way to reduce the number of data records is to aggregate the data records. Data aggregation is a reduction technique that considers all points in the dataset. It preserves enough information from the dataset while combining similar data records into one which summarizes the data records into representative statistical measures. Data aggregation is used in privacy preservation [26] and Wireless Sensor Networks (WSNs) to reduce the volume of data traffic [27]. However, few studies have been conducted to aggregate and summarize big data for data analysis.

One of the early methods to use aggregated data to support advanced analysis and decision making is On-line Analytical Mining (OLAM) [28]. OLAM integrates mining models on top of a multidimensional data cube and OLAP engine. It has been used in many applications to provide efficiency and scalability for data mining and knowledge discovery such as association rules [29]. The multidimensional data cube uses a conceptual hierarchy in each categorial dimension to aggregate data in a multi-resolution format. However, many big data sets are not in relational databases and contain continuous attributes with no concept hierarchy.

A tree index structure is one of the major techniques that can be used to aggregate data at multiple resolutions [30]. Regions in multi-resolution aggregation need to be grouped and linked for quick reference. Many variations of tree structures have been used to save data for fast access such as CF tree used in BIRCH [31], R* tree used in DBSCAN [32, 33], KD-tree [34], aR-tree [35], and R+-tree [36]. However, these trees do not meet all of our requirements such as hosting aggregated data, query independence, simple creation, and incremental updates.

3 Multi-Resolution Tree Structure for Data Aggregation

We are given a dataset of size $N \times d$, where N is the number of data points, d is the number of variables, and $N \gg 2^d$. These variables take continuous numeric values. Continuous data are common such as the data generated from sensors like temperature, air pressure, gyroscope, accelerometer, GPS, gas sensors, water sensors and vital data from patients. Aggregating the dataset into a multi-resolution hierarchical structure is not a straightforward process. Continuous data lacks concept hierarchy.

We propose a hierarchical multilevel grid summarization approach to aggregate the data. Similar method has been used in spatial data mining [37] and robotic mapping [38]. In this method, each dimension in the dataset is divided into a limited number of equal width and nonoverlapping regions (intervals). Each region stores the summarization information of the raw data points falling inside it. We labeled each region by giving it a number to distinguish between them. The multilevel structure aggregates data at multiple granularities, where multiple regions at a lower level are grouped to form one region at the next higher level. We assume each level is summarized into a half number of regions in the immediate beneath level. This method provides a concept hierarchy for continuous data.

The hierarchical multilevel grid can be assumed as a multidimensional data cube. However, regions in multi-resolution aggregation need to be linked for quick references. We propose a tree index structure to host the aggregated data in the multidimensional data cube.

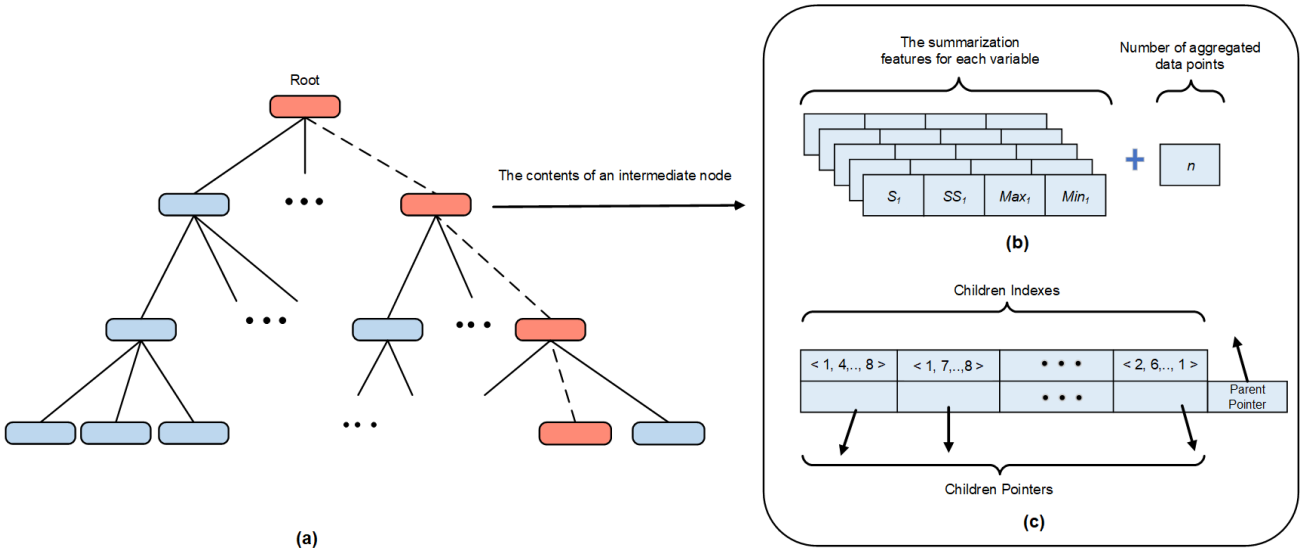


Figure 1: The multi-resolution structure (a) Updating the multi-resolution tree (b) The summarization features (c) The mapping table

The tree is an efficient structure in which summarization features at high levels can be computed from features at low levels without rescanning the original raw data. This is useful for incremental updating. Also, the tree provides quick reference links between its levels to support multiresolution aggregation. Searching the tree structure is faster than searching the massive raw dataset.

All data records are aggregated at each level of the tree. Aggregations are performed on all variables at the same time. This multi-resolution aggregation reduces the number of tree nodes and provides the trade-off between efficiency and accuracy. The user can choose the appropriate level of aggregation. Typically, navigation is a top-down process from the top to lower levels.

Each intermediate node holds a mapping table and a set of summarized statistical measures. The mapping table, Fig.1(c), contains a set of entries that hold node indexes along with their corresponding pointers and is used for linking a node with its parent and children nodes. The mapping tables helps to reduce the I/O operations. For instance, to access a desired child node from its parent, we start by searching the parent's mapping table to check if the index of the desired child is available.

3.1 Tree Node Index and Mapping between Nodes

A tree node is a hyperrectangular area that is specified by an interval in each dimension. Each interval has a unique index number that distinguishes it from other intervals. Consequently, each node has a set of index numbers corresponding to its intervals in all dimensions. We call this set a node index. The node index is used to identify the exact nodes in the tree to be called and uploaded to the main memory.

A mapping function, $f(\{v_1, v_2, ..., v_d\}) = \{name_l\}_{l=1}^{h}$ is used to generate the node index from incoming data point $\{v_1, v_2, ..., v_d\}$, where v_i is the value of variable i, and $name_l$ is the node name at level l. Each name is a tuple of the form $\langle l, I_1, I_2, \ldots, I_d \rangle$, where l is the tree level and I_i is the region number of variable i. The availability of the node names for incoming data point along with mapping tables, which consist of child nodes indexes, helps to assign the incoming data point to its corresponding nodes without searching the tree. Instead, it uploads only the required nodes for updating. See section III (c) for the detailed procedure.

3.2 Sufficient Statistical Measures

Multiple statistical measures can be collected on data points aggregated in a region. The set of statistical measures should be expressive enough to answer most data queries. Since each data mining algorithm needs different statistical measures, we need to find common statistical measures for all or most of the algorithms to ensure that the proposed structure can be used by these algorithms. Extracting sufficient statistical measures of the aggregated data is crucial to provide information to data analysis algorithms. In addition, efficient analysis of streaming data requires incrementally updating of these statistical features.

For each data node, we collect number of data points, sum of data values, sum of squared data values, as well as minimum and maximum values along each variable $(n, S_p, SS_p, Min_p, Max_p)$ in this node. We collected these measures inspired by BIRCH [31]. These five statistical measures are small in size and can be incrementally updated.

Basic statistics like mean, variance, and standard deviation can be incrementally and directly computed from these statistical features. Other features such as variables correlation and variable distribution can also be computed. We assume that the data is uniformly distributed inside each region of the variable.

These measures provide enough information to compute cluster means, sizes, and distances in different modes. In addition, Euclidean, Manhattan, L1 and L2 distances between cluster centers, and average inter-cluster distances can also be computed. This information is useful for clustering algorithms and similarity-based learning. As we will discuss in Sections IV and V on learning algorithms, other measures such as likelihood, conditional probability, and dynamic splitting can be calculated using the collected

statistical measures.

3.3 Implementation and Performance Analysis

The proposed multi-resolution structure is a tree like structure. Its implementation and updating is a top-down process, starting from the root node towards the leaf nodes. Algorithm 1 explains the procedure of implementing and updating the multi-resolution structure. Inputs to the algorithm are the desired number h of levels in the resulted tree structure, expected maximum and minimum values for each variable along with a matrix W of region width for each variable at each level in the tree.

The implementation requires a single scan of the training dataset. To insert a new data record, the algorithm uses the index keys generated by the mapping function and starts at the root node to find an entry that matches the first index key. If the entry is found, the algorithm updates the corresponding node statistical features. In case the entry is not found, the algorithm creates a new entry in the root mapping table to hold the index key. The algorithm then creates a new node, associate it with the newly created entry, and initialize the node statistical features. The algorithm repeats this process until it reaches the leaf node.

The worst performance of the proposed structure occurs when the tree has the maximum number of nodes. We assume that the number of regions for each variable at a level in the tree doubles in the next beneath level. In this case, each node has a maximum of 2^d child nodes. The total number of nodes in the tree with h levels and one root node is:

$$N_{Tree} = 1 + \left\lfloor \sum_{l=1}^{h} \frac{\prod_{p=1}^{d} \left(max_p - min_p\right)/w_{ph}}{2^{(h-l)(d-1)}} \right\rfloor \quad (1)$$

where $w_{ph} \in W$, is the regions width for variable p at level h. The values max_p and min_p are the maximum and minimum values of variable p respectively.

The proposed multi-resolution tree is an efficient structure since it avoids uploading and searching all tree nodes during insertion and updating. Insertion one data points calls only $(h + 1)$ nodes. Empirically for streaming data, the consecutive data points fall within the same range of values which means that they will be aggregated into the same nodes. Therefore, disk caching helps to reduce disk operations. Implementing the multi-resolution tree does not create empty nodes with no data.

To analysis the worst case to implement the multi-resolution structure, we assume the data of N points is scattered all over the regions and no region left empty. In this case, each node has the maximum number of children and thus the multi-resolution structure has the maximum number of nodes. The implementation of such case costs $O\left(Nd(R + (h-1)2^d)\right)$ where R is the number of entries in the root's map table, and $\left((h-1)2^d\right)$ is the cost to search the map tables of visited nodes. Since h is relatively small value, the cost become $O\left(Nd(R + 2^d)\right)$.

Fig. 1 (a) illustrates how Algorithm 1 works. To insert a data record, only the red nodes, one at each level of the tree, are visited and uploaded. The dotted links represent the algorithm traverse and progress from the root to a leaf node.

Algorithm 1: Implementing and Updating the Multi-Resolution Tree

```
Input  : N × d dataset, h number of levels, and W matrix of regions width
Output : T the Multiresolution structure with h levels
foreach O_i ∈ N do
    // The data point O_i is a tuple {v_1i, v_2i, ..., v_di}
    for l ← 1 to h do
        foreach Variable p ∈ d do
            Index[l, p] ← [v_pi − min_p]/W_pl ;
        end
    end
    if T.root is Null then
        Create the root node ;
    else
        CurrentNode ← root ;
        UPDATE(Index, CurrentNode, O_i);
    end
    for l ← 2 to h − 1 do
        UPDATE(Index, CurrentNode, O_i);
    end
end
Procedure UPDATE
  Input Index, CurrentNode, O_i(datarecord)
  Output UPDATED T
  if Index not in CurrentNode.MapTable then
      Create a new entry iEntry in CurrentNode.MapTable;
      iEntry ← Index;
      Create a new node e;
      Create a new entry cEntry in CurrentNode.children;
      cEntry ← e ;
      e.parent ← CurrentNode;
      CurrentNode ← e;
      Initialize CurrentNode statistical features and entries based ;
  else
      CurrentNode ← Child node corresponding to the Index entry
                    in the CurrentNode.MapTable;
      Update CurrentNode statistical features based on O_i values;
  end
  end procedure
```

4 Bayes Data Classification

Class probabilities are essential for Bayesian classifiers, hidden Markov model (HMM), and in defining entropy, gain ratio, and Gini index for decision trees. Naive Bayes is a classification model based on conditional probability. By including class attribute as a regular attribute in building the multi-resolution tree, we can calculate conditional class probabilities from the multi-resolution tree without accessing individual points in the training dataset.

The class attribute, also called the response variable, is a categorical attribute that has a finite number of labels (values). We can divide the class attribute into regions where each region is a unique class label. There is no aggregation applied on the class attribute and it maintains the same set of regions at all levels in the multi-resolution tree. Unlike continuously values variables, no numbers assigned for its regions. Those regions are indexed by their labels.

Given a $N \times d$ dataset with the class attribute $C = \{c_1, c_2, \ldots, c_m\}$, where m is the number of distinct values of class labels, the number of leaf nodes at level h is:

$$N_h = \left\lfloor \prod_{p=1}^{d-1} [\left(max_p - min_p\right)/w_{ph}] \times m \right\rfloor \quad (2)$$

All data points that are aggregated in a leaf node belong to a single class. Aggregating data points from different classes in one node leads to reduce the number of nodes in the multi-resolution tree and then reduce the storage needed for the tree. However, it consumes more memory and increases the number of I/O disk operations. For example, if parent node aggregates data points that belong to two classes A and B, there is a chance that only one child node has the data points of class A and the rest child nodes have data points belong to class B. Consequently, accessing child node of class A requires to upload all child nodes to memory to find that node of class A.

Computation of class probability from the proposed tree structure requires a single scan of the tree. The highest level in the tree maintains the same information of other levels but it is the most aggregated and compact level. The number of child nodes of the root is much less than the number of data points in the training dataset. Thus, the computation of class probabilities from the highest level in the multi-resolution tree consumes less memory and more efficient than working on the very large training datasets.

Given $E = \{e_1, e_2, ..., e_k\}$, a set of child nodes of root in the multi-resolution tree that aggregates dataset with the class attribute $C = \{c_1, c_2, ..., c_m\}$. The probability of a class c_j, $c_j \in C$ and $1 \leq j \leq m$ can be computed by :

$$P(c_j) = \frac{\sum_{i=1}^{k} [n_i \times \mathcal{F}\left(y_i = c_j\right)]}{\sum_{i=1}^{k} n_i} \quad (3)$$

n_i, y_i are the number of aggregated data points and class label of node e_i, $\mathcal{F}$ takes value 1 if $y_i = c_j$ and 0 otherwise.

Naive Bayes, Eq. (4), uses class probability for classification. It requires scanning all data points in the training dataset to classify one unknown point, under the assumption that the variables are independent. Suppose continuous variables are normally distributed, it requires to computes the likelihood for a continuous variable using Eq.(5). Then, finds the estimated class label $\hat{y}$ based on the Bayes theorem that computes class probability $P(C)$ for all classes and assigns the class label c_j with the highest posterior probability to unclassified point x.

$$\hat{y} = \underset{j \in \{1,...,m\}}{\operatorname{argmax}} P(c_j) \prod_{p=1}^{d} P(x_p|c_j) \quad (4)$$

$$P(x_p|c_j) = \frac{1}{\sigma_{pj}\sqrt{2\pi}} e^{\frac{-(x-\mu_{pj})^2}{2{\sigma_{pj}}^2}} \quad (5)$$

The likelihood requires a scan of all points in the training dataset to find the conditional mean, variance and standard deviation to classify a single unknown point. The proposed multi-resolution structure provides an efficient way to compute the conditional mean, variance, and standard deviation of continuous variable p. It requires to scan the nodes of the highest level in the tree only once to compute these conditional features for each class c_j separately, as shown in Table I.

Table 1: Conditional Features Computation for Variable p and class j

Feature		Comments
μ_{pj}	$\frac{S_{pj}}{n_j}$	Conditional mean
${\sigma_{pj}}^2$	$\frac{SS_{p_j} - \frac{\left(S_{p_j}\right)^2}{n_j}}{n_j}$	Conditional variance
σ_{pj}	$\sqrt{{\sigma_{pj}}^2}$	Conditional standard deviation

Consequently, any classification algorithms based on this computation of class probability can work efficiently without loss in accuracy to cope with the large scale data and streaming data. The multi-resolution Naive Bayes algorithm, which accepts the multi-resolution tree as input, has been implemented using Java JDK 13. We have tested the algorithm on eight artificially generated datasets, each set has two continuous variables and binary class label. The multiresolution structured implemented from the training set, 80% of the raw dataset, and the remaining 20% used as a testing set. We have compared the results of applying the multiresolution Naive Bayes on the root level of the multiresolution structure with the results of applying the traditional Naive Bayes algorithm on the raw datasets. Experimental results are given in figure 2. The results show that the multi-resolution Naive Bayes is 25% faster on average than the algorithm that works on raw data sets while keeping the same accuracy of classification.

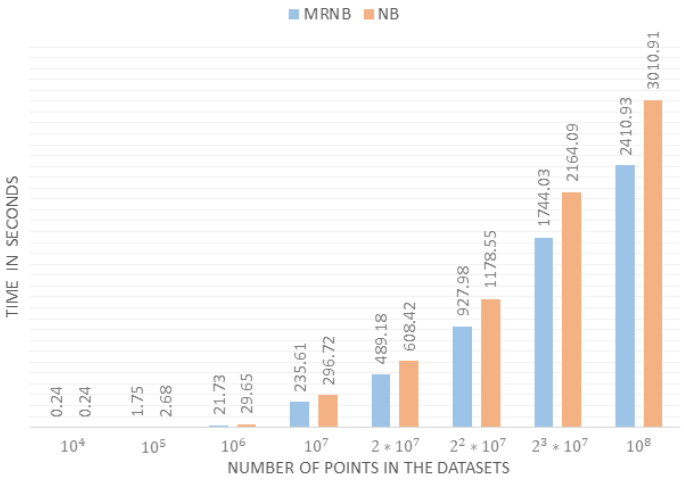


Figure 2: Comparison of the computation time between multiresolution Naive Bayes (MRNB) and raw data Naive Bayes (NB) algorithms for eight datasets

5 Entropy Discretization Overview and Challenges

We present another application to speed up learning algorithms that sort and search the dataset multiple times to produce an optimal solution. In this section, we propose a high speed and incremental technique for entropy discretization of continuous variables using

the multi-resolution structure. The high speed, representative, and incrementally updated structure helps to generate entropy discretization efficiently and effectively for big data and streaming data.

5.1 The Motivation

Several miming and learning logarithms iteratively split the dataset as part of their learning models. Examples of these algorithms are Bayes Belief Network (BBN) classification for continuous data [39], KD-tree used in nearest neighbor (NN) search [40, 41], and conventional decision tree [42] used for classification. The splitting procedure sorts the variable values and then seeks for the best value to split the variable based on a specific criterion. This procedure is repeated for each variable to find the best point and best variable to split the dataset.

The continuous variable has a vast number of values that make sorting and testing of its values during the splitting procedure a complicated task. Entropy discretization converts the continuous variable into a categorical variable by dividing it into a minimal set of non-equal width regions with minimal loss of information. Applying the splitting procedure iteratively on the small number of regions resulted from entropy discretization is more efficient than the iterative scan of the vast number of data points in the dataset. However, entropy discretization is a computationally expensive and memory consuming technique. It needs to search and test all the variable values to divide it into regions. Consequently, entropy discretization is a prohibitive task on big data as well as on streaming data and requires a fast incremental discretization to cope with updating when new data arrives.

5.2 Entropy Discretization Overview and Challenges

Entropy discretization divides the continuous variable based on entropy. Entropy is a measure of information. Higher entropy means less information loss. It is computed by using the frequency-probability distribution in Eq.(6).

$$H(X) = \sum_{x \in X} P(x) \log_2 P(x) \tag{6}$$

where x is a value in the continuous variable X.

Dividing the continuous variable into several regions affects entropy in a way that the entropy increases when the number of regions increases. To decide the optimal number of regions, there is a comprise between keeping information and keeping performance. Choosing a large number of regions seems reasonable to ensure high entropy and better representation of the distribution in such data. However, there is a danger of an over discretization where too many regions can be dispensable because it slows the performance. Hong [43] proposed a method to find the optimal number of regions with minimal loss of information. The relationship between the entropy and the number of regions is curve. If we draw a straight line from the origin of the graph to the maximum entropy and the maximum number of regions, the curve will be always above the line. Hong showed that when the curve is most apart from the line, we reach the optimal number of regions. This point is called the Knee point.

Ellis [39] uses Hong method to discretize continuous variables and feeds them to the Bayes Belief Network algorithm. His method starts with dividing the continuous variable into regions and sorting them. Each region has a distinct value of the continuous variable. The method then merges two adjacent regions with the smallest difference of frequency (number of aggregated data points). The process of merge continues until reaching the maximum height point on the curve (Knee point). The Knee point is reached when the change in the number of regions becomes greater than the change in the entropy of the regions. This relation is explained in Eq.(7) in the next section. However, when the number of distinct values is very big, this method of entropy discretization is computationally prohibitive. The time complexity of this method for d continuous variables is $O(d(m(log_2 m + m^2)$, where m is the number of distinct values. The number of distinct values is very big in some applications like sensor data, human activity recognition, and health care data, where m approaches the number of data points N.

5.3 Proposed Algorithm

We propose a fast entropy discretization technique by using the multi-resolution structure. The proposed technique finds the optimal number of non-equal width entropy regions based on the equal width regions in the multi-resolution structure. The multi-resolution structure has properties such as the hierarchical organization of aggregated data which makes the proposed technique more efficient than working on the massive number of data points in the training dataset to generate the regions of entropy discretization.

The multi-resolution structure discretizes each continuous variable into non-overlapping equal-width regions, each resides on a separate node of the same tree level. The proposed algorithm optimizes this discretization of the variable by merging its regions from two different nodes into one region repeatedly until reaching to Knee point.

Algorithm 2 explains the proposed entropy discretization technique for a dataset of d continuous variables. For each variable, the algorithm starts working on the parents of the leaf nodes (it can easily be adapted to work on any other level except root and leaves). For each parent, the algorithm merges the adjust regions in the sibling leaves. The new resulted regions are saved into a local file associate with the parent node. This pass is called the local mode. Then, in the global mode, the algorithm uses regions from all local files and merge the adjacent regions from nodes belong to different parents. The final result of the global mode is a set of optimal entropy regions that are saved into a global file for each variable, as shown in figure 4.

The region is represented by its boundary (cut point) and the number of its aggregated data points (frequency). This information is obtained from the statistical features of the node that hosts the region. The statistical features Max maximum value of the region is used as a cut point while n the number of aggregated data points inside the node is used as frequency. To deal with multiple regions, the algorithm uses a list to save the cut points of that regions combined with a corresponding list of their frequencies.

The merge procedure in algorithm 2, that is used in both local and global modes, is iterative. It sorts the regions based on their cut points and runs the following steps:

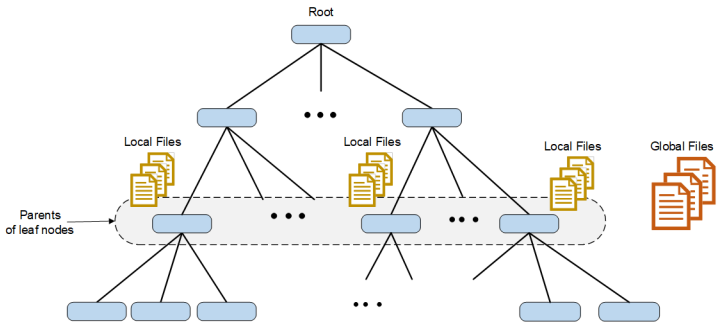


Figure 3: The multi-resolution structure with local and global files, one local file for each variable on every parent node and one global file for each variable

- Compute the entropy, *Entr*, for current K regions using Eq.(6). In this case, x is a region in variable X. The regions frequencies are used to compute the probability.

- Find two adjacent regions with a minimum frequency difference and merge them. The new cut point of the new region is the greatest cut point of the two merged regions while the new frequency for the new region is the sum of the frequencies of the two merged regions.

- Compute the entropy, *newEntr*, for resulted $K-1$ regions after the merge.

- Compute the rate of change in the number of regions to the change in their entropy using Eq.(7).

$$Rate = (newEntr \times K) - (Entr \times (K-1)) \tag{7}$$

The Knee point is proportional to this rate and is reached just before the decrease in this rate. If the rate is decreased, the algorithm stops the merge procedure. Otherwise, the algorithm updates the cut point and frequency lists by replacing the cut points and frequencies of the two merged regions by the cut point and frequency of the new region. The algorithm repeats the last three steps until the rate starts to decrease.

Although the local mode may lead to a little loss in the information due to missing a merge of regions from different parents, the local mode is still important to improve the performance in building and updating entropy discretization. Local mode reduces the sort to only nodes belong to the same parent because these nodes are neighbors to each other and there is a high possibility to merge them. The local mode helps to trigger the global mode only if it is needed, as summarized in Fig 5. After updating the multi-resolution structure due to the arrival of new data points, the local modes are triggered only for the parents of the updated leaves. The local modes for other parent nodes are ignored. If any of the triggered local modes achieves a merge, the global mode will be triggered. Otherwise, the algorithm terminates without global mode because the new data point has not effected the entropy discretization.

Algorithm 2: Build Entropy Discretization

Input : E the set of parent nodes of leaves in the multiresolution tree
Output : *GlobalFiles* set of files, each contains the entropy discretization for a single variable

```
// Main Program
foreach Variable p ∈ d do
    // Local Mode
    foreach e ∈ E do
        Children ← child nodes of e ;
        Cuts ← Max_p values from all Children;
        Freqs ← n values from all Children;
        LCuts, LFreqs ← MergeRegions(Cuts, Freqs);
        LocalFile ← LCuts and LFreqs ;
        AllCuts ← AllCuts ∪ LCuts;
        AllFreqs ← AllFreqs ∪ LFreqs;
    end
    // Global Mode
    GCuts, GFreqs ← MergeRegions(AllCuts, AllFreqs);
    GlobalFile ← GCuts and GFreqs ;
end
```

```
procedure MergeRegions(Cuts, Freqs)
Output : Updated Cuts and Freqs lists after iterative merge of regions
    Sort Cuts and Freqs based on values of Cuts ;
    Initializing the number of regions: K, baseK ← |Cuts|;
    Entr ← calculate entropy for K regions;
    oldRate ← 0 ;
    Flag ← TRUE ;
while (Flag) do
        R1, R2 ← two adjacent regions in Cuts with minimum Freqs difference;
        Merge R1 and R2 into one region ;
        newEntr ← calculate entropy after merge for (K − 1) regions ;
        newRate ← (newEntr × baseK) − (Entr × (K − 1));
    if newRate > oldRate then
        // Apply the merge of R1 and R2
            Update Cuts and Freqs;
            K ← K − 1 ;
            oldRate ← newRate ;
    else
        // Decline the merge of R1 and R2
            Flag ← False;
    end
end
return Cuts, Freqs;
end procedure
```

Finally, the algorithm outputs a set of global files. The global file contains the optimal number of entropy regions that represent the distribution of a single continuous variable. The set of global files can be used to feed the mining and learning algorithms for efficient and effective learning.

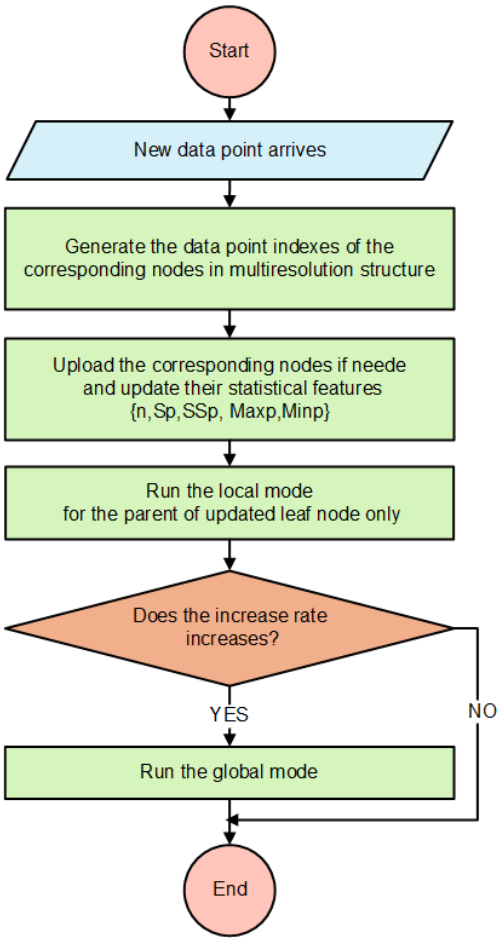


Figure 4: Updating process of multi-resolution structure and entropy discretization

6 Conclusions

In this paper, we have studied fast Bayes classification and entropy discretization of continuous variables using a multi-resolution data aggregation structure for mining and processing of large datasets. The proposed Bayes classification works by getting required conditional probabilities directly from aggregated statistical measures. Entropy discretization works in local and global modes to find the optimal number of non-equal width entropy regions with minimal loss of information based on the equal-width regions of the multi-resolution structure. The multiresolution structure makes the proposed technique more efficient to generate entropy discretization regions than working on the massive number of data points in the big dataset. The resulted entropy discretization regions are saved into files linked to leaf nodes and their parents for fast updates of these regions to cope with streaming data. These files are useful to find the best split point and best split variable for continuous variables that are essential for several learning algorithms that iteratively seek for the optimal variable and value to split the dataset. These algorithms include Bayes Belief Network (BBN) classification for continuous data, KD-tree data structure used in fast nearest neighbor search, and conventional decision trees used for classification.

References

[1] L.-M. Ang, K. P. Seng, "Big Sensor Data Applications in Urban Environments," Big Data Research, **4**, 1–12, 2016, doi:https://doi.org/10.1016/j.bdr.2015.12.003.

[2] B. Fang, P. Zhang, "Big Data in Finance," in S. Yu, S. Guo, editors, Big Data Concepts, Theories, and Applications, 391–412, Springer International Publishing, Cham, 2016, doi:10.1007/978-3-319-27763-9_11.

[3] G.-H. Kim, S. Trimi, J.-H. Chung, "Big-data applications in the government sector," 2014.

[4] L. A. Tawalbeh, R. Mehmood, E. Benkhlifa, H. Song, "Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications," IEEE Access, **4**, 6171–6180, 2016, doi:10.1109/ACCESS.2016.2613278.

[5] H. Zhang, G. Chen, B. C. Ooi, K.-L. Tan, M. Zhang, "In-Memory Big Data Management and Processing: A Survey," IEEE Transactions on Knowledge and Data Engineering, **27**(7), 1920–1948, 2015, doi:10.1109/TKDE.2015.2427795.

[6] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, R. Buyya, "The anatomy of big data computing," Software: Practice and Experience, **46**(1), 79–105, 2016, doi:10.1002/spe.2374.

[7] M. H. ur Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah, S. U. Khan, "Big Data Reduction Methods: A Survey," Data Science and Engineering, **1**(4), 265–284, 2016, doi:10.1007/s41019-016-0022-0.

[8] S. Alwajidi, L. Yang, "Multi-Resolution Hierarchical Structure for Efficient Data Aggregation and Mining of Big Data," in 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 153–159, 2019, doi:10.1109/ICACTM.2019.8776717.

[9] B. H. Brinkmann, M. R. Bower, K. A. Stengel, G. A. Worrell, M. Stead, "Large-scale electrophysiology: Acquisition, compression, encryption, and storage of big data," Journal of Neuroscience Methods, **180**(1), 185–192, 2009, doi:10.1016/j.jneumeth.2009.03.022.

[10] H. Zou, Y. Yu, W. Tang, H. M. Chen, "Improving I/O Performance with Adaptive Data Compression for Big Data Applications," in 2014 IEEE International Parallel Distributed Processing Symposium Workshops, 1228–1237, 2014, doi:10.1109/IPDPSW.2014.138.

[11] A. A. Abdellatif, A. Mohamed, C.-F. Chiasserini, "User-Centric Networks Selection With Adaptive Data Compression for Smart Health," IEEE Systems Journal, **12**(4), 3618–3628, 2018, doi:10.1109/JSYST.2017.2785302.

[12] Q. Han, L. Liu, Y. Zhao, Y. Zhao, "Ecological Big Data Adaptive Compression Method Combining 1D Convolutional Neural Network and Switching Idea," IEEE Access, **8**, 20270–20278, 2020, doi:10.1109/ACCESS.2020.2969216.

[13] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, S. Belfkih, "Big Data technologies: A survey," Journal of King Saud University - Computer and Information Sciences, **30**(4), 431–448, 2018, doi:10.1016/j.jksuci.2017.06.001.

[14] D. M. S. Arsa, G. Jati, A. J. Mantau, I. Wasito, "Dimensionality reduction using deep belief network in big data case study: Hyperspectral image classification," in 2016 International Workshop on Big Data and Information Security (IWBIS), 71–76, 2016, doi:10.1109/IWBIS.2016.7872892.

[15] M. A. Belarbi, S. Mahmoudi, G. Belalem, "PCA as Dimensionality Reduction for Large-Scale Image Retrieval Systems," 2017, doi:10.4018/IJACI.2017100104.

[16] J. A. R. Rojas, M. Beth Kery, S. Rosenthal, A. Dey, "Sampling techniques to improve big data exploration," in 2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV), 26–35, IEEE, Phoenix, AZ, 2017, doi:10.1109/LDAV.2017.8231848.

[17] T. Hasanin, T. Khoshgoftaar, "The Effects of Random Undersampling with Simulated Class Imbalance for Big Data," in 2018 IEEE International Conference on Information Reuse and Integration (IRI), 70–79, 2018, doi:10.1109/IRI.2018.00018.

[18] C. K. Maurya, D. Toshniwal, G. Vijendran Venkoparao, "Online sparse class imbalance learning on big data," Neurocomputing, **216**, 250–260, 2016, doi:10.1016/j.neucom.2016.07.040.

[19] W. Song, P. Liu, L. Wang, "Sparse representation-based correlation analysis of non-stationary spatiotemporal big data," International Journal of Digital Earth, **9**(9), 892–913, 2016, doi:10.1080/17538947.2016.1158328.

[20] G. Cormode, N. Duffield, "Sampling for big data: a tutorial," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '14, 1975, Association for Computing Machinery, New York, New York, USA, 2014, doi:10.1145/2623330.2630811.

[21] M. Malhat, M. El Menshawy, H. Mousa, A. El Sisi, "Improving instance selection methods for big data classification," in 2017 13th International Computer Engineering Conference (ICENCO), 213–218, 2017, doi:10.1109/ICENCO.2017.8289790, iSSN: 2475-2320.

[22] J. K. Kim, Z. Wang, "Sampling Techniques for Big Data Analysis," International Statistical Review, **87**(S1), S177–S191, 2019, doi:10.1111/insr.12290.

[23] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, F. Herrera, "Big data preprocessing: methods and prospects," Big Data Analytics, **1**(1), 2016, doi:10.1186/s41044-016-0014-0.

[24] S. García, J. Luengo, F. Herrera, "Instance Selection," in Data Preprocessing in Data Mining, volume 72, 195–243, Springer International Publishing, Cham, 2015, doi:10.1007/978-3-319-10247-4_8.

[25] R. M. Kaplan, D. A. Chambers, R. E. Glasgow, "Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias," Clinical and Translational Science, **7**(4), 342–346, 2014, doi:10.1111/cts.12178.

[26] Y. Liu, W. Guo, C.-I. Fan, L. Chang, C. Cheng, "A Practical Privacy-Preserving Data Aggregation (3PDA) Scheme for Smart Grid," IEEE Transactions on Industrial Informatics, **15**(3), 1767–1774, 2019, doi:10.1109/TII.2018.2809672.

[27] M. Tamai, A. Hasegawa, "Data aggregation among mobile devices for upload traffic reduction in crowdsensing systems," in 2017 20th International Symposium on Wireless Personal Multimedia Communications (WPMC), 554–560, 2017, doi:10.1109/WPMC.2017.8301874, iSSN: 1882-5621.

[28] J. Han, "Towards on-line analytical mining in large databases," ACM Sigmod Record, **27**(1), 97–107, 1998.

[29] T. Ritbumroong, "Analyzing Customer Behavior Using Online Analytical Mining (OLAM)," 2015, doi:10.4018/978-1-4666-6477-7.ch006.

[30] B. W. F. Pan, D. R. Y. Cui, Q. D. W. Perrizo, "Efficient OLAP Operations for Spatial Data Using Peano Trees," in Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, 28–34, San Diego, California, 2003.

[31] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: an efficient data clustering method for very large databases," in ACM Sigmod Record, volume 25, 103–114, ACM, 1996.

[32] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proceedings 2nd International Conference on Knowledge Discovery and Data Mining, 226–231, 1996.

[33] K. Mahesh Kumar, A. Rama Mohan Reddy, "A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method," Pattern Recognition, **58**, 39–48, 2016, doi:10.1016/j.patcog.2016.03.008.

[34] Y. Chen, L. Zhou, Y. Tang, J. P. Singh, N. Bouguila, C. Wang, H. Wang, J. Du, "Fast neighbor search by using revised k-d tree," Information Sciences, **472**, 145–162, 2019, doi:10.1016/j.ins.2018.09.012.

[35] Y. Hong, Q. Tang, X. Gao, B. Yao, G. Chen, S. Tang, "Efficient R-Tree Based Indexing Scheme for Server-Centric Cloud Storage System," IEEE Transactions on Knowledge and Data Engineering, **28**(6), 1503–1517, 2016, doi:10.1109/TKDE.2016.2526006.

[36] L.-Y. Wei, Y.-T. Hsu, W.-C. Peng, W.-C. Lee, "Indexing spatial data in cloud data managements," Pervasive and Mobile Computing, **15**, 48–61, 2014, doi:10.1016/j.pmcj.2013.07.001.

[37] M. Huang, F. Bian, "A Grid and Density Based Fast Spatial Clustering Algorithm," in 2009 International Conference on Artificial Intelligence and Computational Intelligence, 260–263, IEEE, Shanghai, China, 2009, doi:10.1109/AICI.2009.228.

[38] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, W. Burgard, "OctoMap: an efficient probabilistic 3D mapping framework based on octrees," Autonomous Robots, **34**(3), 189–206, 2013, doi:10.1007/s10514-012-9321-0.

[39] E. J. Clarke, B. A. Barton, "Entropy and MDL discretization of continuous variables for Bayesian belief networks," International Journal of Intelligent Systems, **15**(1), 61–92, 2000, doi:10.1002/(SICI)1098-111X(200001)15:1⟨61::AID-INT4⟩3.0.CO;2-O.

[40] H. Wei, Y. Du, F. Liang, C. Zhou, Z. Liu, J. Yi, K. Xu, D. Wu, "A k-d tree-based algorithm to parallelize Kriging interpolation of big spatial data," GIScience & Remote Sensing, **52**(1), 40–57, 2015, doi:10.1080/15481603.2014.1002379.

[41] B. Chatterjee, I. Walulya, P. Tsigas, "Concurrent Linearizable Nearest Neighbour Search in LockFree-kD-tree," in Proceedings of the 19th International Conference on Distributed Computing and Networking, ICDCN '18, 1–10, Association for Computing Machinery, Varanasi, India, 2018, doi:10.1145/3154273.3154307.

[42] A. Bifet, J. Zhang, W. Fan, C. He, J. Zhang, J. Qian, G. Holmes, B. Pfahringer, "Extremely Fast Decision Tree Mining for Evolving Data Streams," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, 1733–1742, Association for Computing Machinery, Halifax, NS, Canada, 2017, doi:10.1145/3097983.3098139.

[43] S. J. Hong, "Use of contextual information for feature ranking and discretization," IEEE Transactions on Knowledge and Data Engineering, **9**(5), 718–730, 1997, doi:10.1109/69.634751.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 463-468 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Short-Term Dynamic Exchange Rate Model: IFEER Concept Development

Anton Kuzmin*

Department of Data Analysis, Decision-Making and Financial Technologies, the Financial University, Ramsey 125993, Russia

ARTICLE INFO

Article history:
Received: 24 April, 2020
Accepted: 20 July, 2020
Online: 09 August, 2020

Keywords:
Exchange Rate
Model
IFEER
Nonlinear Dynamics
Short-Term Imbalances
Russian Ruble
US Dollar

ABSTRACT

The new model of the short-term exchange rate dynamics was constructed. First of all, the most interesting were the reasons of the deviation from the medium-term equilibrium. The author's IFEER-concept (International Flows Equilibrium Exchange Rate) was used as a base and it was developed. In this study, due to the short-term modeling period the differential approach was applied. The result was an integrated version of the exchange rate dynamics model. The main result of mathematical modeling is a nonlinear multi-factor functional dependence of the exchange rate. The result dynamic functional dependence differs from the previous medium-term dependencies by the type of the internal dynamic function. Economically, this function in the short-term period is responsible for explosive changes in the exchange rate dynamics. The basis for mathematical modeling was the system of fundamental economic factors that affect the dynamics of the exchange rate. The influence of crisis events on the Russian financial market in the short term was studied. The conducted research allowed us to analyze and evaluate the quantitative impact of the short-term effects of the dynamics of the exchange rate of the Russian ruble to the US dollar.

1. Introduction

This paper is an extension of work originally presented in Twelfth International Conference "Management of large-scale system development" (MLSD'2019) [1].

The theory of exchange rate determination is one of the most important components of modern economic theory. Mathematical models of the medium-term dynamics of the exchange rates are widely presented in the works of classics in [2]-[7], and the monetarist school pay considerable attention to modeling the medium-term and the long-term dynamics of exchange rates. The same can be said about DSGE-models in [8].

Medium-term modeling of the exchange rates including the Russian ruble, based on the IFEER-concept (International Flows Equilibrium Exchange Rate concept), was carried out by [9]. This concept is based on mathematical modeling of the foreign trade operations and the capital flows as the most important macroeconomic fundamental determinants of the exchange rate movements. Several studies (for example, in [10]) consider sources of exchange rate fluctuations in other cases.

Against this background, only a few studies are devoted to the short-term exchange rate dynamics. And first of all, it is necessary to note here the works explained in [11-13].

Mathematical modeling of short-term dynamics will reveal the mechanisms of explosive changes in the exchange rate in crisis periods.

This study will build a new model of the short-term exchange rate dynamics. First of all, we will be interested in the reasons for the deviation from the medium-term equilibrium, and these aspects will be formalized mathematically. But in any case, the basis for mathematical modeling will be the fundamental economic factors that affect the dynamics of the exchange rate. Thus we will use the author's IFEER-concept and develop it.

2. Conceptual Framework of Exchange Rate Modelling

In the basic version, discrete flows are considered in the previous author's works [9]. For exchange rates that have market pricing, in a certain period we defined in the *i-th* market transaction:

e_i - the nominal exchange rate,

D_i - the volume of the market transaction in the foreign currency,

R_i - the volume of the market transaction in the national currency.

* Corresponding Author: Anton Kuzmin, Moscow, Russian Federation, a_kuzmin@rambler.ru

From a financial point of view these variables are linked as follows:

$$e_i D_i = R_i. \quad (1)$$

As a result, in the *i-th* transaction:

$$e_i = R_i / D_i. \quad (2)$$

However, in this study due to the short-term modeling period author will also apply a differential approach.

Let's define *R(t)* as a streaming continuously differentiated function of funds in the national currency in the foreign exchange market. Accordingly, *D(t)* is a streaming continuously differentiated function of funds in the foreign currency and *e(t)* is the nominal bilateral exchange rate at time *t*.

By analogy with (1) in the differential form at *t*:

$$e(t)\,\partial D(t)/\partial t = \partial R(t)/\partial t. \quad (3)$$

As a result:

$$\frac{\partial R(t)/\partial t}{\partial D(t)/\partial t} = e(t). \quad (4)$$

By analogy with the previously used discrete IFEER-concept [9] we introduce the weight function *w(t)* for the exchange rate. In the discrete version, the weight function *w(i)* showed the contribution of the particular *i-th* transaction, depending on the volume in the foreign currency:

$$w(i) = \frac{D_i}{\sum_{j=1}^{T} D_j}.$$

In the integral version, the weight function *w(t)* is also based on the volume of funds in the foreign currency in a certain period *T*:

$$w(t) = \left[\frac{\partial D(t)/\partial t}{\int_T \partial D(t)/\partial t\, dt}\right]. \quad (5)$$

Thus we will define the exchange rate as an integral in the certain period *T*:

$$e(T) = \int_T w(t)e(t)dt. \quad (6)$$

Substitute the weight function (5) in the formula (6):

$$e(T) = \int_T w(t)e(t)dt = \int_T \left[\frac{\partial D(t)/\partial t}{\int_T \partial D(t)/\partial t\, dt}\right] e(t)dt =$$

$$= \int_T \left[\frac{\partial D(t)/\partial t\, e(t)}{\int_T \partial D(t)/\partial t\, dt}\right] dt.$$

Let's use properties (3) and (4) and take the denominator beyond the integral:

$$e(T) = \int_T \left[\frac{\partial D(t)/\partial t\, e(t)}{\int_T \partial D(t)/\partial t\, dt}\right] dt = \int_T \left[\frac{\partial R(t)/\partial t}{\int_T \partial D(t)/\partial t\, dt}\right] dt =$$

$$= \frac{\int_T \partial R(t)/\partial t\, dt}{\int_T \partial D(t)/\partial t\, dt}.$$

Now we can disaggregate the streaming functions according to the balance of payments accounts:

$$e(T) = \frac{\int_T \partial R(t)/\partial t\, dt}{\int_T \partial D(t)/\partial t\, dt} = \frac{\int_T \partial R^{CA}(t)/\partial t\, dt + \int_T \partial R^{K}(t)/\partial t\, dt}{\int_T \partial D^{CA}(t)/\partial t\, dt + \int_T \partial D^{K}(t)/\partial t\, dt}.$$

Here funds with the upper index *CA* belong to the current balance, funds with the upper index *K* belong to the capital flows balance.

This formula is the main intermediate result of the IFEER-concept, allowing us to identify foreign trade operations and capital flows as the main factors of the exchange rate movements.

Let's denote:

$$\int_T \partial R^{CA}(t)/\partial t\, dt = I_T,$$

$$\int_T \partial R^{K}(t)/\partial t\, dt = K_T^{-},$$

$$\int_T \partial D^{CA}(t)/\partial t\, dt = E_T,$$

$$\int_T \partial D^{K}(t)/\partial t\, dt = K_T^{+}.$$

From an economic point of view, E_T is the volume of funds as the supply of the foreign currency from export operations, I_T is the demand in the national currency for the foreign currency from import operations. K_T^- is the capital outflow as the demand in the national currency for the foreign currency and K_T^+ is the capital inflow as the supply of the foreign currency.

At this conceptual level, the exchange rate dependence has a dynamic form by *T*:

$$e(T) = e_T = (I_T + K_T^-) \Big/ (E_T + K_T^+) \quad (7)$$

However, for the two involved parties, the nominal exchange rate is determined symmetrically because the country's export is another country's import and the capital outflow is the inflow for the other side. Here the asterisk shows the financial and economic variables of the opposite counterparty in this two-country world:

$$e_T = (E_T^* + K_T^{*+}) \Big/ (I^*_T + K_T^{*-}) = 1 \Big/ e_T^*$$

It should be noted that the dependence (7) is natural from the economic point of view, since, for example, the increase in demand for imports and an increase in capital outflow leads to an increase in the exchange rate:

$$e = e^{*(-1)} = f_e(I^+, (K^-)^+, E^-, (K^+)^-)$$

The upper sign "-" or "+" here and further on this factor shows that the function respectively strictly decreases or increases.

In terms of partial derivatives, for example:

$$\partial f_e(I^+, K^+, E^-, K^-) \Big/ \partial K^+ < 0$$

3. The Two-Country Model: The Level of the Current Account Balance

We will build a two-period model of the exchange rate in periods *t* and *t-1*. Since we are going to use point estimates, we will replace the large *T* as a period index with the small *t* for convenience.

Under these conditions formula (7) will take a more decent look, which has already been used by us:

$$\begin{aligned} e_t &= (I_t + K_t^-) \Big/ (E_t + K_t^+) = \\ &= (E_t^* + K_t^{*+}) \Big/ (I^*_t + K_t^{*-}) = 1 \Big/ e_t^* \end{aligned} \quad (8)$$

One of the most important components of the foreign currency inflow to the domestic market is export revenue. However, its level in the period *t* is determined by decisions of national exporters in the previous period *t-1*:

$$\begin{aligned} E_t = I_t^* = k_E^{/} P_t^* (k_E^{//} (Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\delta})(e_{t-1}^R)^z = \\ = k_E P_t^* ((Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\delta})(e_{t-1}^R)^z \end{aligned} \quad (9)$$

here: k_E is a constant,

P_t^* - the aggregate foreign price level,

Q_t - the total domestic real output,

e^R_{t-1} - the indicator of the international competitive advantages.

The exports of one side in this two-country world are imports of the other side.

The dependence (9) indicates that only the part $k_E^{/}$ of export revenue comes to the domestic currency market. In general, the volume of exports is determined by the part $k_E^{//}$ of the total domestic real output Q_t in the foreign prices P_t^*.

Here intertemporal averaging of the total real output Q_t is used for modeling. Note, that $\frac{1}{x+1} + \frac{x}{x+1} = 1$. In this case, the income at *t* is given more weight than the income at *t-1*. The method of intertemporal averaging of Q_t should not have a significant impact since it is the least volatile factor.

One of the most important factors of the model is the indicator of the international competitive advantages. This is the real bilateral exchange rate in the previous period *t-1*:

$$e^R_{t-1} = e_{t-1} P^*_{t-1} \Big/ P_{t-1} \quad (10)$$

Thus, in formula (9) we take into account the possible different degrees of the export functions responses to changes in the indicator of the international competitive advantages and the total domestic real output. For these purposes, we use various adjustable parameters in degrees: δ, x, z. Their relationships will be shown later.

Functional dependence (9) at the same time also has a natural functional character from the economic point of view with respect to the system of the main determinants of the exchange rate:

$$E_t = f_{E_t}(P_t^{*+}, Q_t^+, Q_{t-1}^+, e_{t-1}^{R+})$$

In accordance with the previous modeling, the demand for the foreign currency from imports is determined completely symmetrically for the other side:

$$I_t = E_t^* = k_I P_t (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\lambda} (e_{t-1}^R)^y. \quad (11)$$

Here: k_I is a constant,

Q_t^* - the total foreign real output.

For the purposes of simulation, it is necessary to impose the restrictions: $x \neq -1$, $x=z-y$.

The properties of the stream functions of the current balance operations:

$$E^*=I,\ I^*=E.$$

Here it is important to note, that the dependencies (10) and (11) we define for the medium-term environment. But we can assume

that these dependencies retain their mathematical factor structure even for the short-term modeling. This is due to the economic inertia of exports and imports in the world economy.

4. The Two-Country Model: the Level of the Capital Flows

Of course, the capital flows are one of the most important factors of the exchange rates dynamics. And the developed author's IFEER-concept allows us to conduct mathematical modeling taking into account the capital flows.

We need to accept the hypotheses about mathematical forms of the basic capital flows dependencies. The capital inflow dependence is an increasing function by the aggregate foreign prices, the indicator of international competitive advantages, and the total domestic real output.

In this context, it is important to note that these functions must be modified for short-term exchange rate modeling in relation to the previously discussed functions in the author's work [9]. Here the constants k becomes functions by time. For the functional dependence of the capital outflow, for example, it is an increasing function in a crisis period. This shows a short-term increase in capital outflow compared to the medium-term equilibrium dynamics.

For the capital outflow, thus, the dependence must satisfy the following condition:

$$K_t^- = f_{K_t^-}(k_{K^-}{}^+(t), P_t^+, Q^*{}_t^+, Q^*{}_{t-1}^+, e_{t-1}^{R\,-})$$

Thus we define the functional dynamic dependence of the capital outflow:

$$K_t^- = k_{K^-}(t) P_t\ (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\rho} (e_{t-1}^R)^y . \quad (12)$$

The capital outflow is a part of the domestic product produced in foreign prices.

In the short-term period in the currency market the crisis phenomena lead to a sharp increase in the capital outflow. Mathematically this is expressed in the strict increase function $k_{k^-}(t)$ by t in the dynamic formula (12):

$$\partial k_{k^-}(t) \big/ \partial t > 0 .$$

At the same time, the functional dynamic dependence of the capital inflow:

$$K_t^+ = k_{K^+}(t) P_t^* (Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\theta} (e_{t-1}^R)^z . \quad (13)$$

Perfectly symmetrical, in the domestic currency market the crisis phenomena leads to a significant reduction of the capital inflows: a strict decrease of the function $k_{k^+}(t)$ by t in the formula (13):

$$\partial k_{k^+}(t) \big/ \partial t < 0 .$$

Here we use the hypotheses about the type of the functions $k_K(t)$, that are not constants for the capital flows in the short-term period. This distinguishes this model from the previously developed author's models.

In accordance with the methodology of modeling dependencies are interrelated:

$$K_t^{*+} = K_t^- = k_{K^-}(t) P_t\ (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\rho} (e^{*R}_{t-1})^{-y} .$$

The properties of the capital stream functions:

$$K^{*-} = K^+ ,$$

$$K^{*+} = K^-$$

Let's substitute the functional dependencies (9)-(13) into the basic formula (8):

$$e_t = e(t, k(t), Q(t), P(t), P^*(t)) =$$

$$= \frac{P_t (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\rho} (e_{t-1}^R)^y (k_I (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\lambda-\rho} + k_{K^-}(t))}{P_t^* k_E (Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\theta} (e_{t-1}^R)^z (k_E (Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\delta-\theta} + k_{K^+}(t))} . \quad (14)$$

Let's introduce the dynamic aggregate function $k(t)$:

$$\frac{(k_I (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\lambda-\rho} + k_{K^-}(t))}{(k_E (Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\delta-\theta} + k_{K^+}(t))} = (k(t))^{x+1} . \quad (15)$$

With (15) in mind, we can rewrite (14) :

$$e_t = \frac{P_t (Q_t^{*\frac{1}{x+1}} Q_{t-1}^{*\frac{x}{x+1}})^{\rho} (k(t))^{x+1}}{P_t^* k_E (Q_t^{\frac{1}{x+1}} Q_{t-1}^{\frac{x}{x+1}})^{\theta} (e_{t-1} \frac{P_{t-1}^*}{P_{t-1}})^x} .$$

Thus, move the member $(e_{t-1})^x$ to the left side:

$$e_t (e_{t-1})^x = \left(k(t) \frac{P_t}{P_t^*} Q_t^{*\rho/x+1} Q_t^{-\theta/x+1} \right) \times$$

$$\times \left(k(t) \frac{P_{t-1}}{P_{t-1}^*} Q_{t-1}^{*\rho/x+1} Q_{t-1}^{-\theta/x+1} \right)^x .$$

Let's make a temporary separation of the modeling variables and re-define $\psi = \rho/x+1$ and $\zeta = \theta/x+1$ for convenience:

$$e(t, k(t), P(t), P^*(t), Q(t)) = k(t) \frac{P_t}{P_t^*} (Q_t^{*\psi})(Q_t^{-\zeta}) . \quad (16)$$

5. Assessment of Short-Term Dynamic Effects

The main result of mathematical modeling is the new nonlinear multi-factor functional dependence of the exchange rate. Here, it

$$k(t).$$

Economically, this function in the short-term period is responsible for explosive changes in the exchange rate dynamics. So, this is the main fundamental difference between this model and the previous author's models [9], etc. In those works, these functions are constants in the medium-term in various modifications.

There is significant stability of $(Q_{t-1}^{\delta/\delta+1} Q_t^{1/\delta+1})$ in the dynamic functional dependence (15) in comparison with other fundamental factors in the short-term period. So, a strict decrease of the internal function $k_{k^+}(t)$ and a strict increase of the internal function $k_{k^-}(t)$ guarantees us a strict increase of the aggregate function $k(t)$ in a crisis period: $\partial k(t)/\partial t > 0$. Mathematically this reflects the reasons for the deviation from the medium-term equilibrium.

Direct economic assessment of the short term impact of the function $k(t)$ is quite difficult. But we will try to do it using the example of the Russian ruble.

Let's use the main result of the author's work [9]. The medium-term dynamic functional dependence of the exchange rate of the Russian ruble against the US dollar is:

$$e_t = k' \frac{P_t}{P_t^*} Q_t^{-\psi}. \quad (17)$$

here: k' is a constant,

ψ - an estimated parameter.

For research and verification of this model we used the Russian CPI (Consumer Price Index) as the model variable *P* (Federal service of state statistics of Russia), the index of the real Russian GDP as the model variable *Q*, the prices of blend crude oil as the model variable *P** (Intercontinental Exchange, US dollars per brent-mix barrel, Bloomberg information agency, Bloomberg Terminal).

For our research the most appropriate is the crisis period 2013 – 2015. This period is associated with the currency crisis and the significant depreciation of the Russian ruble.

The ruble/dollar exchange rate (USD/RUR) is the main rate on the Russian foreign exchange market. The value of the ruble exchange rate in relation to other world currencies is determined through the cross-rate system.

Thus, as the starting point of the research period, December 2013 was selected. At that time, both the international environment and the domestic macroeconomic situation were stable. The exchange rate of the ruble at that moment was *32,73* USD/RUR according to the Central Bank of Russia.

This study suggests using the least squares method with normalization of the exchange rate values. Since the exchange rate changes quite seriously during crisis periods, it is more important to use relative indicators of changes.

Thus, in accordance with numerical simulation, the solution of the parametric minimization problem

$$\left[\min_{\psi} \sum_t \left(\frac{e_t(\psi) - e(\text{nominal})_t}{e(\text{nominal})_t} \right)^2 \right], \quad \psi \geq 0$$

is $\psi = 0,5$. Here *e(nominal)* is the nominal exchange rate USD/RUR, according to the Bank of Russia monthly calculations at the end of each period.

In Figure 1 (monthly data, author's calculations) we present the dynamics of the exchange rate of the Russian ruble against the US dollar in accordance with the nominal exchange rate *e(nominal)* and the result dynamic functional dependence of this research *e(Theor)* (17).

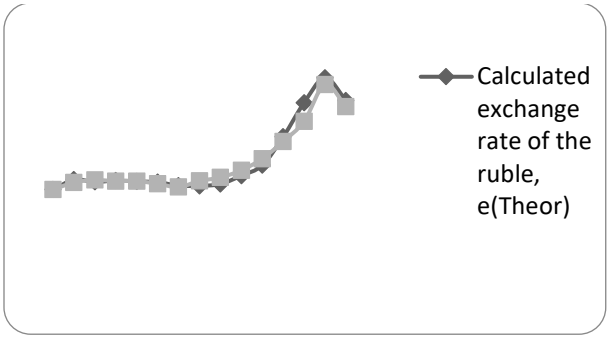


Figure 1: The calculated and nominal exchange rates of the Russian ruble to the US dollar (USD/RUR, monthly data, 2013 - 2015).

The average of absolute normalized deviations of the calculated and nominal exchange rates was 3%, and the average of normalized deviations was 0,3%. As a result, this allows us to conclude that the model quality is high.

This research has shown a high correlation between the exchange rate of the ruble, the inflation, the Russian balance of payments, and other fundamental variables (for example, we can also recommend a work [14]).

Using the methodology, presented in this paper, and the author's work [15] (in Russian) it can be shown, that short-term dynamic functional dependence of the exchange rate of the Russian ruble against the US dollar is:

$$e_t = k'(t) \frac{P_t}{P_t^*} Q_t^{-\psi}. \quad (18)$$

Here, the system of main fundamental economic factors of the exchange rate dynamics in relation to the dependencies (16) and (17) is saved.

In Figure 2 (Moscow Exchange, executions TOM) in the period 2013 – 2015solid line represents the medium-term dynamics of the exchange rate of the Russian ruble against the US dollar. We schematically overlaid data from the author's work [9] and Figure 1. Based on the presented data, we can estimate the cumulative contribution of the function $k'(t)$ to changes in the ruble exchange rate during crisis periods in the short-term in the

amount of 20-25%. The contribution to the crisis period in 2015-2016 can also be estimated at the same amount approximately.

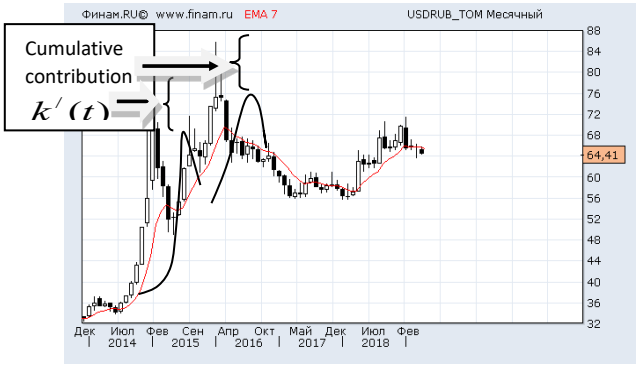


Figure 2: USD/RUR quotes
(Moscow Exchange, monthly data, executions TOM, Japanese candlesticks, 2013 – 2018, Information agency *FINAM*)

6. Conclusion

In this study, the new model of short-term exchange rate dynamics was constructed. First of all, we were interested in the reasons for the deviation from the medium-term equilibrium.

The author used the IFEER-concept as a base and developed it. In this study due to the short-term modeling period, the differential approach was applied. The result was an integrated version of the exchange rate dynamics model. The functional dependencies of the export-import operations and the capital flows were mathematically formalized.

The main result of mathematical modeling is the nonlinear multi-factor dynamic functional dependence of the exchange rate that differs from the previous medium-term functional dependencies by the type of the internal dynamic function. Economically, this function in the short-term period is responsible for explosive changes in the exchange rate dynamics. But in any case, the basis for mathematical modeling was the system of the fundamental economic factors that affect the dynamics of the exchange rate.

The most important limitation of mathematical modeling is the significant difficulty in directly economic assessment of the internal functions $k_{k^+}(t)$ and $k_{k^-}(t)$ that determine short-term capital flows in crisis periods.

But the influence of crisis events on the model in the short term is studied on the Russian financial market at a previous time by assessment of the short-term impact of the function $k(t)$. The conducted research allowed us to analyze and evaluate the quantitative impact of the short-term effects of the dynamics of the exchange rate of the Russian ruble against the US dollar. The estimated indicators allow us to conclude that the Bank of Russia needs to pursue a more active interventionist policy to maintain the stability of the national currency.

At the same time, such economic assessments are individual and specific for each currency and each crisis period.

The developed IFEER-concept allows us to model the exchange rate dynamics in other economic conditions. First of all, this concerns the future research of the long-term dynamics of the main world currencies and the Russian ruble.

Conflict of Interest

The author declares no conflict of interest.

References

[1] A. Kuzmin, "Modeling of Short-Term Exchange Rates Dynamics" in Twelfth International Conference "Management of large-scale system development" (MLSD 2019), Publisher: IEEE, 2019, https://doi.org/10.1109/MLSD.2019.8911067

[2] R.A. Mundell, "Capital Mobility and Stabilization Under Fixed and Flexible Exchange Rates" Canadian Journal of Economics and Political Science, November, 475-485, 1963. https://doi.org/10.2307/139336

[3] A.C. Stockman, "A Theory of Exchange Rate Determination" Journal of Political Economy, August, **88**, 673-698, 1980. https://doi.org/10.1086/260897

[4] P.B. Clark, R. MacDonald, Exchange Rates and Economic Fundamentals: A Methodological Comparison of BEERs and FEERs, IMF Working Paper 98/67, Washington: International Monetary Fund, March, 1998.

[5] P.B. Clark, R. MacDonald, Filtering the BEER a permanent and transitory decomposition, IMF Working Paper 00/144, Washington: International Monetary Fund, 2000.

[6] M. Mussa, "The Theory of Exchange Rate Determination" in Exchange Rate Theory and Practice, J.E.O. Bilson, R.C. Marston (eds.), Chicago: University of Chicago Press, NBER, 13-78, 1984.

[7] L. Killian, "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?" Journal of Applied Econometrics, 14(5), 491-510, 1999. https://doi.org/10.1002/(SICI)1099-1255(199909/10)14:5<491::AID-JAE527>3.0.CO;2-D

[8] M. Obstfeld, K. Rogoff, "Exchange Rate Dynamics Redux" Journal of Political Economy, **103** (3), 624-660, 1995. https://doi.org/10.3386/w4693

[9] A. Kuzmin, "Exchange Rate of the Ruble Modeling" Advances in Systems Science and Applications, **19**(4), 87-93, 2019. https://doi.org/10.25728/assa.2019.19.4.830

[10] I. Chowdhury, "Sources of Exchange Rate fluctuation: empirical evidence from six emerging market countries" Applied Financial Economics, **14**(1), 697-705, 2004. https://doi.org/10.1080/0960310042000243538

[11] R. Dornbush, "Capital Mobility, Flexible Exchange Rate and Macroeconomic Equilibrium" in Recent Developments in International Monetary Economics, E. Claasen, P. Salin (eds.), North-Holland, 261-278, 1976.

[12] R. Dornbush, " Expectations and Exchange Rates Dynamics" Journal of Political Economy, **84**, December, 1161-1176, 1976. https://doi.org/10.1086/260506

[13] R. Dornbusch, "Equilibrium and Disequilibrium Exchange Rates" Zeitschrift fur Wirtschafts- und Socialwissenschaften, Berlin, **102**(6), 573-599, 1982. https://doi.org/10.3386/w0983

[14] L. Strizhkova, "The relationship between inflation, exchange rate and parameters of economic policy (on an example of Russia)" Bulletin of the Institute of Economics of the Russian Academy of Sciences, **5**, 156-176, 2017. (Article in Russian with an abstract in English)

[15] A. Kuzmin, "Modeling short-term dynamics of the ruble exchange rate" Money and credit, **8**, 59-65, 2015. (Article in Russian with an abstract in English)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 464-471 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Extending the Classifier Algorithms in Machine Learning to Improve the Performance in Spoken Language Understanding Systems Under Deficient Training Data

Sheetal Jagdale[1,*], Milind Shah[2]

[1]*Electronic Engineering Department, Shri. L.R Tiwari College of Engineering, Kanakia Park Mira Road (E) Thane Maharashtra, 401107, India*

[2]*Electronics and Telecommunication Engineering Department, Fr. C. Rodrigues Institute of Technology, Sector 9A, Vashi, Navi Mumbai, 400703, India*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 31 October, 2020
Online: 20 November, 2020

Keywords:
Spoken Dialogue System (SDS)
Spoken Language Understanding (SLU)
Deficient Training Data
Rarely Visited Domain
Classifier

ABSTRACT

One of the open domain challenges for Spoken Dialogue System (SDS) is to maintain a natural conversation for rarely visited domain i.e. domain with fewer data. Spoken Language Understanding (SLU) is a component of SDS that converts user utterance into a semantic form that a computer can understand. If we scale SDS open domain challenge to SLU then it should be able to convert user utterance to a semantic form even if less data is available for the rarest visited domain. The SLU reported in literature incorporate classifiers for the task of identifying the domain of user utterance, understanding the intent of the user, and filling slots-value pair. Thus, to address open domain challenges, classifiers in SLU must be robust to scarce training data. This paper presents investigations to improve the performance of SLU to convert user utterance into semantic form even if less training data is available. Eleven classification algorithms from machine learning have experimented under deficient data. The evaluation matrices used are accuracy, f-score, and inter cross-entropy. Comprehensive experimentation is carried out on the two publicly available datasets DSTC2 and DSTC3 were carried out.The accuracy for Support Vector Machine (SVM) , Stochastic Gradient Descent (SGD) and Decision tree are 0.940, 0.960 , 0.955 for DSTC2 and 0.916, 0.900, 0.901 for DSTC3 database respectively. The F-score for SVM, SGD and Decision tree are 0.855, 0.868, 0.849 for DSTC2 dataset and 0.725, 0.715, 0.700 for DSTC3 database, respectively. The ICE for SVM and SGD are 1.191,1.100 for DSTC2 dataset and 3.180,2.999 for DSTC3 database, respectively. The performance of SLU based on SVM and SGD was found to be the best among all. The worst performance in terms of all three evaluation metrics was displayed by SLU incorporating Automatic Relevance Determination (ARD) and Relevance Vector Machine (RVC) classifier.

1. Introduction

Spoken Dialogue System (SDS) interacts with the user in the most natural form of communication. It helps the user to achieve a goal through a series of interactions. SDS is moving to the open domain now. In the open domain, SDS has to maintain natural conversation even if data are scarce. Recently a lot of work is reported in the literature for open-domain Spoken Language Understanding (SLU) where it should be able to convert user utterance to semantic form even if less data is available. SLU reported in literature incorporates classifiers at different levels. The first level classifier in SLU will classify utterance to a domain example restaurant, flight booking, and tourist information. The next level classifier will classify user intend within domain example for example finding a restaurant or flight.

Subsequent levels in SLU should fill the slots such as time, date, etc. Filling slots will allow the flow of conversation to achieve the user goal. Thus, the robustness of the classifier to infer the output even if training data is less is important. There is a need for investigation of the robust classifier for SLU, which can convert user utterance to semantic form even if less training data is available. The work in the paper investigates a robust classifier. This research is broadening of work initially reported in the conference paper [1] The work is extended by incorporating CNN

*Corresponding Author: Sheetal Jagdale, Email: sheetal.jagdale2@gmail.com

for feature extraction and investigating eleven classifiers from literature. Comprehensive experimentation was done on the DSTC2 dataset and the DSTC3 dataset. The evaluation was done on two domain datasets. DSTC2 datasets for the restaurant domain and DSTC3 datasets for the tourist domain were used for evaluation.

Classification is an important task in many applications as image processing, spoken dialogue systems, etc. Choosing the right classifier for an application is a very important task. A lot of work is reported in the literature for investigating a robust classifier for a particular application. The work [2] has predicted error rates of k-neighborhood, SVM, and decision tree classifiers for the application of computer-aided diagnostic (CAD). The evaluation of the classifiers is done on a small training set. Later the work was extended to three tasks in CAD. The task was detecting cancer, identifying high or low-grade and-later the work was extended by adding more features [3] Another work investigates classifiers for text classification [4] The Paper investigates Rocchio Classification, K Nearest Neighbor, Naïve Bayes Classification, SVM Classification, and Regression-Based Classification for the task of classifying text. The paper concludes with a study stating that the right choice of dimension reduction algorithm and classifier can improve the performance of text classification. One more paper [5] refers to the classifier investigation for the task of text classification under limited data. A naive Bayes classifier is also investigated for text classification [6] The work combines the dimension reduction method with a classification algorithm for text classification under limited training data. The study was reported in [7] Paper [8] tackles the issue of imbalance data under training data. They tackle the issue by a stable space for testing data proposed using a metric algorithm. The effect of training size on the neural network classifier is also demonstrated in the literature [9] Recently a work [10] reported classifiers for clinical treatment. The study also includes data augmentation techniques along with classifiers Thus, a lot of classification algorithm investigation for different applications is done. This paper investigates a robust classifier for the slot filling task of SDS.

To improve SLU's performance for classifying utterance to predefined slots, different classification algorithms from machine learning were investigated. The purpose of the work is to:

1) Extend a robust classifier algorithm in machine learning to the semantic decoder in spoken language understanding for the task of converting user utterance into a semantic form of dialogue act and slot-value pair.
2) To demonstrate the impact of data scarcity on SLU incorporating different classifiers.
3) Comprehensive experiment on SLU incorporating eleven classifiers on two public datasets DSTC2 and DSTC3. The metrics used for evaluation are accuracy, f-score, and item cross-entropy.

2. Related Background

2.1. Decision Tree [DT]

DT is a popular classifier and is used in various machine learning applications such as data mining, image processing, statistics, etc. It uses if and then rules and follow the top-down method. In paper [11] decision tree is used for mining applications. A decision tree is used to extract information from huge data. Another work discusses the framework for decision tree algorithm, splitting criteria, and pruning methods [12] The paper also discusses extended decision trees such as obvious tree s decision trees, incremental tree, tree inducers, and fuzzy trees. A decision is also used as an inductive inference system [13] The decision tree algorithm consists of a set of rules called a decision. It extracts knowledge from example. Thus, it resembles the machine learning principle of learning from examples. In a research work presented in paper [14], the IDA decision algorithm is proposed. The ID3 decision algorithm is also discussed and compared with IDA.ID3 algorithm uses entropy for finding the uncertainty of classification output. While IDA uses divergence for classification. As reflected in the results, the IDA algorithm performance displays its effectiveness over ID3. Another paper [15] proposes an improved ID3 algorithm. It incorporates the Taylor series in ID3 to improve the classifier. One of the popular applications of the decision tree is data mining [16]

2.2. Multilayer Perceptron [MLP]

One of the most powerful classifiers used in machine learning is the MLP classifier. It uses a supervised learning and feedforward network. One of the applications reported for MLP is for recognizing a number in the Bengali language [17] For numbers recognizing pre-processing techniques such as feature, the selection is also incorporated. Another application reported is for the inference of liver cirrhosis [18] However, it is very difficult to tune the parameter of a multilayer perceptron. To overcome this difficulty a methodology is proposed in the paper [19] The methodology can find the number of epochs to get the best performance which is optimal.

2.3. Gaussian naïve Bayes

Gaussian naive Bayes classifier is a probabilistic classifier applied to many applications. Bayes theorem is used for classification. One application is discussed in [20] for breast cancer classification. A framework consisting of data pre-processing, feature selection, and classifier is proposed for breast cancer classification.

2.4. Gaussian classifier

Gaussian classifiers are used in many applications such as language identification [21], spectrum recognition [22] and dialogue representation [23] The work in paper [24] discusses the parametric Gaussian classifier. In this work, a Gaussian classifier incorporating a neural network is discussed. The results demonstrate the effectiveness of the parametric approach applied to Gaussian distribution. The Gaussian classifier is also analyzed for limited training data [25] The work discusses the impact of limited data on the decision boundary of a classifier.

2.5. AdaBoost classifier

In literature AdaBoost are used for detecting objects[26] ,detecting faults in gearbox [27] and detecting hands [28] In another paper, AdaBoost is used in mobile intelligent terminals for security level classification [29] Various modified AdaBoost algorithms are were analyzed for classification margin[30]

Further, in the same paper, generalized error for eight variants of AdaBoost was reported. Personalized AdaBoost which is a variant of AdaBoost as reported in the paper has the least generalization error

2.6. Relevance vector machine

Relevance vector machine (RVM) is a probabilistic model used for classification and regression. The prediction of the output of RVM is probabilistic. RVM is used for the classification of hyperspectral images using wavelet kernel [31] As reported in the paper, training speed was improved by RVC incorporating wavelet kernel. Another application was reported for RVM for the detection of clustering in mammogram [32] For mammogram classification, RVM was able to maintain better classification with less computation complexity

2.7. Stochastic gradient (SGD)

It is a linear classifier that incorporates SGD learning. These classifiers are used in classifying emotions in Hindi text [33] In the Hindi emotion classification model, first pos tagging is done on the text and then features are extracted from the text. Extracted features are used to train the SGD classifier or random forest classifier. The results were improved by incorporating reduced features to SGD. Another application for SGD in literature is for the categorization of Bangla text [34] The results for Bangla text classification was improved using SGD.

2.8. Support Vector Machine [SVM]

It is a linear non-probabilistic classifier. It uses Platt's model to predict the output. One of the applications in literature is big data automated selection [35] The quality of data affects the classification performance of SVM. Results in the paper showed that optimizing the SVM parameter can improve the performance. SVM is also used for skin disease classification [36] In this work, the Image was flattened to get an array of pixel intensities. This array of labeled data was used to train SVM. Recently SVM using field programmable gate array is used for knowledge discovery.

2.9. K-nearest neighborhood [KNN]

It is a supervised classifier based on distance measurement. It is widely applied to data mining for the classification task.In recent work[37], KNN is used for saving energy for the home. KNN is used here for appliance classification. The KNN classifiers were able to detect home appliances in a short time. Lot variants of KNN such as dynamic KNN, weight-adjusted KNN and distance weighted KNN are proposed to speed up the learning process [38]

2.10. Random Forest [RF]

It is an algorithm based on a decision tree. In a paper [39], RF is used for Android malware detection applications. The framework proposed in the paper was evaluated on the android feature dataset. RF was used for the classification task of identifying applications as benign or malicious. Misclassification error for classifying malware was least for the random forest. Recently in a paper, RF is used for identifying and classifying applications for network traffic detection. The algorithm used for comparison were random forest, SVM, and logistic regression. RF outperformed all the classifiers in terms of accuracy and f-score.

2.11. Variational Relevance Vector Machine [VRVM]

A variant of RVM is proposed in paper [40] known as variational relevance vector machine. This method is RVM with variational inference. VRVM is applied for RVM multi-dimension array classification. Another application for VRVM is discussed in the literature for the classification of table data such as blocks of image [41] VRVM was found more robust then RVM.

3. Block diagram of SLU

SLU is an important component of the SDS system which converts user utterance to a semantic form consisting of dialogue act and slot value pair. Figure 1 shows the block diagram of SLU.

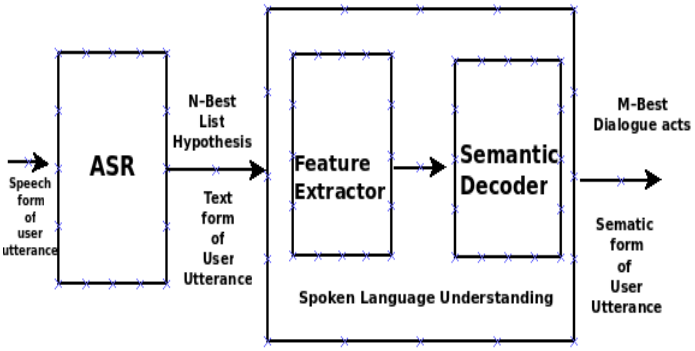


Figure 1: Structure of SLU

SLU plays a very important role in understanding user intent. User utterance is given to the Automatic Speech Recognition (ASR) system which converts speech into text form. The text input is given to SLU. It extracts features from text utterance. The features are fed to a set of classifiers to extract dialogue act and slot-value pair from features. The dialogue act and slot-value pair are combined to form a semantic representation of user utterance. Semantic representation is further analyzed to understand user utterance.

4. Implementation of SLU

Figure 2 represents a semantic tuple classifier for the task of semantic classification which is described in the paper [42]

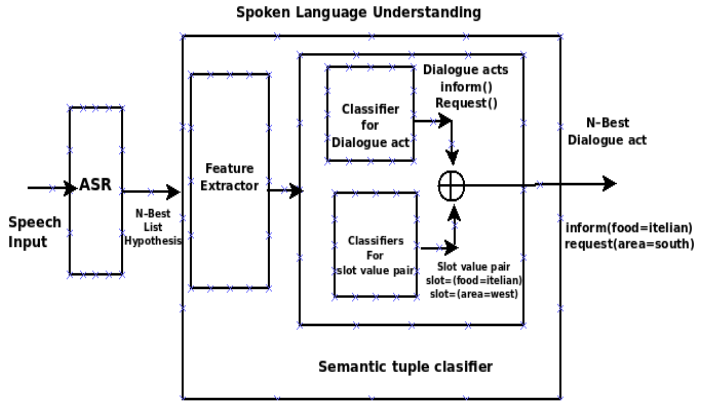


Figure 2: SLU incorporating classifiers for dialogue act and slot-value pair

The speech input is given to ASR. The ASR converts speech input to text form which is the N-best list. The features are extracted from the text. The feature extraction method in paper [42] is extended by incorporating CNN. The deep extracted feature is fed to the classifier. The first classifier classifies user utterance to dialogue act such as to inform (), request (), and bye (). This will help to understand the goal of the user. The next set of classifier identifies the slot-value pair for user utterance. The semantic form of user utterance consists of dialogue act and slot- value pair. Next, eleven classifiers from machine learning are investigated for dialogue act classification and to find slot–value pair. The classifiers investigated on SLU are support vector machine, stochastic gradient descent classifier, AdaBoost classifier, decision tree, relevance vector classifier, random forest, *k*-nearest neighbor classifier, multi-layer perceptron classifier, automatic relevance determination classifier, Gaussian naive Bayes classifier, and variational relevance vector machine. The experimentation was performed with a different amount of training data. The performance of SLU was studied and compared incorporating different classifiers under limited training data on DSTC2 and DSTC3 databases. The experimentation for measuring robustness on limited training data was conducted by training the model on 25% training data and then testing the model on entire testing data. Later experimentation was repeated for 50% training data, 75% of training data, and 100% training data. The evaluation was done on metric Inter Cross-Entropy (ICE), accuracy, and F-score.

5. Results and Discussion

5.1. Datasets

Experimentation is done on two public datasets DSTC2 and DSTC3. The DSTC2 dataset for the restaurant domain. It consists of 1765 dialogues for training data, 441 validation data and 1117 dialogue for testing data. The DSTC3 dataset is for the tourist domain. It consists of a total of 2265 dialogues. The training data used for the experiment is 1265 dialogues and 250 dialogues for validation. The testing data for the experiment is 750 dialogues.

5.2. Evaluation metrics

The framework is evaluated with three metrics: Inter Cross-Entropy (ICE), accuracy, and F-score. The ICE is cross-entropy between the probability density from the confidence and the optimal density given by delta functions at the correct values [43] In the paper, it was demonstrated ICE gives a consistent performance ranking for both the confidence scores and the overall correctness of the system. A lower ICE metric indicates good dialogue confidence performance. Accuracy and F-score are in range 0 to 1 and higher values indicate good performance. ICE metric is highly correlated with the performance of SDS. Accuracy and F-score will reflect the performance of the SLU model. Thus for the evaluation of model in the experiment is done with ICE, f-score, and accuracy.

5.3 Baseline Method

The framework used for experimentation is an extension of the semantic decoder [42] The decoder is extended by incorporating CNN for feature extraction. The extended framework experimented with incorporating eleven different classifiers from machine learning. The framework was evaluated on DSTC2 and DSTC3 datasets. The framework was also investigated for robustness for less training data.

5.4 Experimental setting 1

The first set of experimentation was carried out on the DSTC2 dataset. The experiment was conducted on a deep semantic decoder by incorporating 11 different classifiers. For each classifier, the robustness was evaluated by training the model with different quantity of training data and then testing using the entire set. The different quantity of training data was 1765 dialogues (100% training data), 1323 dialogues (75% data), 882 dialogues (50% training data), and 441 dialogues (25% training data). Testing of SLU for all different quantity of training data was done on 1117 testing dialogues. The classifiers investigated on SLU are SVM, SGD, AdaBoost, DT, RVC, MLP, RF, *k*-neighbourhood, ARD, Gaussian naive Bayes, and VRVM.

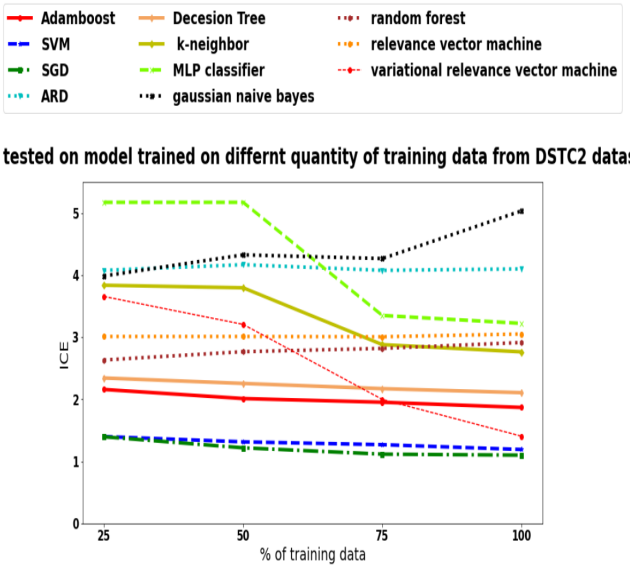


Figure 3: ICE for a model trained on 25% training data, 50% training data, 75% training data, and 100% DSTC 2 training data.

Figure 3 shows the experimental results for ICE measured on the dSTC2 dataset for different quantity of training data. Results show ICE is directly proportional to training data. Almost all classifier ICE was high when training data was less. This indicates the overall quality of distribution was least for 20% of training data and most when 100% training data. Their set of classifiers displaying the best performance are SVM, SGD, and VRVM at 100% data. SVM and SGD displayed the least variation in ICE at the different quantity of training data. The variation in VRVM and MLP was most for different quantity of training data. The set of classifiers displaying the worst performance are ARD, RVC, and MLP. The classifier which had an overall good performance for all quantity of training data and full data is SGD.

The testing accuracy results of the experiment on the DSTC2 dataset are shown in Figure 4. The measurement accuracy is for correct classification of user utterance to semantic form consisting of dialogue act and slot value pair. The classifiers whose topmost semantic hypothesis was correct are SVM, VRVM, decision tree, and SGD. These classifiers were also able to predict the semantic hypothesis for testing data even if the training data was less.

VRVM was not able to maintain its performance for limited training data. Thus SVM, decision tree, and SGD were able to maintain its performance even if training data was scarce. The classifiers displaying the least performance are ARD, RVC, and random forest. The VRVM's testing accuracy measured for 20% training data is 0.827 and for training 100% data it is 0.939. Thus, it is not able to maintain its performance under a scarcity of data. Similarly, MLP and k– neighborhood classifiers were also not able to maintain its performance under data scarceness.

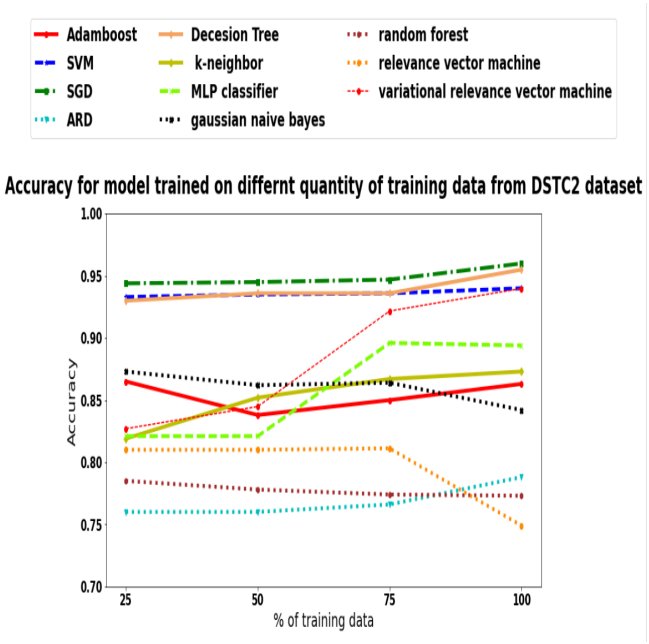


Figure 4: Testing data accuracy for a model trained on 25% training data, 50% training data, 75% training data, and 100% DSTC 2 training data.

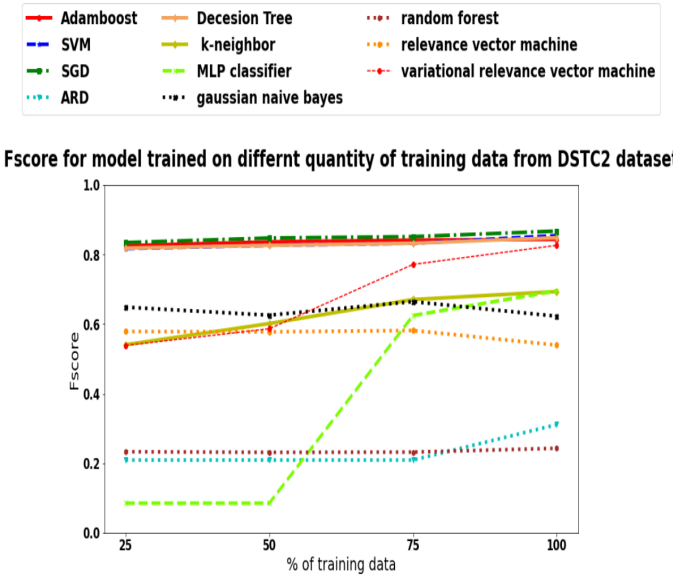


Figure 5: Testing data F-score for a model trained on 25% training data, 50% training data, 75% training data, and 100% DSTC 2 training data.

F-score for testing data DSTC2 is presented in Figure.5. A high value of the F-score will indicate a low false-negative and positive value of the semantic hypothesis. A high F-score was displayed by SLU incorporating Decision Tree, SVM, VRVM, and SGD. SLU's robustness for limited data was increased with these classifiers. VRVM again failed to show robustness for limited data. Low F-score was displayed by random forest and RVC. SLU incorporating MLP also had a lot of variation in F-score for different quantity of training data.

5.5. Experimental setting 2

The second set of experiments was carried out on DSTC3. The different quantity of training data was 1265 dialogues (100% training data), 925 dialogues (75% training data), 600 dialogues (50% training data), and 300 dialogues (25% training data). The testing data consist of 750 dialogues. The eleven classifiers included in an experiment setting 1 is also included in experiment setting 2 as well.

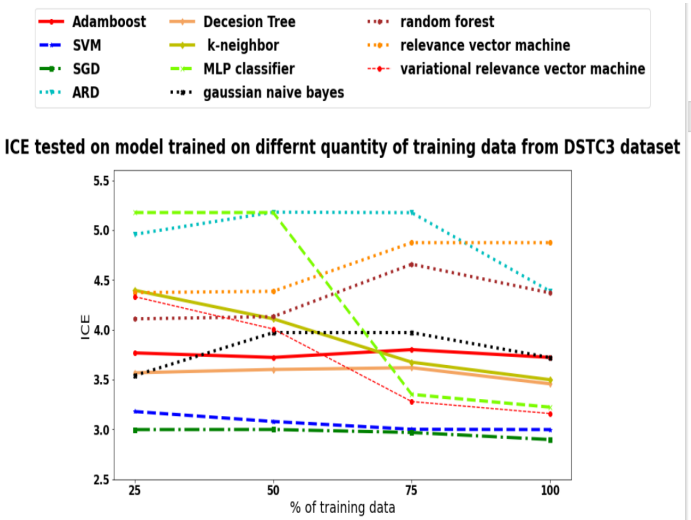


Figure 6: ICE for a model trained on 25% training data, 50% training data, 75% training data, and 100% DSTC 3 training data.

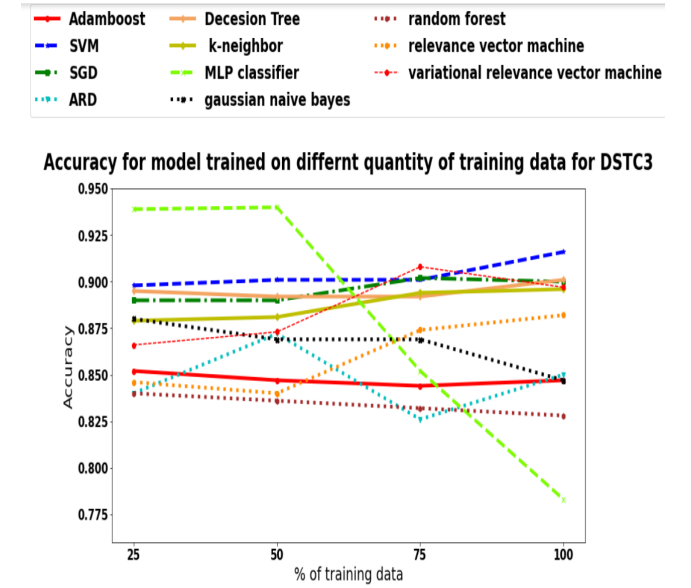


Figure 7: Testing data accuracy for a model trained on 25% training data, 50% training data, 75% training data, and 100% DSTC 3 training data

Figure 6 shows the ICE readings for the different quantity of training data for DSTC3 datasets. The results are similar as found for the DSTC2 experiment. The SLU incorporating the set of

classifiers displaying good results are SVM, SGD, and VRVM. Variation in ICE for different quantity of training data was minimum for SVM and SGD. Variation in ICE was more in VRVM. The SLU incorporating a set of classifiers displaying the worst results are ARD, RVC, and RF. The maximum variation in ICE was displayed by MLP, RVC, and K-neighbourhood.

Testing accuracy for the SLU model for different quantity of training data for the DSTC3 dataset is shown in Figure 7. Again, the results are in line with DSTC2 results. The classifiers which were able to classify user utterance to semantic form most accurately were SVM, SGD, and decision tree. The classifiers performing well under data scarcity are SVM, SGD, and decision tree. The worst performance in terms of data scarcity was showed by the MLP classifier. In terms of accuracy worst performance was shown by ARD, RF, and RVC.

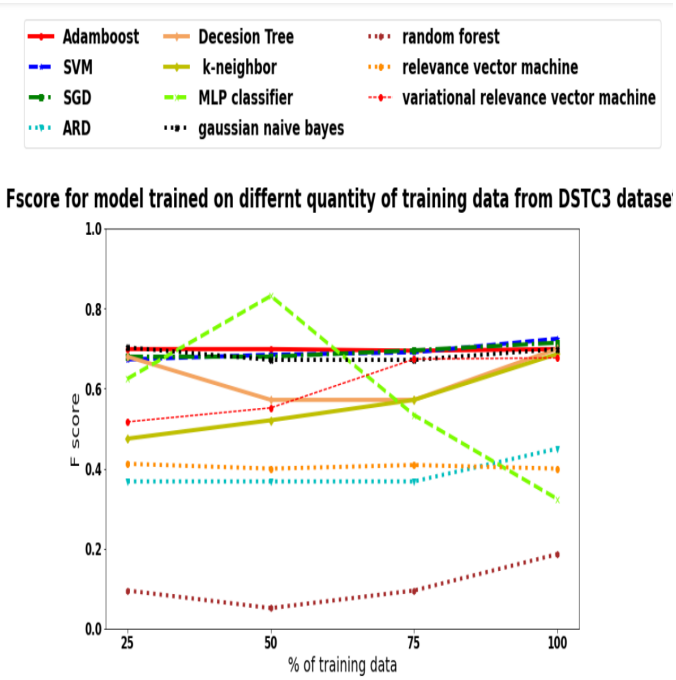


Figure 8: Testing data F-score for a model trained on 25% training data, 50% training data, 75% training data, and 100% DSTC 3 training data.

The results for SLU model performance on the DSTC3 dataset are shown in Figure 8. The results are again similar to the DSTC2 dataset. F-score was high for SLU when incorporated SGD, SVM, VRVM, and decision tree. Low F-score was displayed by SLU when incorporated RVC, RF, and ARD.

Variation in ICE for SGD and SVM when SLU is trained for least and highest train data is 0.296 and 0.209 for the DSTC2 dataset and 0.190 and 0.190 for the DSTC3 dataset, respectively. The least variation in ICE was found for SVM and SGD. Thus, the set of classifiers showing robustness under limited training data for ICE metrics are SVM and SGD. Variation in ICE for MLP and Gaussian naive Bayes when SLU is trained for least and highest train data is 1.95 and 1.45 for DSTC2 and 2.251 and 0.18 for DSTC3 dataset, respectively. The highest variation in ICE was found for MLP and Gaussian naive Bayes. Thus set of classifiers showing the worst result for ICE are MLP and Gaussian naive Bayes.

Table 1: Results on DSTC2 dataset

S.No	Classifier	ICE	Accuracy	F-score
1	Adamboost	1.861	0.863	0.848
2	SVM	1.191	0.940	0.855
3	SGD	1.100	0.960	0.868
4	ARD	4.02	0.788	0.311
5	Decision Tree	2.107	0.955	0.849
6	k-neighbor	2.765	0.873	0.694
7	MLP classifier	3.224	0.894	0.696
8	gaussian naive Bayes	5.03	0.842	0.623
9	random-forest	2.916	0.773	0.243
10	RVC	3.053	0.799	0.540
11	VRVM	1.405	0.939	0.827

Table 1 shows the results for the DSTC2 dataset. Table 2 shows results for the DSTC3 dataset for the tourist domain. Comparison of accuracy, F-score and ICE from table 1 and table 2 can be studied.

Table 2: Results on DSTC3 dataset

S.No	Classifier	ICE	Accuracy	F-score
1	Adamboost	3.723	0.847	0.699
2	SVM	3.180	0.916	0.725
3	SGD	2.999	0.900	0.715
4	ARD	4.380	0.850	0.450
5	Decision Tree	3.458	0.901	0.700
6	k-neighbor	3.499	0.896	0.687
7	MLP classifier	3.780	0.783	0.324
8	gaussian naive Bayes	3.721	0.847	0.699
9	random forest	4.179	0.828	0.186
10	RVC	4.375	0.884	0.400
11	VRVM	0.311	0.897	0.677

The results from table 1 and table 2 are analyzed to find the classifier performing best and worst. The lowest ICE results are obtained by SLU incorporating SGD and SVM when evaluated on both DSCT2 and DSTC3 datasets. The highest accuracy is obtained by SLU incorporating SVM and SGD for DSTC2 and DSTC3. The highest F-score is obtained by SLU incorporating SVM and SGD for DSTC2 and DSTC3.

6. Conclusion

The work extends the classification algorithm in machine learning to the semantic decoder in SLU. Experimental results on both the DSTC2 database for the hotel domain and the DSTC3 database for the tourist domain had similar findings. The SLU's robustness for data scarcity was measured by three evaluation

metrics: accuracy, ICE, and F-score. The SLU model was trained at 25% data (least training data), 50% data, 75% data, and 100% data (highest training data). The work indicates three important findings from the results obtained from experimentation. The first finding is that both the databases are displaying variation in SLU performance for different classifiers indicating SLU performance is influenced by the selection of proper classifiers. The second finding is that there is variation in ICE for least and maximum training data for all the classifiers indicating ICE is highly influenced by the quantity of training data used to train the SLU model. The third finding is the amount of variation in ICE displayed by each classifier is different indicating the robustness of SLU to work under data scarceness is influenced by the right choice of the classifier.

Variation in ICE, accuracy and F-score for SLU incorporating SVM is 0.209, 0.007, 0.035 for DSTC2 database and 0.190, 0.018, 0.052 for DSTC3 database, respectively. Variation in ICE, accuracy, and F-score for SLU incorporating SGD is 0.296, 0.016,0.033 for DSTC2 database and 0.190, 0.010, 0.035 for DSTC3 database, respectively. Thus, the overall performance of SLU, maintaining good performance for all the three evaluation metrics, is by incorporating SVM and SGD. The overall low performance is displayed by SLU incorporating ARD, RVC, MLP, Gaussian naive Bayes, and RF.

References

[1] S. Jagdale, M. Shah, "Evaluation of Stochastic Gradient Descent and Spoken Language Understanding," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), 1–4, 2019, DOI:10.1109/ICNTE44896.2019.8946073

[2] A. Torres-Rodríguez, D. Morillón-Gálvez, D. Aldama-Ávalos, V.H. A. Basavanhally, S. Doyle, A. Madabhushi, "Predicting classifier performance with a small training set: Applications to computer-aided diagnosis and Rutgers, The State University of New Jersey Department of Biomedical Engineering," Biomedical Engineering, **1**, 229–232, 2010, DOI:10.1109/ISBI.2010.5490373

[3] A. Basavanhally, S. Viswanath, A. Madabhushi, "Predicting classifier performance with limited training data: Applications to computer-aided diagnosis in breast and prostate cancer," PLoS ONE, **10**(5), 1–18, 2015, DOI:10.1371/journal.pone.0117900.

[4] V.K. Vijayan, K.R. Bindu, L. Parameswaran, "A comprehensive study of text classification algorithms," 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017, 2017-Janua, 1109–1113, 2017, DOI:10.1109/ICACCI.2017.8125990

[5] G. Forman, I. Cohen, "Learning from little: Comparison of classifiers given little training," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3202, 161–172, 2004, DOI:10.1007/978-3-540-30116-5_17.

[6] A.S. Altheneyan, M.E.B. Menai, "Naïve Bayes classifiers for authorship attribution of Arabic texts," Journal of King Saud University - Computer and Information Sciences, **26**(4), 473–484, 2014, DOI:10.1016/j.jksuci.2014.06.006

[7] S. Pournouri, S. Zargari, B. Akhgar, "An Investigation of Using Classification Techniques in Prediction of Type of Targets in Cyber Attacks," Proceedings of 12th International Conference on Global Security, Safety and Sustainability, ICGS3 2019, 202–212, 2019, DOI:10.1109/ICGS3.2019.8688266.

[8] N. Wang, X. Zhao, Y. Jiang, Y. Gao, "Iterative metric learning for imbalanced data classification," IJCAI International Joint Conference on Artificial Intelligence, 2018-July, 2805–2811, 2018, DOI:10.24963/ijcai.2018/389

[9] G.M. Foody, M.B. McCulloch, W.B. Yates, "The effect of training set size and composition on artificial neural network classification," International Journal of Remote Sensing, **16**(9), 1707–1723, 1995, DOI: 10.1080/01431169508954507.

[10] S.S. Yadav, S.M. Jadhav, "Deep convolutional neural network-based medical image classification for disease diagnosis," Journal of Big Data, 6(1), 2019, DOI: 10.1186/s40537-019-0276-2.

[11] H. Sharma, S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," International Journal of Science and Research (IJSR), **5**(4), 2094–2097, 2016, DOI:10.21275/v5i4.nov162954

[12] N. Patel, S. Upadhyay, "Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA," International Journal of Computer Applications, **60**(12), 20–25, 2012, DOI: 10.5120/9744-4304.

[13] J.R. Quinlan, "Induction of decision trees," Machine Learning, **1**(1), 81–106, 1986, DOI: 10.1007/bf00116251.

[14] P.L. Tu, J.Y. Chung, "A new decision-tree classification algorithm for machine learning," Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 1992-Novem, 370–377, 1992, DOI:10.1109/TAI.1992.246431

[15] H. Sharma, S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," International Journal of Science and Research (IJSR), **5**(4), 2094–2097, 2016, DOI:10.21275/v5i4.nov162954.

[16] H. Sharma, S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," International Journal of Science and Research (IJSR), 5(4), 2094–2097, 2016, DOI:10.21275/v5i4.nov162954

[17] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, D.K. Basu, "An MLP based approach for recognition of handwritten 'Bangla' numerals," Proceedings of the 2nd Indian International Conference on Artificial Intelligence, IICAI 2005, (March), 407–417, 2005.

[18] M. Abdar, N.Y. Yen, J.C.S. Hung, "Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees," Journal of Medical and Biological Engineering, **38**(6), 953–965, 2018, DOI: 10.1007/s40846-017-0360-z.

[19] T. Windeatt, "Ensemble MLP classifier design," Studies in Computational Intelligence, **137**(5), 133–147, 2008, DOI:10.1007/978-3-540-79474-5_6.

[20] B.M. Gayathri, C.P. Sumathi, "An Automated Technique using Gaussian Naive Bayes Classifier to Classify Breast Cancer," International Journal of Computer Applications, **148**(6), 16–21, 2016

[21] R. Cordoba, L.F. D'Haro, F. Fernandez-Martinez, J.M. Montero, R. Barra, "Language identification using several sources of information with a multiple-Gaussian classifier," International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007, **3**(May 2014), 1913–1916, 2007

[22] P. Guo, Y. Jia, M.R. Lyu, "A study of regularized Gaussian classifier in high-dimension small sample set case based on MDL principle with application to spectrum recognition," Pattern Recognition, **41**(9), 2842–2854, 2008, DOI:10.1016/j.patcog.2008.02.004.

[23] M. Morchid, R. Dufour, P.M. Bousquet, M. Bouallegue, G. Linares, R. De Mori, "Improving dialogue classification using a topic space representation and a Gaussian classifier based on the decision rule," ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing - Proceedings, 126–130, 2014, DOI:10.1109/ICASSP.2014.6853571.

[24] H. El Sorady, A. Shoukry, S. Bassiouny, "Improving the parametric Gaussian classifier using neural networks," National Radio Science Conference, NRSC, Proceedings, 1996, DOI:10.1109/nrsc.1996.551118.

[25] C. Lee, E. Choi, B. Baek, C. Yoon, "Analyzing the Gaussian ML classifier for limited training samples," International Geoscience and Remote Sensing Symposium (IGARSS), **5**(C), 3229–3232, 2004, DOI:10.1109/igarss.2004.1370389

[26] Y.G. Zhao, F. Zheng, Z. Song, "Hand Detection Using Cascade of Softmax Classifiers," Advances in Multimedia, 2018, 2018, DOI:10.1155/2018/9204854

[27] D. Song, C. Lu, J. Ma, "Gearbox fault diagnosis based on VMD-MSE and AdaBoost classifier," Vibroengineering Procedia, **14**, 120–125, 2017, DOI:10.21595/vp.2017.19242.

[28] M.L.R. Chandra, B.V. Kumar, B. Sureshbabu, "IoT enabled home with smart security," 2017 International Conference on Energy, Communication, Data Analytics, and Soft Computing, ICECDS 2017, 1193–1197, 2018, DOI:10.1109/ICECDS.2017.8389630.

[29] F. Wang, D. Jiang, H. Wen, H. Song, "AdaBoost-based security level classification of mobile intelligent terminals," Journal of Supercomputing, 75(11), 7460–7478, 2019, DOI:10.1007/s11227-019-02954-y.

[30] S. Wu, H. Nagahashi, "Analysis of generalization ability for different AdaBoost variants based on classification and regression trees," Journal of Electrical and Computer Engineering, 2015, 2015, DOI:10.1155/2015/835357.

[31] C.H. Zhao, Y. Zhang, Y. Wang, "RVM classification of hyperspectral image based on wavelet kernel function," 2011 International Conference on Multimedia Technology, ICMT-2011, 3647–3650, 2011, DOI:10.1109/ICMT.2011.6002159.

[32] L. Wei, Y. Yang, R.M. Nishikawa, M.N. Wernick, A. Edwards, "Relevance vector machine for automatic detection of clustered micro calcifications," IEEE Transactions on Medical Imaging, **24**(10), 1278–1285, 2005, DOI:10.1109/TMI.2005.855435.

[33] K. Pal, B. V. Patel, Emotion Classification with Reduced Feature Set SGD Classifier, Random Forest, 2020, DOI: 10.1007/978-981-15-6648-6.

[34] F. Kabir, S. Siddique, M.R.A. Kotwal, M.N. Huda, "Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier," Proceedings - 2015 International Conference on Cognitive Computing and Information Processing, CCIP 2015, 0–3, 2015, DOI:10.1109/CCIP.2015.7100687.

[35] J. Ding, J. Wang, X. Kang, X.H. Hu, "Building an SVM Classifier for Automated Selection of Big Data," Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017, 15–22, 2017, DOI:10.1109/BigDataCongress.2017.12.

[36] N.V. Kumar, P.V. Kumar, K. Pramodh, Y. Karuna, "Classification of Skin diseases using Image processing and SVM," Proceedings - International Conference on Vision Towards Emerging Trends in Communication and Networking, ViTECoN 2019, 1–5, 2019, DOI:10.1109/ViTECoN.2019.8899449.

[37] T.D. Nguyen, T.D. Do, M.H. Le, N.T. Le, W. Benjapolakul, "Appliance Classification Method Based on K-Nearest Neighbors for Home Energy Management System," 2019 1st International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics, ICA-SYMP 2019, 53–56, 2019, DOI:10.1109/ICA-SYMP.2019.8646278.

[38] S. Taneja, C. Gupta, K. Goyal, D. Gureja, "An enhanced K-nearest neighbor algorithm using information gain and clustering," International Conference on Advanced Computing and Communication Technologies, ACCT, 325–329, 2014, DOI:10.1109/ACCT.2014.22.

[39] M.S. Alam, S.T. Vuong, "Random forest classification for detecting Android malware," Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-IThings-CPSCom 2013, 663–669, 2013, DOI:10.1109/GreenCom-iThings-CPSCom.2013.122.

[40] S. Hwang, M.K. Jeong, "Robust relevance vector machine for classification with variational inference," Annals of Operations Research, **263**(1–2), 21–43, 2018, DOI: 10.1007/s10479-015-1890-9.

[41] D. Kropotov, D. Vetrov, L. Wolf, T. Hassner, "Variational relevance vector machine for tabular data," Journal of Machine Learning Research, **13**(2001), 79–94, 2010.

[42] M. Henderson, M. Gasic, B. Thomson, P. Tsiakoulis, K. Yu, S. Young, "Discriminative spoken language understanding using word confusion networks.," SLT, 176–181, 2012, DOI:10.1109/SLT.2012.6424218.

[43] B. Thomson, "Statistical methods for spoken dialogue management," Ph.D. Cambridge University, (December), t, 2009, DOI: 10.1007/978-1-4471-4923-1_1.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 469-477 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Spot Toyota: Design and Development of a Mobile Application for Toyota's Promotion Actions to the Young Audience

Nuno Martins[1,*], Joel Enes[2]

[1]*Polytechnic Institute of Cavado and Ave / ID+, 4750-810 Barcelos, Portugal*

[2]*Polytechnic Institute of Cavado and Ave, 4750-810 Barcelos, Portugal*

ARTICLE INFO

Article history:
Received: 21 April, 2020
Accepted: 29 May, 2020
Online: 20 June, 2020

Keywords:
Mobile App
UI & UX Design
Digital Design
Communication Design
Toyota

ABSTRACT

This project aims to demonstrate the importance that digital media can have to the development of loyalty programs, namely in creating empathy and proximity relationships between brands and their target audience: young people. This study consisted of the creation of a digital platform for Toyota Portugal, named Spot Toyota, to communicate actions promoted by the car brand, especially during music festivals. With the support of advertising agency Caetsu, this mobile application was developed to bring the brand closer to a younger audience – festival fans – with potential interest in two Toyota fleet car models: Aygo and C-HR. Through strategies typical of loyalty programs, such as the awarding of vouchers or coupons, the accumulation of points or winning prizes, a system was produced with the main focus on attracting users to the platform in a continuity perspective. The working process of this investigation resulted in the design of a smartphone application, based not only on the analysis of other examples present in the market but also on the understanding of crucial subjects such as loyalty programs, UX and UI design, application of personas models, creation of wireframes and workflows, and development of usability tests.

1. Introduction

This project arose from a proposal, from the Portuguese advertising agency Caetsu, to develop a digital application (app) for the Toyota car brand.

This app, entitled "Spot Toyota", was originally created in July 2018, by an internal department of Toyota Portugal. The main purpose of the app was to create connections of proximity and empathy between the Toyota brand and the younger audience (often at music festivals). Thus, it was intended that this app worked as a loyalty product and not as a direct way of promoting or selling cars.

The study initially focused on searching mobile applications in the automotive sector. However, since no considerable supply was found in this sector, other platforms such as Cartão Continente, MB Way, McDonald's and Galp EvoDrive, designed under the same assumptions as Spot Toyota, were taken into account. In addition to the consumer-perceived value, other key concepts for research were also analysed, such as loyalty programs, usability, UX and UI design.

After understanding the problem, the work moved on to the platform design phase. In this sense, essential methods and guidelines were put into practice in the process of designing the mobile application. Realizing the value that this solution could provide to the market was one of the first steps. Then, the project advanced to the creation of *persona* models in order to understand the pattern of the platform users.

Subsequently, the phase of developing information architecture began, which would be fundamental to the creation of the app *wireframes* and *workflows*. In this step, the screen sketches were conceived, distributing the graphic elements according to the importance and hierarchy of the selected information. To the design, were observed the user actions in the intended tasks accomplishment.

The platform creation evolved to the interface design phase.

Following the brand guidelines, the design of the screens began with a more considered and rigorous focus on visual issues, the

*Corresponding Author: Nuno Martins, nunomartins.com@gmail.com

graphic elements distribution, the interface communication, the color selection, shape, and information. This process aimed to create a credible product and offer the best user experience. Finally, the phase of usability tests was initiated, the main purpose of which was to detect problems in the interface and develop alternatives to improve its performance.

2. General and specific objectives

The main objective of this project arose from a real need on the part of Toyota Portugal to create a digital platform that would help the brand get closer to one of its target segments, the young audience. In this sense, this platform was created for an audience potentially interested in two models of Toyota's fleet in the Portuguese market: Toyota Aygo and Toyota C-HR.

Consideration was given to developing a mobile application (app) as a privileged channel to communicate for the brand's actions with the defined target, in order to capture a maximum of leads; to support lead conversions in effective customers; and to retain both customers (buyers) and fans of the brand (not necessarily buyers, but individuals capable of influencing their close circle).

As a result of the introduction and implementation of RGPD (General Data Protection Regulations), Toyota was forced to efface the contact data the company used to communicate brand information and actions. The app to be developed, thereby, in addition to trying to regain lost contacts, would mainly serve to establish and consolidate a regular empathy with the public.

The goal was not only to develop a conceptual study of a mobile app through good design practices, but also to build a tangible product, with added value, that could be implemented and be a reference in this automotive market [1].

To achieve the general objective (sustainable development of an app), a set of specific objectives was addressed:

- Identifying, understanding and systematizing the goals proposed by Toyota for the app: Through the meetings held with Caetsu, the company responsible for monitoring the project, a detailed analysis of the work context was prepared. For the app development, it became necessary to understand and systematize all the raw information, transforming it into raw material to start the project.
- Listing the added-value offers that Toyota can make through the app (discounts, points, coupons, among others): As the mobile app had well-defined objectives, it was essential to understand the type of functionalities and specific offers that it would have to contemplate. In order for the brand to be successful in approaching its target audience, the app would have to be perceived as an offering of added value to its users. From this point of view, the focus shifted to examining loyalty program techniques, such as discounts, points and coupons, and designing attractive means to encourage the use of the app.
- Designing an app that contains relevant and useful information and functionalities for the user, with an appealing and intuitive interface, culminating in a high degree of usability satisfaction.

3. Methodologies

In an initial phase, an in-depth inquiry of the problem was carried out through relevant data collection and analysis.

A firststage of information gathering was carried out in a meeting with the Caetsu team, the agency responsible for monitoring the Toyota project, in which the main objectives, needs and expectations concerning the app were discussed. Correspondingly, schemes and mental maps, which defined the essential factors to conceive the platform's contents, were produced. It became crucial to understand the typology of mobile application offerings in the automotive sector, but above all, other applications built on the same assumptions as Spot Toyota. Analysing these examples, we were able to understand the range of features they offer as a reference for the work to be developed in the app under study. Having said that, it was essential to understand what is a loyalty program, how it works, and how the app could use these strategies to bring the user closer to the platform.

In a second phase, and after this analysis of the state of the art, the work was developed with special focus on interface design, experience and usability heuristics.

In the first stage of this second phase, the *personas* study began, which focused on the construction of profiles based on real users. Mirroring the target, the *personas* represented the various types of users that can use the developed app. Then, the project proceeded to the information architecture phase. The content present in the app was mapped, as well as the paths and links between the various stages of interaction within the platform. This work would lead to the construction of the app *wireframes* and *workflows*. The project sketches were drawn using basic representations to facilitate the perception and thinking of the interface division. In this follow-up, the *workflows* were fundamental to organize the *wireframes*, and to perceive step by step (in an illustrated way) the user's navigation intentions.

Later, the project evolved to the graphical construction of the interface. Respecting the Toyota brand identity standards, the screens design began. The work focused on information hierarchy, readability, color and contrast. On this step, there was a rigorous focus on visual issues in order to create a credible product and offer the best possible user experience.

Finally, based on user feedback, the usability testing phase was carried out. Through methods to test and observe user behavior, it was possible to detect interface problems and develop enhanced alternatives. These improvements meant a better user's perception, and easier and faster navigation within the contents.

After this process of app development, data collection, review and feedback from users, started the implementation of the programming language, a task that was assigned to the Caetsu team. Once finished, the app became available on the marketplaces to be installed and used according to the defined purposes.

4. State of the Art

4.1. Loyalty programs

Loyalty programs, in general, are sales techniques that guarantee consumers benefits from their purchasing behavior, while allowing sellers and service providers to build a long-term relationship with the customer [2]. These consumer benefits can take the form of discounts, cash, bonuses or even access to special

services such as in the case of magazines and newspapers that offer exclusive content for loyal readers.

With the development and investment of the technology industry in the smartphone and mobile applications market, a favorable environment was created for the growth and evolution of loyalty programs. Mobile applications provide a highly segmented channel for two-way communication, becoming powerful drivers of customer loyalty.

"Mobile applications are the main channel for most, and it is clear that they present themselves to businesses as an unprecedented opportunity to interact directly with customers. They are easily accessible through an icon on the main screen; they do not require URL's and currently provide the best experience on mobile devices." [3]

Spot Toyota, despite not presenting itself as a service aimed at converting sales or with the objective of becoming exclusively a loyalty program, sought important guidelines and tools for creating empathy with its users in these marketing techniques. It was in this context that it became essential to leverage all the advantages of a mobile application (app) to boost and develop direct communication channels with the defined target audience. For instance, with a mobile application, the process of sending individual and segmented messages, known as "push notifications" [1] is simpler. With a great distance between the company and the user, this conversation mode can make a difference in terms of gaining customer loyalty.

With people spending much of their time on smartphones and applications, companies have the opportunity to increase their customer loyalty by offering a personalized application experience and making the customer feel valued [3].

4.2. Apps in the automotive sector

After an analysis focused on the offer of mobile applications in the automotive industry, especially in the Portuguese market and in Google Play (Android) and App Store (iOS), it was understood that mobile applications can be divided into three major typologies: Lifestyle; Remote Control; and Utilities.

Some brands that can be considered in the Premium segment, such as Volvo, BMW and Mercedes, make a digital magazine available to their customers in a mobile application format. With a refined and high-quality graphic presentation, they approach a well-defined target, giving access to exclusive news about the brand, and offering a wide range of information about services and accessories, inspirational articles, films, photography and links to explore their vehicles universe.

In the second typology, it is possible to include an entire range of applications aimed at controlling the car digital system itself. In these cases, after pairing the smartphone with the vehicle, certain levels of control access are given to the system: it is possible to access and monitor the GPS and select options on the music player, among some functionalities. An example in this regard is the Volkswagen Media Control app.

At the last typology, we can inscribe the automotive industry applications with a great presence in stores, distributed by countries or markets. This is the class of applications for car daily control, in which it is possible to: monitor the vehicle condition; manage duration, distance and consumption of trips; and record the entire maintenance plan, miles and guarantees.

Although applications with the purpose of increasing loyalty and offering value to attract new customers are recurrent in the digital market, we may declare can be in that the automotive sector has not yet adopted this strategy, addressing all its offer to existing vehicle owners. Given the potentially disruptive advantages granted by modern-day technology, such approach can be considered a misuse.

4.3. Reference Apps in the area of loyalty

In the universe of mobile applications, there is a wide range of examples to observe. However, this part of the research has not focused on the automotive sector because, as we have seen before, there has not been much supply in this area so far. Therefore, the approach focused on other business sectors in Portugal with models that could serve as guidance examples for Spot Toyota.

The "Cartão Continente" app, belonging to one of the major Portuguese hypermarkets' chains, has more than 500K installations in Google Play (Google Applications Store) and it is a reference app in consumer loyalty. In a clear graphic approach and adequately mirroring the brand identity, it offers a set of tools that capture the user's interest and promote its use. It makes available numerous discount coupons and offers when buying products and services. Similarly, the "Galp EvoDrive" app, from a fuel supplier, is identified by the constant providing of discount coupons to users in order to attract them to the shops. A relevant feature of this application is the possibility of sharing discount codes between users. This way, the brand benefits through P2P (person-to-person) advertising and invites further usage of the app.

With a graphic interface based on a colorful chromatic palette, which can be assumed to be dedicated to the young audience, the McDonald's app stands out. It also makes use of loyalty tools, such as discount coupons on certain products and the accumulation of points for purchases made.

In the services context, "Via Verde", from a company that provides an electronic toll collection system, presents another type of benefit to consumers. In this case, the attribution of points and discounts is not made through what the brand sells, but through advantages in its entire network of travel and hotel partners.

As another example within this particular selection of reference mobile applications, is "MB Way", an app that offers a product with interesting tools for attracting users. Although it is designed mainly to facilitate money transfers, it uses playful interfaces (gamification), creating an environment for the users to engage with recreational features and game dynamics, both individually and in connection with other users. According to Cook [4], when customers are driven to participate, or to act in some way, they are more likely to feel that they will be rewarded for their action,

[1] "Push Notification" is a notification that the user receives through a smartphone application, tablet or browser without requesting it.

thereby reinforcing the interaction. The "MB Way" is comprehensive when it comes to loyalty techniques, however, certain usability issues are evident in terms of the clarity of perception regarding all possible interactions.

Regarding the interfaces of the applications presented, it is possible to observe a careful and contained use of the most expressive and complex graphic elements, creating a simplified and intuitive language. Mostly designed on clear backgrounds, the interfaces show concern towards clarity in terms of comprehending all features and interactions. Due to the large volume of information and the fact that the success of these applications depends on guarantying a good user experience, it can be considered that all these platforms are designed with special focus on usability and accessibility issues. As Norman [5] states, "Good design requires good communication, especially from the machine to the person, indicating what are the possible actions, what is happening and what is about to happen".

With these examples, we understood how the strategies and functionalities that these applications use to enhance the loyalty of their users are put into practice. We can consider that the component of discount coupons and accumulation of points became a common and transversal practice in almost all applications analysed, highlighting the importance of rewarding the user for using the app. In accordance, Spot Toyota used these proven models in the market as benchmarks to design its own features set.

5. App development

This project was developed from an already existing product that was still basic, with a strong need for development.

After discussing with Caetsu and Toyota Portugal managers what was intended for the app, a first prototype was developed. The prototype contemplated all the functionalities and needs discussed in the first meeting between the research team and the Caetsu and Toyota Portugal managers.

Although it was developed on a poorly sustained basis, with a lack of definition of *workflows* and critical paths, the first prototype was fundamental for the second meeting with Caetsu and Toyota. It was the basis for discussion towards defining the design and functionalities of the final project.

The prototype presentation contributed to a more objective discussion about the app's problems and to the suggestion of new improvement ideas, namely: the design of a graphic interface with lesser red tone; the change of the typographic font to the official Toyota font (the Toyota Text Font); a greater presence of the Toyota Spot logo on the various screens; the redesign of some tab bar icons as well as some buttons.

After this second meeting, the entire app structure design was started. After collecting all the first prototype inputs, the final *workflows* were designed, the *personas* created, the use cases studied and the functionalities to contain were listed. Then, the redesign phase of the *wireframes* has started, based on the corrections discussed.

To improve and understand the interaction with the platform, usability tests using real users were developed. With the collection of their feedback, it was possible to apply the necessary adjustments and changes to evaluate and verify the user experience and interface design, thus solving most usability problems encountered [6-10].

It is important to note that all changes and decisions made during this whole process were constantly communicated to those responsible for monitoring the project, with the final validation, as well as the definition of other necessary corrections or changes [11,12].

5.1. Personas

When creating a digital product such as an app or a website, it is essential to know who will use it. In this sense, and to anticipate the problems that may arise from its use, it is necessary to understand the user and his needs [13].

The *Personas* are fictional characters, created based on a market analysis. They represent the various types of users who can have a similar comportment when using the service or product developed.

This process of idealization leads us to understand the needs, experiences, behaviors and goals of users [14]. *Personas* simplify the design process, in the perspective that all processes are developed on the basis of ensuring the best user experience for the identified target audience [15].

All this survey of the user's characteristics, needs and objectives is translated into relevant information, which helps us to calculate the route the user takes on the platform.

Through the development of the *personas*, we try to create empathy with the target users and correspond to their needs, which are tested in prototypes. The creation of these profiles is based on real users, i.e. all the conversion of needs into effective functionalities of the application derives from the real needs of users [16].

The development of the *personas* for this project (Table 1) has gone through study and contact with different people who fit the target audience of Spot Toyota. The insights provided by Caetsu and Toyota Portugal, who had already gathered detailed information about the defined target market, along with a series of informal conversations with recurring festival attendees took by the researchers, made it possible to embody the *personas* portraits.

Profiles were constructed based on the same interests and assumptions: a young, active audience, who had already participated in music festivals or had shown interest in them [13].

It was added some personal information that summarized what matters in the *persona* who relates to the product.

To add a demographic profile describing the personal and professional background, the user environment, and the psychographics, like interests, motivations, and pain points, was also important.

The following step was to attribute the end goal, the motivation factor that inspires the action. This determines what the *persona* wants or needs to fulfil.

To conclude, it was created a scenario to describe how the *persona* interacts with the product in a particular context to achieve his end goals.

Table 1: Analysis grid of different *personas*.

RAQUEL 18 years old	**PEDRO** 25 years old	**JOANA** 32 years old	**TIAGO** 20 years old	**ABEL** 26 years old
Fan of indie rock music; about to enter university education; loves cats.	Addicted to live concerts; intern in a marketing studio; fan of promotions.	Loves summer festivals; loves photography ; works in an architecture studio.	Camping enthusiast; studying computer engineering; motorsports fan.	Enjoys a good concert; Guitar teacher; Environmentalist.
Use case: Raquel, for financial reasons, participates in all competitions and activities in which she can win tickets to concerts and festivals. Uses apps and groups in social networks to be up to date with all the news.	**Use case:** João adheres to all services that earn him points and gets rewarded for it. If there are no associated costs, he installs all the applications that may bring him advantages in the short or medium term.	**Use case:** Joana is addicted to new technology and all that is practical and time efficient. She uses digital versions of tickets for concerts, trips, and offers.	**Use case:** Tiago, as a fan of the automotive world, ardently follows brands and their activities, to have the possibility to see new car models first hand, and to be able to go for test-drives.	**Use case:** Ricardo, not a big fan of digital media, uses some applications that his friends impose upon him to validate codes that they forward. Since he is indifferent to them, he ends up helping them.
Objectives: - Follow news and campaign releases; - to receive notifications to the second of all news.	**Objectives:** - Earn points to be able to convert into Toyota prizes. - Be rewarded for daily use of the app.	**Objectives:** - Use festival coupons to enjoy Toyota rides; - Participate in brand actions just for the pleasure of the challenge.	**Objectives:** - Be informed of Toyota brand actions; - Know what actions will be undertaken where, and which campaigns will be associated with automobiles.	**Objectives:** - Validate codes sent by friends; - Send codes for friends to validate.
Frustrations: - Being afraid of losing some important information or launching a hobby. - Having to look for information on scattered platforms.	**Frustrations:** - Having difficulty converting points into advantages; - Knowing that if would require a long time for any reward to be effectuated.	**Frustrations:** - No quick and intuitive access to coupons; - Not being able to access and utilize through smartphone.	**Frustrations:** - Access outdated information; - Unclear and disorganized information.	**Frustrations:** - Time-consuming and complicated process; - Need to go through complex registration processes in order to access the app.

5.2. Information architecture

After studying the competitors and analysing the *personas*, we moved on to the project information architecture phase. This process became decisive for the final product development, since it was the main stage of content organization for the app.

Based on the first conversations with Caetsu, and taking into account the existing product (Spot Toyota in browser version), the design and structuring of the flowchart began.

In order to meet the brand's purposes, it was fundamental to set up reward methods and techniques. The intention was to encourage the use of the app and consequently connect the user to the environment that the brand had imagined for the project. This system was developed based on the good references in the market and the cases analysed in the state of the art section of this article.

Thus, based on the research, it was defined that the points would be awarded as per the following:

- **For the daily use of the app:** once a day, and after the user enters the app, points are automatically generated and added to the balance of the user presented in his/her personal area.
- **For the use of vouchers or coupons:** upon the brand's interest in the actions where it made itself present, vouchers or coupons can be awarded to the user, which, after use and validation, turn into more points.
- **By invitation:** in the personal area of the app, there is a section called "Invite friends". Here, and by sharing a code with other contacts, a user can add points after at least one of those contacts has validated the code sent.
- **By inserting a code:** in precisely the same place mentioned in the previous point, there is another section entitled "Insert promotional code". If the user has received a sharing code from a contact or by some promotional action of the brand, he/she gains points after the insertion and validation by the app of that same code.

In addition to these point attribution methods, another specific system was developed for the project: through the interaction with the app, the user progresses on a scale of five levels. On reaching the last stage, he/she is rewarded and goes back to the initial step. With this process, the user is able to add points always with a concrete objective.

We concluded that flowchart had translated into an added value for the perception of the whole project hierarchy and organization, as well as a fundamental work base for the construction of the corresponding *wireframes*.

5.3. Wireframes

The next phase of the app construction was the wireframe design. We can define this concept as a low fidelity representation in the initial design phase of a certain visual product. It is the outline of the structure of a digital project, using basic representations, such as lines or circles, in order to facilitate the ideation of the interface division. Normally, *wireframes* are used to arrange the content and functionality of a page, reflecting the user needs and paths of interaction. Because it is a sketch, this type of work allows the solutions presented to be subjected to constant changes, until a suitable result of the app is obtained in the final interface design [17,18].

For this project, although the product would eventually be made available cross-platform (including Android), we chose to develop the app in iOS, the mobile operating system for iPhones

and iPads from Apple Inc. The choice was made in line with the convenience and experience of the researcher responsible, because it was the system with which he was best familiar. The knowledge of its usability standards, through the daily use of the platform, thus made the project construction process more conscious and reliable.

5.4. Workflows

Workflows are sequences of steps a user goes through to accomplish a certain task. In this type of project, *workflows* are fundamental for the *wireframe's* organization, in order to understand, step by step, the user's navigation intentions. This is the flow that is established for the tasks' realization. For the *workflows* development we use the Sketch software for aiding in the rapid prototyping of the various screens, and testing all the interaction between *wireframes* and *workflows*.

In this sense, we highlighted some of the critical concerns of the *personas* created, in order to perceive the applicability of the designed flows and how these minimized the expected constraints. Based on the cases of Joana, Pedro and Ricardo (Table 1), we developed the *workflows* related to the tasks that could be most relevant to their concerns.

For example, in Joana's case, one of her frustrations was the fear of not having access to the coupons quickly and intuitively. Since the beginning of the project, and being one of the main features of this app, the access to vouchers and coupons would have to be as upfront and evident as possible, creating a high degree of user satisfaction.

To understand how this *workflow* works, we present the following example: "Toyota Rides" is one of the activities practiced by Toyota at summer festivals. Through the fleet of cars available at festivals, hitchhikers are given rides between points of interest and the concert venue. This app also appears to speed up all registration and validation processes at time of enjoying the ride. In this case, after getting into the Toyota car, Joana needs to access the mobile phone and show the travel coupon in order to get the ride.

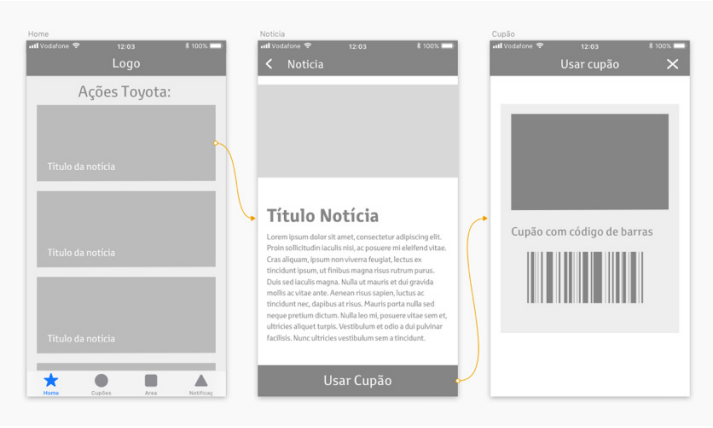


Figure 1: *Workflow* vouchers/coupons 01

To increase the assertiveness in carrying out this task, we have created, for most of the cases, two possibilities to access vouchers and coupons. During a summer festival, such as the Vodafone Paredes de Coura, there are brand activities that matter to the target market, such as the "Toyota rides". As illustrated in Fig 1, the user can access the details in the news concerning the hitchhiking, and from there is redirected to the coupon associated with the action. However, and as a faster and more direct way of accessing, the user always has a direct button to the vouchers and coupons section available in the lower tab bar (Fig. 2). From here, all available coupons are displayed, and the corresponding details can be accessed just by clicking on "Use Coupon".

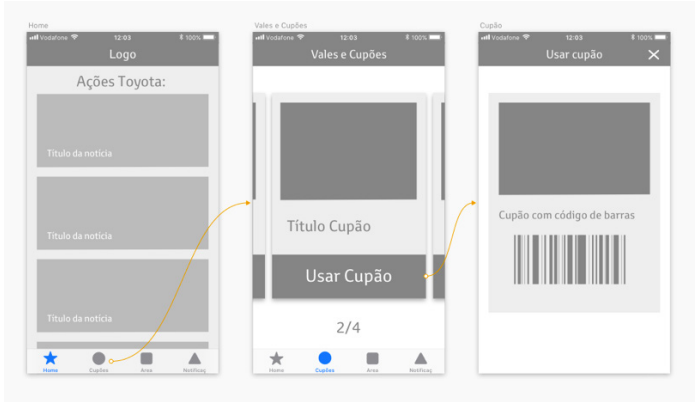


Figure 2: *Workflow* vouchers/coupons 02

This is an app aimed at gaining and maintaining user loyalty. The associated system of validation and acquisition of points, management of vouchers, and the awarding of rewards provides the app with a strong playful aspect, and is intended to contribute further towards connecting present and potential future users with the Toyota brand.

5.5. Interface design

After the *wireframes* and *workflows* design, we moved on to the rigorous design phase of the interface. The screens were designed with a set of graphic and usability concerns, with the goal of creating a credible product that reflected the values of the Toyota brand, and which could provide a best user experience. According to Tidwell [8], users do not trust amateur platforms where there are no efforts taken to create a pleasant appearance.

Since there was already a Spot Toyota version (but only a browser version), a number of aspects had to be taken into account for the restructuring of the entire platform, including the "Spot Toyota" logo (Fig. 3), regarding which it was decided to be kept unchanged.



Figure 3: Spot Toyota logo.

The graphical approach was developed based on the standards and guidelines defined by Toyota. Although this is an independent product from the car sales business, we have endeavored from the outset to maintain visual consistency with Toyota's identity in

Portugal. To achieve this goal, we worked with the guidelines provided in the new Toyota brand visual identity manual [19,20].

For the color system, the work was based on the reference colors defined by Toyota's identity manual: "Primary Color" (grey "PMS Cool Grey 5") and "Highlight Colors" (white, red "PMS 186", and dark grey "PMS CG 11"). In addition to the color, the icons (Fig. 4) adapted for the platform were also conceptualized. We selected those that we considered as having a greater relevance with the app environment. The icons were thus kept as simple line drawings, that were easy to read and carried a minimalist and youthful visual language.



Figure 4: Application icons.

For the Toyota Display and Toyota Text typefaces (original typefaces designed exclusively for Toyota), specific indications in the Toyota brand's visual identity manual were considered.

The layouts design was shaped based on these visual elements, such as icons, color and text. It was a constantly evolving process, however, at each step, Caetsu's approvals were took into account, alongside with conclusions from user tests; and strategies that were considered appropriate for the target audience [21,22].

The first designed layouts had a higher incidence of the color red. With the work development, red was replaced with other colors of the pre-defined chromatic palette. With this change, it was possible to create greater harmony and clarity on the screens, particularly in those containing images, simplifying the interface graphic environment and minimizing the use of unnecessary contrasts.

In this follow-up, and through constant articulation with Caetsu's perspectives, changes were implemented that would end up significantly altering the graphic aspect. Although the structure has been closed since the *wireframe* phase, it was necessary to optimize certain graphic aspects in order to achieve the best result.

As we can see in figures 5 and 6, the tab bar was one of the elements in which the color was changed from red to light grey.

As it was felt that the Spot Toyota logo needed further prominence, a bar in the same tone was also implemented, increasing the contrast considerably. This change ended up highlighting the menu titles in the nav bar in the same way.

Regarding the homepage, it was concluded that the news/action cards needed to highlight more their button functions. With these elements redesign, an attempt was made to remove any doubt that these areas were clickable, pushing this action.

Other changes focused mainly on the emphasis given to certain titles or text fields, the fonts weight on certain buttons, chromatic coherence in the "call to action" style, as well as certain improvements in content.

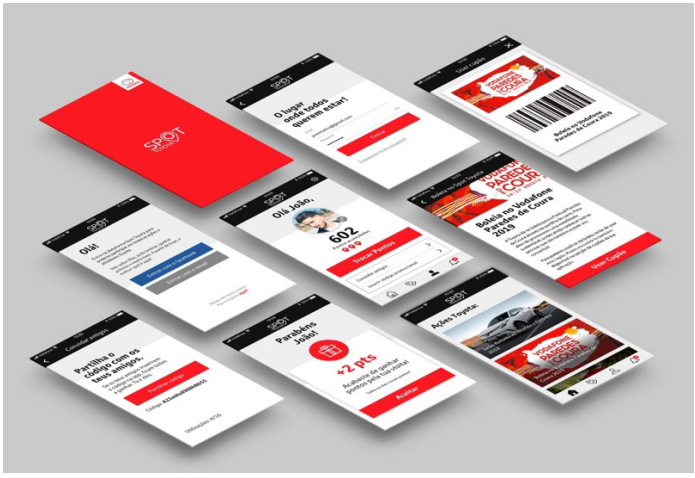


Figure 5: Final version screens.

In figure 6, we present through the example of the "home page" screen, the result of the changes made in the interface graphics.

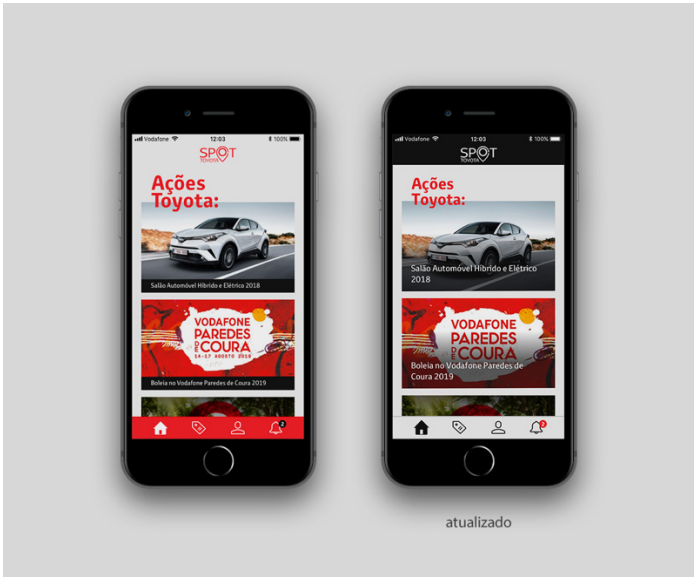


Figure 6: The first version (left) and the final version (right) of the “home page” screen.

5.6. Prototyping

The entire prototyping process proved essential for the communication between the entities involved in the project, Caetsu and Toyota. Since the first meeting, the prototype was an indispensable tool for the evaluation of the work developed and of what would be the real functioning of the mobile application (app).

For the realization of the prototype, we used the Marvel web application. By mapping the clickable zones per screen, we assign the function, the type of interaction and the animation we want for the target screen. Once this process was concluded, based on the *workflows* developed, we had a prototype capable of transmitting what was intended for the final app, and therefore a fundamental tool for testing with users [23,12].

Besides this type of prototyping, some simple animations were created to illustrate the app dynamics. Using these dynamic

images, it became easier to present to the programming team what we wanted (Fig. 7).

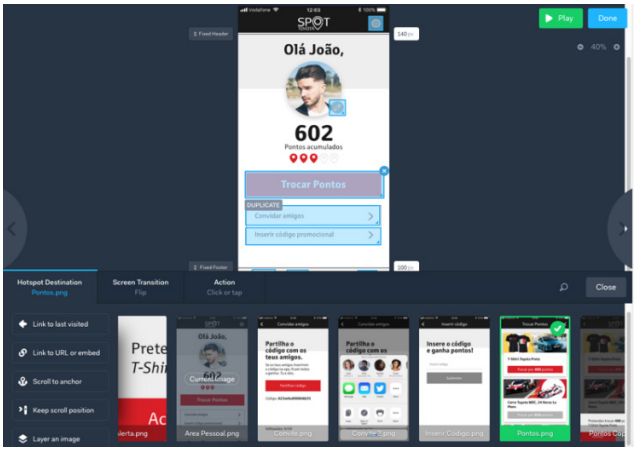


Figure 7: Marvel Platform. Connections between screens.

5.7. Usability tests

Usability tests were developed beginning from the first phase of interface designs. In conjunction with Caetsu, and as the project evolved, decisions were made based on constant user feedback. This process allowed the app to reach a final phase with considerable optimization of the user experience and interface, as well as a reduced number of errors and inconsistencies to be detected at a later stage.

For concluding usability tests with greater accuracy in the analysis method, ten users with an approximate profile to the project's target audience were selected: young users, between the ages of 17 and 32, with some connection degree to the music festival environment.

In this process, user reactions were observed throughout their interaction with the app. For this procedure, we used the last developed online prototype.

For the tests, a set of different tasks to be performed by the test users was drawn up.

Task list:

1. Log in to the app until you find Toyota shares/news;
2. Access the coupon "Hitchhiking on Vodafone Paredes de Coura 2019";
3. Exchange 400 points for a black Toyota T-shirt;
4. Enter a promotional code;
5. Access the news "Hitchhiking on Vodafone Paredes de Coura 2019";
6. Edit your data profile.

After the presentation of these tasks, we asked users to indicate the difficulties encountered, suggestions for improvements, questions, and finally a personal evaluation of the app through a short inquiry form.

According to the observation made to the participants, we concluded that everyone was able to accomplish the proposed tasks. It was also found that the time taken to perform the tasks was identical, with minor differences justifiable by the different levels of users' digital literacy. As expected, we depict these results has the outcome of constant proximity to users during the project development.

In order to assess the level of these same users' satisfaction, we also used, as mentioned above, a short inquiry form. This was provided after completing the usability tests and enquired about the interaction quality and the information appearance, alongside the usefulness and the global appreciation.

The usability tests were extremely important as they allowed us to analyse and evaluate the project with regard to navigation issues, understanding and interface involvement. It proved to be a fundamental process to improve the app usability.

Based on the collected feedback, a high degree of overall satisfaction was obtained, both in terms of the app usage and perception of its added value. These assessments validated and helped to optimize content structuring, information organization, aesthetic and graphical approach, as well as the functionalities developed to capture the users' interest.

6. Conclusions

The main objective of this research was to design a mobile application (app) capable of stimulating a closer relationship between the Toyota brand and its target audience, young people. Thus, through this app, it was intended to explore the potential of technology and digital communication for the loyalty programs development.

Based on the identification and analysis of the state of the art, it was possible to establish objectives that led to the conception of a totally differentiating and pioneering product in the automotive sector. Developed on a real need, and already tested in a previous version, this app aims to create value for Toyota, with effective implementation capacity in the market.

It became essential to understand all the needs, objectives and goals that Caetsu and Toyota Portugal had for the project. Only through this multidisciplinary work, reflected upon the objective articulation of ideas and communication between the entities involved, was it possible to successfully carry out this research project.

The methodology, once properly systematized, made it possible to explore and understand the context of use and the user, as well as Toyota's intentions and expectations. We used several methods throughout the platform development process, including the understanding of UX and UI design, the application of the *personas* method, the design of information architecture and the creation of *wireframes* and *workflows*.

The implementation of Spot Toyota is expected to contribute effectively to a targeted and closer communication between the brand and its audience. From a selective list of features, users will be given access to exclusive brand information; rewards will be given for the use of the app; unique features will be made available for participation in brand actions, and finally, access will be given to a set of tools for awarding points that can be exchanged for products or services.

For Toyota, it will provide the opportunity to regain contacts lost by the entry into force of RGPD; increase its contact database;

develop lead capture strategies; get to know better its target audience in terms of habits, behaviors and needs; and finally establish strong and lasting relationships with its target group.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This research was supported by the Polytechnic Institute of Cavado and Ave [IPCA] and the Research Institute for Design, Media and Culture [ID+].

References

[1] N. Martins, D. Brandão, D. Raposo (Eds.), Perspectives on Design and Digital Communication: Research, Innovations and Best Practices, Cham: Springer, 2021 (in press). DOI: 10.1007/978-3-030-49647-0. ISBN: 978-3-030-49646-3.

[2] M. Erbschloe, "Customer Loyalty Programs". *Salem Press Encyclopedia*. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89163635&site=eds-live. [Accessed: 23-Jul-2019]

[3] J. Oliver, "Forward Thinking Accounting Firms Turn to Apps to Boost Client Loyalty." *Credit Control*, 39(1/2), 30–32. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=heh&AN=128028330&site=eds-live [Accessed: 5-Nov-2019]

[4] W. Cook, "Five Reasons Why You Can't Ignore Gamification". Incentive, 187(1), 22–23, 2013. https://doi.org/10.1248/bpb.23.762

[5] E. Jun, H. Liao, A. Savoy, L. Zeng, and G. Salvendy, "The design of future things", by D. A. Norman, basic books, New York, NY, USA. Hum. Factors Man., 18: 480-481, 2008. https://doi.org/10.1002/hfm.20127

[6] C. Vargas-Irwin, J.P. Donoghue "Automated spike sorting using density grid contour clustering and subtractive waveform decomposition", Journal of Neuroscience Methods, vol. 164, 1-18, 2007. https://doi.org/10.1016/j.jneumeth.2007.03.025

[7] S. Krug, Don't Make Me Think! In K. Whitehouse & L. Brazieal (Eds.), Don't Make Me Think! A Common Sense Aproach to Web Usability (Second). Berkeley: New Riders, 2006.

[8] A. Dickinson, J. Arnott, S. Prior, and A. Dickinson, 'Methods for human-computer interaction research with older people', *Behav. Inf. Technol. is*, vol. 26, no. 4, 343–352, 2007.

[9] P. Morville & L. Rosenfeld, "Information Architecture for the World Wide Web". In IEEE Transactions on Professional Communication (Third, Vol. 43), 2006. https://doi.org/10.1109/tpc.2000.826425

[10] R. A. Virzi, "Refining the test phase of usability evaluation: How many subjects is enough?" Special Issue: Measurement in human factors. Human Factors, 34(4), 457–468, 1992.

[11] E. Stevens, "What Is User Experience (UX) Design? Here's What You Need To Know", 2018 [Online]. Available: https://careerfoundry.com/en/blog/ux-design/what-is-user-experience-ux-design-everything-you-need-to-know-to-get-started. [Accessed: 8-May-2019]

[12] J. Nielsen, "10 Usability Heuristics for User Interface Design", *Nielsen Norman Group*, 1995. [Online]. Available: https://www.nngroup.com/articles/ten-usability-heuristics. [Accessed: 14-Oct-2019].

[13] Raven L. Veal. "How to define a User Persona" [Online]. Available: https://careerfoundry.com/en/blog/ux-design/how-to-define-a-user-persona. [Accessed: 29-Mar-2019]

[14] R. Dam & T. Siang. "Personas: A Simple Introduction". *Interaction Design Foundation*. [Online]. Available: https://www.interaction-design.org/literature/article/personas-why-and-how-you-should-use-them. [Accessed: 28-Mar-2019]

[15] N. Martins & T. Araújo, "The contribution of design in supporting the pregnancy process: The study of a mobile application towards a more informed relationship between the pregnant woman and the healthcare professional" in ARTECH 2019 - Proceedings of the 9th International Conference on Digital and Interactive Arts, Braga, 2019. Ed: ACM. pp. 191-197. https://doi.org/10.1145/3359852.3359857.

[16] Interaction Design Foundation, "What are User Personas?", 2018. [Online]. Available: https://www.interaction-design.org/literature/topics/user-personas. [Accessed: 11-Jul-2019]

[17] Guilizzoni, P. "What Are Wireframes?" *Balsamiq Wireframing Academy*, 2019. [Online]. Available: https://balsamiq.com/learn/resources/articles/what-are-wireframes. [Accessed: 22-Jul-2019]

[18] J. Bernal, G. Flores, R. Vargas, "Diseñas: The web app for learning design terms for deaf students" in DIGICOM 2019 - 3rd International Conference on Design and Digital Communication: Proceedings, Barcelos, 2019. Ed: IPCA. pp. 315-324. https://digicom.ipca.pt/docs/DIGICOM2019-Proceedings.pdf. ISBN: 978-989-54489-5-1.

[19] J. Tidwell, Designing Interfaces. In M. Treseler & R. Monaghan (Eds.), Animal Genetics (Second Edi, Vol. 39). Sebastopol: O'Reilly Media, 2011.

[20] Toyota. Novo Manual de Identidade Visual da Marca Toyota. Portugal, 2012.

[21] J. K. Liker, D. Meier, The Toyota Way Fieldbook: A Practical Guide for Implementing Toyota's 4Ps. New York, London: McGraw-Hill, 2006.

[22] S. Kujala, R. Mugge, T. Miron-Shatz, "The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service" International Journal of Human-Computer Studies, Vol. 98, 51-61, 2017 https://doi.org/10.1016/j.ijhcs.2016.09.011.

[23] C.P. Jansen, S.F. Donker, D.P. Brumby, A.L. Kunc "History and future of human-automation interaction", International Journal of Human-Computer Studies, vol. 131, 99-107, 2019. https://doi.org/10.1016/j.ijhcs.2019.05.006

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1233-1246 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design and Development of Electronic Sensor and Monitoring System of Smart Low-cost Phototherapy Light System for Non-Invasive Monitoring and Treatment of Neonatal Jaundice

Paul Cabacungan[1,*], Carlos Oppus[1,2], Gregory Tangonan[1], Nerissa Cabacungan[3], John Paul Mamaradlo[1], Neil Angelo Mercado[1]

[1]*Ateneo Innovation Center, Ateneo de Manila University, Quezon City, 1108, Philippines*

[2]*Electronics, Computer, and Communications Engineering Department, Ateneo de Manila University, Quezon City, 1108, Philippines*

[3]*Ateneo Grade School, Ateneo de Manila University, Quezon City, 1108, Philippines*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 28 September, 2020
Online: 24 October, 2020

Keywords:
Phototherapy
Neonatal Jaundice
Yellow Detection
Irradiance
Sensor and Monitoring System

ABSTRACT

This paper showcases our previous and continuously improving development at Ateneo Innovation Center (AIC) and partners in designing and further enhancing the existing Low-cost Phototherapy Light System (LPLS) and Improved Low-cost Phototherapy Light System (ILPLS) to the new Smart Low-cost Phototherapy Light System (Smart LPLS) with non-invasive jaundice monitoring for newborns with Neonatal Jaundice (NNJ). Developing this tool will help determine the intensity of yellowish color in infants and can monitor NNJ in a non-invasive way. The system is envisioned to be integrated with Mobile or Near Cloud as part of Smart Nursing Station together with other hospital equipment for monitoring, collection, and management of medical records and services. Its solar-power features for off-grid and remote deployments were also explored. This contribution is an extension of the Intelligent Sensors and Monitoring System for Low-cost Phototherapy Light for Jaundice Treatment that was presented in the International Symposium on Multimedia and Communication Technology (ISMAC) in 2019.

1. Introduction

This contribution is an extension of the previous development [1-3] of the Ateneo Innovation Center (AIC) in the development of a phototherapy system for the treatment of infants with jaundice. The system has been proven to be reliable and effective, with over 30 patients treated [2, 3] during clinical investigation. Our previous efforts described that this unit has been developed at far less cost and may achieve similar results like the commercially available units.

Neonatal jaundice (NNJ) occurs in 25% to 60% of full-term newborns and 80% of preterm newborns in the first two weeks of life, as studied by Cecelia Henny-Harry and her research partner [4]. It is a benign transient physiological event in the majority of newborns but can cause irreversible brain damage and kernicterus in some infants if the serum bilirubin levels are high. Various mechanisms involved in producing this 'physiological' increase in total serum bilirubin include increased production of bilirubin due to lysis of red blood cells, decreased ability of liver cells to clear bilirubin and increased enterohepatic circulation. Any condition that further increases bilirubin production or alters the transport or metabolism of bilirubin increases the severity of physiological jaundice [5].

The main indicator of NNJ is the presence of high levels of bilirubin. Among the effects of high levels of bilirubin caused by rapid RBC metabolism is the increased levels of carbon monoxide in the body, which sometimes causes neurologic dysfunction [6]. Adults may also contract jaundice by having elevated serum bilirubin levels. Jaundice in adults is often an indicator of underlying condition, such as hepatitis or liver disease, and injury [7]. The recurrence of jaundice or its persistence can have astounding and long-standing effects on the health and overall welfare of the patient.

Some factors that increase the likelihood of contracting jaundice are genetics, hospital setting, and social-economic status. In addition, persons of African, Mediterranean, Middle Eastern, or Southeast Asian descent are more likely to contract jaundice [8]. With this in mind, we can infer that Filipinos belonging to lower-income households will be more susceptible to this disease, and so

*Corresponding Author: Paul Cabacungan, pcabacungan@ateneo.edu

are communities with little to no access to sufficient medical services.

Phototherapy is the most frequently used treatment when serum bilirubin levels exceed physiological limits. To initiate phototherapy without delay is the most important intervention for infants with severe hyperbilirubinemia. Phototherapy uses blue light to convert bilirubin into water-soluble photo-products that can bypass the hepatic conjugating system and be excreted without further metabolism. The clinical response to phototherapy depends on the efficacy of the phototherapy device, as well as the infant's rates of bilirubin production and elimination [9]. Phototherapy, despite being the recommended standard treatment for hyperbilirubinemia in newborns, may not be accessible to many people most especially in developing countries. This is due to its high procurement and maintenance cost ranging from PHP 150,000 to PHP 350,000 (USD 3,100 to USD 7,000) [10]. In the Philippines, for rural clinics and maternity hospitals to be accredited by Philippine Health Insurance Corporation (PhilHealth), a tax-exempt Government Corporation attached to the Department of Health that administer the National Health Insurance Program [11], they must possess an operating phototherapy system for jaundiced infants.

For more efficient operations, clinics and hospitals need a detailed digital record of consultation or confinement such as demographic data of the patient, identity of attending medical staff, medical and laboratory procedures and results, among others. In the case of NNJ, rural hospitals will need chemical reagents and a hematology analyzer to administer actual bilirubin counting of blood samples. Access to comprehensive patient digital records and to needed reagents and medical devices is usually problematic, especially in remote areas. When an infant is suspected to be suffering from NNJ, the doctors resort to physical inspection to assess the existence or degree of jaundice.

In 2016, Ateneo Innovation Center together with its partner pediatricians conducted a two-year pilot study and clinical investigation to evaluate the capability of the LPLS in lowering bilirubin levels among healthy term infants diagnosed with indirect hyperbilirubinemia admitted at the nursery, OB ward, and pediatric ward of a tertiary hospital in Metro Manila. [2,3]. Figure 1 shows the LPLS administered by a pediatrician during the clinical investigation, after passing through all the necessary protocols required by the ethics committee.



Figure 1: Infant diagnosed with indirect hyperbilirubinemia underwent phototherapy using the LPLS prototype in a tertiary hospital in Metro Manila, Philippines

The clinical investigation revealed that 16% of the total number of patients completed treatment after 24 hours of phototherapy under the LPLS and were classified to be in the low risk zone while 36% completed treatment after 48 hours. The total bilirubin significantly decreased by 16.5% ($p = 0.0001$) from baseline bilirubin levels after 24 hours. There was a significant 29.9% mean percentage of reduction in bilirubin level after 48 hours. The proportion of subjects in the high-risk zone during baseline to 24th hour went down significantly from 80% to 28% ($p = 0.0003$). In the same manner, comparing baseline to 48th hour, the percentage of subjects in the high-risk zone went down from 80% to 9.5% ($p = 0.0001$). Figure 2 shows the graphical representation of the results.

After the clinical investigation, there were no reported occurrences of rebound hyperbilirubinemia. No patient experienced any complication while on treatment. This makes the system safe and effective in lowering total serum bilirubin among healthy term infants with indirect hyperbilirubinemia [3].

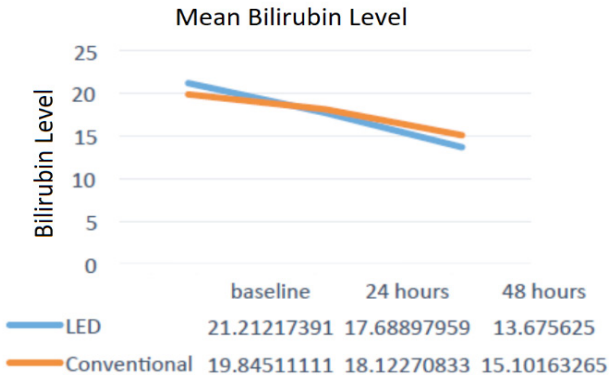


Figure 2: Mean TSB levels in both groups at different period [3]

Ateneo Innovation Center developed further the Low-cost Phototherapy Light System (LPLS), shown in Figure 3, with the integration of sensors making it an Improved Low-cost Phototherapy Light System (ILPLS) as seen in Figure 4. ILPLS was designed to have the same functions, to provide the same treatment, and to keep its low cost, as the original prototype.

The main components of the ILPLS were the eight pieces of 7-watt LED blue spotlight, which were more robust and had the same cost as the 3-watt bulbs of the LPLS. The wavelength was within the 400 - 500 nanometers (peak at 460nm) which are specifically used for administering phototherapy [12]. The measured irradiance was three times stronger than that of the 3-watt bulb and its unit exceeded the total irradiance of the previous unit. The bulbs were positioned alongside each other at 13 centimeters apart (from center to center). This engineering design made the light beams overlap by 43% resulting in a more intensified lighting. There is an 11% intersection of light beams coming from the bulbs positioned across each other, which are 22 centimeters apart. The 22-centimeter distance was an engineering trade-off to give way to the fans that provide fail-safe mechanisms. During the 120-hour operation, the ILPLS bulbs performed in an acceptable temperature range of 35.8 to 39.0 °C, with ambient average temperature of 32.5 degree Celsius. No issues were encountered at all. An additional 10-watt computer fan was also installed.



Figure 3: Low-cost Phototherapy Light System (LPLS)



Figure 4: Improved Low-cost Phototherapy Light System (ILPLS)

The sensor and monitoring system were also introduced and successfully integrated into the ILPLS without affecting the unit's optimal functioning. The camera system is intended to provide digital records of the patient's ID, time stamp, unit tag and audio tags by medical personnel. The smart camera was able to transmit images and sounds within the 10-meter radius in a laboratory setting, despite the presence of walls, demonstrating that data collection at nearby stations is possible. The location of the phototherapy unit is crucial to readily interface with nursing stations and hospital medical records. This approach is scalable and useful to situations where multiple phototherapy units are operated at the same time. It can easily record the data of the ongoing phototherapy and supplement the nurses' records. Four light dependent resistors (LDRs) were installed to ensure emission of appropriate light intensity. The two-in-one sensor was also able to measure temperature and humidity inside the light box. Alarms would signal readings beyond threshold, adding safety features to the system.

Ateneo Innovation Center acknowledges the long-standing problem of the people in remote and off grid communities who have little or no access to this device. Through the efforts of this organization, professors, students, and different organizations continue to collaborate and work together in developing innovative designs and ideas that provide solutions with impact on the betterment of society. This paper presents another Ateneo Innovation Center's development that aims to address a need in remote settings - the Smart Low-cost Phototherapy Light System (SLPLS) with non-invasive jaundice monitoring for neonatal jaundice.

Shown in this paper is how we are developing new sensor and monitoring systems that enhance the performance of our low-cost phototherapy light systems and facilitate data collection and sharing with patients' families and the stakeholders. Having a digital record to determine the progress and success of treatment, doctors can monitor and assess the progress of the treatment in a more scientific way. We are developing yellow jaundice sensors that can monitor the treatment of the patient and can provide a possible basis for detection. We explore different color sensor experiments, using accessible devices and open source applications that can be integrated into our latest prototype. We also demonstrate optical techniques for color measurement that can be a tool for obtaining a digital record of the degree of jaundice. Lastly, we deal with an important aspect of the development - solar powering of the phototherapy system for 24/7 performance in case of power outage or blackouts and remote deployment. To further prove that this system is effective enough to be used as an alternative than that of a commercial unit, we have yet to conduct further tests with actual human subjects that will be supervised by medical personnel.

2. System Overview

This paper describes the most recent Smart LPLS prototype with further improvements to its electronic sensor and monitoring system. We also present non-invasive methods that will help determine the intensity of yellowness in infants as a way of detecting NNJ. The Smart LPLS is composed of eight 7-Watt LED bulbs at the top and the additional ten 10-watt LED tube light pieces at the bottom, as seen in Figure 5. This makes the system more robust compared to the 3-watt and 7-watt LED bulbs used in previous prototypes presented earlier. These additional T5 LED tube lights are 90% more efficient than traditional light bulbs, 60% compact than fluorescent bulbs, and have a lifespan of 30,000 hours.

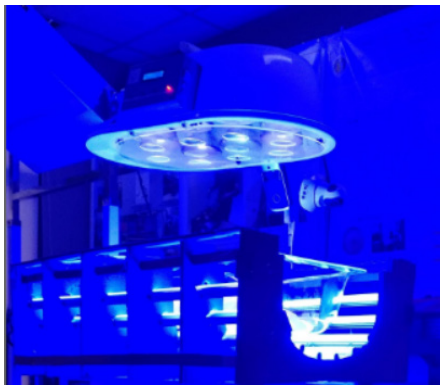


Figure 5: New Smart Low-cost Phototherapy Light System

To easily monitor the system's performance, an automated system that stores data about the device's light intensity and bulb temperature and records ambient humidity was installed. An alarm system that activates the warning lights if sensor readings are in critical level and if the bulbs need to be replaced are built into the system. A time setting feature is also included to manually set the time of operation and automatically turn-off the device. The system diagram for the smart sensor and monitoring system can be seen in Figure 6.

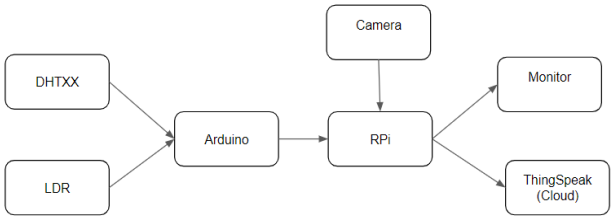


Figure 6: System diagram for the smart sensor and monitoring system

The DHT22 sensor, which can measure both humidity and temperature, is placed inside the system. Meanwhile, the photoresistors or the light dependent resistors (LDRs) are placed near the phototherapy light to monitor the light intensity. Both are connected to the Arduino which will send data to the Raspberry Pi later. While the Arduino is feeding sensor data to the Raspberry Pi, a camera may also be connected to it. The Processing IDE will process the data and display everything in one window through the monitor where people can see if everything is as expected. Measured data like temperature, humidity, light intensity, and images are sent to ThingSpeak IoT platform, which can be used to send data to the cloud from any Internet-enabled device. Users can configure actions and alerts based on real-time data and unlock the value of data through visual tools.

The non-invasive jaundice treatment monitoring feature of the Smart LPLS aims to provide an alternative platform for physicians to conduct a non-invasive method to check how severe the jaundice of an infant is while on treatment. Blood analysis is currently used to determine bilirubin level in the blood. This process is invasive, and the results may take some time. Physical inspection may also be done but the accuracy of results depends heavily on the skill of the medical practitioner performing the inspection. In our experience, provincial hospitals often lack the necessary reagents to carry out these tests, necessitating an alternate measurement. Developing a tool that will determine the intensity of yellowish color in infants can give immediate and accurate detection of NNJ in a non-invasive way. Also, commercial phototherapy systems and bilirubinometers are expensive and rural hospitals might not have the financial resources to afford these. The sensor will be helpful in gauging and monitoring if the phototherapy treatment is sufficient and successful in treating the jaundice in infants. If phototherapy was not enough, blood transfusion might be needed. The system may store collection of readings and activate a warning light indicator if sensor readings that are related to bilirubin level go beyond the set ceiling, indicating that the condition cannot be managed through phototherapy light and a more aggressive treatment approach such as blood transfusion is needed. The same automation system can also activate a different warning light if sensor readings go below a set floor, indicating the full treatment of NNJ. This means that the infant can be removed from phototherapy light exposure.

In *Table 1*, we compared the prototypes with a commercially available unit. What sets the Smart LPLS apart from the previous prototypes are the top and bottom irradiation. This results in a 360° light exposure, without the need to tilt the baby every two hours. Moreover, the irradiance of top-mounted lights is three times stronger than the minimum required irradiance of 30 µW/cm²/nm for NNJ treatment. In addition, the bottom-mounted lights provide an irradiance of 34 µW/cm²/nm, leading to an expectation for faster treatment. Despite the installation of more lights, the Smart LPLS has two fans with air vents to provide safety mechanisms.

The Smart LPLS for off-grid deployment with a total unit cost of PHP 38,000.00 or around USD 780.00 uses energy efficient LED bulbs and tubes having a full width half maximum wavelength range of 439.66 nm to 467.64 nm that is within the curing wavelength of 400-500 nm, irradiance of 34µW/cm²/nm, with top & bottom lights still within the minimum wavelength, additional air vents, sensors and monitoring system, and with integrated non-invasive yellow detection feature. We have listed our estimates of the baseline system costs, however, due to time constraints and the current setup of the research laboratory, the systematic optimization in relation with its cost discussing the reliability, longevity, and efficiency is of pending approval by medical authorities will be discussed in future publications.

Table 1: Comparison of Smart LPLS with its Predecessors and Commercial Unit

Parameters	**LPLS**	**ILPLS**	**Smart LPLS**	**Commercial Unit**
Light Bulb	Twenty 3watt LED bulbs	Eight 7-watt LED spotlights	Eight 7-watt LED and Ten 10-watt LED tubes	Five 20-watt Fluorescent
Wavelength	462.1 nm to 476.4 nm	437.8 nm to 458.48 nm	437.8 nm to 458.48 nm; and 439.66 nm to 467.64 nm	(5 white and 2 blue)
Irradiance at 30 cm. average distance	84 µW/cm² /nm	90 µW/cm² /nm	90 and 34 µW/cm² /nm	70 µW/cm² /nm
Total Power	75 watts	90 watts	240 watts	400 watts
Ventilation	one fan	two fans	two fans with air vents	air-cooled
Sensor and Monitoring System	none	present	present	none
Integrated Non-invasive yellow monitoring	none	none	present	none
Unit Cost	Php20k	Php35k	Php38k	Php150k

Since 2015, the Ateneo Innovation Center has been studying and exploring the Near Cloud or Mobile Cloud, recognizing its powerful computing and caching capabilities in the context of disaster resilience, education, and healthcare. The concept of caching is powerful especially in a telecom infrastructure-less environment. This technology uses a combination of a plug computer, preloaded content, and key web applications on terabyte drives at fiber-like speeds. It addresses limited bandwidth and connectivity by providing valuable preloaded content and acting as a gateway to the Internet [13-15].

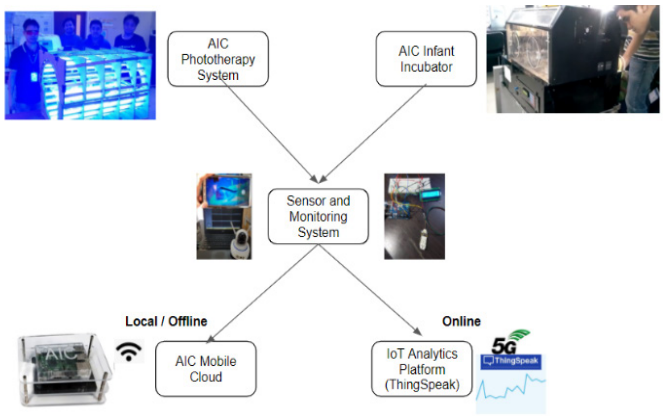


Figure 7: Smart LPLS as part of the envisioned Smart Nursing Station

We foresee this system to be part of a Smart Nursing Station, which diagram is in Figure 7, wherein a visual and auditory monitoring system is implemented to remotely oversee infants from the nurse station. The cameras will capture image and recordings of the infant that will serve as logs and be written in baby's medical records. It also sends alerts and data to the medical doctor or staff if there are any updates. Medical records and images of patients can be easily archived and retrieved if stored locally in the Near Cloud. Having the patient's data in the cloud also promotes interoperability and seamless transfer of data among the various stakeholders and speeds up healthcare delivery. This therefore closes the Medical Data Loop and promotes 5G and mobile cloud architectures in academic and business settings as well as for applications in sustainable and resilient communities [16].

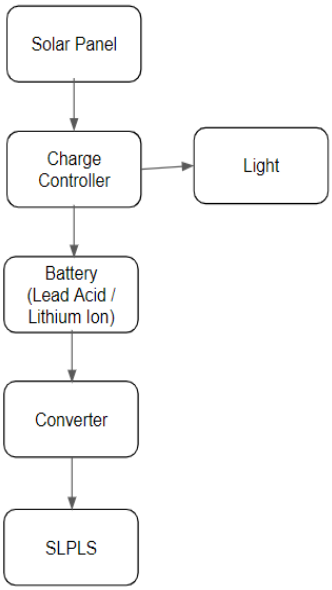


Figure 8: Off grid setup for Smart LPLS

In our previous study, we were able to run the system for 10 hours powered by two 50-ampere-hour commercial lead acid car batteries that were pre-charged by two 50-watt solar panels [1]. In this study, we continue to explore ways to make the system energy-efficient to be easily supported by solar powering. Ultimately, we hope to deploy the Smart LPLS in remote areas using the set-up shown in Figure 8.

3. Review of Related Literature

Total body phototherapy exposure (TBPE) is a new and effective technique to avoid the risks of invasive exchange blood transfusion (EBT) in neonates with jaundice. TBPE consists of top- and bottom-mounted lights to provide whole body phototherapy light exposure. Ogah Emeka Onwe and his research team in 2019 used an imported phototherapy light unit that has an average spectral irradiance of 34 µW / cm^2 / nm and 50 µW/cm^2/nm in the top and bottom lights, respectively, and a peak wavelength of 455 - 470 nm. It was reported that out of 82 severely jaundiced cases, all TBPE cases were successfully discharged except for two mortalities due to related exchange blood transfusion complications [17].

In general, the same authors found that TBPE is an efficient and a good non-invasive alternative to EBT intervention [17]. It is also affordable for resource-constrained settings. But some limitations were encountered during the study. The frequent occurrences of unannounced power failures at the hospital facility might have adversely affected the overall treatment time. Also, the identification of qualified subjects for inclusion in the study was limited due to poor record keeping.

In determining jaundice condition in infants, the practice of visual assessment is quite subjective and needs a lot of experience from the medical practitioner to determine severity of jaundice condition, as stated in The Journal of Pediatrics by Arieh Riskin and his co-researchers [18]. Thus, medical practitioners can use devices which can measure the levels of jaundice in infants without blood analysis. Infants under treatment can still have increased bilirubin levels which is very dangerous when not recognized immediately. On the other hand, infants with decreased bilirubin levels can only be verified upon examining the blood. In the absence of total bilirubin serum (TBS) test and hematology analyzer, a database of infant's yellowish skin color across varied skin tones may be helpful in the visual assessment of the patients' condition.

Methods of digitally sensing color pigment from reflected light may offer a simple and objective way of determining the existence or severity of jaundice condition during visual inspection. Current procedure of detecting jaundice in infants, such as newborn screening (NBS) test, is done by physicians through visual inspection and blood analysis to measure the bilirubin level. The study of Hafizon Baharuddin and his group in 2010 presented an infrared sensor that detects bile pigments, but blood extraction is still needed to assess the bilirubin level [19].

A method for measuring jaundice using non-contact optical devices at the conjunctiva of human subjects was described by Nabarun Polley and his colleagues in 2015 [20]. The light from the source is transmitted through the six excitation fibers of the excitation arm and incident on the subject (conjunctiva). The diffused light is collected by the detection fiber and transmitted through the detection arm to the spectrograph. The spectral response corresponding to the conjunctiva is processed and generated in the laptop computer as shown in Figure 9 below.

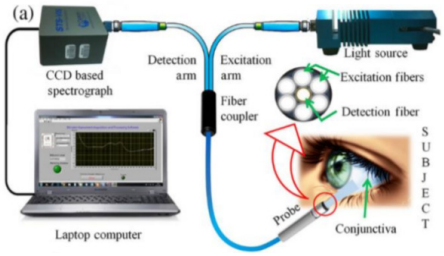


Figure 9: Use of non-contact optical devices at the conjunctiva of human subject

In the figure below, the comparative spectral response of a normal volunteer and a jaundice patient using the wavelength calibration of their setup has been established with a He-Ne laser (632.8 nm), fluorescent lamp, and emission/absorption of a number of dyes including aqueous bilirubin solution are shown. A distinct difference in the spectral appearance is visible; the contribution of yellow pigment deposited in the conjunctiva of the jaundice patients is higher compared to the normal volunteer. An absorption spectrum of aqueous bilirubin solution is also included in Figure 10 as reference.

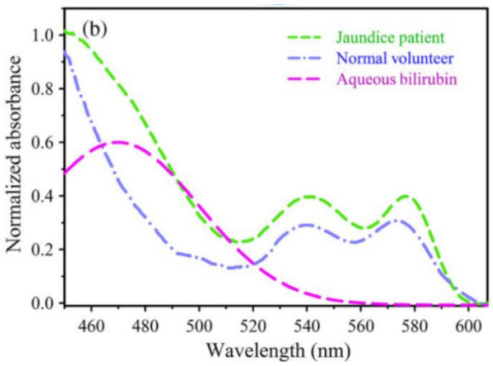


Figure 10: The comparative spectral response of conjunctiva of a normal volunteer and jaundice patient

They have also developed a non-invasive and practical device for measurement of bilirubin levels. While serum bilirubin measurements are still required for precise diagnosis, the proposed device shows the potential that reduce frequent blood sampling. This procedure may be difficult to use knowing that newborns sleep most of the time and their eyes are small and sensitive.

4. Results and Discussion

This is a design and engineering study with no infants or humans involved, with no cross-reference with real jaundice patients and no photos during treatment and corresponding bilirubin measurements were taken due to ethics concerns and time constraints. This study was partially funded by Ateneo de Manila University, where the authors are affiliated with, so there is no conflict of interest to be declared.

4.1. Increased Surface Exposure

The main component of the phototherapy treatment is the blue light which is provided by the blue LEDs. Using LEDs also reduces the risks of complications such as progressive skin damage brought about by conventional ultraviolet light used by standard phototherapy light treatments. To ensure the efficiency and effectiveness of the device, light intensity, the distance from the subject of phototherapy, the area of exposure, the duration of the treatment and the amount of bilirubin present in the body of the subject must be taken into consideration.

The blue light source in the Smart LPLS is coming from the top-mounted spotlights and bottom-mounted tube-type LEDs. The Smart LPLS has a wavelength that peaks at 253.99 nm and has a full width at half maximum (FWHM) range of 439.66 nm to 467.64 nm. Its irradiance was measured from an oval grid with a length of 60 cm and width of 40 cm from 35 different data points. The average distance of the measuring device (RS Pro) from the phototherapy light mimics the distance of the infant patient from the phototherapy light which is about 30 cm. The measured irradiance is 34 $\mu w/cm^2/nm$, which is also above the minimum required irradiance of 30 $\mu w/cm^2/nm$.

The additional set of phototherapy lights installed under the transparent bassinet also acts as the frame and stand of the whole system. The existence of bottom-mounted lights provides additional irradiation that may shorten treatment time.

4.2. Monitoring of Neonatal Jaundice Patient

One of the significant features of the Smart LPLS is the color sensor to monitor the treatment progress of the infant with jaundice based on the intensity of skin yellowness.

We are convinced in the importance of having a patient monitoring process when the phototherapy light system is deployed in remote areas without reagents and hematology analyzer. The cost of transcutaneous bilirubinometer (TcB) is similarly prohibitive making it difficult to find in remote areas.

This paper presents the use of inexpensive electronic sensors and devices that can detect even small color changes in the infant's skin by image processing. A low-cost method of skin tone simulation that is digitally generated and classified to detect change in yellow pigments can also address the restrictive financial resources in the rural areas. The sample size is limited only to the number of images we have conducted in our initial tests. Though the tests exhibit promising results, the data gathered are patient specific since we covered only three colors of skin.

4.2.1. Profiling the Sensors

For the non-invasive jaundice monitoring feature of the Smart LPLS, we utilized HD web camera and Raspberry Pi, Arduino, and color sensors, and RGB phone app to monitor patients non-invasively. We implemented the yellow detection feature by quantifying the reading of different sensors based on their profile.

4.2.2. Creation of Yellow Swatches

Due to time and ethical constraints, especially during this COVID-19 pandemic, yellow level detection tests using different sensors were done using printed yellow swatches, not using human subjects. We created a standard swatch of yellow that varies from 0% to 100% with increments of 10%. We also tested the swatches with plastic film to compare it with different opacities. We are envisioning that doctors can use this plastic film or transparency swatches by putting it on a non-jaundice skin (with similar tone as the infant's skin) and comparing it with the

level of yellowing in a jaundiced infant's skin. The jaundiced infant's condition can be initially verified through blood test and correlate with the increments of yellowness represented in the swatches. The image or data can then be evaluated by the supervising doctor remotely or check the history of treatment of the baby.

RGB color picker was used to make yellow swatches that will be detected by the sensor and the camera. It is based on the color theory that all visible colors can be made using the additive primary colors of red, green, and blue. Hue (H), saturation (S), and value (V) are the main color properties that allow us to distinguish between different colors.

Hue is described by the dominant wavelength and is the first item we refer to when adding in the three components of a color and it refers to a color having full saturation. value refers to the lightness or darkness of a color. Value on the other hand indicates the quantity of light reflected. Saturation lastly defines the brilliance and intensity of a color [21]. In color theory, a tint is a mixture of a color with white, which reduces darkness, while a shade is a mixture with black, which increases darkness. Both processes affect the resulting color mixture's relative lightness [22].

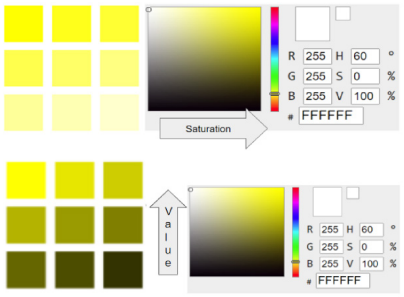


Figure 11: Different levels of Saturation (S) and Value (V) to obtain yellow swatches

Changing the S while maintaining H and V in constant values will give us different saturation levels of yellow. Also, shades of yellow in RGB have R and G values at 255. As the transparency decreases, B value increases from 0 to 255. At 100% transparency, B is at its maximum value at 255, while 0% transparency is at minimum value at 0.

4.2.3. TCS3200/TCS230 RGB Color Sensor

In building the intelligent system, codes were developed to be used for the microcontrollers, sensors, and actuators. Calibration and testing of the system were also conducted to ensure the efficiency of the system. To jumpstart the tests using a detector to gauge the yellowness level of the skin, we used the TCS3200 / TCS230 RGB color sensor that has an array of photodiodes with 4 different filters and is mostly used for color recognition projects and can detect a wide variety of colors based on wavelength [23]. By selectively choosing the photodiode filter's readings, we were able to detect the intensity of the yellow and other different colors. This sensor has a current-to-frequency converter which converts its readings into a square wave with frequency that is proportional to the light intensity of the chosen color [23].

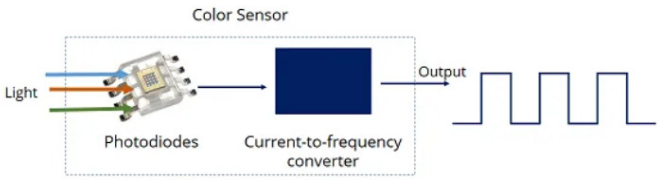


Figure 12: TCS3200 / TCS230 color sensor [23]

The sensor was connected to the Arduino and was able to receive serial data RGB values ranging from 0 - 255. This sensor can accurately reproduce colors which were monitored by getting output from the Arduino. Processing3 IDE was used to verify by sight the similarity of the actual color and sensed color by taking in the serial data from the sensors, and visualize the color seen by the device by using the RGB value attained, and displaying it on a computer screen.

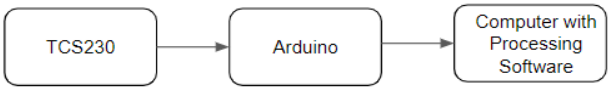


Figure 13: Block diagram for basic TCS3200 / TCS230 sensor test

Initially, we used a clear filter and set the control pins S2 and S3 to HIGH and LOW, respectively. We calibrated and enclosed the color sensor in a box to prevent ambient light from contaminating the sensor and concentrate the light reflected to the sensor.

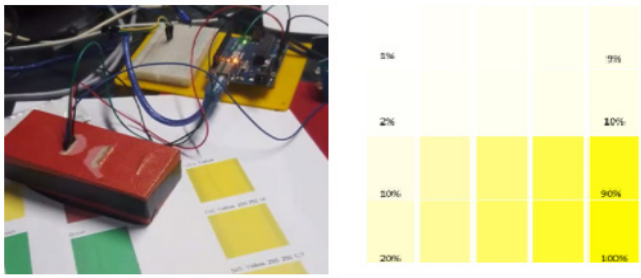


Figure 14: Enclosed TCS3200 / TCS230 and color swatches

Running the serial monitor will provide values that are dependent on the selected frequency-scaling and the condition of ambient light. The R, G, and B values differ due to the different sensitivity of each photodiode type, as seen in *Figure 15*. To represent the detected colors with the RGB Model with values from 0 to 255, we will use the *map()* function to map or convert the readings to values from 0 to 255.

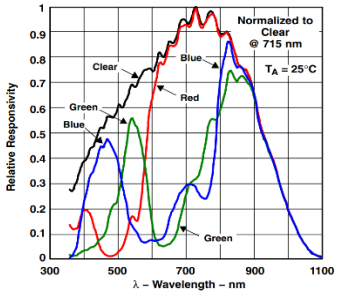


Figure 15: TCS3200 / TCS230 spectral responsivity diagram

We profiled and checked the sensor to set baseline data. The graph in Figure 16 shows that the sensor can detect colors normalized to different values. The sensor is more responsive to more colors.

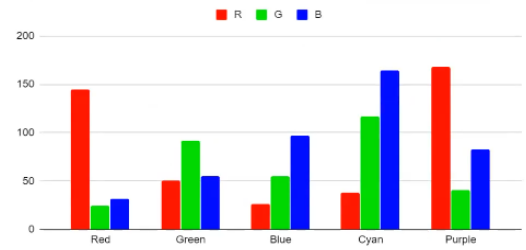


Figure 16: TCS3200 / TCS230 color profiling

We interpreted the theoretical RGB values for different shades of yellow as seen in Figure 17. Both R and G values are at 255 while B is at 255 if the color is at 0% opacity or 100% transparency. Meanwhile, B is at 0 when the yellow swatch is at 0% transparency. Once transparency increases, B increases as well, at a linear rate. Using this information, we compared the graph obtained by the blue values of the sensor, to the true-blue values of yellow swatches.

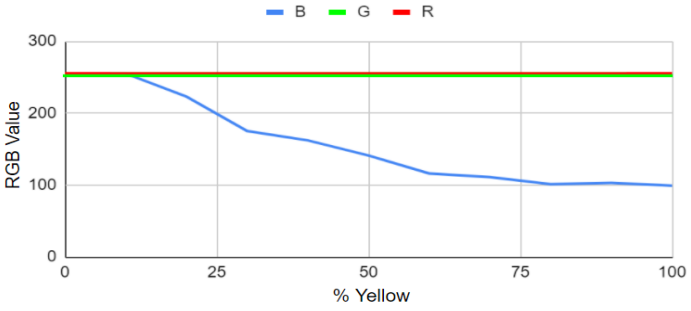


Figure 17: Theoretical and expected RGB values with different yellow values

As expected, a downward curve is shown in Figure 18 with 0% to 60% yellow of sensor values in line with the theoretical values, but it started to deviate from the theoretical value at around 60% yellow. We found that the sensor was able to accurately determine the B values of the swatches ranging from 40% to 100% transparency. The yellow pigment absorbs blue light and it allows red and green light to pass through or reflect from its surface.

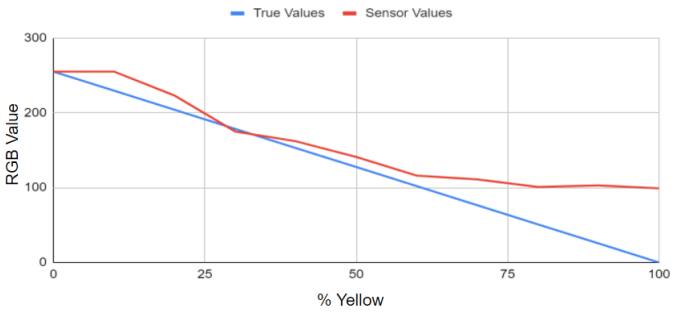


Figure 18: Theoretical vs sensor values of B values using without color filter

We also applied various color filters to the TCS3200 sensor and saw any changes or improvements in performance. In Figure 19, the experimental setup using the color filter of the sensor was shown. The built in LED of the sensor as the light source will shine light in the yellow swatches. Light will be bounced off the color filter that will be detected by the TCS3200 / TCS230 sensor which will generate a pulse width modulation data will proceed to the Arduino and be sent in the serial monitor.

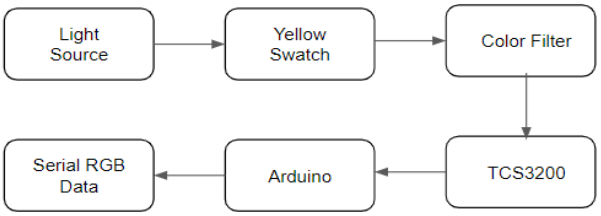


Figure 19: Experimental setup using TCS3200 / TCS230 sensor with filter

In Figure 20, a blue filter was placed in between the swatches and the sensor. For the blue level to be detected by the photodiode, we set the control pins S2 and S3 to LOW and HIGH, respectively. Green and red are not in constant 255 because of the filter and blue, as expected, the graph also shows a downward curve. It can be observed that blue dips in accuracy past 30% and none of the values follow the ideal relationship between blue values and yellow percentage.

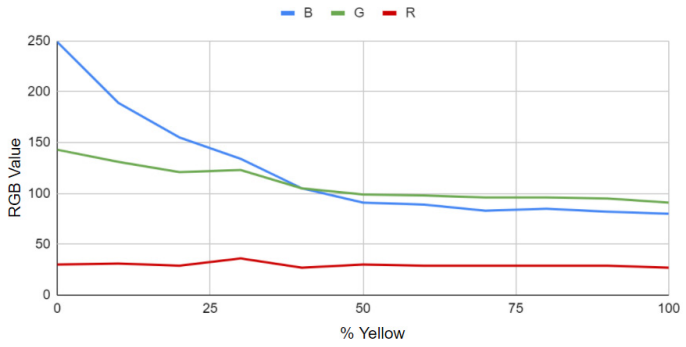


Figure 20: TCS3200 / TCS230 sensor using blue filter

None of the values are constant, but blue seems to be decreasing in a similar curve as it did with the other test. Looking at the results from the blue filter and no filter tests, blue values from the sensor seem close to the true values from 0% - 60%.

4.2.4. ISL29125 RGB Color Sensor with IR Filter

We also conducted tests and determined the performance of the ISL29125 sensor without filters and compared it to the true RGB values and each other. ISL29125 is a low power, high sensitivity, RGB color light sensor with an I2C interface and operates at 3.3V TTL and 56 uA typical supply. Its state-of-the-art photodiode array provides an accurate RGB spectral response and excellent light source to light source variation. It has integrated IR filter and compensation and outputs the actual light intensity with analog-to-digital converter and is designed to reject IR in light sources that allows the device to operate from well-lit to dark rooms. All the features of the device are controlled by the registers. The device is designed for operation under dark glass cover which significantly attenuates visible light and passes the infrared light without much attenuation. The device has an on-chip passive optical filter designed to block or reject most of the

incident infrared. This sensor is often used in ambient light color detection and / or correction [24].

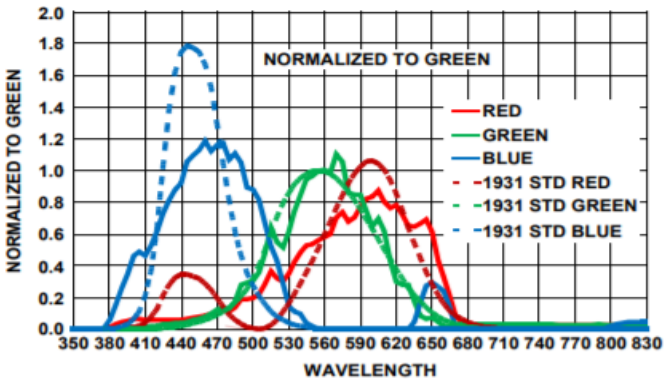


Figure 21: ISL29125 responsivity [24]

Housing for the sensor was created to keep out any ambient light from interfering with the received data. Using the ISL29125 with I2C interface can be both an advantage and disadvantage as it operates at 3.3V TTL and a logic converter is needed if Arduino is used since the datasheet specifies 4V as absolute max voltage while using less pins. An LED in series with a 510 Ohm resistor was used because there is no built-in light for the sensor.

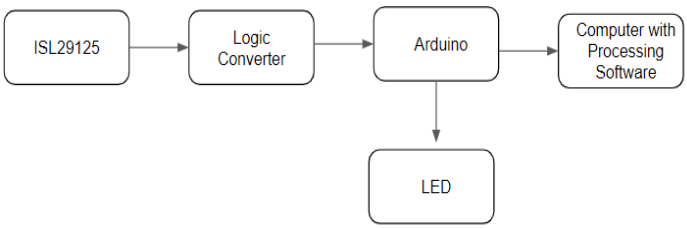


Figure 22: Block diagram for ISL29125 color sensor test

The sensor has three configurable settings. At CONFIG1, mode can be picked that determines what colors does the sensor samples, light intensity range from 375 lux to 10,000 lux depending on lighting condition, choose between 12 or 16 bit ADC precision depending on speed and precision, and sets the behavior of the interrupt pin if to be used as an input that triggers sensor sampling. CONFIG2, on the other hand, provides the option to set the level and range of active infrared compensation and lastly CONFIG3 that sends an interrupt to the Arduino once RGB values reach a certain threshold. In this test, we used the default configuration by enabling all the RGB, set the light intensity range to 10,000 lux, 16-bit ADC, high IR compensation on the lower range, and no interrupts.

We also profiled and checked the sensor to set baseline data. The graph in *Figure 23* shows that the ISL29125 sensor can detect colors normalized to different values. The sensor has slower deviation to colors as compared to the TCS3200.

The ISL29125 was used in determining the RGB values in different yellow swatches and it shows in Figure 24 that Red is constant at 255. Green, on the other hand, showed a downward trend from 232 at 25% yellow to 184 at 100% yellow. Blue, as expected, followed a downward trend as well from 161 at 25% to 63 at 100% yellow.

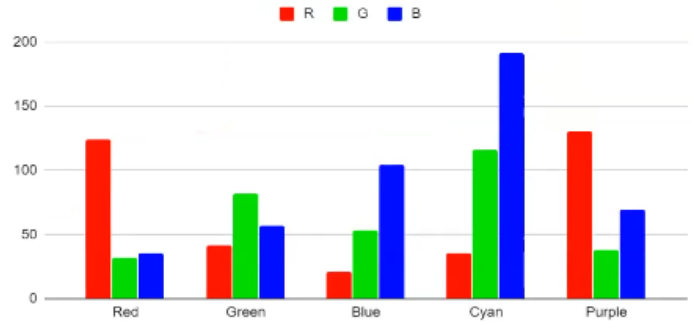


Figure 23: ISL29125 color profiling

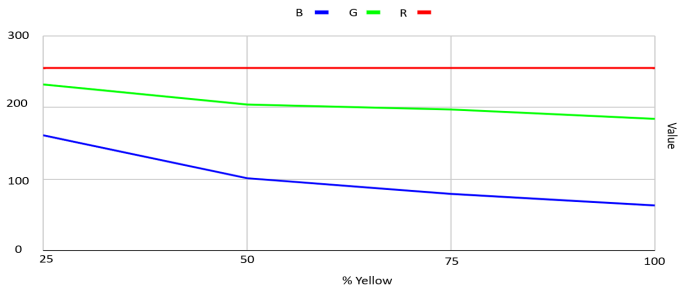


Figure 24: Test using ISL29125 sensor

Instead of using a white LED to illuminate the subject or surface, we used a red LED to test the reflectance of the surface for a specific color. In theory, red should reflect some, if not all, the red LED's light and sensor readings should be relatively higher compared to those that absorb instead of reflecting. This was seen in the data where the sensor read Red from colors where the red values were maxed out.

As seen in Tables 2 and 3, white, red, yellow, and purple swatches reflected red as expected. Values correspond to their measured RGB values and by looking at their full measured Red data of the swatches with red illuminant, white has the highest red value measured tied with yellow, then followed by purple, and lastly red. This order is also seen when measuring the colors using a white illuminant where White has the highest red value with red reading the lowest. Whether a white or red illuminant is used, the measured values are proportional in that their order from 'most red' to 'least red' is the same regardless of the illuminant.

Table 2: ISL29125 with red LED illuminant

Swatch Color	Red Value
White (#FFFFFF)	68
Red (#FF0000)	44
Yellow (#FFFF00)	66
Purple (#FF00FF)	58
Black (#000000)	0
Cyan (#00FFFF)	0
Green (#00FF00)	0
Blue (#0000FF)	0

Table 3: ISL29125 with white LED illuminant

Swatch Color	Red Value
White (#FFFFFF)	255
Red (#FF0000)	124
Yellow (#FFFF00)	255
Purple(#FF00FF)	130

The other colors had 0 color red reflected as detected by the sensor because by theory, they did not reflect red. This was done only with one red LED and compared to the white LED used in the earlier tests. Since the highest measured value is only 68 for red illuminant as to 255 of the white illuminants, the red LED might be too dim compared to the white illuminant. For further tests, it is recommended to install more LEDs for illumination since its dimness may not reflect the whole swatch or sample picture. It may be possible that the other colors do not reflect red but realistically they still could reflect a little red due to the color accuracy of the printers. Other colors may also be tried as well such as the suggested pulsing of Red and Green for Yellow. Proper calibration is also needed as the sensor's datasheet specifies very specific methods and processes to calibrate.

We also conducted an IR compensation test using the ISL29125 sensor which provides a programmable active IR compensation which allows fine tuning of residual infrared components from the output, which allows optimizing the measurement variation between differing IR-content light sources. There are two ranges set by B7:0 in the compensation register. In the library they are labeled as Offset and No Offset. B5:0 configures the specific value within the ranges. The preset values are HIGH, MED, LOW, but any 6-bit value can be set as the compensation. This combined with the offset gives 128 possible values for the active IR compensation. *Figure 25* shows a typical system measure for both IR Comp Adjust and IR Comp Offset.

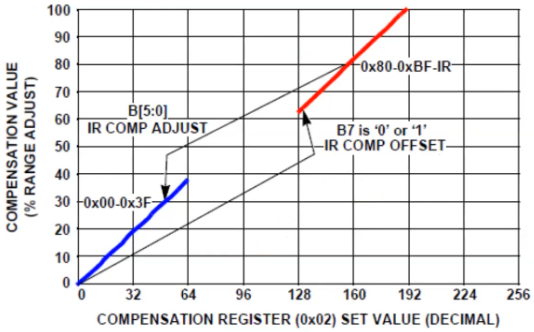


Figure 25: IR compensation set [24]

Using yellow swatch in this IR compensation test, Red and Blue have constant values as seen in Figure 26. When testing on colors, only the green color varied. With the highest IR compensation, the value of green is at its lowest. On the other hand, if there is no IR compensation set, green is at its highest.

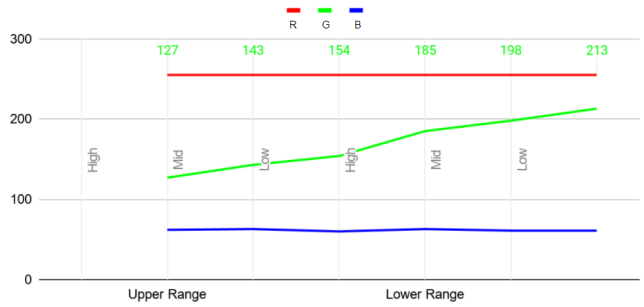


Figure 26: IR compensation using ISL29125 sensor

To test the difference on how the IR compensation works, the sensor was used on the same swatch while changing the IR modes between measurements. The measured values are seen in Figure 27. Since this printed color swatch was used, the most accurate value is the one with no offset / IR compensation.



Figure 27: IR compensation using ISL29125 sensor

In these series of tests using sensors, both TCS3200 and ISL29125 perform similarly and the deciding factor would be whether configuring the sensor is needed. The power draw of the ISL29125 may be considered as it can be set to standby / power down mode to consume much less power. Another consideration is the footprint of the two sensors, the TCS230 is bigger and has 4 LEDs mounted. From the documentation, the ISL29125 is designed for operation under dark glass cover which could be an option and the purpose of the IR configuration.

4.2.5. Web Camera

For yellow color detection using a camera, we obtained image samples for different yellow levels using enclosed camera set-up and gathered its RGB values and used the weighted sum to quantify color level. Using off-the-shelf web camera is a low-cost platform in detecting the yellow levels of NNJ among infants.

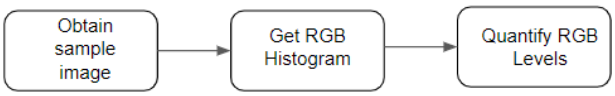


Figure 28: Flowchart for yellow color detection using camera

For this test, we used A4Tech PKS 730G (480p) Web Camera with 16MP camera sensor which supports up to 480p and with built-in LED. The camera was enclosed in a cup to prevent stray light from being detected by the camera. The camera could be connected to a laptop or low-cost Raspberry Pi computer via USB. The captured photo was then loaded to Processing software to

obtain the histogram of the image, generate a CSV file and quantify the R,G and B values using weighted average technique.



Figure 29: Test setup for yellow color detection using camera

Sample images from different levels of yellow opacity (0% most transparent; 100% most opaque) were obtained using the camera and the RGB values, using weighted average of the RGB histogram, were recorded as shown in *Figure 30*.

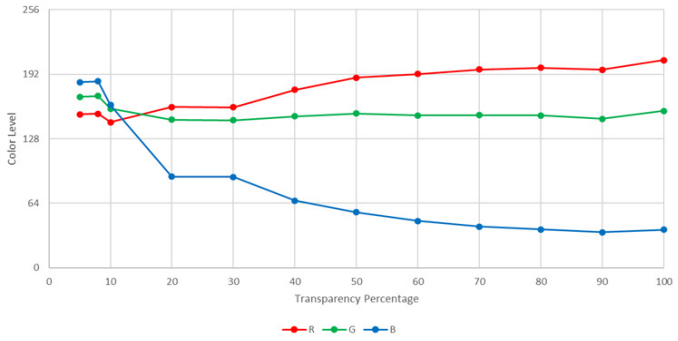
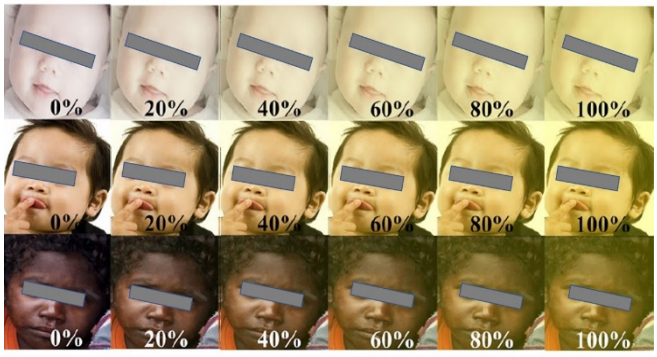


Figure 30: Different yellow levels for A4Tech PKS 730G

Theoretically, the red and green levels must remain constant and the blue level decreases as the yellow opacity level increases. From Figure 30, it could be observed that the red and green values started to stabilize at 50% yellow while values of blue followed a downward curve.

We expect fewer blue values as the blue color is absorbed by yellow, like how the phototherapy treatment works. To simulate the condition of a patient with NNJ, pictures of children with white, brown, and black skin tones were taken. The image used for the white skin tone swatch was from *pexels.com*, a free stock photo sharing website [25] while for the brown and black skin tone photo was from the photo collection of the author.



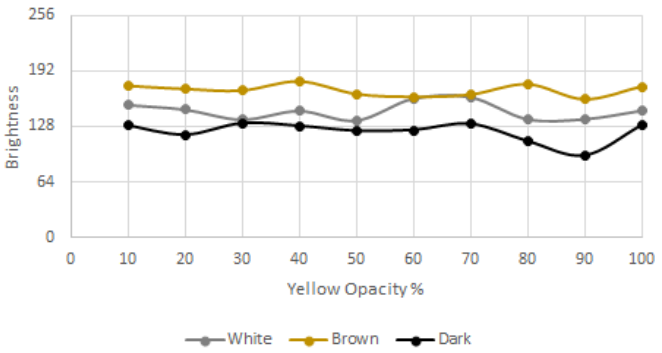


Figure 32: Red level for different baby skin types with yellow overlay

The yellow level corresponded to the opacity level of the yellow color. The swatches were built using Adobe Photoshop and were printed on a transparency sheet. The forehead area was chosen because face pictures were more available, and the said area had a strong correlation with serum bilirubin, as stated by Gwendoline Chimhini and her research collaborators [26].

In Figure 32, the brown skin type had a higher red level detected by the camera due to the brown color having red as the dominant color (brown RGB = 150, 75, 0). This test did not consider the reddening of the baby, particularly the white baby, due to blood or other sickness.

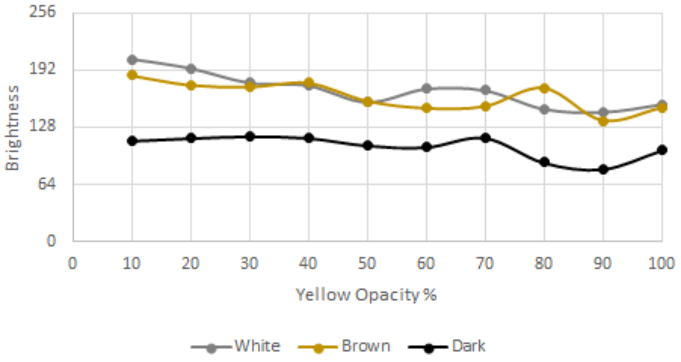


Figure 33: Blue level for different baby skin types with yellow overlay

In Figure 33, the blue levels detected for white and brown babies with yellow transparency overlaid were increasing as the yellowing of the baby decreased. This was due to the yellow color absorbing the blue. The decrease in yellow increased reflection of the blue level. However, for the black baby, no significant change in blue level was observed.

The improved electronic sensor and monitoring system complements the ability of doctors in observing and managing the condition of infants with NNJ and assists the doctors and patients in having a digital copy of medical records. In the future, we will explore and record various skin tones and nail colors of Filipino infants in photos as reference. The detection of the intensity of yellowness will be tested by placing different shades of yellow pigment over varied skin tones or nail colors.

4.2.6. Smartphone RGB Color Application

We also explored the use of Color Meter app, a simple color meter free mobile application that allows users to pick live colors and displays RGB color on the screen and the hexadecimal

(HTML) color code that is used in graphics and web applications that can detect changes in RGB values. We used a Samsung S9+ smartphone with $f/1.5$ and $f/2.4$ (Dual Aperture) 12-megapixel camera and to explore the RGB values at zero and 100% yellowness.

The changes in RGB values are detected upon overlay of different levels of yellowness (0% to 100% yellow) on varied skin tones represented by white-, brown- and dark-skinned photos of infants, as seen in *Figure 34*. The graph shows the different amount of reflectance of RGB colors upon overlay of zero percent yellow transparency and one hundred percent yellow transparency on three skin tones.

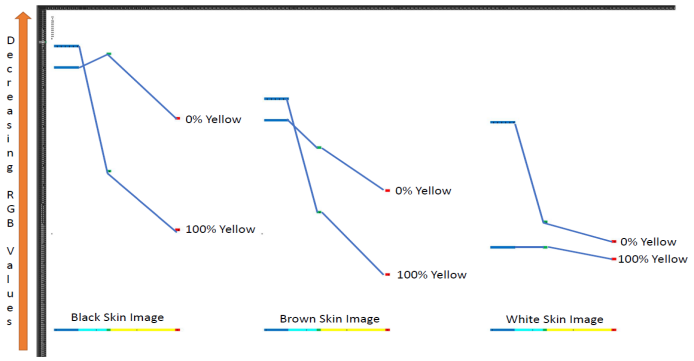


Figure 34: RGB values that correspond with zero percent and one hundred percent yellowness in three skin tones

Note that we flipped the graph by showing the lower values as you go up. In doing so, the behavior of the lines mimics the graph presented by Nabarun Polley and his team in 2015 on the absorption rate of colors in conjunctiva of infants with and without jaundice.

In summary, the color experimentation revealed the following results in Table 4 below.

Table 4: Changes of RGB values from zero to 100% yellow

Skin Tone	Blue Color Values	Red/Green Color Values
White	Decreased	Decreased
Brown	Decreased	Increased
Dark	Decreased	Increased

Zero and 100% yellowness show differences in scores of RGB. With white skin tones, all RGB scores went down in 100% yellowness indicating greater reflectance of light in jaundiced skin. In both brown and dark skin tones, Red and Green values increased while Blue score decreased when yellowness moved from zero to 100%. Thus, jaundiced babies with brown- and dark-skin tone tend to reflect Red and Green colors more than the non-jaundiced infants.

Varied percentages of yellowness in between zero and 100% bring about changes in the RGB scoring. Detection of smaller increments of yellowness will be further explored. We are also looking for a spectroscopic approach related to the study of Nabarun Polley and his team which will be reported soon.

4.3. Mobile or Near Cloud for Smart Nursing Station

The smart sensor and monitoring system of the Smart LPLS can be integrated with the Mobile or Near Cloud as part of the Smart Nursing Station for medical use cases. The main component of this Near Cloud System is the Raspberry Pi, a low-cost, credit card sized microcomputer that offers various interfaces. Raspbian, a Debian-based operating system, is burned to the micro SD card for the Raspberry Pi to function. For easy deployment, the system should be portable and have the necessary processing power for its essential functions.

Since content are cached in the Near Cloud node, the medical doctors and nurses as Mobile or Near Cloud clients can upload, download, or stream prepositioned health and medical data remotely in their own most convenient time using their Wi-Fi enabled devices. To observe security measures, administrators can identify users and set restrictions. Server-side encryption feature encrypts files stored on the server and files on remote storage. When encryption is enabled, all files are encrypted and decrypted by the software application and saved on the remote storage. Also, Wi-Fi Protected Access version 2 or WPA2, a method of securing your network with the use of the optional Pre-Shared Key authentication, is offered by the node. Time-based one-time passwords was also enabled for the system to add another layer of access security to the system.

This technology complements the off-grid setup of the Smart LPLS as it is readily deployable, works with or without internet connection, and can be powered by using a mobile power bank or DC adapter. It is of great value in remote health facilities for patient monitoring of select few doctors who work in barrios lacking in medical equipment and supplies, as featured by Juan Alfonso Leonardia and his co-proponents [27]. Its architecture features load-sharing and can be interconnected with other Near Cloud nodes via an inner wireless mesh network using B.A.T.M.A.N-adv routing protocol.

4.4. Improving Power Storage for Off-grid Deployment

Phototherapy treatment usually requires exposure of the infant patient for an extended period. However, deployment of a phototherapy unit in remote areas can present issues on power supply which is either intermittent or totally off grid. To address this problem, we designed the smart LPLS to be more energy-efficient, which makes provision of solar powering less costly and more feasible. We tested and compared the performance of lithium ion, motorcycle, and car batteries. The use of car and motorcycle batteries in powering a system in off grid deployment had been studied in a recent publication by Paul Cabacungan and his co-researchers [28].

In this paper, an experimentation on the discharging rate of the 12-volt lead-acid battery and 120 lithium-ion batteries (all of them with a lifespan of 3-5 years) was conducted using a seven-watt LED bulb. The lead-acid battery was connected to a 7-watt LED

bulb while the lithium-ion battery that was uniformly configured (Li-ion UC), i.e. with uniform number of battery cells in parallel, was attached to the same bulb with converter. A lithium-ion battery that was reconfigured (Li-ion RC), i.e., 45-30-45 batteries in parallel, used a charge controller instead of a converter. Results are shown in Table 5 below.

Table 5: Comparison of lead acid and lithium-ion batteries

	Ampacity	**Volume (sq.cm.)**	**Weight (kg.)**	**Duration of Discharging**
Lead Acid	50 AH	8840	15	26 hours
Li-ion UC	40 AH	3500	3	5 hours
Li-ion RC	40 AH	3500	3	17.75 hours

Results showed that lead-acid batteries took longer to discharge compared to lithium-ion. However, lithium-ion has its advantages in remote deployment - being twice smaller and 5 times lighter. Moreover, when batteries are discharged below their manufacturer-stated end-of-life discharge voltage, cells are permanently damaged, preventing it from ever being recharged. Lithium-ion batteries require low maintenance because each cell, once damaged, can be individually replaced without having to replace all the batteries in the system.

We have also done initial power calculations for a set of 10 conventional 60 cm LED strips with 36 LEDs each. 3 LEDs draw 20 milliamperes at 12 V to consume a total of 28.8 watts, about 70% less power consumption than the T5 LED. In using LED strips, an overall power consumption is estimated to decrease from 240 watts (SPLS) to 90 watts (SPLS off-grid), which already includes power usage of the charge controller, sensor, and monitoring system.

For future innovations, another prototype with top and bottom-mounted phototherapy lights that will operate on DC can be explored to eliminate the power consumed by the DC to AC converter and the T5 LED tubes' internal AC to DC converters. This prototype can use LED strips with transparent outer cover, which is expected to provide greater irradiance compared to T5 LED with translucent cover that limits light intensity.

5. Conclusion and Future Works

We presented an improved architecture and design for the current system, making the Smart LPLS a viable option for rural and underserved areas with limited resources to treat newborn patients with jaundice. The integration of skin color monitoring aids medical doctors by providing a non-invasive method of possible detection and monitoring of jaundice condition. Several approaches are being taken as we describe how we are developing a complementary system of transparencies that doctors can use in the field with photos to determine the severity of the yellowing and will aid in monitoring the progress of the treatment. We acknowledge the limitation of the current results on skin color monitoring, having tried the transparencies only on images with three specific kinds of skin color. We cannot claim a generalized conclusion yet, thus, we recommend that for future development, we can increase the sample size and cover a wider variety of skin colors. Also, aside from the usual placing of sensors on the forehead and chest, we explore putting it on other parts of the body such as nails, palm or sole.

The robust LED spotlights mounted on top and the additional T5 LED tube lights at the bottom offer more powerful and faster jaundice treatment due to greater irradiance. Its low power consumption and solar powering provision make it easily deployable even in remote off grid health systems. The Mobile Cloud System provides easy access to data which can supplement the patient's existing records.

The initial tests conducted in several patients exhibit results that are promising with treatment outcome and safety features comparable to commercial units available in the market today. However, we have yet to conduct further tests with actual human subjects, under the supervision of medical personnel to generalize results. To sum up, the Smart Low-cost Phototherapy Light System integrated with additional functions can still be made more energy-efficient, appropriately deployable in off-grid locations and yet maintain its affordable cost.

Acknowledgment

The prototypes built by Ateneo Innovation Center went under extensive research acceptance tests in public hospitals in Metro Manila by our collaborators – Dr. Vanessa Marie Calabia, Dr. Viel Bagunu, Dr. Ma. Lucila Perez, and Dr. Jeremie De Guzman. We thank them for their efforts and valuable feedback. We are grateful to our researchers, Mr. Alfonso Agustin, Mr. Alfonso Cuezon, Mr. Reymond Cao, Mr. Kerwin Caballas, Mr. Lawrence Ibarrientos, Mr. Lemuel Emman Tan and the interns and OJTs who worked in th research center for the past five years. In their capacities, they gave their valuable assistance in the different stages of fabricating the prototypes. We are indebted to the University Research Council (URC) of the Ateneo de Manila University for granting financial support to come up with this paper and to explore further developments of the Smart LPLS. Our gratitude goes to the Ateneo Innovation Center for serving as a hub for the phototherapy light system developments since its initial conceptualization.

References

[1]. P. Cabacungan, C. Oppus, J. De Guzman, G. Tangonan, I. Culaba and N. Cabacungan, "Intelligent Sensors and Monitoring System for Low-cost Phototherapy Light for Jaundice Treatment," 2019 International Symposium on Multimedia and Communication Technology (ISMAC), Quezon City, Philippines, 1-6, August 2019, doi:10.1109/ISMAC.2019.8836133.

[2]. V. Calabia, M. Perez, G. Tangonan, P. Cabacungan, I. Culaba, and J. De Guzman. "Bilirubin lowering effect and safety of a prototype low cost blue light emitting diode (LED) phototherapy device in the treatment of indirect hyperbilirubinemia among healthy term infants in a tertiary government hospital: a pilot study," arXiv preprint arXiv:2008.08875, August 2020.

[3]. V. Bagunu and M. Perez, "Comparative study of Prototype Low-Cost Light Emitting Diode Phototherapy Device versus Conventional Fluorescent Phototherapy in the Treatment of Indirect Hyperbilirubinemia among Term Infants in a Tertiary Government Hospital," Unpublished, August 2018.

[4]. C. Henny-Harry and H. Trotman, "Epidemiology of neonatal jaundice at the University Hospital of the West Indies," The West Indian Medical Journal,

61(1), 37-42, January 2012.
[5]. P. Kumar, D. Chawla, and A. Deorari, "Light-emitting diode phototherapy for unconjugated hyperbilirubinemia in neonates," Cochrane Database of Systematic Reviews, **12**(7969), December 2011, doi:0.1002/14651858.CD007969.pub2.
[6]. S. Moerschel, L. Cianciaruso, and L. Tracy, "A practical approach to neonatal jaundice," American Family Physician, **77**(9), 1255-1262, May 2008.
[7]. M. Fargo, S. Grogan, and A. Saguil, "Evaluation of Jaundice in Adults," American Family Physician, **95**(3), February 2017.
[8]. E. Gurley, A. Halder, P. Streatfield, H. Sazzad, T. Nurul Huda, M. Jahangir Hossain, and S. Luby, "Estimating the burden of maternal and neonatal deaths associated with jaundice in Bangladesh: possible role of hepatitis E infection," American Journal of Public Health, **102**(12), 2248-2254, December 2012, doi:10.2105/AJPH.2012.300749.
[9]. Iowa Department of Public Health, "Managing Newborn Hyperbilirubinemia and Preventing Kernicterus Progeny,"http://www.idph.state.ia.us/hpcdp/common/pdf/perinatal_newsletters/progeny_june2013.pdf. Accessed: August 2020
[10]. World Health Organization International, "Bilirubinometer - Core medical equipment - Information," https://www.who.int/medical_devices/innovation/bilirubinometer.pdf. Accessed: August 2020.
[11]. PhilHealth, "Agency's Mandate And Functions," https://www.philhealth.gov.ph/about_us/mandate.html#gsc.tab=0. Accessed: August 2020.
[12]. H. Vreman, R. Wong, and D. Stevenson, "Phototherapy: current methods and future directions," Seminars Perinat, **28**, 26-33, October 2004, doi: 10.1053/j.semperi.2004.09.003.
[13]. J. dela Cruz, N. Libatique and G. Tangonan, "Design of a Disaster Information System using Mobile Cloud Wireless Mesh with Delay Tolerant Network," 2019 IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA, 1-8, October 2019, doi: 10.1109/GHTC46095.2019.9033450.
[14]. N. Mercado, Design and Demonstration of Near Cloud System for Digital Education and Disaster Resiliency, MS ECE thesis, Electronics, Computer and Communications Engineering, Ateneo de Manila University, Quezon City, Philippines, May 2020.
[15]. J. Mamaradlo, Design and Demonstration of a Mobile Cloud System for Smart Transportation System Use Case, MS ECE thesis, Electronics, Computer and Communications Engineering, Ateneo de Manila University, Quezon City, Philippines, May 2020.
[16]. J. Mamaradlo, N. Mercado, N. Libatique, G. Tangonan, R. Solis, V. Rodriguez, B. Dingel, C. Pineda, and C. Lopez, "University campus 5G testbed and use case deployments in the Philippines," SPIE Proceedings, Broadband Access Communication Technologies XIV, **11307**, January 2020, doi:3. 10.1117/12.2549903 2020.
[17]. O. Onwe, O. Ezeanosike, C. Obu, O. Daniyan, and H. Amadi, "Economic cost savings with the use of total body phototherapy for the treatment of severe neonatal jaundice in Nigeria," Journal of Pediatrics and Neonatal Care, **9**(6), December 2019, doi: 10.15406/jpnc.2019.09.00401.
[18]. A. Riskin, A. Tamir, A. Kugelman, M. Hemo, and D. Bader, "Is Visual Assessment of Jaundice Reliable as a Screening Tool to Detect Significant Neonatal Hyperbilirubinemia?," The Journal of Pediatrics, January 2008, doi: 10.1016/j.jpeds.2007.11.003
[19]. H. Baharuddin, M. Sulong, A. Joret, T. Rahman, and N. Ismail, "Bile Pigments Detection via IR Sensor," Proceedings of EnCon2010 3rd Engineering Conference on Advancement in Mechanical and Manufacturing for Sustainable Environment, Kuching, Sarawak, Malaysia, April 2010.
[20]. N. Polley, S. Saha, S. Singh, A. Adhikari, S. Das, B. Choudhury, and S. Pal, "Development and optimization of a noncontact optical device for online monitoring of jaundice in human subjects," Journal of Biomedical Optics, **20**(6), June 2015, doi: 067001
[21]. Learn Leight Cotnoir, "Hue, Value, Saturation," http://learn.leighcotnoir.com/artspeak/elements-color/hue-value-saturation/. Accessed: August 2020
[22]. P. Mollica, Color Theory: An essential guide to color-from basic principles to practical applications (Artist's Library), Walter Foster Publishing, 17, January 2013.
[23]. Random Nerd Tutorials, "Guide for TCS230/TCS3200 Color Sensor with Arduino," https://randomnerdtutorials.com/arduino-color-sensor-tcs230-tcs3200/. Accessed: August 2020.
[24]. Renesas Electronics, "ISL29125 | Light-to-Digital Sensors," https://www.renesas.com/us/en/products/sensors/ambient-light-sensors/light-to-digital-sensors/device/ISL29125.html. Accessed: August 2020.
[25]. Pexels.com, "Photo of Couple with Baby Sitting," https://www.pexels.com/photo/photo-of-couple-with-a-baby-sitting-3995914/. Accessed: August 2020.
[26]. G. Chimhini, S. Chimhuya, and V. Chikwasha, "Evaluation of transcutaneous bilirubinometer (DRAEGER JM 103) use in Zimbabwean newborn babies," Maternal Health, Neonatology and Perinatology, **4**(1), April 2018.
[27]. J. Leonardia, H. Prytherch, K. Ronquillo, R. Nodora and A. Ruppel, "Assessment of factors influencing retention in the Philippine National Rural Physician Deployment Program," BMC Health Services Research, **12**(411), November 2012.
[28]. P. Cabacungan, G. Tangonan, and N. Cabacungan, "Water - Electricity - Light System (WELS): Technology Innovations," 2020 International Journal of Recent Technology and Engineering (IJRTE), **8**(6), 3061-3068, March 2020, doi: 10.35940/ijrte.F8103.038620

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 469-476 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Sound of Trust: Towards Modelling Computational Trust using Voice-only Cues at Zero-Acquaintance

Deborah Ooi Yee Hui[1], Syaheerah Lebai Lutfi[1,*], Syibrah Naim[2], Zahid Akhtar[3], Ahmad Sufril Azlan Mohamed[1], Kamran Siddique[4,*]

[1]*Universiti Sains Malaysia, School of Computer Sciences, Penang, 11800, Malaysia*

[2]*Woosong University, Technology Department, Endicott College of International Studies, Daejeon, 34606, Republic of Korea*

[3]*University of Memphis, Tennessee, 38152, United States of America*

[4]*Xiamen University Malaysia, Sepang, 43900, Malaysia*

ARTICLE INFO

Article history:
Received: 16 May, 2020
Accepted: 26 June, 2020
Online: 09 August, 2020

Keywords:
True
Voice
Artificial neural network
Zero acquaintance

ABSTRACT

Trust is essential in many interdependent human relationships. Trustworthiness is measured via the effectiveness of the relationships involving human perception. The decision to trust others is often made quickly (even at zero acquaintance). Previous research has shown the significance of voice in perceived trustworthiness. However, the listeners' characteristics were not considered. A system has yet to be produced that can quantitatively predict the degree of trustworthiness in a voice. This research aims to investigate the relationship between trustworthiness and different vocal features while considering the listener's physical characteristics, towards modelling a computational trust model. This study attempts to predict the degree of trustworthiness in voice by using an Artificial Neural Network (ANN) model. A set of 30 audio clips of white males were obtained, acoustically analyzed and then distributed to a large group of untrained Malaysian respondents who rated their degree of trust in the speakers of each audio clip on a scale of 0 to 10. The ANOVA test showed a statistically significant difference of trust ratings across different types and intensities of emotion, duration of audio clip, average fundamental frequencies, speech rates, articulation rates, average loudness, ethnicity of listener and ages of listener ($p < .01$). The findings conclude that Malaysians tend to trust white males who talk faster and longer, speak louder, have an f0 between 132.03Hz & 149.52Hz, and show a neutral emotion or rather stoic (arousal<.325). Results suggest that Indians are the most trusting Malaysian ethnic group, followed by Bumiputera from East Malaysia and then followed by Malays. Chinese are the least trusting Malaysian ethnic group. The data was fed into an ANN model to be evaluated, which yielded a perfect percentage accuracy (100%) in degree of trustworthiness 39.70% of the time. Given a threshold of two-point deviation, the ANN had a prediction accuracy of 76.86%.

1. Introduction

Trust is a fundamental part of various commercial and interpersonal relationships, especially wherever risk, uncertainty, or interdependence exist. Human trust is influenced by perception, intuition, environment and many other factors to the human centred scheme that is generally defined in any communication systems, natural and synthetic environment in everyday human interaction. Investigating state of the art in theoretical studies in trust and extending the finding as a base to symmetrically model and simulate trust in synthetic agents is crucial, especially in this digital era. In [1] it is stated that an abundance of information about other people can be determined through only a quick observation of their expressive behaviour. Humans can gauge trustworthiness in zero acquaintance situations [1], [2]. Zero acquaintance is a situation in which perceivers make judgments about targets they are given no opportunity to interact with [3]. Several other studies have been conducted previously with regards to zero acquaintance

*Corresponding Authors: Syaheerah Lebai Lutfi & Kamran Siddique, Emails: syahherah@usm.my, kamran.siddique@xmu.edu.my

[4]–[7], with each experiment utilizing thin slices that last from 5 seconds to 30 seconds each. In a successful study by Mohammad and Kiritchenko [8], only short 10-second audio clips were used to detect the effects of prosodic features of voice on personality. Therefore, the 10s length is enough to capture personality impressions at zero acquaintance while lightening the load on the listeners.

In recent years, many studies have shown the significance and importance of voice in perceived trustworthiness. These studies have analysed the effect of various vocal cues (e.g., pitch, speech rates, and fundamental frequency (f0)) on trustworthiness. However, prior research did not consider the characteristics of the listener (the rater/ trustor). It is important to include the characteristic of the listener as well, since trust is a concept that involves two parties [9].

Recently, chatbots or robots have started taking over the roles of teachers and tutors [10]. One such example of a chatbot would be Jill Watson, who was created to handle forum posts by students [11]. However, chatbots communicate primarily via text, which would be inconvenient visually impaired students or students who suffer from dyslexia. Furthermore, [12] indicated that a chatbot's suggestions are not fully trusted by humans. Since voice has a significant impact on perceived trustworthiness, people might trust the chatbots more if the chatbots had a trustworthy voice. Therefore, this research aims to determine the specific characteristics displayed in a "trustworthy" voice. When it comes to chatbots, many studies have been conducted to investigate the aptitude and efficiency of the chatbot applications, ignoring the voice characteristics. One interesting study in the potential of having chatbot imitate human ability during a conversation is by [13]. The outcome of the study showed that the "uncanny valley" effect (where people become uncomfortable to a lifelike robot to a certain point) is lowered by using psychophysiological responses in simple chatbox developments compared to a multifaceted one.

A prediction model would be useful to measure the degree of trustworthiness of the chatbot's synthetic voice before releasing it to the public. There has yet to be a system that can quantitatively predict the degree of trustworthiness in a voice.

This research aims to investigate the relationship between vocal cues and trustworthiness at zero acquaintance, and to predict the degree of trustworthiness of a voice by using an Artificial Neural Network (ANN) model. However, there are several constraints.

Firstly, the chosen audio clips only consist of white male speakers and the accents of the voices are therefore only limited white accents. Secondly, the survey does not consider the listener's emotional state at the point of answering. Thirdly, due to time and financial constraints, this survey was only open to Malaysians. Hence, all the findings of this study are only valid and applicable among Malaysians.

2. Background and Related Work

In recent years, the significance of voice in perceived trustworthiness has been investigated in many previous studies [14]–[18]. Many papers have also investigated the effect of individual vocal cues on trust at zero acquaintance.

In [19] it is noted that pitch has a small role in detecting cheating probability of infidelity among committed romantic partners, but it does not represent the entire picture. Hughes and Harrison [19] did not consider other acoustical measures or any combination of different acoustic measures in the research.

The authors in [20] utilized data-driven voice computational modelling to study the relationship between acoustics and perceived trustworthiness in the single word "hello". The findings demonstrate a strong acoustical basis for perceived voice personality. However, this research did not take the listener characteristics into consideration at all.

In [18] it is investigated the effect of trustworthiness impressions on vocal expressive and person characteristics and how the relationship can be explained by acoustical profiles. In [18] it is indicated that positive speaker valence but not arousal, consistently predict greater perceived trustworthiness. Furthermore, female as compared with male speakers and younger voices compared with older were perceived as more trustworthy. However, [18] did not consider the ethnicity of the listener. Ethnicity of listeners is vital as interpersonal trust is largely influenced by culture [21]–[23].

A computer based, acoustic profile of Steve Job's tone of voice was created by Niebuhr et al. [24] Findings indicate that all the melodic features of charisma could be found in his tone of voice. However, this study is purely from the author's assumption of Steve Jobs as a charismatic speaker, and the author did not take into consideration the perception of the general public. Furthermore, this study only investigated the perception of charisma in a voice and not trustworthiness.

With regards to evaluating the voice of an artificial intelligence (A.I.) instructor, [10] indicated that students who perceive themselves as being "older" trust the older A.I. voice more than a younger A.I. voice. Although this study did consider the listener's age, they did not consider other acoustic or prosody characteristics of the A.I. instructor's voice.

There also exists some degree of dissension regarding the direction of association between fundamental frequency (f0) and trustworthiness. On one hand, a deception study indicated that participants consistently raised their f0 when lying [25] while another study indicated the opposite [26]. Yet another further study found no differences in f0 or other acoustic characteristics when analyzing deceptive and truthful messages [27]. This situation highlights the fact that there exist inconsistent results from research regarding the direction of association between fundamental frequency (f0) and trustworthiness.

Research has also used prosody features to measure trustworthiness. In [28] the authors used the investment game to assess the effect of prosody characteristics on trust attributes. Although higher investments (higher trust) were associated with faster articulation rate and higher pitch, this finding might have been influenced by game behavior and game turn, both of which have the highest effect on investments (trust).

Evidently, previous research has only considered the impact of several different vocal cues on trust at zero acquaintance. This research aims to provide a more comprehensive analysis of the effect of vocal cues on trust at zero acquaintance by including more

vocal features. The specific acoustic characteristics that will be investigated in this research are (a) Mean f0 (b) Min f0 (c) Max f0 (d) Mean Intensity/Loudness (e) Min Intensity (f) Max Intensity (g) Articulation Rate (h) SD of f0 (i) Speech Rate.

Most of the researches mentioned did not consider the characteristics of the listener (the rater). When the characteristics of the listener were considered, the ethnicity of the listener was not taken into consideration. It is important to include the characteristics of the listener as well, since trust is a concept that involves two parties [9]. This study will, at the very least, take the physical characteristics of the listener into consideration e.g. age, gender, English proficiency, and ethnic groups. In addition to that, this study will also consider the arousal and valence of the emotion displayed by the speaker.

3. Recent Algorithms Used to Measure / Predict Trust

Very few studies have attempted to quantitatively measure the degree of trustworthiness. Korovaiko and Thomo [29] have attempted to predict trust from online review ratings by using Random Forest models and Support Vector Machines, and [30] have tried predicting trust and distrust between two strangers in social networks by using an inference algorithm. Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

ANN models have also been used to create a broker-assisting trust and reputation system and to investigate the potential determinants of relationship quality in relationship marketing [31, 32]. Although the research conducted by [32] is closely related to this research, it considers trust as one of the measurements of relationship quality instead of the feature to be predicted.

Lee et al. [33] used a Trust Hidden Markov Model to predict whether an individual is willing to behave cooperatively (trust) with their novel partner based on gestural cues. Although Lee et al. [33] does use artificial intelligence to predict trust at zero acquaintance, this research investigates the effect of gesture and not voice on interpersonal trust.

Evidently, there has been no prior research using artificial intelligence to predict the degree of trustworthiness in a person's voice. Therefore, this study aims to predict the degree of trustworthiness in a person's voice by using an ANN model that will take both the acoustic characteristics of the speaker's voice and the physical characteristics of the listener into consideration.

A typical artificial neural network model (ANN) was used in this research. The ANN is a feedforward neural network with an input layer of source neurons, one or more hidden layer(s) of computational neurons, and an output layer of computational neurons. The input layer receives input signals and redistributes these signals to all neurons in the hidden layer. The hidden layer detects the features and represent the features in the input patterns by manipulating the weights of the neurons. The output pattern of the entire network is established by the output layer. The ANN is also capable of supervised learning through the backpropagation learning algorithm. In other words, the ANN is able to learn from its mistakes and improve the accuracy of its predictions automatically [34].

The use of a feedforward backpropagation ANN model was due to several reasons: a) capable of supervised training, b) minimization of training errors, c) robustness to noisy data, d) high computational speed across large input datasets. The main disadvantage is the many variables and the training time which must be considered when constructing an ANN.

In this experiment, a large number of variables must be considered when predicting the degree of trustworthiness, where the exact relationship between each variable on trustworthiness still remains a mystery. As such, the use of an ANN model is the most suitable approach to represent such a complex relationship and adequately predict the degree of trustworthiness in a voice.

4. Materials and Methods

The methodology undertaken in this research is similar to the methodology undertaken by Hughes and Harrison [19], Imhof [35] and Smith et al. [36]. Participants were informed about the study procedure and given the option to opt out of the online survey at any point of time.

4.1. Participants

This study was conducted in Universiti Sains Malaysia. Malaysia has a multicultural population that consist of 3 main ethnic groups; Chinese, Malay and Indian. There also exist other minority ethnic groups from East Malaysia, such as Kadazans, Bajaus and Dayaks [37], [38]. Since these minority ethnic groups share the bumiputera (native/indigenous) status with the Malays, we will consider this minority ethnic groups as "Other Bumiputera".

In total, the participants of this study consisted of 86 Malaysian individuals (44 men and 42 women) of which 40 were Chinese, 21 were Malay, 15 were Indian and only 10 were from Other Bumiputera. The participants had a mean age of 24.337 years (SD 5.517, 17-53). 64 of them (74.42%) were students while 22 (25.58%) were non students. The participants were also asked to rate their own English proficiency. 12 (13.95%) participants considered their English to be Highly Proficient, 35 (40.70%) satisfactory, 19 (22.09%) proficient, 15 (17.44%) modest, 2 (2.33%) expert and 3 (3.49%) limited command of English.

4.2. Stimuli

Six labelled audio clips were sourced from the One-Minute Gradual-Emotion Recognition (OMG-Emotion) dataset [39]. The dataset comprises of 567 YouTube videos, which were divided into clips of varying length and labelled based on utterances. At least five independent subjects annotated each utterance using an arousal/valence scale and emotion tags based on the six universal emotions from Ekman (fear, disgust, happiness, sadness, surprise and anger) [40].

The 6 audio clips to be used in this research were chosen based on the following conditions: a) each audio clip must be about 10s long, b) it must be from a white male speaker and c) there must be a balanced number of audio clips representing each different emotion.

The audio clips were digitally manipulated with Praat into a higher and lower pitch and speech rate version. The degree to which it was manipulated was based from previous research

conducted by [41], [42]. This manipulation results in a total of 30 audio clips being formed (6 original and 6*4 modified clips), which will be used for this research.

The following acoustic characteristics of each audio clip was extracted with Praat and tabulated: (a) Mean f0 (b) Min f0 (c) Max f0 (d) Mean Intensity/Loudness (e) Min Intensity (f) Max Intensity (g) Articulation Rate (h) SD of f0 (i) Speech Rate.

4.3. Procedure

86 Malaysian respondents were asked via an online survey to listen to the audio clips and rate their degree of trust in the speakers of each audio clip on a scale of 0 to 10 (0 being absolutely untrustworthy and 10 being most trustworthy). Respondent details such as whether they were Malaysian or non-Malaysian, Student or nonstudent, Gender, Age, Ethnicity and English proficiency were also collected in the survey. The responses were recorded and tabulated.

Each individual response was mapped to the acoustic characteristics and the emotion annotations of the respective audio clips as obtained from the data set. After data cleaning, the total number of valid participants were reduced to 74 (N=2220).

Analysis of Variance (ANOVA) was used to evaluate the null hypothesis that there is no statistically significant difference between each vocal cue or rater characteristic on the trustworthiness ratings. Each vocal cue or rater characteristic was used as the dependent variable and the trust ratings were used as the independent variable. The Welch test was used when the vocal cues violated the assumption of homogeneity of variance.

Since arousal and valence are closely related, a Multivariate Analysis of Variance (MANOVA) test was conducted with arousal and valence as the dependent variables and trust ratings as the independent variable. However, the ANOVA and MANOVA test does not tell us exactly how the mean varies between different groups of the data. Therefore, a means plot of each vocal cue on trust rating was carried out.

The data was split into train, validation and test sets. The train set trained the neural network and the validation set verified the prediction accuracy. Finally, the model was run on the test set to determine the model performance on unfamiliar data.

4.4. Artificial Neural Networks

A feedforward backpropagation ANN was used in this research. The model aims to obtain a low mean-squared-error (MSE) as well as a high percentage of low-deviation predictions. The network was built with Keras API. The model was compiled with Adam optimizer with learning rate = 0.0005, $\beta 1 = 0.9$, $\beta 2 = 0.999$, =None, decay=0.0003 and amsgrad=False. The loss function used in this neural network is categorical cross-entropy, with mean squared error and accuracy as the metrics for checkpointing.

After many simulations, the architecture of the ANN that produced the lowest mean squared error is as follows:

1. Input layer: Dense,28 input neurons, ReLu activation.
2. Hidden layer: Dense, 40 neurons, ReLu activation.
3. Dropout layer: Dropout rate=0.2
4. Hidden layer: Dense, 30 neurons, ReLu activation
5. Dropout layer: Dropout rate=0.4
6. Hidden layer: Dense, 16 neurons, ReLu activation Output layer: Dense, 11 output neurons, Softmax activation

The data underwent feature selection with the score function f_classif from SelectKBest prior to being used for ANN. Findings indicate that only the maximum f0, average f0 and minimum f0 had insignificant results ($p > .05$) and hence these features were removed from the data set. The input layer of the ANN has 28 nodes for each of the acoustic characteristics, emotion annotations as well as listener characteristics. The output layer has 11 nodes for predicting the degree of trust on a scale of 0 to 10. Softmax activation was used in the output layer for multiclass classification.

The model was fit with the training data set (n=1506) and run over 600 epochs, with batch size of 12 and the validation data set (n=377) was simultaneously used to check for overfitting. EarlyStopping callback was used to ensure minimal overfitting by monitoring accuracy. Finally, the ANN was tested with the testing dataset.

The predictions obtained were recorded and the results were evaluated by measuring the deviation of the predicted trust ratings from the original. In this study, the threshold for acceptable deviation in predictions is arbitrarily set to 2. This method of model evaluation has not previously been used in any prior research. The use of a novel evaluation approach is due to several reasons.

Firstly, prior research that uses Artificial Intelligence to predict trust considers trust as a binary variable. However, Yulin [43] indicates that trust is a spectrum. To simplify and bin the responses into either trustworthy or untrustworthy would significantly reduce the richness of the data and the subsequent results. Hence, for this study, the author maintains the view of trust as a spectrum instead of binary variables.

By setting the threshold to 2, predictions that are less than or equal to 2 marks away from the actual rating would then also be considered as accurate. As the original purpose of the ANN model is to measure the degree of trustworthiness of a chatbot's synthetic voice, it is more realistic to provide an acceptable margin of error, rather than ensuring that the model predicts a specific number.

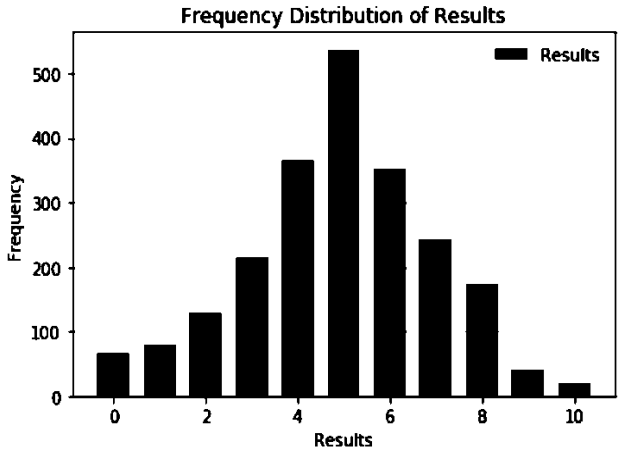


Figure 1 Frequency Distribution of Results.

5. Results

A graphical view of the frequency distribution of the results are shown in Figure 1, It is observed that the results are normally distributed around the mean (μ = 4.90, SD = 2.052) with skewness = -0.228 and kurtosis = -0.060. From Figure 1, it is observed that the majority of trust ratings (n = 537) were 5, and the two extreme trust ratings (0 and 10) were less commonly chosen (n = 68, n=19) respectively.

5.1. The Relationship between Vocal Cues and Trustworthiness

There was a statistically significant difference between arousal of emotion, valence

of emotion, type of emotion, duration of audio clip, speech rate, articulation rate, average loudness, ethnicity of rater and age of rater on trust ratings ($p < .01$). A more detailed view of the results is shown in Table 1.

Table 1: ANOVA / Welch test results between each individual vocal cue on trust ratings

Vocal Cue	Type of Test	df1	df2	F	p	ω
Duration	Welch	10	311.15	10.73	0	0.75
Speech rate	ANOVA	1	10	4.46	0	0.14
Articulation rate	ANOVA	1	10	5.95	0	0.16
Average Loudness	Welch	10	306.56	4.77	0	0.58
Arousal	ANOVA	1	10	3.36	0	0.12
Valence	Welch	10	307.34	2.41	0.009	0.4
Age of rater	Welch	10	327.13	7.88	0	0.7

There was also a statistically significant difference across trust ratings when considered jointly on arousal and valence. Pillai's Trace = .028, F (20, 4358) = 3.147, p = .000, partial η2 = .014. The average f0 was split into 4 groups using the first, second and third percentile as cutoff points. The groups were: Low f0 (M=107.03, SD=6.87, n=511), LowMedium f0 (M=125.23, SD=4.52, n=584), Medium-High f0 (M=139.03, SD=5.57, n=511), and High f0 (M=191.19, SD=29.58, n=584). The average f0 was used as the independent variable and the trust ratings were used as the dependent variable. Likewise, the emotion and ethnicity of rater were also used as the independent variable. The results are as shown in Table 2.

Table 2 ANOVA / Welch test results between trust ratings and the groups of each characteristics

Characteristic	Type of Test	df1	df2	F	p	ω
Emotion	Welch	5	76.69	19.05	0	0.74
Ethnicity of rater	Welch	3	833.733	33.29	0	0.75
Average *f0*	ANOVA	1	3	12	0	0.13

The means plot indicate that Malaysians trust more when the white male speaker a) speaks for a longer period (12s) than a shorter period (10.5s), b) speaks faster (170wpm and/or 3.6 syllables per second) than slower (153wpm and/or 3.1 syllables per second), c) has a louder voice (roughly about 56dB), d) has an average f0 which is between 132.03Hz and 149.52Hz, e) shows a neutral emotion or less emotion (arousal<.3250). Finally, findings among Malaysians indicate that Indians are the most trusting ethnic group with trust ratings of (5.5< μ <5.75), followed by "Other Bumiputera" (5.0< μ <5.25) and then Malay (μ <4.75). Chinese are the least trusting Malaysian ethnic group. Graphical representations of the means plots are as shown in Figure 2 to Figure 9.

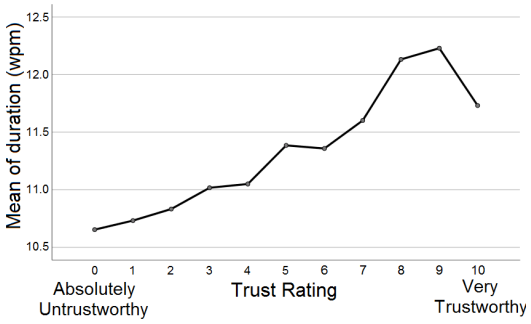


Figure 2 Means plot of duration across trust ratings.

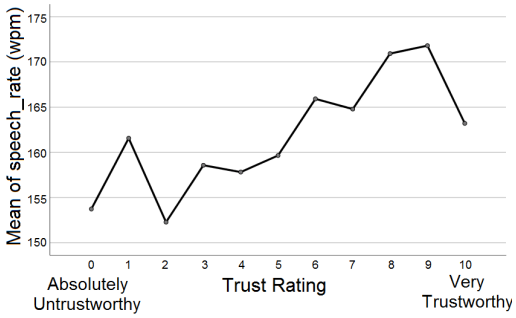


Figure 3 Means plot of speech rate across trust ratings.

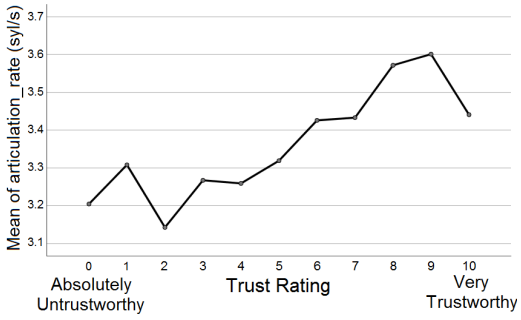


Figure 4 Means plot of articulation rate across trust ratings.



Figure 5 Means plot of loudness across trust ratings.

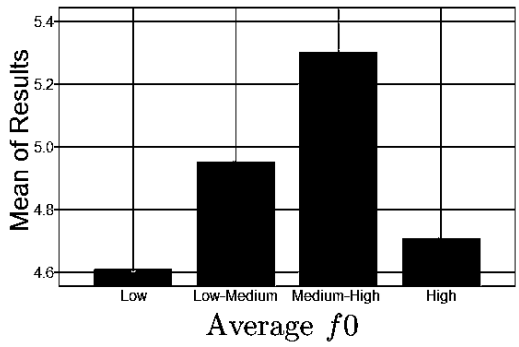


Figure 6 Means plot of f0 across trust ratings.

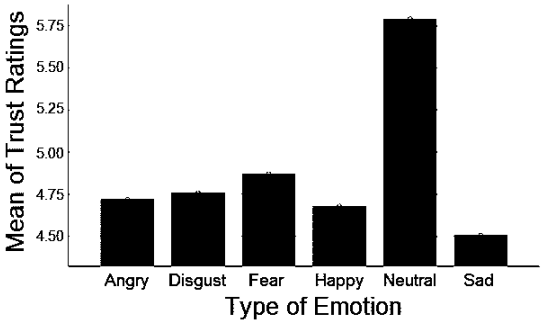


Figure 7 Means plot of emotion across trust ratings.

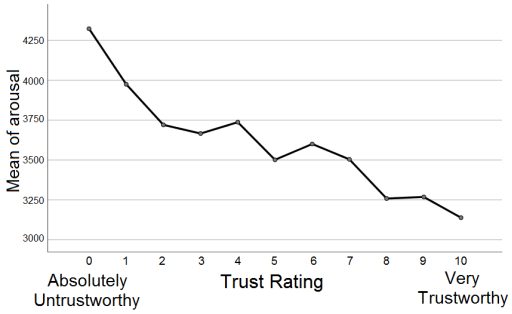


Figure 8 Means plot of arousal across trust ratings.

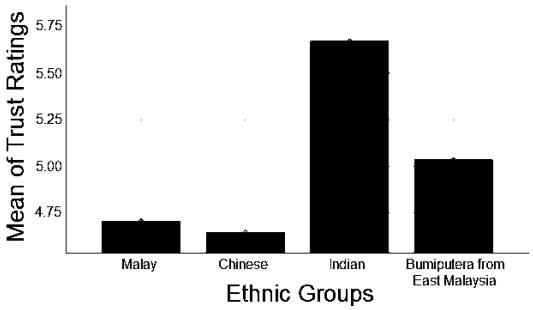


Figure 9 Means plot of ethnic groups across trust ratings.

5.2. Artificial Neural Network

The results of the ANN predictions are as shown in Figure 10, where "freq" represents frequency of occurrence and "difference" represents difference between the predicted and original value. It indicates how much deviation has happened between the predicted results and the original trust rating, and how often that deviation has occurred. Results indicate that the majority of predictions (76.86%) have an acceptable difference (the difference is less than or equal to 2), and only a few predictions (23.14%) are very far off the mark (having a difference of more than 2).

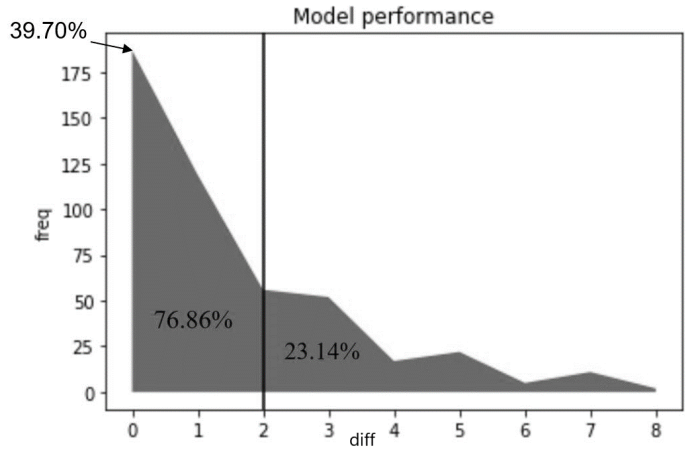


Figure 10 The cumulative percentage performance of ANN.

5.3. Discussion and Conclusion

The accurate predictions of the ANN clearly signify a relationship between vocal cues and trustworthiness. Contrary to previous related research, findings indicate that Malaysians tend to trust white male voices the most when the average f0 of the voices fall between 132.03Hz and 149.52Hz. This could potentially explain the reason behind the degree of dissension regarding the direction of association between different vocal features. It is not a direct association but rather, it is a parabolic curve with its peak when the average f0 is between 132.03Hz and 149.52Hz.

Findings prove that Malaysians tend to trust people who talk faster and for a longer time versus slower and shorter. This could be attributed to knowing more of the context of the audio clip and then judging the perceived trustworthiness according to the speaker's knowledge of the topic. Results also suggest that Indian Malaysians trust more than the other ethnic groups in Malaysia, while Chinese Malaysians trust the least. The underlying reasons for this relationship are still unknown, and it could potentially be investigated in future work.

Findings show that Malaysians tend to trust people with louder voices. This relationship could be due to the clarity of the audio clip - it is easier to identify the words being said when the voice is louder. Hence, clarity of voice could potentially be used as another variable when carrying out future work.

With regards to the ANN, 39.70% of the testing data set has been perfectly predicted by the neural network. With the threshold set at 2, the percentage accuracy of the neural network has risen to 76.86%. With such high trust prediction accuracy, the neural network model is potentially capable of fulfilling its original purpose.

In conclusion, there was a statistically significant difference between the arousal of emotion, valence of emotion, type of emotion, duration of audio clip, average f0, speech rate,

articulation rate, average loudness, ethnicity of rater and age of rater on trust ratings ($p < .01$).

The findings from this research also conclude that Malaysians tend to trust white males who a) talk faster and longer vs slower and shorter, b) speak louder, c)have an f0 between 132.03Hz and 149.52Hz, d) show a neutral emotion or rather stoic (arousal<.325). Furthermore, results prove that Indians are the most trusting Malaysian ethnic group, followed by Bumiputera [1] from East Malaysia and followed by Malays. Chinese are the least trusting Malaysian ethnic group.

Lastly, the degree of trustworthiness has been quantitatively measured and successfully predicted by using an ANN model. The ANN model perfectly predicts the trustworthiness of a voice to an accuracy of 39.70% and given a threshold of two-point deviation, had a prediction accuracy of 76.86%. This Neural Network model will hopefully prove invaluable when used to predict the degree of trustworthiness in a chatbot's synthetic voice.

Although these conclusions have been derived and proven to the best of the authors' abilities, there is still room for further validation. As this research is the first of its kind, especially pertaining to the use of a Neural Network model to predict the degree of trustworthiness in a voice, there is a need for more researchers to replicate or potentially improve the neural network model. Future researchers investigating this topic could possibly expand the variety of audio clips to encompass both male and female voices and people from various races. Future work could potentially include listener's emotional state and clarity of the audio clip as a factor in perceived trustworthiness. Lastly, this research would bring more value if future researchers are able to obtain responses from an international audience.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the Research Management Center, Xiamen University Malaysia under the XMUM Research Program Cycle 3 (Grant XMUMRF/2019-C3/IECE/0006). The authors also thank Universiti Sains Malaysia for the partial funding of this work from the grant no. 304/PKOMP/6315137 and 1001/PKOMP/8014001.

References

[1] N. Ambady, M. A. Krabbenhoft, and D. Hogan, "The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness," J. Consum. Psychol., **16**(1), 4–13, 2006. https://doi.org/10.1207/s15327663jcp1601_2.

[2] M. Stirrat and D. I. Perrett, "Valid Facial Cues to Cooperation and Trust: Male Facial Width and Trustworthiness," Psychol. Sci., **21**(3), 349–354, 2010. https://doi.org/10.1177/0956797610362647.

[3] D. A. Kenny and T. V West, "Zero acquaintance: Definitions, statistical model, findings, and process," First impressions, 129–146, New York, Guilford Press, 2008.

[4] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness.," J. Pers. Soc. Psychol., **64**(3), 431, 1993. https://doi.org/10.1037/0022-3514.64.3.431.

[5] N. Eisenkraft, "Accurate by way of aggregation: Should you trust your intuition-based first impressions?," J. Exp. Soc. Psychol., **49**(2), 277–279, 2013. https://doi.org/10.1016/j.jesp.2012.11.005.

[6] M. A. Hecht and M. LaFrance, "How (Fast) Can I Help You? Tone of Voice and Telephone Operator Efficiency in Interactions 1," J. Appl. Soc. Psychol., **25**(23), 2086–2098, 1995. DOI: 10.1111/j.1559-1816.1995.tb02389.x.

[7] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," 32, Academic Press, 201–271, 2000. https://doi.org/10.1016/S0065-2601(00)80006-4.

[8] S. M. Mohammad and S. Kiritchenko, "Using nuances of emotion to identify personality," Proc. ICWSM, 2013. https://arxiv.org/abs/1309.6352.

[9] H. H. Höhmann and E. Malieva, "The concept of trust: Some notes on definitions, forms and sources," Trust Entrep.), 7–23, 2005.

[10] C. Edwards, A. Edwards, B. Stoll, X. Lin, and N. Massey, "Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions," Comput. Human Behav., 90, 357–362, 2019. https://doi.org/10.1016/j.chb.2018.08.027.

[11] A. Goel, B. Creeden, M. Kumble, S. Salunke, A. Shetty, and B. Wiltgen, "Using watson for enhancing human-computer co-creativity," in 2015 AAAI Fall Symposium Series, 2015.

[12] A. Følstad, C. B. Nordheim, and C. A. Bjørkli, "What makes users trust a chatbot for customer service? An exploratory interview study," in International Conference on Internet Science, 194–208, 2018. https://doi.org/10.1007/978-3-030-01437-7_16.

[13] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," Futur. Gener. Comput. Syst., 92, 539–548, 2019. https://doi.org/10.1016/j.future.2018.01.055.

[14] J. Fenwick, L. Barclay, and V. Schmied, "Chatting: an important clinical tool in facilitating mothering in neonatal nurseries," J. Adv. Nurs., 33(5), 583–593, 2001. https://doi.org/10.1046/j.1365-2648.2001.01694.x.

[15] N. Bos, J. Olson, D. Gergle, G. Olson, and Z. Wright, "Effects of Four Computer-mediated Communications Channels on Trust Development," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2002), 135–140. https://doi.org/10.1145/503376.503401.

[16] L. Qiu and I. Benbasat, "Online Consumer Trust and Live Help Interfaces: The Effects of Text-to-Speech Voice and Three-Dimensional Avatars," Int. J. Human–Computer Interact., **19**(1), 75–94, 2005. https://doi.org/10.1207/s15327590ijhc1901_6.

[17] S. Greenspan, D. Goldberg, D. Weimer, and A. Basso, "Interpersonal Trust and Common Ground in Electronically Mediated Communication," in Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, 2000), 251–260. https://doi.org/10.1145/358916.358996.

[18] A. Schirmer, Y. Feng, A. Sen, and T. B. Penney, "Angry, old, male--and trustworthy? How expressive and person voice characteristics shape listener trust," PLoS One, **14**(1), e0210555, 2019. https://doi.org/10.1371/journal.pone.0210555.

[19] S. M. Hughes and M. A. Harrison, "Your Cheatin' Voice Will Tell on You: Detection of Past Infidelity from Voice," EPsychol., **15**(2), p. 1474704917711513, 2017. https://doi.org/10.1177/1474704917711513.

[20] P. Belin, B. Boehme, and P. McAleer, "The sound of trustworthiness: Acoustic-based modulation of perceived voice personality," PLoS One, **12**(10), 1–9, 2017. https://doi.org/10.1371/journal.pone.0185651.

[21] P. M. Doney, J. P. Cannon, and M. R. Mullen, "Understanding the influence of national culture on the development of trust," Acad. Manag. Rev., **23**(3), 601–620, 1998. https://doi.org/10.2307/259297.

[22] D. Gefen and T. H. Heart, "On the need to include national culture as a central issue in e-commerce trust beliefs," J. Glob. Inf. Manag., **14**(4), 1–30, 2006. https://doi.org/10.4018/jgim.2006100101.

[23] S. Zaheer and A. Zaheer, "Trust across borders," J. Int. Bus. Stud., **37**(1), 21–29, 2006. https://doi.org/10.1057/palgrave.jibs.8400180.

[24] O. Niebuhr, J. Voße, and A. Brem, "What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice," Comput. Human Behav., 64, 366–382, 2016. https://doi.org/10.1016/j.chb.2016.06.059.

[25] L. Anolli and R. Ciceri, "The voice of deception: Vocal strategies of naive and able liars," J. Nonverbal Behav., **21**(4), 259–284, 1997.

[26] M. Zuckerman, R. S. DeFrank, J. A. Hall, D. T. Larrance, and R. Rosenthal, "Facial and vocal cues of deception and honesty," J. Exp. Soc. Psychol., **15**(4), 378–396, 1979. https://doi.org/10.1016/0022-1031(79)90045-3.

[27] C. Kirchhübel and D. M. Howard, "Detecting suspicious behaviour using speech: Acoustic correlates of deceptive speech--An exploratory investigation," Appl. Ergon., **44**(5), 694–702, 2013. https://doi.org/10.1016/j.apergo.2012.04.016.

[28] I. Torre, L. White, and J. Goslin, "Behavioural mediation of prosodic cues to implicit judgements of trustworthiness," 2016. https://doi.org/10.21437/SpeechProsody.2016-167.

[29] N. Korovaiko and A. Thomo, "Trust prediction from user-item ratings," Soc. Netw. Anal. Min., **3**(3), 749–759, 2013. https://doi.org/ 10.1007/s13278-013-0122-z

[30] T. DuBois, J. Golbeck, and A. Srinivasan, "Predicting trust and distrust in social networks," in 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, 2011, 418–424. https://doi.org/10.1109/PASSAT/SocialCom.2011.56.

[31] B. Zong, F. Xu, J. Jiao, and J. Lv, "A broker-assisting trust and reputation system based on artificial neural network," in 2009 IEEE International Conference on Systems, Man and Cybernetics, 2009, 4710–4715. https://doi.org/10.1109/ICSMC.2009.5346098.

[32] D. Bejou, B. Wray, and T. N. Ingram, "Determinants of relationship quality: an artificial neural network analysis," J. Bus. Res., **36**(2), 137–143, 1996. https://doi.org/10.1016/0148-2963(95)00100-X.

[33] J. J. Lee, B. Knox, and C. Breazeal, "Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions," in 2013 AAAI Spring Symposium Series, 2013. http://hdl.handle.net/1721.1/69244.

[34] L. Ekonomou, "Greek long-term energy consumption prediction using artificial neural networks," Energy, **35**(2), 512–517, 2010. https://doi.org/10.1016/j.energy.2009.10.018.

[35] M. Imhof, "Listening to Voices and Judging People," Int. J. List., 24(1), 19–33, 2010. https://doi.org/10.1080/10904010903466295.

[36] B. L. Smith, B. L. Brown, W. J. Strong, and A. C. Rencher, "Effects of Speech Rate on Personality Perception," Lang. Speech, **18**(2), 145–152, 1975. https://doi.org/10.1177/002383097501800203.

[37] M. N. Ismail, S. S. Chee, H. Nawawi, K. Yusoff, T. O. Lim, and W. P. T. James, "Obesity in Malaysia," Obes. Rev., **3**(3), 203–208, 2002. https://doi.org/10.1046/j.1467-789X.2002.00074.x.

[38] V. Swami and A. Furnham, "Self-assessed intelligence: Inter-ethnic, rural--urban, and sex differences in Malaysia," Learn. Individ. Differ., 20(1), 51–55, 2010. https://doi.org/10.1016/j.lindif.2009.11.002**.**

[39] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-Emotion Behavior Dataset," in Proceedings of the International Joint Conference on Neural Networks, 2018, 2018-July. https://doi.org/10.1109/IJCNN.2018.8489099.

[40] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," J. Pers. Soc. Psychol., **17**(2), 124–129, 1971. https://psycnet.apa.org/doi/10.1037/h0030377.

[41] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and W. J. Barry, "Modelling personality features by changing prosody in synthetic speech," 2006. https://doi.org/10.22028/D291-25920.

[42] H. Quené, "On the just noticeable difference for tempo in speech," J. Phon., 35(3), 353–362, 2007. https://doi.org/10.1016/j.wocn.2006.09.001.

[43] G. Yulin, "The Spectrum of Trust and Distrust," Jiangsu Soc. Sci., 1, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 472-480 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Reliability Improvement of Radial Distribution System by Reconfiguration

Srividhya. P[1], Mounika. K[1], Kirithikaa. S[1], Narayanan. K[1,*] Gulshan Sharma[2], Girish Ganesan. R[3], Tomonobu Senjyu[3]

[1]*EEE Department, SASTRA Deemed University, Thanjavur, 613401, India*

[2]*Department of Electrical Power Engineering, Durban University of Technology, Durban, 4001, South Africa*

[3]*Power and Energy System Control Laboratory, University of the Ryukyus, Okinawa, 903-0213, Japan*

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 20 October, 2020
Online: 20 November, 2020

Keywords:
SAIDI
SAIFI
CAIDI
CAIFI
Reconfiguration

ABSTRACT

In a radial distribution system, component failure leads to service interruption and subsequent disconnection of a few load points. An abnormal variation in the voltage levels and high power loss gives rise to reliability issues. In this work, feeder reconfiguration is performed to improve reliability, minimize voltage deviation and power loss. A Binary Particle Swarm Optimization based search function handles the reconfiguration. The goal of this search function is to identify the optimal tie switches which minimize the multi-objective function. The search algorithm based on Binary Particle Swarm Optimization is utilized to handle the reconfiguration which identifies the optimal tie switches for minimizing the multi-objective function. The effect of the various algorith parameters on the convergence rate is studied. The proposed algorithm has been tested on Standard IEEE 33 and 69 test systems and the obtained results show a boost in the system reliability.

1 Introduction

Electric distribution system (EDS) are exponentially evolving due to the increasing demand and the changing sensitivity of customers loads. An interruption in electric power supply to the customers is caused by many events such as equipment failure, faults at substations, trees and much more. These interruptions cause a severe impact in the load equipment apart from the obvious financial losses and inconvenience. It is proclaimed that the failures in the distribution system causes nearly 80% of the interruptions. The customers of the electric utility desires an interruption free reliable power supply at the end of the day, emphasizing the need for calculating the reliability of EDS.

Reconfiguration is opening and closing of sectionalizing and tie switches. The former is normally closed whereas the latter is normally open switches. The alteration of the status of the switches under normal conditions or during contingencies results in change in topology of the feeders. It reduces the cost of implementation of various protective devices while satisfying the constraints such as maintaining radial condition of the system and feeding all the buses. It has become a practice nowadays to implement distributed generators along with reconfiguration to obtain better results [1]-[3]. But implementing Distributed Generation (DG) increases the overall cost of the EDS. Implementing reconfiguration will ensure that the customer does not face outages by rerouting the required power thereby enhancing reliability of the system. Voltage fluctuation of can be endured only for a brief stretch. If not, the impairment instigated in the equipment connected to the EDS would be high [4]. So reconfiguration is used to maintain the same within limits. Reconfiguration contributes to proper distribution of load and hence lower magnitude of current in the branches thereby reducing the power loss. This increases the loadability of the feeder thus contributing to the enhancement of reliability of the system.

There are several documented ways to solve for the optimal system reconfiguration, namely, heuristic, meta-heuristic and deterministic methods. Deterministic models hold no uncertainity and the output only depends on the intial state of the system and the value of the system parameters. Heuristic techniques are often greedy algorithms pursuing a local optimum instead of a global optimum even though they are problem dependent. Meta-heuristic

*Corresponding Author: Narayanan. K, EEE Department, SASTRA Deemed University, Thanjavur, Tamil Nadu, India. Mob: +91-9790258910 Email: narayanan.mnit@gmail.com

approaches are independent of the problem and not greedy, allowing them to perform and in-depth search of the workspace to get an optimum solution. In most cases, the parameters of interest are power loss, voltage profile and reliability. From the extensive literature cited in this work it is seen that employing meta-heuristic methods are techniques has been a recent trend in the research field.

In [5]-[14] an objective function has been developed consisting of both reliability indices and loss in models with and without a DG while [15]-[18] involve loss reduction and voltage profile improvement. Since 1980, Genetic algorithm(GA) which is classified as an evolutionary algorithms is widely used as an optimization tool [5], [8], [19] and [20]. But it is a slow algorithm [19] as it requires decent sized population and a number of generations to see good results whereas Particle Swarm Optimzation (PSO) entails small quantity of ciphering time not counting the better accuracy in results in contrast to GA. In [8] an Enhanced GA is presented that can resolve the reconfiguration of the feeder proficiently and sturdily with fewer iteration and time. A novel Micro genetic algorithm with reduced number of efficient solution which works well for multi-objective functions is presented in [6].

The technique shown in [7] adopts an Analytic Hierarchy Process for decision making with multiple criteria analysis. In the work presented in [9], Dynamic programming is used to find minimal cutsets and a search based technique for reconfiguration to attain the ideal set of switches has been articulated for RDS using DG. In [11] reconfiguration is solved using a novel Imperialist Competitive Algorithm which intensifies the exploration ability.In [14] the randomness in the reliability data is taken into account and has introduced a robust approach. In [3] a CRO algorithm to deal with complex network reconfiguration with DG present is taken. Different versions of Particle Swarm Optimization (PSO) is used in [10], [15], [17], [19], [21]–[24] where [15] and [23] takes another step by hybridizing PSO with Ant Colony Optimization and Evolutionary Programming respectively, for reconfiguration for various objectives. This algorithm provides better results faster in comparison to the other algorithms.

The contribution of this paper is the enhancement of reliability indices by performing an optimal reconfiguration of the distribution system using a modified version of PSO. The objective function involves minimization of Reliability index SAIDI, Voltage Deviation and Power loss. Four other reliability indices are calculated as well. The influence of certain parameters in PSO are determined using a MATLAB simulation environment and the results prove the validity of the proposed concept.

2 Problem Formulation

Reliability is the ability of a distribution system to provide uninterrupted power supply to customer [25]. Reliability indices are the benchmark parameters for evaluating reliability of the distribution system and performance of the utility. There are five standard reliability performance indices: System Average Interruption Frequency Index (SAIFI), System Average Interruption Duration Index (SAIDI), Customer Average Interruption Duration Index (CAIDI), Average System Availability Index (ASAI) and Average System Unavailability Index (ASUI) which are briefed below.

2.1 Reliability Indices

SAIFI is the average frequency of interrupt occurences that a customer experiences,

$$SAIFI = \frac{\sum_i \lambda_i \times N_i}{N_i} f/yr \tag{1}$$

where,

λ_i - failure rate(failure/year) of ith load point

N_i - Number of customers at point i.

SAIDI is the average duration of interrupt a customer experiences,

$$SAIDI = \frac{\sum_i U_i \times N_i}{N_i} h/yr \tag{2}$$

where,

U_i - product of failure rate and repair rate of ith load point.

Failure rate (λ_i) and Repair rate (r_i) are called load point indices and are calculated using Cutset Approach as the systen is a series-parallel system.

CAIDI is the average interruption time a consumer is affected by interruption per year. It is related to SAIFI and SAIDI indices.

$$CAIDI = \frac{SAIDI}{SAIFI} \tag{3}$$

ASAI represents the percentage of time during a year that the average customer have power.

$$ASAI = \frac{N \times 8760 - \sum_i U_i \times N_i}{N \times 8760} \tag{4}$$

It is the average system unavailability index is the contrary of ASAI and is given by,

$$ASUI = 1 - ASAI \tag{5}$$

2.2 Voltage Deviation and Power Loss

A stable voltage profile alongside low power loss is important as it directly benefits the electric utility operation and subsequently increases customer satisfaction. For this reason, it is embedded into the objective function. Voltage deviation and power loss can be calculated as follows,

$$V\_DEV = \sum_{i=2}^{N} (1 - |V_i|)^2 \tag{6}$$

$$P\_LOSS_i = \frac{R_i(P_i^2 + Q_i^2)}{|V_i|^2} \tag{7}$$

$$P\_LOSS = \sum_{i=1}^{N} k_i P\_LOSS_i \tag{8}$$

where,

P_LOSS_i is the power loss of branch i, P_LOSS is the total power loss, P_i and Q_i are the real power and reactive power of branch i, R_i and is the branch resistance, V_i is the voltage at bus i, N is the total number of branches and k_i is the status of branch i (1=close, 0=open)

2.3 Constraints

2.3.1 Failure and Repair rate constraints

$$\lambda_{i,min} \leq \lambda_i \leq \lambda_{i,max} \tag{9}$$

$$r_{i,min} \leq r_i \leq r_{i,max} \tag{10}$$

2.3.2 Real and Reactive power constraints

$$P_{G,min} \leq P_G \leq P_{G,max} \tag{11}$$

$$Q_{G,min} \leq Q_G \leq Q_{G,max} \tag{12}$$

2.3.3 Voltage constraint

$$V_{i,min} \leq V_i \leq V_{i,max} \tag{13}$$

2.3.4 Radial topology constraint

$$\varphi(L) = 0 \tag{14}$$

$\lambda_{i,min}$, $r_{i,min}$, $\lambda_{i,max}$ and $r_{i,max}$ are the limits of the values of failure and repair rate. $P_{G,min}$ and $P_{G,max}$ are the vectors of lowest and highest available real power generation whereas $Q_{G,min}$ and $Q_{G,max}$ are the vectors of lowest and highest available reactive power generation. $V_{i,min}$ and $V_{i,max}$ are the limits of the branch voltages and $\varphi(L)$ is the number of loops in the system.

2.4 Objective Function

$$Minf = w_1 \times \frac{SAIDI}{SAIDI_T} + w_2 \times \frac{V\_DEV}{V\_DEV_T} + w_3 \times \frac{P\_LOSS}{P\_LOSS_T} \tag{15}$$

Since reliability improvement is the major objective of the work, w_1 is given higher value such that,

$$w_1 + w_2 + w_3 = 1 \tag{16}$$

As the objectives have different dimension they're conveter to per-unit values. $SAIDI_T$,V_DEV_T and P_LOSS_T are the initial case values of SAIDI, Voltage deviation and Power Loss respectively. Thus multiple objectives are unified into a single objective resulting in reduced optimization load. The system condition before reconfiguration is taken as the base.

3 Proposed Method

This section presents the basics of PSO and the formulation an intelligent search method depending on PSO. All the feasible configurations of the distribution system that satisfy the objective function is presented.

3.1 Particle Swarm Optimization

PSO is a meta heuristic population based technique with its genesis from the observation of social behaviour associated with birds flocking. It can be applied to both continuous and discrete systems. In the presented problem, the switches are chosen from a discrete set for different bus systems and therefore discrete PSO is applied in this case.

The algorithm develops a population of particles each of which are a viable solution for the problem. Initially each particle is assigned a random position, velocity and objective function, X_i, V_i and F_i respectively. The evaluation is carried out using positional coordinates and population size as the input. Each particle has a follow up of its position and the best value acquired so far and is stored in P_{best}. . In the meantime the overall best value produced so far by any particle is stored as G_{best}.

$$w = w_{max} - \frac{w_{max} - w_{min}}{iter_{max}} \times iter \tag{17}$$

The interia of the particles has an impact on the future velocities of the particles. Experimentally it is observed that decreasing the value of inertia over the iterations causes exploration and exploitation to occur linearly. Number of particles in the swarm is represented by n. Smaller value takes large number of iterations to converge whereas high value cause a rise in computational complexity and time consumed. So a suitable value is chosen.

$$v_{ij}^{t+1} = w \times v_{ij}^t + c_1 r_{1j}^t [P_{best} - x_{ij}^t] + c_2 r_{2j}^t [G_{best} - x_{ij}^t] \tag{18}$$

where,

v_{ij}^t - particle velocity

x_{ij}^t - particle position

c_1 - cognitive parameter

c_2 - social parameter c_1 and c_2 represents the weights of the stochastic acceleration term that pulls the particles towards P_{best} and G_{best} respectively.

r_1 and r_2 are random values in the range [0,1].

The particles coordinates are updated using the sigmoidal equation,

$$s_{ij}^t = \frac{1}{1 + e^{-v_{ij}^{t+1}}} \tag{19}$$

The *ciel* of the function converts the continuous to discrete values. It is further modified according to the problem at hand to choose the values from a particular set of values. In our work, the set of values corresponds to the switches that are to be opened for reconfiguration from the array of switches from each loop of the bus system if the tie lines are closed.

3.2 Algorithm

1. Add Load bus data, line data maximum iterations and switch value matrix. Initialize population, position, random velocities, P_{best}, G_{best}.
2. Calculate inertia, update velocity and particles coordinates.
3. Check radiality constraint and calculate fitness. Power loss and voltage deviation are calculated from load flow solution. Reliability is calculated using the concept of cutsets.

Backward Forward Sweep Algorithm is used as the load flow solution.

4. Update P_{best} if the fitness function is lesser than the previous value.
5. Update G_{best} if the fitness function is lesser than the previous value.
6. If there is no criteria match, update particle velocity using the previous velocity, distance to P_{best} and distance to G_{best} in the equation
7. Update particle position from provided switches.
8. Terminate if the maximum iterations is reached. Go to step 3. The flowchat of the proposed algorithm is shown in Figure 1.

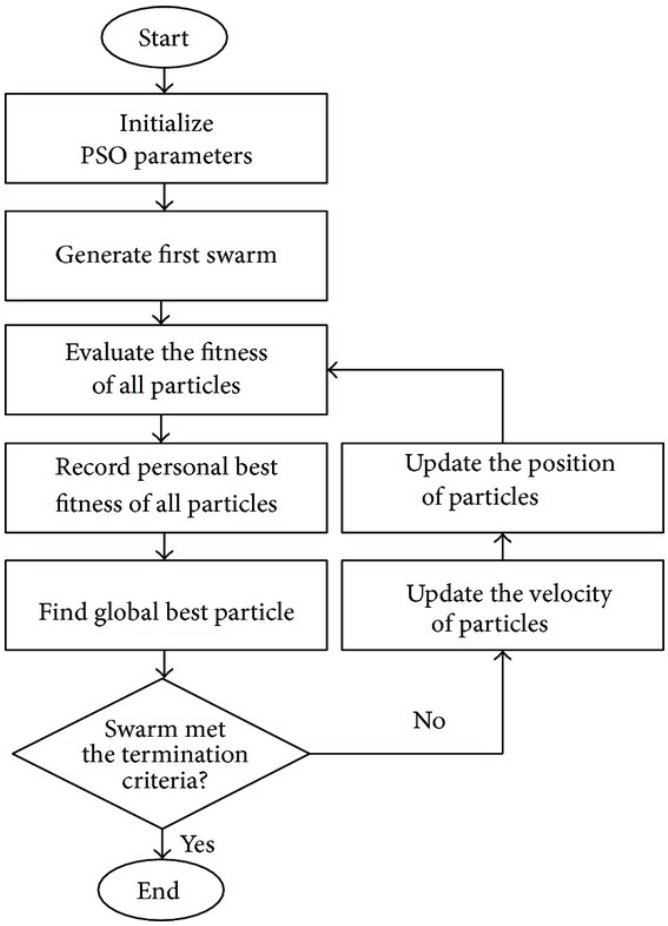


Figure 1: Flowchart of Proposed PSO

4 Results and Discussion

The IEEE 33 and 69 bus system data is presented in [10]. The failure rate and repair rate for the same is presented in [20]. The proposed method is implemented on the test cases and and the effectiveness is tested for the given objective function. Each load point is assumed to have one customer. Four different cases is made out of the objective function presented in Eq.15 with different weightage to each objective but with heighest weightage to SAIDI.

1. Case 1: $w_1 = 1.0$, $w_2 = 0$ and $w_3 = 0$
2. Case 2: $w_1 = 0.75$, $w_2 = 0$ and $w_3 = 0.25$
3. Case 3: $w_1 = 0.75$, $w_2 = 0.25$ and $w_3 = 0$
4. Case 4: $w_1 = 0.5$, $w_2 = 0.25$ and $w_3 = 0.25$

4.1 33 Bus System

It has a voltage rating of 12.7kV with a constant power load of (3.715 +j2.3) MVA. It is generally used to evaluvate the distribution reconfiguration problem. Figure 2 shows the single line diagram for a 33 bus system. In the base configuration, the solid lines represent the NC switches (1 through 32) and the dotted lines represed the NO switches (33 through 37).

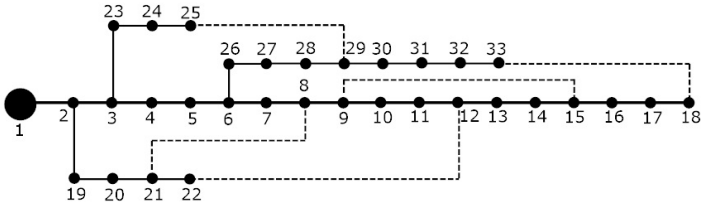


Figure 2: 33 bus system before reconfiguration

4.1.1 Case 1

The main objective of this case is to minimise the reliability index, SAIDI, which is used to find how reliable the system is. The algorithm is trained using this test case in order to find optimal solutions in the upcoming cases for the multi-objective function. For comparison purposes, the values of the indices of the base case is calculated.

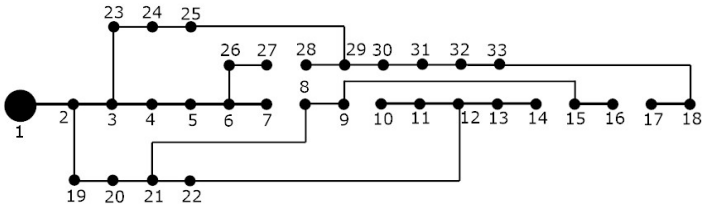


Figure 3: 33 bus system after reconfiguration (Case 1)

Figure 3 shows the bus system post reconfiguration. In this case a drop of 0.8769 p.u for duration index which is a 40% reduction from the original value making the system highly reliable. The Table 1 shows the best possible values, w_{min}, w_{max} and population size. A value of population swarm between 40 and 50 reduces the convergence time. The fitness function converges in the 6_{th} iteration when population is randomized and in the 8_{th} iteration when inertia is randomized.

Table 1: PSO parameters - 33 bus system

Parameter	Value
w_{min}	0.2
w_{max}	0.9
Population	48

The Voltage deviation and power loss have shown a reduction of 24% and 55% for the reconfigured system even without giving weightage to the same as shown in Table 2. Power loss and voltage deviation has shown drastic reduction in its values in spite of not giving weightage to them in this case thus concluding on the fact that all the three objectives are interrelated.

4.1.2 Case 2

The values of the parameters, inertia weight (minimum and maximum) and population of the swarm, are fixed for cases 2, 3 and 4. From Table 2 it is clear that with a 0.5% sacrifice of SAIDI, the voltage deviation and power loss values have been lowered by additional 4% and 2% respectively in comparison with the results of case 1. As shown in Figure 4 the system is radial post reconfiguration.

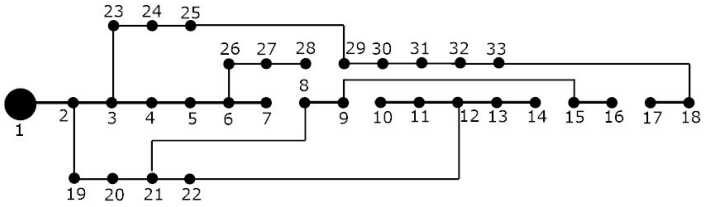


Figure 4: 33 bus system after reconfiguration (Case 2)

4.1.3 Case 3

It was found that the results of Case 3 is identical and hence the single line diagram is not repeated. The loss is reduced by an additional 4% when compared to case 1 while giving up only a 0.5% in SAIDI value with respect to case 1.

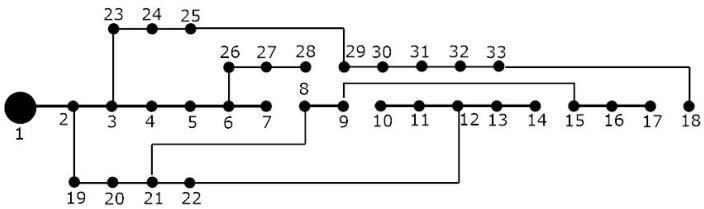


Figure 5: 33 bus system after reconfiguration (Case 4)

4.1.4 Case 4

It can be seen that a further reduction of 4kW power loss in comparison to the previous two cases. Voltage deviation has also reduced by an extra 2% when compared to the past two cases along with the betterment of minimum voltage in the system. Hence it is proved that the system is highly reliable with power loss and voltage deviation improved drastically. Figure 5 displays the bus system configuration after reconfiguration. Due to space issues, Voltage profile and Convergence graph of the final case is alone shown in Figure 6 and 7 to picturize the improvement in the voltage profile and the efficacy of the proposed method.

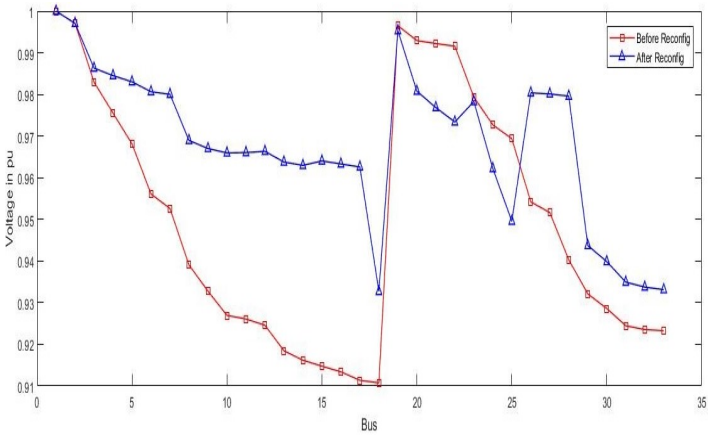


Figure 6: Voltage Profile pf 33 Bus System (Case 4)

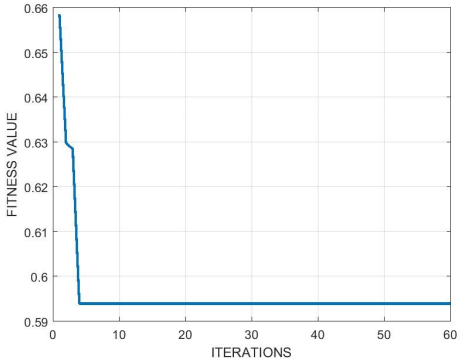


Figure 7: Convergence graph of 33 Bus System (Case 4)

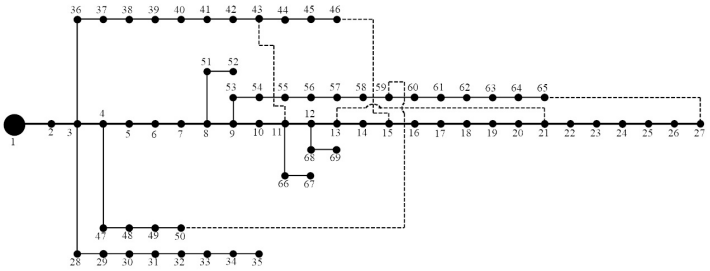


Figure 8: 69 bus system before reconfiguration

4.2 69 Bus System

The proposed topology is tested on a IEEE Standard 69 Bus system. This system has a rated voltage of 12.7kV.NO switches are 69 through 73 as in Figure 8. The reliability indices before reconfiguration are recorded by applying the algorithm to the base case of the 69 bus system. Other than the zero injection buses, each load

Table 2: Consolidated Results of 33 Bus System

INDICES	BASE CASE	CASE 1	CASE 2	CASE 3	CASE 4
TIE SWITCHES	33,34,35,36,37	7,9,14,16,27	7,9,14,16,28	7,9,14,16,28	7,9,14,17,28
SAIFI	2.5741	2.2675	2.258438	2.258438	2.308125
SAIDI	2.2291	1.3522	1.362656	1.362656	1.405469
CAIDI	0.86597	0.59633	0.603362	0.603362	0.608922
ASAI	0.99975	0.99985	0.999844	0.999844	0.99984
ASUI	0.00025446	0.00015436	0.000156	0.000156	0.00016
V_DEV (%)	2.8945	1.3137	1.20724	1.20724	1.137742
P_LOSS (kW)	202.7142	154.2636	150.2211	150.2211	146.3066
V_MIN (p.u)	0.9131	0.9276	0.9291	0.9291	0.9327

point is assumed to have one customer. The Inertia and Population are randomized separately and run hundreds of iterations to arrive at the best values. The values obtained as shown in Table 3 further implicate that w_{min}, w_{max} should be chosen with a difference of 0.5 for accurate results and reaffirms that population value within [40,50] holds the best results.

4.2.1 Case 1

Apart from SAIFI and SAIDI other indices such as CAIDI, ASAI and ASUI are found for base case in order to compare the results of all the cases. The switches that are to be opened post reconfiguration are {14, 18, 56, 64, 69}. When the switches are compared with the failure and repair rate, it is found that the branches with maximum failure rate are opened so that the values of the indices drop according to the Eq. 1 and Eq. 2. Table 4 shows that there is a major cutback of 0.7431 p.u which is 28.8% for SAIFI and 0.3804 which is 22.28% for SAIDI. Values of all the indices are shown as well. The power loss and voltage deviation is determined for the reconfigured system. Loss before and after reconfiguration presents a substantial drop of 47.11%. Voltage deviation has reduced by 71.2%. Figure 9 represents the 69 bus system after reconfiguration maintaining radial topology and its load evened out.

4.2.2 Case 2

In this case duration and voltage deviation are optimized. SAIDI plummeted by a 21% while voltage deviation which was given a weightage of 0.25 in the objective function dropped by 2.49 p.u which is nearly 74% with respect to base improving the voltage profile substantially and thereby improving overall reliability of the power system. Power loss for the above configuration was 119.6465kW which an improvement from the previous case highlighting the impact of better voltage profile on power loss. Figure 10 represents the 69 bus system after reconfiguration maintaining radial topology.

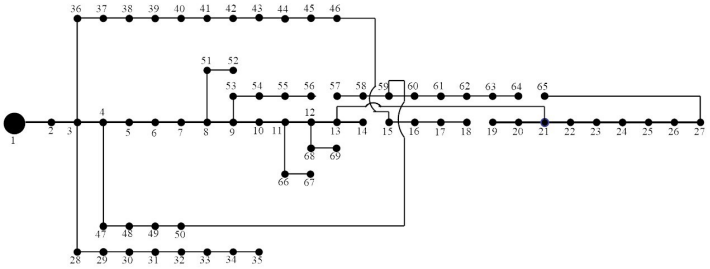


Figure 9: 69 bus system after reconfiguration (Case 1)

Table 3: PSO parameters - 69 bus system

PARAMETER	VALUE
w_{min}	0.2
w_{max}	0.7
Population	46

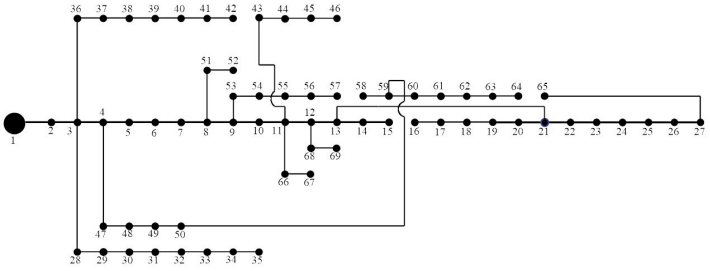


Figure 10: 69 bus system after reconfiguration (Case 2)

4.2.3 Case 3

This case involves optimizing SAIDI and Power loss with a weightage of 0.75 and 0.25 respectively. Table 4 shows the result after

reconfiguration and it is found that loss has reduced to 106.7573kW even though SAIDI raised by a value of 0.04. The voltage deviation calculated after reconfiguration was 0.8958 which is just a 0.02 p.u rise while the loss has reduced considerably. The minimum value of voltage has improved due to minimized power loss. This helps the sensitive equipment from falling a prey for low voltage conditions and hence improves the reliability of the system. Figure 11 represents the 69 bus system after reconfiguration maintaining radial topology.

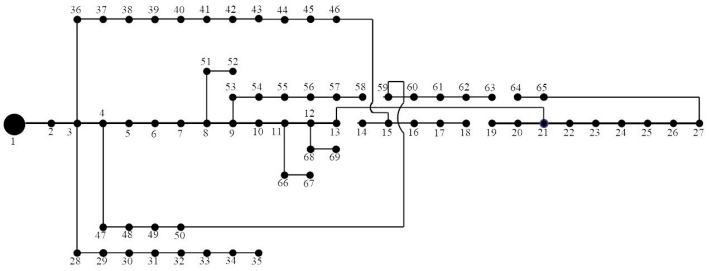


Figure 11: 69 bus system after reconfiguration (Case 3)

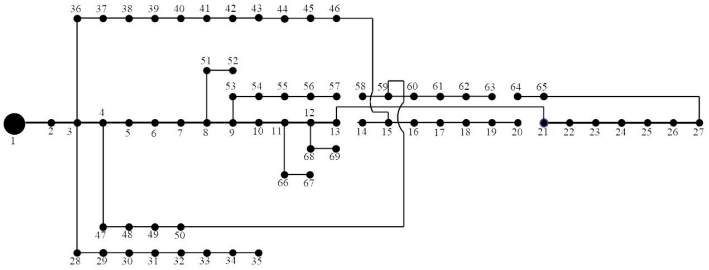


Figure 12: 69 bus system after reconfiguration (Case 4)

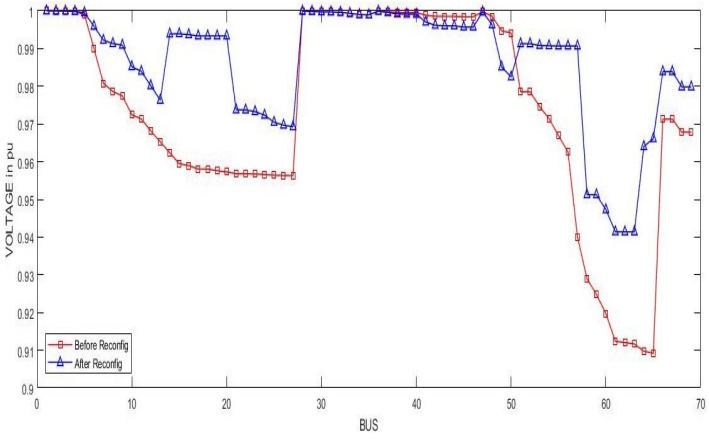


Figure 13: Voltage Profile of 69 Bus System (Case 4)

4.2.4 Case 4

This case includes all the three objectives with equal weightage to voltage deviation and power loss of 0.25 and reliability index SAIDI holding higher importance with weightage of 0.5. It is clearly seen from Table 4 that all three objective functions have their values reduced from case 3 giving us the best possible combination of switches under the above condition. The minimum voltage stays the same as previous case but overall voltage profile has improved. In comparison to the base case there is a drop of 20.4%, 73.45% and 52.57% in SAIDI, Voltage deviation and Power loss respectively as shown in Table 4. Figure 12 sketches the 69 Bus System with loads distributed and following the radial topology constraint. Figure 13 shows the pictorial representation of the improvement in voltage profile while Figure 14 supports the claim of high convergence rate in the proposed method.

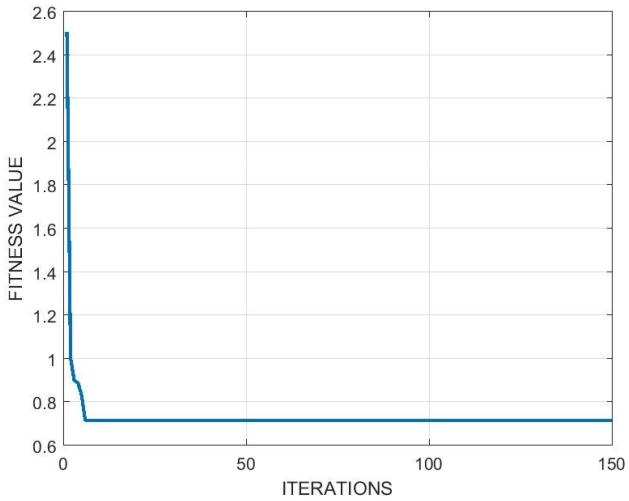


Figure 14: Convergence graph of 69 Bus System (Case 4)

5 Conclusion

This work presents a method for reliability improvement of Radial Distribution System through feeder reconfiguration. It is constructed based on Particle Swarm Optimization. The radial topology is conserved using the graph theory function in MATLAB. The load point indices such as failure rate and repair rate are calculated using Cut-set concept. The reliability indices are calculated using theoretical formulae. Power loss and Voltage deviation is computed by utilizing Backward-Forward Sweep algorithm.

The Inertia weight and Population values in the proposed algorithm clearly influences the accuracy of the results and the time to taken to obtain the same as it regulates the exploration and exploitation of the search space.

In single objective function, the switch selection for reconfiguration solely depends on the values of the load point indices (λ_i & U_i) and hence switches with highest values of the indices are opened. This results in cutting down the number of load points in the longest branch and distributing them to other branches.

In multi objective function, the switch combination chosen not only depends on the failure rate and repair rate of the load points but also on the voltage and power loss. This results in minor increase in the SAIDI and other reliability indices but it improves the voltage profile aiding the loads connected the system. It takes some demand of the utility system due to reduced power losses.

Table 4: Consolidated Results of 69 Bus System

INDICES	BASE CASE	CASE 1	CASE 2	CASE 3	CASE 4
TIE SWITCHES	69,70,71,72,73	14,18,56,64,69	15,42,57,64,71	13,18,58,63,69	13,20,57,63,69
SAIFI	2.5778	1.834623	1.85446	1.912598	1.8446
SAIDI	1.7071	1.326789	1.34073	1.381665	1.358867
CAIDI	0.6622	0.723194	0.722973	0.722402	0.736673
ASAI	0.999746	0.999849	0.999847	0.999842	0.999845
ASUI	0.000254	0.000151	0.000153	0.000158	0.000155
V_DEV (%)	3.3742	0.9717	0.8777	0.8958	0.895569
P_LOSS (kW)	225	119.6465	119.6465	106.7573	106.7085
V_MIN (p.u)	0.9092	0.9299	0.9299	0.9414	0.9414

References

[1] K. Jasthi, D. Das, "Simultaneous distribution system reconfiguration and DG sizing algorithm without load flow solution," IET Generation, Transmission Distribution, **12**(6), 1303–1313, 2018, doi:10.1049/iet-gtd.2017.0338.

[2] J. Wen, Y. Tan, L. Jiang, K. Lei, "Dynamic reconfiguration of distribution networks considering the real-time topology variation," IET Generation, Transmission Distribution, **12**(7), 1509–1517, 2018, doi:10.1049/iet-gtd.2017.1304.

[3] P. Sharma, N. Gupta, K. R. Niazi, A. Swarnkar, "Investigation of network reconfiguration on the reliability and performance of distribution systems using CRO," in 2017 6th International Conference on Computer Applications In Electrical Engineering-Recent Advances (CERA), 80–85, 2017, doi: 10.1109/CERA.2017.8343305.

[4] K. Muttaqi, A. Dt Le, G. Ledwich, M. Negnevitsky, "Minimising Voltage Deviation in Distribution Feeders by Optimising Size and Location of Distributed Generation," Australian Journal of Electrical and Electronics Engineering, **3**, 2007, doi:10.1080/1448837X.2007.11464155.

[5] M. M. Legha, M. Mohammadian, M. Darini, "Loss Reduction and Reliability Improvement with Optimization of DGs Placement Using Genetic Algorithm - A Case Study on the Electrical Distribution Network of North Kerman Area," International Journal of Scientific & Engineering Research, **5**(2), 2014.

[6] J. E. Mendoza, M. E. Lopez, C. A. C. Coello, E. A. Lopez, "Microgenetic multiobjective reconfiguration algorithm considering power losses and reliability indices for medium voltage distribution network," IET Generation, Transmission Distribution, **3**(9), 825–840, 2009, doi:10.1049/iet-gtd.2009.0009.

[7] A. P. C. Mello, D. P. Bernardon, L. L. Pfitscher, L. N. Canha, V. J. Garcia, M. Sperandio, "Reconfiguration of distribution network considering distributed generation and multivariables criteria," in 2013 48th International Universities' Power Engineering Conference (UPEC), 1–6, 2013, doi: 10.1109/UPEC.2013.6714870.

[8] D.-L. Duan, X.-D. Ling, X.-Y. Wu, B. Zhong, "Reconfiguration of distribution network for loss reduction and reliability improvement based on an enhanced genetic algorithm," International Journal of Electrical Power & Energy Systems, **64**, 88 – 95, 2015, doi:https://doi.org/10.1016/j.ijepes.2014.07.036.

[9] P. Pavani, S. N. Singh, "Reconfiguration of radial distribution networks with distributed generation for reliability improvement and loss minimization," in 2013 IEEE Power Energy Society General Meeting, 1–5, 2013, doi: 10.1109/PESMG.2013.6672876.

[10] S. Elsaiah, M. Benidris, Y. Tian, J. Mitra, "A Comprehensive Analysis of Reliability-oriented Optimal Distribution System Reconfiguration," in 2018 IEEE Industry Applications Society Annual Meeting (IAS), 1–9, 2018, doi: 10.1109/IAS.2018.8544624.

[11] M. Esmaili, M. Sedighizadeh, M. Mahmoodi, M., "Reconfiguration of distribution systems to improve reliability and reduce power losses using Imperialist Competitive Algorithm," Iranian Journal of Electrical and Electronic Engineering, **13**, 2017, doi:10.22068/IJEEE.13.3.6.

[12] A. Kavousi-Fard, M.-R. Akbari-Zadeh, "Reliability enhancement using optimal feeder reconfiguration," Neurocomputing, **106**, 1–11, 2013, doi:10.1016/j.neucom.2012.08.033.

[13] N. G. Paterakis, A. Mazza, S. F. Santos, O. Erdinç, G. Chicco, A. G. Bakirtzis, J. P. S. Catalão, "Multi-Objective Reconfiguration of Radial Distribution Systems Using Reliability Indices," IEEE Transactions on Power Systems, **31**(2), 1048–1062, 2016, doi:10.1109/TPWRS.2015.2425801.

[14] J. C. López, M. Lavorato, J. F. Franco, M. J. Rider, "Robust optimisation applied to the reconfiguration of distribution systems with reliability constraints," IET Generation, Transmission Distribution, **10**(4), 917–927, 2016, doi:10.1049/iet-gtd.2015.0558.

[15] M. Amin Heidari, "Optimal network reconfiguration in distribution system for loss reduction and voltage-profile improvement using hybrid algorithm of PSO and ACO," CIRED - Open Access Proceedings Journal, **2017**(1), 2458–2461, 2017, doi:10.1049/oap-cired.2017.1007.

[16] I. A. Quadri, S. Bhowmick, D. Joshi, "Multi-objective approach to maximise loadability of distribution networks by simultaneous reconfiguration and allocation of distributed energy resources," IET Generation, Transmission Distribution, **12**(21), 5700–5712, 2018, doi:10.1049/iet-gtd.2018.5618.

[17] S. R. Tuladhar, J. G. Singh, W. Ongsakul, "Multi-objective approach for distribution network reconfiguration with optimal DG power factor using NSPSO," IET Generation, Transmission Distribution, **10**(12), 2842–2851, 2016, doi: 10.1049/iet-gtd.2015.0587.

[18] Z. Gong, Q. Chen, K. Sun, "Novel methodology solving distribution network reconfiguration with DG placement," The Journal of Engineering, **2019**(16), 1668–1674, 2019, doi:10.1049/joe.2018.8521.

[19] B. Vasudevan, K. Sinha, A., "Reliability improvement of reconfigurable distribution system using GA and PSO," Electrical Engineering, **100**, 1–13, 2017, doi:10.1007/s00202-017-0580-9.

[20] N. Gupta, A. Swarnkar, K. R. Niazi, "Distribution network reconfiguration for power quality and reliability improvement using Genetic Algorithms," International Journal of Electrical Power & Energy Systems, **54**, 664 – 671, 2014, doi:https://doi.org/10.1016/j.ijepes.2013.08.016.

[21] S. Elsaiah, M. Benidris, J. Mitra, "Reliability improvement of power distribution system through feeder reconfiguration," in 2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 1–6, 2014, doi:10.1109/PMAPS.2014.6960676.

[22] K. B. Kela, L. D. Arya, "Reliability Optimization of Radial Distribution Systems Employing Differential Evolution and Bare Bones Particle Swarm Optimization," Journal of The Institution of Engineers (India): Series B, **95**, 231–239, 2014, doi:10.1007/s40031-014-0094-z.

[23] M. F. Sulaima, N. H. Shamsudin, H. I. Jaafar, W. M. Dahalan, H. Mokhlis, "A DNR and DG Sizing Simultaneously by Using EPSO," in 2014 5th International Conference on Intelligent Systems, Modelling and Simulation, 405–410, 2014, doi:10.1109/ISMS.2014.75.

[24] I. I. Atteya, H. Ashour, N. Fahmi, D. Strickland, "Radial distribution network reconfiguration for power losses reduction using a modified particle swarm optimisation," CIRED - Open Access Proceedings Journal, **2017**(1), 2505–2508, 2017, doi:10.1049/oap-cired.2017.1286.

[25] P. C. Sekhar, R. A. Deshpande, V. Sankar, "Evaluation and improvement of reliability indices of electrical power distribution system," in 2016 National Power Systems Conference (NPSC), 1–6, 2016, doi:10.1109/NPSC.2016.7858838.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 481-488 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

American Sign Language Recognition Based on MobileNetV2

Kin Yun Lum[*,1], Yeh Huann Goh[1], Yi Bin Lee[1]

[1]*Department of Mechanical Engineering, Faculty of Engineering and Technology, Tunku Abdul Rahman University College, Kuala Lumpur, 43200, Malaysia*

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 20 October, 2020
Online: 20 November, 2020

Keywords:
Sign Language
Convolutional Neural Network
Depthwise Separable Convolution
MobileNet
Transfer Learning

ABSTRACT

Sign language is a form of communication language designed to link a deaf-mute person to the world. To express an idea it requires the use of hand gestures and body movement. However, the bulk of the general population remain uneducated to understand the sign language. Therefore, a translator is required to facilitate the communication. This paper wishes to extend the previously proposed Convolutional Neural Network (CNN) model for predicting American Sign Language with a MobileNetV2-based transfer learning model. The latter model effectively generalized on a dataset which is around 18 times larger with 5 additional groups of hand signs. Over 98% of the recognition accuracy had been reported. Because of its relatively fewer parameters and less intensive computational operations compared to other deep learning architectures, the model was also ideal to be implemented on mobile devices. The model will serve as the key to deploying a sign language translator software on smartphone to enhance communication efficiency between the deaf-mute person and the general public.

1 Introduction

This paper is an extension of work originally presented in conference IEEE International Conference on Signal and Image Processing Applications (ICSIPA) 2019 [1]. The work was enhanced in the following aspects:

1. Dataset: Previous study used only 4800 images with 24 groups of self-collected images except for letters J and Z, since both signs require movement. Furthermore, all images were taken from a fixed distance, which is not generally the case when it comes to actual application. The sign performer can appear at different distances from the smartphone camera. In this paper Kaggle's dataset was used. Each sign has variations in capturing distance and brightness. There are a total of 87,000 images with 29 classes. In addition to the 26 classes belonging to letter A-Z, there are an additional 3 classes for SPACE, DELETE and NOTHING. As per the description of the dataset, these 3 classes are found to be useful in real time application and classification.
2. Method: Previous proposed shallow network failed to generalize well with this new ~18x larger dataset. Instead, transfer learning using MobileNetV2 was implemented. Unlike other deep learning models, the MobileNetV2 gives high prediction accuracy without penalizing too much on computational cost and memory. This met our original intention to have this model deploy as a mobile application in the future.

Spoken language binds a significant section of the population. The spoken language would not, however, support the deaf and mute population. Statistics from the World Health Organization (WHO) shows there are at least 5% or 466 million of people suffer from hearing impairment [1]. An individual with hearing impairment may have a distorted speech or may not speak at all. This creates a barrier to contact between them and the society.

Fortunately, sign language comes to help for this special group of people. It is a kind of visual communication language that uses a combination of hand gesture, facial expression and body posture. It helps the hearing-impaired communities to express their feelings and addresses the general communication issues within their communities.

The sign language is completely different from spoken language. It has its own syntax and the way it expresses itself. For the general population this may be difficult to understand and practice [2].

[*]Corresponding Author: Kin Yun Lum, Department of Mechanical Engineering, Faculty of Engineering and Technology, Tunku Abdul Rahman University College, Kuala Lumpur, Malaysia, lumky@tarc.edu.my

Moreover, most general public are illiterate to sign language. Unless they are surrounded by a deaf-mute person, most appear to be uninterested and disregard the meaning of sign language. This presents a real barrier of contact between the deaf-mute community and the rest of society, as a issue which has yet to be completely resolved.

Innovations of technology in the area of sign language recognition try to break down this communication barrier. To date, it addresses the issues mainly with two different approaches: 1) Sensor based system 2) Vision based system [3].

The sensor based system is a wearable device which utilizes various sensors to capture motion, spatial positions and velocity of the hand. For example, cyber glove is a wearable device that employs flex sensor to detect fingers movements and Inertial Measurement Unit (IMU) sensor to track the hand motions [4]. There are also invasive and non-invasive electromyography (EMG) wearable sensors system to measure human muscle's electrical pulses and collect the bio-signal to detect fingers movements [5, 6]. These wearable system can be costly, cumbersome, inconvenient, heavy and uncomfortable for daily use.

Vision based system has recently become increasingly popular because there is no need to attach sensors physically to the human. There are attempts to use Leap Motion Controller [2] and Microsoft Kinect which have depth sensor [4, 7, 8] to construct the sign language recognition system. However, such technology is not widely accessible and it can be expensive to the users. Furthermore, the systems were implemented using desktop or laptop computer, which is impractical because of its weight and size.

Since hand sign gestures are perceived through vision, computer vision can be a subject of profound interest. One of the solutions is to build a vision based sign language recognition system with the commonly available smartphone camera. Vision-based systems are historically prone to reliability issues. Background noise, colours, and lighting are greatly varied under actual environment. This would result in a lower rate of detection. The latest advance in computer vision and machine learning technologies has shown considerable progress in the classification of images [9, 10]. In the competition for image classification, the algorithm called Convolutional Neural Networks (CNNs) and its derivative networks such as AlexNet, GoogleNet and ResNet achieved more than 90% accuracy [11]. Conventional machine learning relies on manual features extraction. Instead of creating complex handcrafted features, CNNs will automate the feature construction process. That would significantly reduce human error and increase the accuracy of detection.

There are several studies using CNN for the identification of American Sign Language (ASL) but the findings are not satisfactory [12]. Furthermore, while some deep networks may offer high classification accuracy, a large number of parameters and heavy mathematical operations limit their application in smartphones.

Our aim is to develop a computer vision system that can use mobile camera to recognize sign language, so that it can be used everywhere and anywhere. Previous attempt to create a CNN network to recognize 24 classes of sign language has demonstrated a detection accuracy of 95% experimentally. This paper extends the work to draw on a much larger dataset in order to enable greater generalization. The larger dataset contains 5 more classes than the previously used dataset. It included all 26 alphabets and another 3 classes for the sign of SPACE, DELETE and NOTHING to make the whole sign language model more complete. In addition, this dataset has also images which were captured at different distance from the camera, which may reflect a real-life application more closely. Transfer learning using MobileNetV2 was implemented to further improve the model accuracy to more than 98%.

The remainder of this paper is structured as follows: the relevant works will be discussed in Section 2. Followed by the section 3 concept and theory that will address the different CNN architectures and put emphasis on MobileNet, which is the transfer learning model to be implemented in this paper. Dataset description and preprocessing will be described in the methodology in Section 4. Then we will detail our model architecture and setup of experiments. The section ends by showing some evaluation metrics that we used to measure the performance of our model. Section 5 addresses experimental findings with discussions. Lastly, Section 6 ends with the conclusions of this paper with some suggestions for future study.

2 Related Works

The works related to vision based sign language recognition system can be categorized into two, which are non-CNN and CNN approaches. The non-CNN approaches were mainly the traditional methods which involved manual feature extraction and classification. The involved classification algorithms were k-nearest neighbors (KNN) and Neural Network [13, 14]. These approaches normally had lower reported prediction accuracy. On the other hands, the CNN approaches that had better reported results were either proposing their custom CNN models [7, 8, 15] or implementing transfer learning [11, 12, 16, 17]. There was also literature which compared both their custom CNN model and transfer learning [18].

For custom CNN models, Vivek Bheda et. al. used a deep convolutional network to classify ASL with alphabets and digits. The proposed network had 3 cascaded convolutional layers and 1 max-pooling layer. There are two hidden layers with dropout before connecting to the output layer. Accuracy of 82.5% had been reported [7].

Ameen. S. et. al. separated the image and depth information from Microsoft Kinect input and used two convolution layers to perform feature extraction on each of them. Both extracted feature maps were combined in the second stage of convolution before passing to the classifier. The reported precision and recall were 82% and 80% respectively [15].

Similar architecture was used by Pigou. L. et. al. to extract hand features and upper body features separately from Microsoft Kinect dataset. Each CNN was three layers in depth. They had reported an accuracy of 91.7% [8].

For transfer learning, Das, A. et. al. used Inception v3 on custom processed static gesture images and obtained an average validation accuracy of 90% [16]. Alashhab. S. et. al. tested on VGG16, VGG19, ResNet, Xception, Inception V3, MobileNet and SqueezeNet. They were dealing with hand gestures detection instead of sign language. There were 5 classes of hand gestures used. The highest accuracy of transfer learning usign MobileNet was 99.45% [11]. Kania, K.et. al. build a network to recognize ASL with reference to the Wide Residual Networks. Data augmentation was used to increase the number of training sets and reported a high-

est accuracy of 93.3% [12]. Garcia and Viesca used GoogLeNet architecture to recognize 24 classes of sign language with only 70% or reported accuracy [17]. As far as our literature, there was only one work to compare custom CNN model and transfer learning done by Bousbai, K. and Merah, M. Their training set had 1815 coloured images with black background collected from five volunteers. Their custom CNN had a reported efficiency of 98.9%, slightly surpassed the transfer learning model with 97.06% accuracy [18].

In general summary, transfer learning can perform better than most of the custom CNN model when come to pure image recognition. Custom model is only needed usually to cater for the extra input like depth information. Transfer learning model specially designed based on pure images dataset like ImageNet might not doing as good as the custom model. Other than this reason, transfer learning is a more appropriate solution to develop a well performing model. Therefore, it is being used in this paper to enhance the results of our previous work.

3 Concept and Theory

3.1 Convolutional Neural Network (CNN) Architecture

Convolutional Neural Network (CNN) has proven to be very effective to handle image classification problems. It is in use by many industry leaders such as Amazon, Google and Facebook [8]. The convolution layers in the CNN perform multiple discrete convolutions on the input with defined number of trainable weights filters. The filters' weights were updated during training. After applying filtering on each channel, it will capture the image features such as lines, edges, colors and other visual elements. The deeper the network goes, more complex features like shapes and patterns can be generated. Combining all these features together with the activation functions can produce feature maps which can then pass to the ordinary neural network classifier to correctly classify the object.

Due to significant progress has been achieved in image classification, there were various CNN architectures introduced such as VGG16, ResNet and Inception Net. These models were trained and verified on large-scale annotated datasets like ImageNet and proven to be effective. Instead of constructing the CNN from scratch, leverage on the transfer learning on these architectures can significantly reduce the time required to produce a working model for a particular problem. Moreover, since these models were trained on ImageNet which consists of 1.2 million categorized natural images of 1000+ classes. Using the trained weights as a starting point will definitely help the model to converge faster as compared to training from scratch. As discussed by Shin, H. C. et. al, even though there is a disparity between the ImageNet and our sign language images, the deep architecture which is trained comprehensively can as well generalize on other specific domain problem with only minor fine-tuning training [19]. Among different CNN architectures, MobileNet was employed in this paper because it is an efficient network architecture designed to run on any computationally limited platform.

3.2 MobileNetV1

MobileNet had been introduced as a small networks and it is optimized for latency to match the restricted resource application such as in mobile devices, robotics and self-driving vehicles. These application platforms are not as powerful as a general purpose computer, but still requires comparable recognition accuracy under a timely computation.

Unlike other network architectures use shrinking or compression to reduce its size and in return to improve on computational costs, MobileNet was completely rebuilt using depthwise separable convolutions. Depthwise separable convolution is a form of factorized convolutions which decomposes a conventional convolution into a depthwise convolution and a pointwise convolutoin. This factorization ends up reducing computation cost and model size drastically. The rationale behind this is to reduce the costly convolution process due to multiple multiplication as explained below [20].

Examine a standard convolution process, let an input image, F of dimension $D_W \times D_H \times M$ where D_W and D_H correspond to the spatial width and height of the input image and M is the number of input channels (depth). There are N numbers of square filters, K with size of $D_K \times D_K \times M$ where $D_K$is the spatial dimension of the kernel. An example of standard convolution with an input matrix with channel, $M = 3$ and N number of 3×3 kernels is illustrated in Figure 1a. All the convolution processes discussed in this section are assumed to have stride of one and padding. This standard convolution between F and N numbers of K will have a computational cost of:

$$J_c = N\left[(D_W \times D_H) \times (D_K \times D_K \times M)\right] \tag{1}$$

The MobileNet addresses this issue by splitting the convolution process into two steps, namely depthwise convolution and pointwise convolution. The whole process is called depthwise separable convolution. In the depthwise convolution process, instead of using kernel with depth M, single layer kernels are applied on each input channel as showed in Figure 1b. This results in a computational cost of:

$$J_d = M\left[(D_W \times D_H) \times (D_K \times D_K)\right] \tag{2}$$

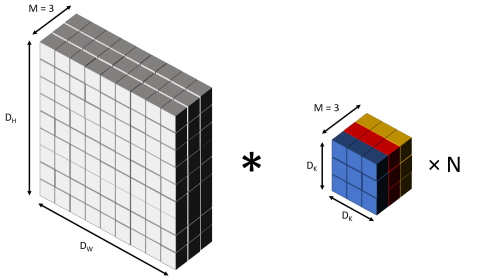
Then, pointwise convolution is performed on the output of the above depthwise convolution feature map. Pointwise convolution is being done by N number of 1×1 kernels with depth of M. This process is also called linear combination since the output matrix is of summation of linear scaling multiples. An pictorial illustration of this process is shown in Figure 1c. The computational cost of the pointwise convolution is:

$$J_p = N\left(M \times D_W \times D_H\right) \tag{3}$$

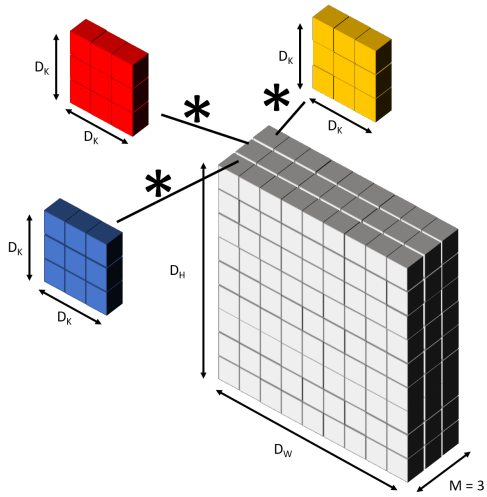
The total computational cost of a complete depthwise separable convolution is $J_{dp} = J_d + J_p$. Comparing to the standard convolution, there is a reduction factor of:

$$J_r = \frac{J_{dp}}{J_c} = \frac{1}{N} + \frac{1}{D_K^2} \tag{4}$$

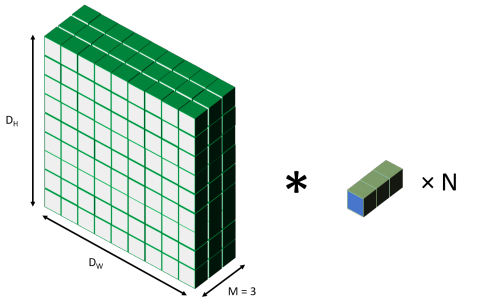
For a depthwise separable convolution with 3×3 kernels used, the computational cost can be 8 to 9 times lesser with only minor trade-off in accuracy [20].



(a) Standard convolution



(b) Depthwise convolution



(c) Pointwise convolution

Figure 1: Comparison between (a) Standard convolution (b) Depthwise convolution and (c) Pointwise convolution

3.3 MobileNetV2

MobileNetV2 is an improved version of MobileNetV1. MobileNetV1 or most deep convolution network uses the Rectified Linear Unit (ReLU) activation function defined by

$$f(x) = \begin{cases} x, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{5}$$

The ReLU activation function disregards any values smaller than 0. This non-linear transformation is argued to result in the loss of information, especially on input with a lower number of channels. ReLU may have less impact on input with lots of channels because the missing information in particular activation might still be preserved by other channels [10].

To tackle this issue, MobileNetV2 uses a narrow -> wide -> narrow approach in its inverted residual block. The low-dimensional input is first expanded by using a pointwise 1×1 convolution to produce a higher dimensional space which can cater for the information loss caused by the ReLU activation. Then, spatial filtering using depthwise convolution with ReLU activation was implemented in this higher dimensional feature map. Finally, another pointwise convolution was performed to project back to a lower dimensional output feature map. For the last step, linear activation instead of ReLU was used to preserve more information when encoding to a lower dimensional output map. This idea was called linear bottleneck by the authors. The non-linear transformation only happened in the internal expanded higher dimensional space inside the block. When come to lower dimensional output, linear transformation was used [10].

Besides, a skip connection similar to residual block is added between the input and output of the inverted residual block to allow gradient flow during backpropagation. This approach is essential to build a deep network. One last minor adjustment is the use of ReLU6 in this inverted residual block which capped the maximum output to 6 as defined by:

$$f(x) = min(max(0, x), 6) \tag{6}$$

The complete inverted residual block discussed in MobileNetV2 is shown by Figure 2.

4 Methodology

4.1 Dataset

Figure 3 shows the American Sign Language (ASL) for all alphabets. In the previous article, dataset was produced by smartphone camera captures [1]. There were a collection of 4800 images with 24 classes of gestures. The signs for character J and Z were excluded due to dynamic. All the images were taken from at a fixed camera distance with little lighting variation as displayed in Figure 4.

The challenge of developing a usable sign language recognition model lies in the variation in size, position, and shapes of the input image [15]. In real life applications, the sign language performer might stands at different distance relative to the smartphone camera. Provided data augmentation which is widely adopted in training CNN with limited dataset, which introduces more variations with scaling, rotating, flipping, shifting, shearing, etc. However, this image augmentation failed to reproduce the realistic perspective variations as discussed by Wenjin Tao et. al [4].

To extend the work, the dataset from Kaggle named "ASL Alphabet" was used [21] . Each sign has variations in distance capture and illumination. An example of sign B was shown in Figure 5. There are a total of 87,000 images in 29 classes. Besides the 26 classes belong to the letter A - Z, there are three additional classes for SPACE, DELETE and NOTHING. As per dataset description, these 3 classes are found to be helpful in real time application and classification.

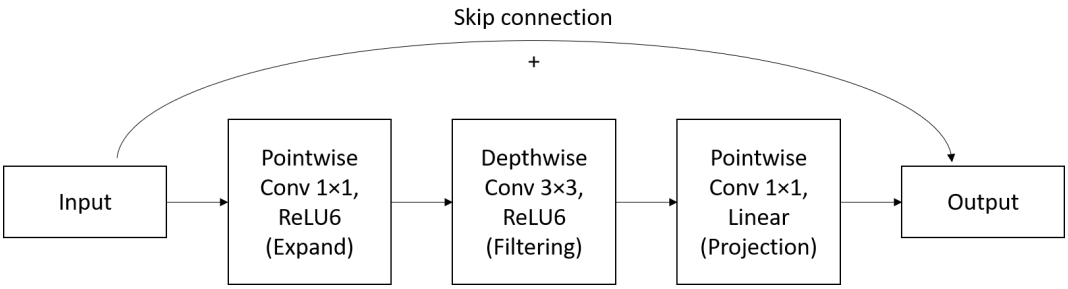


Figure 2: Inverted residual block

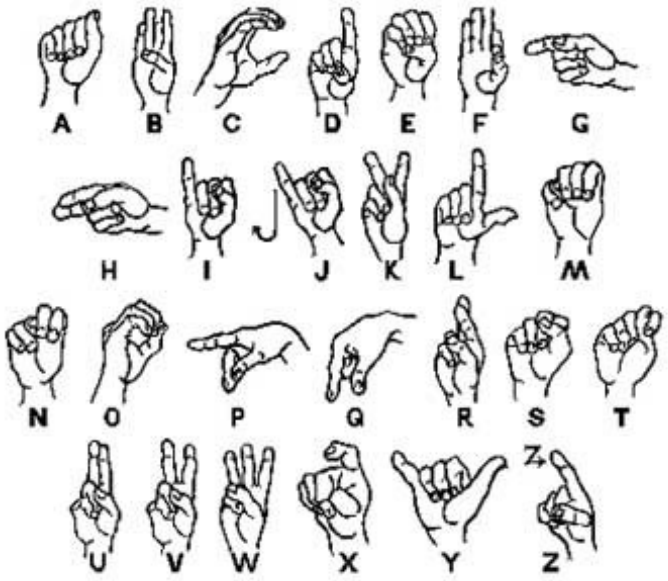


Figure 3: American Sign Language for all alphabets



Figure 4: Example self-captured dataset from previous article. From left to right (top) A, B, C and (bottom) D, E, F [1]

4.2 Model Architecture

The new dataset in this paper is approximately 18 times larger than the previous dataset. The previous proposed CNN network which had only two cascaded convolution layers failed to generalize to this dataset. The model was therefore being reconstructed using transfer learning with MobileNetV2. This CNN architecture was optimized for mobile devices to give high prediction accuracy while keeping the parameters and mathematical operations low. Compared to MobileNetV1, the MobileNetV2 is about 30 - 40% faster when tested on Google Pixel phone due to 2 times fewer mathematical operation and 30% lesser parameters. Not only that, it outperforms MobileNetV1 in terms of recognition accuracy [18].

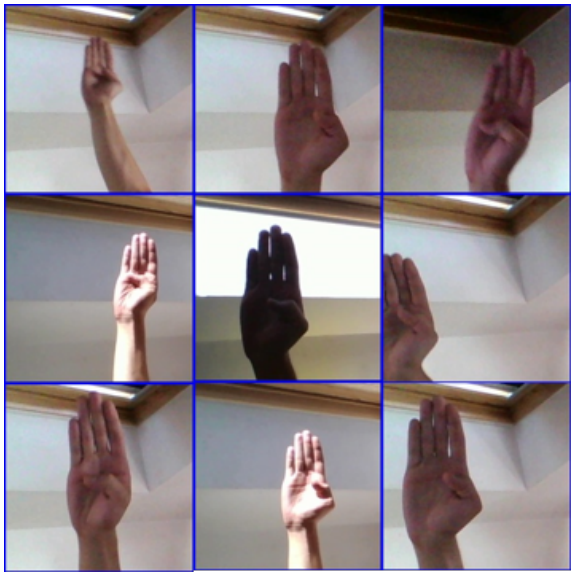


Figure 5: Example sign for letter 'B' under different capturing distance and lighting condition

All 29 classes had been used in this work. The images were resized to 224×224 pixels as suggested in the article by Samer Alashhab. et. al. [11]. Among different input sizes, 224×224 pixels gives best performance by experiment. The inputs were shuffled randomly and normalized. The images were divided into training and validation sets using a ratio of 80:20. The training set has 69,600 images while the validation set has 17,400 images. In view of the larger dataset, no data augmentation was carried out.

The built architecture is shown in Figure 6. The MobileNetV2 was preloaded with weights trained by ImageNet. Average pooling layer was added to the MobileNetV2 output feature map. Continued with a dense neural network layer of 1000 neurons. The activation function for this dense network was standard ReLU. A 50% dropout layer was added after that to prevent overfitting. Final layer was the dense layer that provides classification for 29 classes of images. It had a softmax activation function as defined by:

$$\sigma(z)_j = \frac{e_j^z}{\sum_{k=1}^{K} e_k^z} \tag{7}$$

The model was compiled using categorical crossentropy loss function. The model had about 3.5 millions training parameters. The optimizer used was RMSprop with a learning rate of 1×10^{-4}. Since the sign language images are differ from the ImageNet dataset, all the layers were unfreezed and retrained. Experiments were all conducted on a computer with Intel Core i5 CPU, 12 GB RAM and a NVIDIA GeForce GTX 1060. All of the steps mentioned in this section were performed using Keras deep learning framework with TensorFlow as the backend.

4.3 Performance Metrics

The performance metrics used to evaluate the build model are defined below, where TP stands for True Positive (correct prediction), TN is True Negative (correct rejection), FP is denoted as False Positive (incorrect prediction) and FN is defined as False Negative (incorrect rejection).

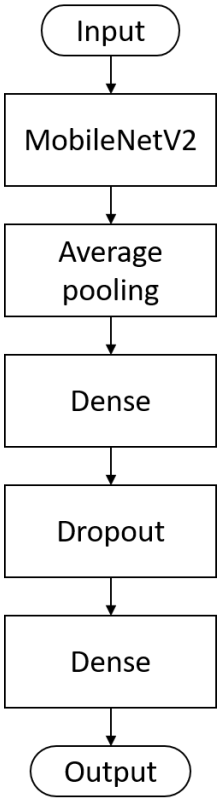


Figure 6: Model architecture

1. Accuracy - the overall correct prediction of the model

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

2. Precision - ratio of correct prediction over all the predicted values of a particular class.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

3. Recall - ratio of correct prediction over all predicted values that are supposed to be positive.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

4. F1-Score - weighted average of precision and recall. It tells the balance between precision and recall.

$$F1\ score = 2\left[\frac{Precision \times Recall}{Precision + Recall}\right] \tag{11}$$

5 Results and Discussion

The highest overall accuracy obtained by this model was 98.67%. The confusion matrix is shown in Figure 7. For each of the sign, 28 out of 29 were recorded with accuracy higher than 95%. The highest single class accuracy was 100% and the lowest was 85.67%. Lowest accuracy was recorded for the letter Q. There were only 514 out of 600 images were correctly predicted. The rests were all being predicted as P. This was believed to be the similarity existed in these two signs, where the index finger and thumb are both pointed out.

Looking at the precision, recall, and f1-score for each class as shown in Table 1, most of them obtained a score of 95% and above. The precision of letter P and recall of letter Q was lower due to the above mentioned similarity. Hence it also lowered the f1-score for these two classes. The second lowest recall was the sign for letter X which scored 92.83%. Some of them were being wrongly classified as either S or U due to the similar rounded fist shape.

Table 2 shows the comparison with some previous works. To have a fair comparison, only those works implementing transfer learning were compared. In the comparision, there was only one study done by Alashhab. S. et. al. had slightly better accuracy then this work, but they were not working with ASL and there were only 5 classes of hand signs used. Other than that, this work showed better results than the prior works. For a similar MobileNetV2 transfer learning architecture for ASL, there is also a marginally improvement over the work did by Bousbai, K and Merah, M.

Table 1: Precision, Recall, F1-score for each class

	Precision	Recall	F1-score
A	0.9677	1.0000	0.9836
B	0.9933	0.9917	0.9925
C	1.0000	1.0000	1.0000
D	1.0000	1.0000	1.0000
E	0.9693	1.0000	0.9844
F	0.9950	1.0000	0.9975
G	1.0000	0.9850	0.9924
H	0.9709	1.0000	0.9852
I	1.0000	0.9800	0.9899
J	1.0000	0.9633	0.9813
K	1.0000	1.0000	1.0000
L	1.0000	1.0000	1.0000
M	0.9771	0.9950	0.9860
N	0.9817	0.9850	0.9834
O	1.0000	0.9950	0.9975
P	0.8721	0.9883	0.9266
Q	1.0000	0.8567	0.9228
R	0.9934	1.0000	0.9967
S	0.9769	0.9850	0.9809
T	0.9983	0.9783	0.9882
U	0.9707	0.9933	0.9819
V	1.0000	0.9917	0.9958
W	0.9868	1.0000	0.9934
X	0.9893	0.9283	0.9579
Y	0.9983	1.0000	0.9992
Z	0.9983	0.9983	0.9983
delete	0.9967	1.0000	0.9983
nothing	1.0000	1.0000	1.0000
space	1.0000	1.0000	1.0000

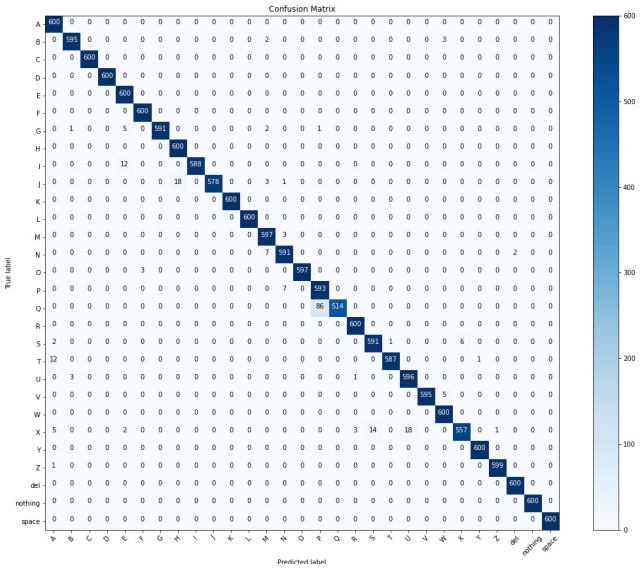


Figure 7: Confusion Matrix

Table 2: Comparison of this work and previous works

Authors	**Description**	**Accuracy**
Das, A. et. al., 2018	Transfer learning using Inception v3 on custom dataset	90.0%
Alashhab, S. et. al., 2018	Transfer learning on multiple architecture like VGGNet, ResNet, etc. on custom 5 classes of hand gestures.	99.45% (MobileNet)
Kania, K.et. al., 2018	Transfer learning using Wide Residual Networks with data augmentation on ASLs	93.3%
Garcia and Viesca, 2016	CNN on 24 ASLs with GoogLeNet tranfer learning	70%
Bousbai, K. and Merah, M., 2019	Compare custom CNN model and transfer learning using MobileNetV2 on ASLs	97.06% (MobileNetV2)
Our previous work, 2019	**Custom CNN on 24 ASLs using phone camera**	**95%**
This work	**Transfer learning using MobileNetV2 on 29 classes of ASLs**	**98.67%**

6 Conclusion and Future Work

This paper presented a model that was able to recognize American Sign Language with 98.67% accuracy. It is an extension of the work originally presented in ICSIPA 2019. Besides the improvement of the model accuracy, the following additional highlights are listed:

1. A much bigger dataset was used with additional 5 classes to make the ASL more complete. That allowed a better generalization of the model. In addition, this distance varying images given in the dataset could also make the model less sensitive to the difference in distance between the ASL users and the camera.

2. Even with the deep transfer learning network used, we still maintained our original intention to make the model suitable for running on mobile devices in order to use the smartphone camera for the convenience of users.

The model was less computational burdening since it had fewer parameters and lesser mathematical computations. With the implementation of the model as an smartphone app, it is believed to be able to improve the quality of communication between the deaf-mute person and the general public.

To offer more usability, this work can be continued with real-time video based sign language recognition. To reduce the effects of ambient noise, there is a need for real-time video processing, such as region of interest segmentation and hand tracking. Apart from that, the image occlusion has yet to be investigated. This can be a challenging issue when part of the performing signs is blocked.

Acknowledgment I wish to acknowledge Tunku Abdul Rahman University College for sponsoring this research.

References

[1] L. Y. Bin, G. Y. Huann, L. K. Yun, "Study of Convolutional Neural Network in Recognizing Static American Sign Language," in 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 41–45, IEEE, 2019, doi:10.1109/ICSIPA45851.2019.8977767.

[2] T.-W. Chong, B.-G. Lee, "American sign language recognition using leap motion controller with machine learning approach," Sensors, **18**(10), 3554, 2018, doi:10.3390/s18103554.

[3] M. A. Jalal, R. Chen, R. K. Moore, L. Mihaylova, "American sign language posture understanding with deep neural networks," in 2018 21st International Conference on Information Fusion (FUSION), 573–579, IEEE, 2018, doi: 10.23919/ICIF.2018.8455725.

[4] W. Tao, M. C. Leu, Z. Yin, "American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion," Engineering Applications of Artificial Intelligence, **76**, 202–213, 2018, doi:10.1016/j.engappai.2018.09.006.

[5] M. J. Cheok, Z. Omar, M. H. Jaward, "A review of hand gesture and sign language recognition techniques," International Journal of Machine Learning and Cybernetics, **10**(1), 131–153, 2019, doi:10.1007/s13042-017-0705-5.

[6] J. Wu, L. Sun, R. Jafari, "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors," IEEE journal of biomedical and health informatics, **20**(5), 1281–1290, 2016, doi: 10.1109/JBHI.2016.2598302.

[7] V. Bheda, D. Radpour, "Using deep convolutional networks for gesture recognition in American sign language," arXiv preprint arXiv:1710.06836, 2017.

[8] L. Pigou, S. Dieleman, P.-J. Kindermans, B. Schrauwen, "Sign language recognition using convolutional neural networks," in European Conference on Computer Vision, 572–578, Springer, 2014, doi:10.1007/978-3-319-16178-5_40.

[9] P. Molchanov, S. Gupta, K. Kim, J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 1–7, 2015, doi: 10.1109/CVPRW.2015.7301342.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.

[11] S. Alashhab, A.-J. Gallego, M. Á. Lozano, "Hand gesture detection with convolutional neural networks," in International Symposium on Distributed Computing and Artificial Intelligence, 45–52, Springer, 2018, doi:10.1007/978-3-319-94649-8_6.

[12] K. Kania, U. Markowska-Kaczmar, "American Sign Language Fingerspelling Recognition Using Wide Residual Networks," in International Conference on Artificial Intelligence and Soft Computing, 97–107, Springer, 2018, doi: 10.1007/978-3-319-91253-0_10.

[13] R. Hartanto, A. Kartikasari, "Android based real-time static Indonesian sign language recognition system prototype," in 2016 8th International Conference on information Technology and Electrical Engineering (ICITEE), 1–6, IEEE, 2016, doi:10.1109/ICITEED.2016.7863311.

[14] H. M. Ewald, I. Patil, S. Ranmuthu, "ASL fingerspelling Interpretation," University of Stanford, Reports, 2016.

[15] S. Ameen, S. Vadera, "A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images," Expert Systems, **34**(3), e12197, 2017, doi:10.1111/exsy.12197.

[16] A. Das, S. Gawde, K. Suratwala, D. Kalbande, "Sign language recognition using deep learning on custom processed static gesture images," in 2018 International Conference on Smart City and Emerging Technology (ICSCET), 1–6, IEEE, 2018, doi:10.1109/ICSCET.2018.8537248.

[17] B. Garcia, S. A. Viesca, "Real-time American sign language recognition with convolutional neural networks," Convolutional Neural Networks for Visual Recognition, **2**, 225–232, 2016.

[18] K. Bousbai, M. Merah, "A Comparative Study of Hand Gestures Recognition Based on MobileNetV2 and ConvNet Models," in 2019 6th International Conference on Image and Signal Processing and their Applications (ISPA), 1–6, IEEE, 2019, doi:10.1109/ISPA48434.2019.8966918.

[19] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE transactions on medical imaging, **35**(5), 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[21] Akash, "ASL Alphabet, https://www.kaggle.com/grassknoted/asl-alphabet," 2018 (Accessed July 1, 2020).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 488-498 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Neural Network-Based Fault Diagnosis of Joints in High Voltage Electrical Lines

Marco Bindi[*,1], Igor Aizenberg[2], Riccardo Belardi[1], Francesco Grasso[1], Antonio Luchetta[1], Stefano Manetti[1], Maria Cristina Piccirilli[1]

[1]*University of Florence, Department of Information Engineering, Firenze 50139, Italy*

[2]*Manhattan College, Department of Computer Science, NY,a 10471, USA*

ARTICLE INFO

Article history:
Received: 10 June, 2020
Accepted: 20 July, 2020
Online: 09 August, 2020

Keywords:
High-Voltage transmission line
Complex neural network
Smart monitoring
Fault diagnosis
Electrical joint modeling
Testability analysis

ABSTRACT

In this paper a classification system based on a complex-valued neural network is used to evaluate the health state of joints in high voltage overhead transmission lines. The aim of this method is to prevent breakages on the joints through the frequency response measurements obtained at the initial point of the network. The specific advantage of this kind of measure is to be non-intrusive and therefore safer than other approaches, also considering the high voltage nature of the lines. A feedforward multi-layer neural network with multi-valued neurons is used to achieve the goal. The results obtained for power lines characterized by three and four junction regions show that the system is able to identify the health state of each joint, with an accuracy level greater than 90%.

1. Introduction

In the management of a long electrical transmission line, fault location represents an important aspect, because it allows to reduce recovery times and increase network availability [1]. Currently, high voltage lines are protected by devices which allow to identify the fault position and put out of service the non-functioning part of the grid. For example, the distance protection employed on overhead transmission lines uses Intelligent Electronic Devices (IEDs) to estimate the impedance between the protection point and the failure point by measuring the line voltage and current [2]. In this way it is possible to obtain the distance from the fault, exploiting the correspondence between conductor impedance and length of the line section. In the last years new protection devices have been introduced, based on traveling wave detection; they allow to measure the impulsive signals generated by failures with greater precision [3]. However, these methods allow to locate the failure only once it has occurred [4].

In this paper, an extension of a work originally presented at the International Conference on Soft Computing & Machine Intelligence (ISCMI 2019) [1], a system for monitoring the health status of the joints [1], [5] is presented. The system allows a preventive maintenance, so increasing the availability of the line. The joints (or junction regions) are the connection points between two different parts of the same phase conductor, and they represent one of the most stressed parts of overhead electrical lines [6], [7]. The proposed method exploits the equivalent lumped circuit of conductors and junction regions [8], [9]. In particular, the model of the elementary section of the network consists in the connection of the equivalent circuits of joint and conductor. The model of the whole line is obtained by the cascade connection of several elementary sections. The health status of the joint is obtained by measuring the network frequency response. To this aim it is necessary to establish a nominal working range for each electrical parameter of the line model. Any departure from this interval can be interpreted as a symptom of a possible fault. The deviation from the nominal condition is obtained through a comparison between the theoretical frequency response and measurements taken at several frequencies. The results of this comparison are traced back to the corresponding variation of the joint parameters. This operation is executed by a neural network able to classify the severity of the joint degradation. The selected neural network is a feedforward Multi-Layer network with Multi-Valued Neurons (MLMVN) whose training phase is carried out using the theoretical frequency response simulated on *SapWin IV* and MatLab®. In this way it is possible to obtain a smart monitoring approach capable of classifying the measurements obtained through the PLC (Power Line Communication) systems [10]. This type of equipment is usually integrated on high and medium voltage networks to use the phase conductor as a channel for the

* Corresponding Author: Marco Bindi, 50139 Via S. Marta 3, Firenze, Italy, m.bindi@unifi.it

transmission of information. Since the communication method is based on the modulation of the carrier wave, PLC systems can be used to generate signals with different frequencies obtaining measurements of the line transfer function [11].

The paper is organized as follows. Section 2 describes the model of conductors and joints. Section 3 is focused on the preliminary steps of the procedure, while section 4 is dedicated to the simulation procedure. Finally, Sections 5 and 6 report some specific application cases and conclusions, respectively.

2. Line Modelling

The main physical components of a line are junction regions and conductor stretches. The derivation of the equivalent circuits of these components represents the starting point for the line modelling. It is then necessary to establish the relationships between fault mechanisms and variation of electrical parameters. Only the joint degradation is considered in this paper and two different failure mechanisms are introduced: oxidation process and partial breakage of the joint structure. In this way it is possible to obtain three different intervals for the values of the electrical parameters, corresponding to the nominal situation, to the oxidation condition and to the presence of partial breaking [1].

2.1. Conductor modelling

In this paper an ACSR (Aluminum Conductor Steel Reinforced) with a diameter of 22.8 mm is used (Figure 1a). The equivalent model of the conductor (Figure 1b) contains four electrical elements and is called "canonical π-model".

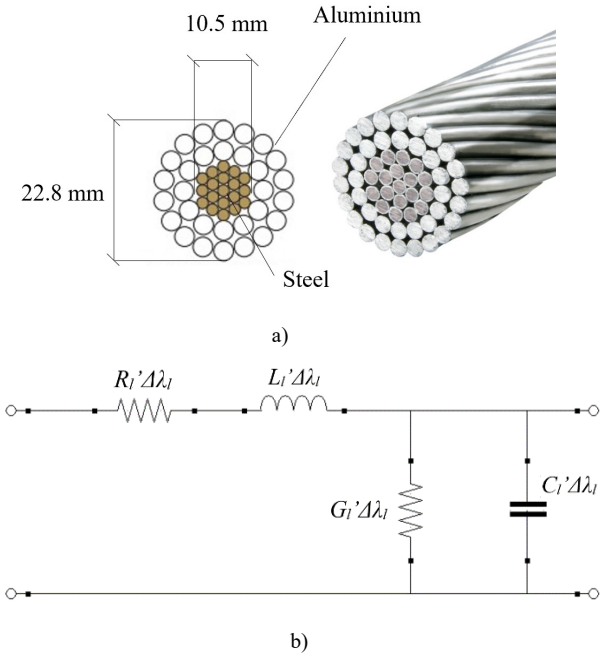


Figure 1 – a) Physical structure of the conductor; b) Lumped circuit of the conductor.

It is commonly used to study wave propagation and power flow analysis [12], [13]. The term $\Delta\lambda_l$ represents the conductor length, while the values of the parameters R_l', L_l', G_l', C_l' depend on the mechanical characteristics of the phase conductor; all these parameters are quantities per unit of length [14].

2.2 Joint modelling

Junction regions represent the connection points between two different parts of the same phase conductor, and they must guarantee the electrical continuity along the network [15], [16]. Bolted joints are used close to the pylons due to their high mechanical strength [17]. However, no available literature exists that describes their electrical behaviour; an acceptable alternative consists in modelling them with the better-known model of the solder joint, for which some information is available about its degradation mechanisms [18], [19]. The physical structure of the joint is shown in Figure 2a.

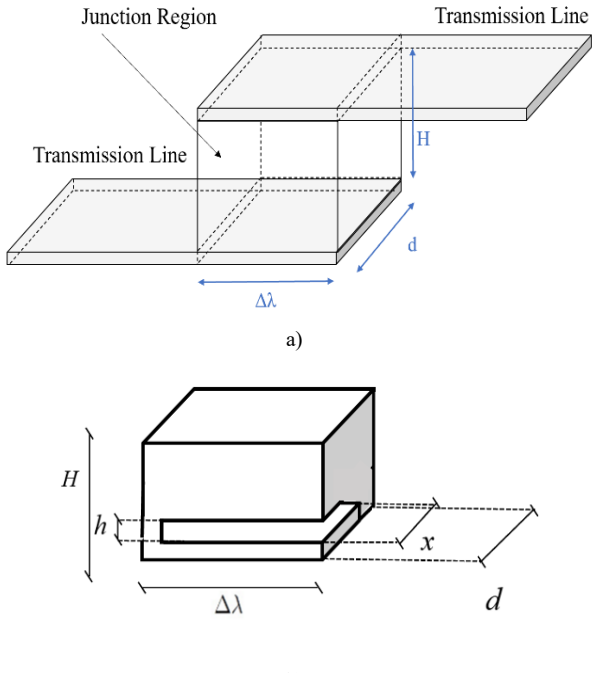


Figure 2: a) Physical structure of the junction region; b) Braking parameters.

Δλ is the length of the junction region, d is its width and H its height. The characteristic parameters of the breaking mechanism are shown in Figure 2b where x represents the crack width and h is its height. The physical model and the equivalent lumped circuit of the joint are shown in Figure 3a and Figure 3b, respectively. In the equivalent circuit, the value of R_{sj} and L_{sj} in nominal conditions can be calculated by $R_{sj} = \frac{\rho\, H}{\Delta\lambda\, d}$ and $L_{sj} = \frac{\mu_0\, \mu_r\, H}{2\pi}\left[\ln\frac{2H}{\Delta\lambda+d} + 0{,}5\right]$ respectively, while the capacitance value is nominally zero. If the frequency value used for the measurements exceeds the network frequency, it is necessary to consider the skin effect. For this reason, the penetration depth of the current (a frequency dependent parameter, calculated by $\delta = \frac{1}{\sqrt{\pi\sigma\mu_0\,\mu_r f}}$) is introduced and (1) is used to evaluate the resistance R_{sj}.

$$R_{sj} = \frac{\rho\, H}{2\delta\,(\Delta\lambda + d - 2\delta)} \tag{1}$$

In oxidation conditions the resistance changes its value due to the increase of resistivity, while the partial breakage mechanism induces variations on the electrical parameters of the joint. In this work the consequentiality between oxidation mechanism and partial breakage process is assumed.

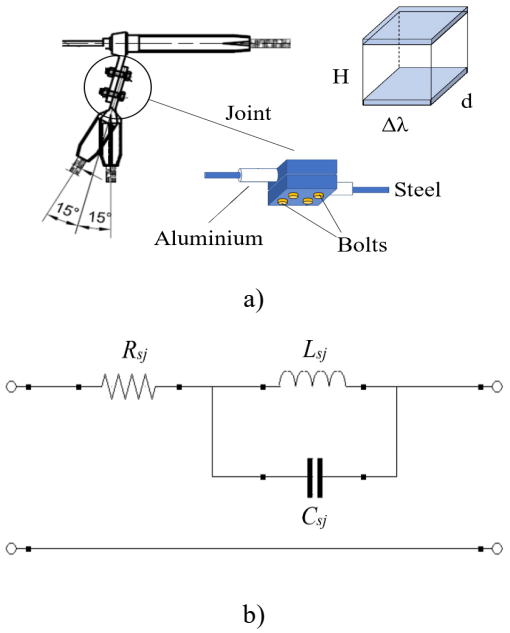


Figure 3 – a) Physical model of the joint; b) Lumped circuit of the joint.

The formulas (2), (3) and (4) are obtained by introducing the typical breaking mechanism of solder joints and describing the variations of R_{sj}, L_{sj} and C_{sj} as functions of the crack width x and the crack height h [18]. Actually, only the R_{sj} and L_{sj} variations are considered, because the sensitivity of the C_{sj} variation is low.

$$R_{sj} = \begin{cases} \rho\left[\frac{H-h}{2\delta(\Delta\lambda+d-2\delta)} + \frac{h}{2\delta(\Delta\lambda+d-x-2\delta)}\right] & (d-x) \geq 2\delta \\ \rho\left[\frac{H-h}{2\delta(\Delta\lambda+d-2\delta)} + \frac{h}{\Delta\lambda\,(d-x)}\right] & (d-x) < 2\delta \end{cases} \quad (2)$$

$$L_{sj} = \frac{\mu_0\,\mu_r\,H}{2\pi}\left[\ln\frac{2(H-h)}{(\Delta\lambda+d)} + 0{,}5\right] + \frac{\mu_0\,\mu_r\,H}{2\pi}\left[\ln\frac{2H}{(\Delta\lambda+d-x)} + 0{,}5\right] \quad (3)$$

$$C_{sj} = \frac{\varepsilon_0\varepsilon_r\,x\,\Delta\lambda}{h} \quad (4)$$

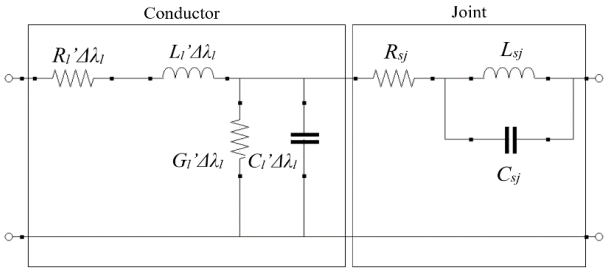


Figure 4: Elementary section of the line.

3. Problem outline

Once the equivalent lumped circuits of joint and conductor have been built, it is possible to describe the whole network by connecting in cascade the elementary line sections shown in Figure 4. Each line section between two pylons is represented by an elementary section and each joint model corresponds to the pylon position. A simplification is used in this case because there should be two joints and a short section of the conductor at the pylon. This is not a problem for the theoretical basis of the procedure presented in this paper, but it should be taken into account in practical applications [4, 1].

The next step of the procedure consists in the simulation of the model. This phase is needed for testability analysis, frequencies selection and neural network training, as shown in the following.

3.1. Testability analysis

It goes without saying that an a-priori knowledge of how many and which parameters are actually identifiable is quite useful in any fault diagnosis procedure, since it allows to correctly select the measurement test points. Testability analysis is exactly aimed at providing this kind of information [20, 21, 22]. It consists of two steps, testability evaluation and ambiguity groups identification. The testability gives the solvability degree of the fault diagnosis equations [23], i.e. it allows establishing the number of parameters whose value can be determined, starting from measurements taken on a given number of test points. The maximum possible value of testability is the number N of parameters present within the circuit under test. This maximum value is often not reached; this means that, in case of fault, there are parameters that are indistinguishable one to each other. In other words, it is possible to identify a set that contains the faulty component, but it is not possible to exactly identify the faulty one inside the group. The ambiguity groups are all those groups constituted by sets of indistinguishable parameters. The determination of the ambiguity groups gives information about which circuit parameters are identifiable.

In the procedure presented in this paper, testability analysis is a preliminary phase, aimed at discovering which and how many parameters of the line model are identifiable. This operation is executed starting from the network function, which acts as reference for the measurements, taken at selected frequencies. Consequently, it is necessary to simulate the line model in order to determine its network function.

3.2. Optimal frequency selection

As previously stated, the fault identification is the result of a comparison between the measurements at the test point and the nominal amplitude response, evaluated at the same frequencies. As shown in [24], measurement errors and manufacturing tolerances, which affect the result in a similar manner, depend on the choice of the frequencies. A proper selection allows minimizing the effect of these errors, increasing the probability to detect the fault. A method for carrying out this selection is presented in [24] and, in this work, it is used to obtain one test frequency. However, three octaves of the selected frequency are also used in the simulation procedure in order to make the system more robust and better train the complex neural network. It is worth pointing out that in [24] the use of more than one frequency has the scope to increase the number of equations, so that their number is at least as high as that of the unknown parameters. In the procedure presented in this paper there is no parameter identification, so there are no fault diagnosis equations to solve. Nevertheless, it is necessary to determine the network function of the line model in order to apply the procedure in [24], based on the network function derivatives.

3.3. Neural network training

The diagnosis method proposed in this paper uses neural networks, which do not require the derivation of fault equations.

The analysis of the circuit model is however necessary for the training of the neural network, as it will be shown in the next section. The used intelligent classifier is a complex neural network, that presents excellent performances compared to other machine-learning techniques. It is based on a feedforward multilayer neural network with multi-valued neurons (MLMVN), characterized by a derivative free learning algorithm [25], an alternative algorithm based on the linear least square (LLS) methods [26] to reduce the high computational cost of the original backpropagation procedure, a soft margin method [27, 28] in order to make the network a good classifier. Further details will be given later in the application description.

4. Simulation Procedure

All the fundamental steps in the classification of the joint conditions are presented in this paragraph. First it is necessary to define the main characteristics of the network, the type of conductor and the physical size of the joints. Referring to the technical data sheets of the ACSR shown in Table 1, it is possible to obtain the value of the DC electrical resistance per unit of length R_l' at ambient temperature (293K) and, consequently, the value of the resistivity $\rho_l = \frac{R_l' \pi \left(\frac{D}{2}\right)^2}{\Delta\lambda_l}$. The conductor characteristics are extracted from [17], while the capacitance C_l' and the inductance L_l' are calculated assuming the transposition of the phase conductors. The formulas used in this case consider the number of conductors, their relative position and their diameter:

$$L_l^{'} = 0{,}46 \log\left(\frac{\Delta^{'}}{(D/2)}\right) = 1.28 \text{ mH/km} \tag{5}$$

$$C_l^{'} = \frac{0{,}024}{\ln\frac{\Delta^{'}}{(D/2)}} = 3.74 \text{ nF/km} \tag{6}$$

where Δ' is the equivalent distance between the conductors ($\Delta^{'} = \sqrt[3]{\Delta_{12}\Delta_{23}\Delta_{31}}$). The physical size of the joint and the nominal value of the DC electrical resistivity are extracted from [15] and shown in Table 1. The joint resistivity can be calculated through the equation $\rho_{sj} = \frac{R_{sj}\, d\, \Delta\lambda}{H}$. In nominal conditions the inductance value L_{sj} is $1.5\,\mu H$ and the capacitance C_{sj} is calculated by (4) assuming very low values for the breaking parameters ($h = 0{,}001H$ and $x = 0{,}001d$). In this way, it is possible to take into account the parasitic effects and maintain the same circuit topology for the nominal condition and the partial breaking condition. Therefore, the nominal value of the joint capacitance is $0.01\ pF$.

4.1. Tolerance on the conductor electrical parameters

As previously stated, the conductors are assumed not faulty, hence they are set to their nominal values. Their resistance R_l' depends on temperature and is subjected to the skin effect, on its time depending on the frequency f_m used for the measurements. The penetration depth of the current δ_l is given by the standard relation $\delta_l = \frac{\sqrt{\rho_l}}{\sqrt{\pi\, \mu_l\, f_m}}$, where ρ_l and μ_l are respectively the resistivity of the conductor and the relative magnetic permeability. In order to obtain the correct value of the resistance, a parameter $K_0 = \frac{\left(\frac{D}{2}\right)}{2\,\delta_l}$ is introduced, which is calculated as the ratio between the radius of the conductor and the penetration depth of the current. The electrical resistance of the conductor at ambient temperature (293K) can be obtained by multiplying R_l' for a coefficient K so defined:

$$\begin{cases} K = \left[1 + \frac{K_0^4}{3}\right] & (K_0 < 1) \\ K = \left[0{,}25 + K_0 + \frac{3}{K_0}\right] & (K_0 > 1) \end{cases} \tag{7}$$

Then the final form of the electrical resistance is obtained by $R_{lfm}^{'} = KR_l^{'}$. As well known, the resistivity of a conductor changes its value with respect to the temperature. This means that there is a tolerance around the nominal resistance due to the conductor temperature. The definition of the resistance range is very important for verifying the performance of the monitoring system, since the neural network considers all the conductor components fixed at their nominal values, but the measurements of the frequency response change according to their real values. Therefore, the energy balance equation of transmission line is introduced [29, 30]:

$$C_m v\pi A_t \frac{dT_c}{dt} = Rl'^{(1+\alpha(T_c-293))I^2} + \alpha_s Q_s A_s - \alpha_c (T_c - T_\alpha) A_s 5{,}67\alpha_r \left[\left(\frac{T_c}{100}\right)^4 - \left(\frac{T_\alpha}{100}\right)^4\right] A_r \tag{8}$$

where: T_c is the temperature of the conductor; T_a is the air temperature surrounding the conductor; C_m is the specific heat of the conductor materials; v is the density of the conductor materials; A_t is the conductor section; A_s is the surface area receiving illumination with unit of length; A_c is the convection surface area; Ar is the radiation surface area per unit of length; α_c is the coefficient of convective heat; α_r is the radiation factor of the conductor materials; α_s is the sunshine absorption rate of the

Table 1: Characteristic of the line conductor and bolted joint

ASCR 22.8	Line Voltage [kV]	Outer Diameter D [mm]	Wire diameter (Al) [mm]	Number of wires (Al)	Wire Diameter (Steel) [mm]	Number of wires (Steel)	R_l' [Ω/Km]
	132	22.8	3.5	26	2.8	7	0.12
Bolted joint	H [m]		Δλ [m]		d [m]		R_{sj} [μΩ]
	0.038		0.008		0.008		60

Table 2: Characteristics of the conductor materials

C_m [J/(kg K)]	ν [kg/m^3]	A_t [m^2]	A_s [m^2/m]	A_r [m^2/m]	A_c [m^2/m]	α_s	α_c [W/(kg/m^3)]	α_r
0.88 e-3	2.7 e3	$\pi(D/2)^2$	$\pi(D/2)$	πD	πD	0.6	5.8	0.2

conductor; Q_s is the radiation intensity of the sun and sky. Within the equation there is the term R_l' because it produces heat due to the Joule effect. It is necessary to observe that this is the value of the DC resistance considered equal to the resistance value at 50 Hz. The characteristics of the conductor material are extracted from [29] and reported in Table 2.

The expression of temperature variations as a function of time is obtained by solving the energy balance equation. In order to solve the equation, the initial conditions of the conductor, the radiation intensity Q_s and the air temperature T_α must be fixed. Typical values of Q_s and T_α, based on the measurements obtained in Italy in 2011, are used in this work. In the following, Q_{smax} represents the average of the maximum radiation levels for a specific month and $T_{\alpha max}$ that of the maximum air temperatures. The average of the minimum temperature levels T_{min} is considered as the mean temperature in the night period and, for this reason, it is associated with a zero-radiation level. For simplification, the average values for each season are considered in Table 3. Once the load current has been fixed, the incidence of the environmental conditions on the resistance value can be obtained.

Table 3: Average values of solar radiation and temperature for each season

	Winter	Spring	Summer	Autumn
Q_{smax} [W/m^2]	392	753	750	357
$T_{\alpha max}$ [K]	281	295	301	284
Q_{smin} [W/m^2]	0	0	0	0
$T_{\alpha min}$ [K]	275	286	292	279

For example, if in summer the conductor temperature is calculated using (Q_{smax};$T_{\alpha max}$), the maximum resistance error in the night period is about 16%. Considering the interval 0÷500 A for the load current and setting the environmental conditions, it is possible to obtain the incidence of the load on the resistance value. For example, in summer, the influence of the load current is about 7%. The worst situation for each season is presented in Table 4 and the errors are calculated assuming the maximum difference between the presumed and the real current. Similarly, the difference between the assumed and actual environmental conditions is maximized.

Table 4: Worst error for each season

Selected load current [A]	Real load current [A]	Error %			
		Winter	Spring	Summer	Autumn
500	0	17.34	25.16	24.92	16.1

In this case the maximum error is made in spring and its percentage value is about 25,16%. By calculating the temperature with intermediate starting conditions between (Q_{smax};$T_{\alpha max}$) and (Q_{smin};$T_{\alpha min}$) and fixing the current at 250 A, it is possible to consider a 10÷12% variation range for the conductor resistance. Where (Q_{smax};$T_{\alpha max}$) and (Q_{smin};$T_{\alpha min}$) are selected and real environmental condition respectively. The same variation range is extended to each electrical parameter of the conductor.

4.2. Circuit designing on SapWin IV

The simulator *SapWin IV* (Symbolic Analysis Program for Windows) [31, 32] is used for executing the circuit analysis and obtaining the symbolic network function of the model. The line admittance calculated at the starting point of the line is the network function. The first step of the fault procedure is the testability analysis, carried out by the program LINFTA [20] exploiting the *SapWin IV* simulation. Despite the fact that only the joint parameters are variable, all the electrical components are taken into account in the testability analysis. The circuits characterized by 2, 3, 4 and 5 elementary sections have been tested and the corresponding testability values are always maximum. The second step consists in the determination of the optimum set of frequencies. A program associated with the package *SapWin IV* performs this operation [24]. In this phase many circuit simulations are needed, at various different frequencies. The availability of the network function in symbolic form, as that provided by *SAPWIN*, is fundamental for keeping the processing times within reasonable limits. A single simulation is in fact sufficient, because, once the symbolic network function is available, it can be directly used to derive the response at all the desired frequencies. In the third step of the fault procedure the symbolic network function extracted from *SapWin* is processed on MatLab® to obtain the training samples. Through a MatLab® script, it is possible to set the conductor parameters at the numerical values and generate the training samples by varying the electrical components of each joint in its fault classes. Each conductor parameter is generated randomly in an interval of 10% around its nominal value, so taking into account the effects of environmental conditions and load current.

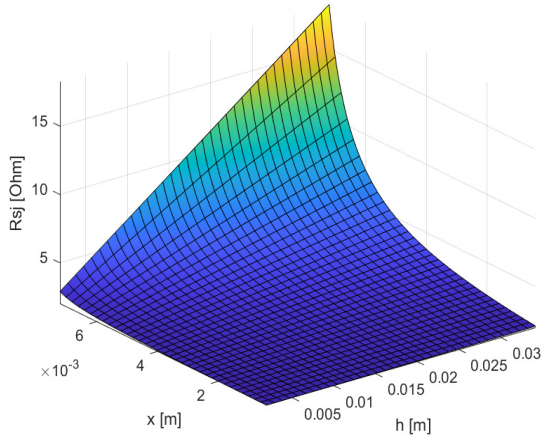
4.3. Fault classes and neural network setup

Three possible health states are considered for each junction region: nominal condition, structure oxidation and partial breakage. In order to obtain the correct classification of the working conditions by using the neural network, it is necessary to set the corresponding intervals for each electrical parameter. As mentioned above, the nominal value of the joint resistance is obtained from [17], while those of the inductance and capacitance have been calculated. Using the results presented in [17], it is also possible to define the resistance values in case of presence of the oxidation process (Table 5). These values represent the DC resistances of the joint in each operating condition and the

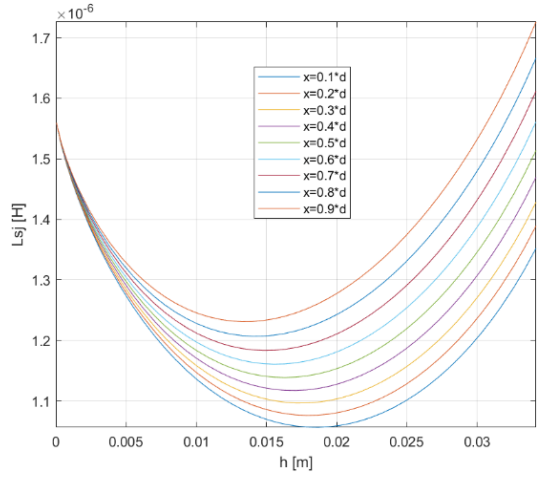
Table 5: Values of the joint resistance in the different oxidation conditions

Nominal condition	Low oxidation	Medium oxidation	High oxidation
$R_{sj} = 60\ \mu\Omega \div 2.5\ m\Omega$ $\rho_{sj} = 10^{-7} \div 4 \cdot 10^{-6}\ \Omega$ m	$R_{sj} = 2.5\ m\Omega \div 5\ m\Omega$ $\rho_{sj} = 4 \cdot 10^{-6} \div 8.5 \cdot 10^{-6}\ \Omega$ m	$R_{sj} = 5\ m\Omega \div 100\ m\Omega$ $\rho_{sj} = 8.5 \cdot 10^{-6} \div 1.7 \cdot 10^{-4}\ \Omega$ m	$R_{sj} = 100 m\Omega \div 2\ \Omega$ $\rho_{sj} = 1.7 \cdot 10^{-4} \div 3.4 \cdot 10^{-3}\ \Omega$ m

corresponding resistivities. The resistance values are given by (1) and we can note that the incidence of the skin effect decreases as the oxidation process increases. In fact, the oxidation process increases the resistivity of the material and, consequently, also increases the depth of the current penetration. Therefore, the skin effect is not relevant in the oxidation conditions, while it could introduce a false positive error in the nominal conditions.



a)

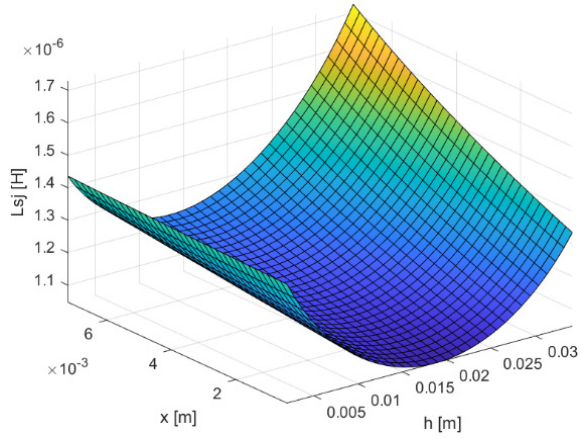


b)

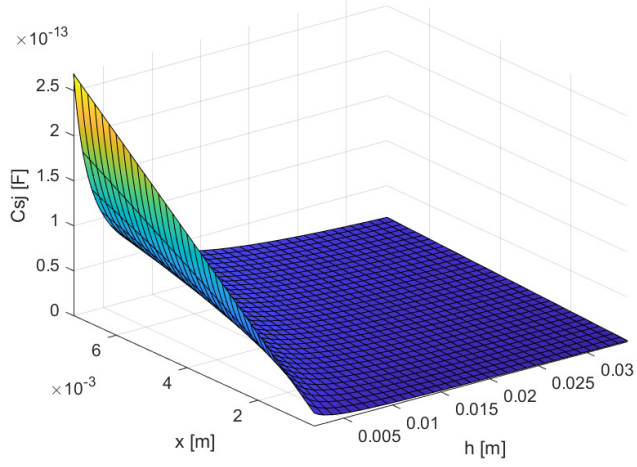
Figure 5: a) Resistance of the joint with respect to the crack height and crack width; b) Inductance of the joint with respect to the crack height and crack width

The skin effect can be also neglected in correspondence with the partial damage of the joint structure since the consequentiality between oxidation and breaking mechanism has been assumed. For this reason, only the second relation presented in (2) is used to describe the variation of the joint resistance with respect to the size of the crack (Figure 5a). In this case the resistivity $\rho_{sj} = 3{,}4 \cdot 10^{-3}\ \Omega m$ is used and $\delta = d/2$. The resistance of the joint $R_{sj} = 2\ \Omega$ represents the upper limit in the oxidation conditions. The inductance variation with respect to the size of the crack is presented in Figure 5b and Figure 6a. According to the inductance value previously calculated, a tolerance of 0.2 µH is used to define the nominal conditions.

As shown in Figure 5b the value of L_{sj} leaves the interval of nominal conditions for h and x included between 10% and 65% of H and $\Delta\lambda$ respectively.



a)



b)

Figure 6: a) Inductance of the joint with respect to the crack height and crack width; b) Capacitance of the joint with respect to the crack height and crack width

Since it is reasonable to consider equal percentage variations of x and h, since 65% of breakage represents a very high level of degradation, the interval chosen for the braking conditions is $L_{sj} = 1 \div 1.3\ \mu H$ neglecting the size of the crack for which the inductance returns in the nominal interval. Concerning the capacitance of the joint, its variation with respect to the size of the crack is shown in Figure 6b. Unfortunately, the sensitivity of the

Table 6: Fault classes

Joint Parameters	Nominal Condition	Oxidation Condition	Breaking Condition
R_{sj}	$60\ \mu\Omega \div 2.5\ m\Omega$	$2.5\ m\Omega \div 2\ \Omega$	$> 2\ \Omega$
L_{sj}	$1.5 \pm 0.2\ \mu H$	$1.5 \pm 0.2\ \mu H$	$1 \div 1.3\ \mu H$
C_{sj}	$0.01\ pF$	$0.01\ pF$	$0.01\ pF$

line frequency response with respect to the capacitance is low and, consequently, the value of C_{sj} is fixed to the nominal value.

Starting from these results the fault classes are defined. They are shown in Table 6. Once the equivalent lumped circuit has been obtained and the value of each electrical parameter has been defined, the complex neural network must be adapted in order to classify the health state of the joints.

The inputs of the neural network are the complex values corresponding to the network function at the selected frequencies. The network is characterized by two layers of neurons and two output neurons for each junction region (Figure 7).

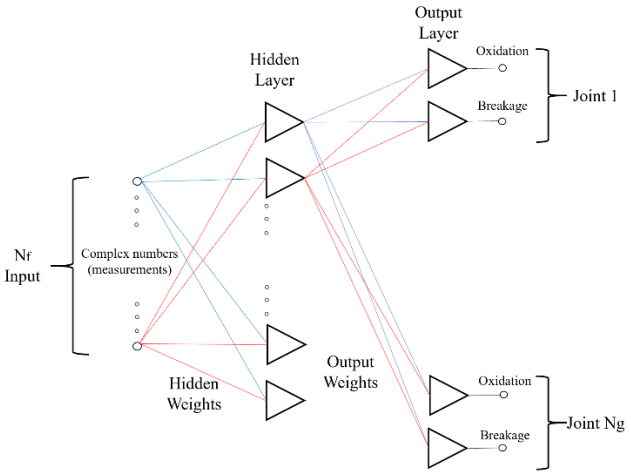
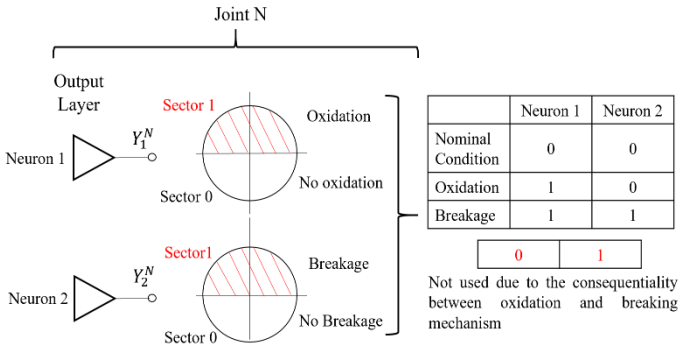


Figure 7: Structure of the neural network

Each output neuron divides the complex plane in two different sectors (Figure 8) representing the nominal conditions and the presence of the failure mechanism.



	Neuron 1	Neuron 2
Nominal Condition	0	0
Oxidation	1	0
Breakage	1	1

0	1

Figure 8: Output combinations for each neuron

The first output neuron of each joint takes into account the oxidation process, while the second one considers the partial breaking mechanism. When the output level of both neurons is low (sector zero), the corresponding junction region is within the nominal conditions. If the first output is high (sector one) and the second one is low (sector zero), the combination represents the oxidation of the joint. If each output presents a high level (sector one), the corresponding junction region has partial damage. The last combination is not taken into consideration due to the hypothesis of the consequentiality between oxidation and partial breaking mechanism.

This paper uses a complex neural network implemented on MatLab® version R2019b, but the whole code can also be used in previous versions of this software. Once the number of joints N_g has been set, there are $N_c = 3^{N_g}$ possible output combinations, since each junction region can be oxidized, broken or fully functional.

4.4. Neural network training and testing

The complex neural network used in this work, like other perceptron neural networks, has a good learning speed and allows to significantly reduce the computational cost by having a derivative free learning algorithm. Furthermore, the complex nature of this network is very suitable for dealing with electrical quantities and usually it is possible to obtain a better generalization capability in comparison with other neural networks and neuro-fuzzy networks. The learning rule is shown in (9):

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \frac{\alpha_r}{(n+1)|Z_k|}(D - Y)\bar{\boldsymbol{X}} \tag{9}$$

where $\boldsymbol{w}_{k+1}$ is the vector of the corrected weights, α_r is the learning constant, n is the number of the inputs, Z_k is the weighted sum, D is the desired output, Y is the actual output and $\bar{\boldsymbol{X}}$ is the conjugate complex vector of the inputs. Considering N_{rs} random samples for each combination, the matrix of the data sets presents $N_t = N_{rs}N_c$ rows. Each row contains N_f magnitude measurements and N_f phase measurements of the same network function, with N_f number of test frequencies.

$$\begin{bmatrix} |M_1^1|_{f_1} & |\phi_1^1|_{f_1} & \cdots & |M_1^1|_{f_{N_f}} & \cdots & O_1^1\,O_2^1 & \cdots \\ |M_2^1|_{f_1} & |\phi_2^1|_{f_1} & \cdots & |M_2^1|_{f_{N_f}} & \cdots & O_1^1\,O_2^1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\ \vdots & \vdots \\ |M_{N_{rs}}^{N_c}|_{f_1} & |\phi_{N_{rs}}^{N_c}|_{f_1} & \cdots & |M_{N_{rs}}^{N_c}|_{f_{N_f}} & \cdots & O_1^1 O_2^1 & \cdots \end{bmatrix} \tag{10}$$

Equation (10) shows the structure of the matrix containing the data sets. For example, $|M_2^1|_{f_1}$ represents the second measure of magnitude corresponding to the first combination made at the frequency f_1, $|\phi_2^1|_{f_1}$ represents the second measure of phase corresponding to the first combination made at the frequency f_1 and O_2^1 is the second output of the first joint. Initially the weights of the neural network are randomly chosen and, subsequently, for each sample the error between the desired output and the actual one is

calculated. The procedure shown in [26] is used to obtain the correction of the weights. This method uses a part of the total samples N_a, while the remaining part is used for the testing procedure. The backpropagation formulas calculate the error for each weight of the hidden layer and, then, the system shown in (11) is obtained.

$$\begin{bmatrix} 1 & X_1^1 & \cdots & X_{N_f}^1 \\ 1 & X_1^2 & \dots & X_{N_f}^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_1^{N_a} & \dots & X_{N_f}^{N_a} \end{bmatrix} \begin{bmatrix} \Delta W_0 & \Delta W_0 & \dots & \Delta W_0 \\ \Delta W_1^{11} & \Delta W_1^{21} & \cdots & \Delta W_1^{N_J 1} \\ \vdots & \vdots & \cdots & \vdots \\ \Delta W_{N_f}^{11} & \Delta W_{N_f}^{21} & \dots & \Delta W_{N_f}^{N_J 1} \end{bmatrix} = \begin{bmatrix} \delta_{11}^1 & \delta_{21}^1 & \cdots & \delta_{N_J 1}^1 \\ \delta_{11}^2 & \delta_{21}^2 & \dots & \delta_{N_J 1}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{11}^{N_a} & \delta_{21}^{N_a} & \dots & \delta_{N_J 1}^{N_a} \end{bmatrix} \quad (11)$$

For example, X_1^2 is the first line input corresponding to the second sample, ΔW_1^{21} represents the correction for the weight between the first input and the second neuron of the first layer, δ_{21}^2 is the error on the second neuron of the first layer corresponding to the second sample. The values of the corrections are calculated through the Q-R decomposition and the same procedure is used for the output layer on the system shown in (12). The weight adjustment is repeated until the error satisfies the stopping criteria indicated in [27]. Therefore, there are two different stopping criteria: the first concerns the maximum number of classification errors that can be committed during the test phase and this means that the operator can set the maximum number of outputs that can be outside the desired sector . The second stopping criterion is chosen according to the rules of the soft margin technique and consists of a tolerance range on the distance between the value of each output and the bisector of the desired sector.

$$\begin{bmatrix} 1 & Y_{11}^1 & \cdots & Y_{N_J 1}^1 \\ 1 & Y_{11}^2 & \dots & Y_{N_J 1}^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Y_{11}^{N_a} & \dots & Y_{N_J 1}^{N_a} \end{bmatrix} \begin{bmatrix} \Delta W_0 & \Delta W_0 & \dots & \Delta W_0 \\ \Delta W_1^{12} & \Delta W_1^{22} & \cdots & \Delta W_1^{N_m 2} \\ \vdots & \vdots & \cdots & \vdots \\ \Delta W_{N_J}^{12} & \Delta W_{N_J}^{22} & \dots & \Delta W_{N_J}^{N_m 2} \end{bmatrix} = \begin{bmatrix} \delta_{12}^1 & \delta_{22}^1 & \cdots & \delta_{N_m 2}^1 \\ \delta_{12}^2 & \delta_{22}^2 & \dots & \delta_{N_m 2}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{12}^{N_a} & \delta_{22}^{N_a} & \dots & \delta_{N_m 2}^{N_a} \end{bmatrix} \quad (12)$$

5. Results

All the fundamental steps in the classification of the joint conditions are presented in this paragraph to evaluate the performances of the proposed diagnostic method. The main objective of these simulations is to show that the system is able to identify the fault classes presented in Table 6, so confirming the validity of the presented prognostic approach for the predictive maintenance of high voltage electrical lines. A branch of electrical network containing three junction regions is the first configuration considered for the simulations. In this case, since the conductor length between two consecutive joints is 300 meters, the diagnostic system analyses approximately 900 meters of network. The analytical procedure shown in the paragraph 3.2 is used to obtain the optimal test frequency (f_m) and, starting from this value, also the next three octaves are considered to realize the dataset. This means that four signals are used to measure the line frequency response and, consequently, the network function magnitude and phase are calculated for each test frequency. The equivalent line admittance is used in this work. It corresponds to the ratio between the output current and the input voltage measured at the starting point of the network. From the theoretical point of view, the main steps to obtain the measurements are the equivalent circuit simulation on SapWin IV and the dataset generation on MatLab®. The first step is the equivalent circuit simulation, which allows to obtain the symbolic formula of the line network function. For this reason, a cascade of three elementary sections (Figure 4) is realized on SapWin IV. The joint electrical parameters are kept symbolic, while the conductor components are set to their nominal values (as already said, taking into account the tolerances). Once the equivalent line admittance has been obtained, a specific MatLab® code is used to generate the dataset. The main task of this code is to calculate the analytical formulas of magnitude and phase for each test frequency; subsequently the neural network inputs are calculated for each health state of the line by replacing the symbolic parameters with their numerical values. Since there are 3 junction regions in the first simulation (Ng = 3), 27 different combinations of the health state of the joints are obtained ($N_c = 3^{N_g} = 27$) and, for each of them, 100 random samples (N_{rs} =100) are generated to complete the dataset. This means that each electrical parameter of the joints is randomly selected within one of the intervals shown in Table 6. The choice of 100 random samples is the result of some tests made to obtain a good compromise between the speed of processing and the representativeness of the dataset; it is not an optimized choice but it could, using specific techniques. Finally, the MatLab® script uses the organization of the outputs shown in section 4 to associate the exact health state of the joints with each sample. Table 7 summarizes the situation for the first simulation. Finally, it is necessary to select the characteristics of the neural network, i.e. number of hidden neurons, initial values of the weights, learning rate, backpropagation procedure, computational method used for the weight adjustment. The last two are described in paragraph 4, while the fixed parameters of the neural network are selected through a heuristic approach. In particular, the number of hidden neurons is chosen through some tests to obtain good performances and a good generalization capability. A specific MatLab® application is used to facilitate the selection, because it

Table 7: Dataset parameters for the simulation with three junction regions

Ng	Hidden Neurons	Output Neurons	Initial values of the weights	Training samples	Test samples	Learning rate	Classification rate
3	70	6	5	80%	20%	1	0.84074

allows to obtain a very fast evaluation of the neural network performances by modifying the most important parameters.

The index used to evaluate neural network performances is called classification rate, and it is defined as the ratio between the number of correctly classified samples and the total number of samples used in the test phase. Indeed, only a percentage of the dataset is used for training the neural network, while the other samples are used to verify the correct functioning of the system. Table 8 presents the best neural network configuration for the first simulation and the corresponding value of the classification rate. The classification rate shown in Table 8 represents the global index to evaluate the performance of the diagnostic system and it is obtained by considering the classification errors made on 540 samples. This means that 2160 samples are used to modify the weight values during the training phase, while the remaining 540 samples are used to verify the results during the test phase.

Table 8: Complex neural network configuration for the simulation with three junction regions

Ng	Nrs	Nc	Nt	Dataset Input	Dataset Output	f_m [kHz]
3	100	27	2700	8	6	35

It must be observed that each neuron has a specific classification rate greater than 0.84074 and this means that the performance of the system on a single junction region is much better than the global one. Table 9 shows the results obtained for each neuron and each junction region in the first configuration. Therefore, considering each joint separately, there is a probability range of 91÷95 % that the health state is correctly classified. It is not possible to establish a standard relationship between the classification rate for each pair of neurons and for the corresponding joint, because the number of errors for each junction region depends on the oxidation neuron, the rupture neuron or both.

Table 9: Results obtained for each joint and each neuron in the simulation with three junction regions

Fault class	Neuron	Classification Rate for each neuron	Joint	Classification Rate for each joint
Oxidation	1	0.9741	1	0.9470
Partial Breakage	2	0.9796		
Oxidation	1	0.9685	2	0.9430
Partial Breakage	2	0.9977		
Oxidation	1	0.9444	3	0.9130
Partial Breakage	2	0.9778		

To get a more accurate evaluation of the system performance it is necessary to introduce the cross-validation method. In this case, the previously described training phase is repeated 5 times changing the learning data and the testing data. Therefore, the global classification rate is calculated on 2700 samples and represents the most reliable index. Cross-validation method is presented in Table 10 where the maximum and minimum classification rates represent the best and the worst result obtained at the end of each learning phase.

Table 10: Results obtained through the cross-validation method in the simulation with three junction regions

Fault class	Neuron	Joint	Maximum Classification Rate	Classification Rate
Oxidation	1	1	0.97037	0.9648
Partial Breakage	2		0.98889	0.9793
Oxidation	1	2	0.96481	0.9533
Partial Breakage	2		0.98333	0.9778
Oxidation	1	3	0.95000	0.9411
Partial Breakage	2		0.99259	0.9837
Global Classification Rate			0.8167	

The same simulation procedure is used to evaluate the health state of a line with four junction regions. In this case 6480 samples are used to modify the weight values during the training phase, while the remaining 1620 samples are used to verify the results during the test phase. Table 11 summarizes the situation and shows the classification rate used to choose the neural network set up.

Table 11: Dataset and complex neural network set up for the simulation with four junction regions

Ng	Nrs	Nc	Nt
4	100	81	8100
Hidden Neurons	Output Neurons	Initial values of the weights	Training samples
190	8	5	80%
Dataset Input	Dataset Output	f_m [kHz]	Test samples
8	8	80	20%
Classification rate			
0.70247			

Table 12: Results obtained through the cross-validation method in the simulation with four junction regions

Fault class	Neuron	Joint	Maximum Classification Rate	Classification Rate
Oxidation	1	1	0.97716	0.9700
Partial Breakage	2		0.98889	0.9830
Oxidation	1	2	0.93025	0.9188
Partial Breakage	2		0.98333	0.9757
Oxidation	1	3	0.94383	0.9367
Partial Breakage	2		0.98210	0.9781
Oxidation	1	4	0.91667	0.9046
Partial breakage	2		0.98333	0.9793
Global Classification Rate			0.6899	

Also, in this case the cross-validation method is used to carry out the correct evaluation of the diagnostic system (Table 12). As mentioned above, by repeating the training phase 5 times, it is possible to evaluate all 8100 samples during the test phase. As shown in Table 12, the results obtained for each neuron are still excellent, while the global classification rate is lower than the first configuration. Obviously, increasing the number of the joints reduces the reliability of the complete system, which must be considered as a connection of four classifiers (one for each junction region). Therefore, if the main objective of the diagnostic system was to classify the exact combination of the health state of the joints, the system reliability is about 70%.

On the other hand, the working condition of each junction region can be classified with an accuracy level higher than 90%. This result is obtained considering each pair of output neurons separately (Table 13).

Table 13: Results obtained for each joint in the simulation with four junction regions

Joint	Classification Rate for each joint
1	0.9538
2	0.9006
3	0.9157
4	0.8854

The results presented in Table 13 can be used to compare the performance of the complex neural network with that of the other classifiers. Table 14 shows the comparison with one of the most used classifiers based on quadratic SVM (Support Vector Machine).

Table 14: Comparison between complex neural network and SVM classifier in the simulation with four junction regions

Joint	Classification for each joint with complex neural network	Classification rate for each joint with quadratic SVM
1	0.9538	0.9670
2	0.9006	0.8500
3	0.9157	0.8530
4	0.8854	0.8480

6. Conclusion

A diagnostic system based on a complex neural network has been presented and the procedure to obtain the state classification of the electrical joints has been verified through some different simulations. The results obtained for power lines characterized by three and four junction regions show that the system is able to identify the health state of each joint, with an accuracy level greater than 90%. The actual limit for the diagnostic method is represented by the line length, because the global classification rate obtained for a line with five joints is lower than the previous ones (Tables 15-16). Since the results obtained for the fifth junction region are not good, the system can be only used in a real application for small branches of the network. For this reason, future developments will concern the creation of a macroscopic method of locating failure mechanisms: the section of the non-functioning line could be localized through classification methods based on load flow measurements and after, the system presented in this work could be integrated to identify the junction region in the worst condition. In this way it would be possible to analyze a great variety of networks. Finally, a further development of the prognostic method could concern the analysis of the errors. The division between false positive and false negative errors is currently calculated, but no operation results from this. In the future, the method could be modified to minimize false negatives, which are the most dangerous errors for a diagnostic system

Table 15: Dataset and complex neural network configuration for the simulation with five junction regions

Ng	Nrs	Nc	Nt
5	100	243	24300
Hidden Neurons	Output Neurons	Initial values of the weights	Training samples
280	10	5	80%
Dataset Input	Dataset Output	f_m [kHz]	Test samples
8	10	80	20%
Classification rate			
0.098971			

Table 15: Results obtained for each neuron in the case of five junction regions

Fault class	Neuron	Joint	Specific Classification Rate
Oxidation	1	1	0.9383
Partial Breakage	2		0.9885
Oxidation	1	2	0.7043
Partial Breakage	2		0.7648
Oxidation	1	3	0.7342
Partial Breakage	2		0.8438
Oxidation	1	4	0.7792
Partial breakage	2		0.9023
Oxidation	1	5	0.6780
Partial Breakage	2		0.6926

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Smart Energy Lab, University of Florence and Dept. of Information Engineering, University of Florence.

References

[1] M. Bindi, F. Grasso, A. Luchetta, S. Manetti, M. C. Piccirilli, "Smart Monitoring and Fault Diagnosis of Joints in High Voltage Electrical Transmission Lines," in 2019 6th International Conference on Soft

Computing & Machine Intelligence (ISCMI), Johannesburg, South Africa, 40-44, 2019. doi: 10.1109/ISCMI47871.2019.9004307

[2] Y. D. Liu, G. H. Sheng, Z. M. He, X. Y. Xu, X. C. Jiang, "Method of fault location based on the distributed traveling-wave detection device on overhead transmission line," in 2011 IEEE Power Engineering and Automation Conference, Wuhan, 136-140, 2011. doi: 10.1109/PEAM.2011.6134925

[3] X. Dong, S. Wang, S. Shi, "Research on characteristics of voltage fault traveling waves of transmission line", in 2010 Modern Electric Power Systems, Wroclaw, 1-5, 2010. doi: 10.1109/MEPS16993.2010

[4] L. Xun, L. Shungui, H. Ronghui, A. Jingwen, A. Yunzhu, C. Ping, X. Zhengxiang, "Study on accuracy traveling wave fault location method of overhead line — Cable hybrid line and its influencing factors," in 2017 Chinese Automation Congress (CAC), Jinan, 4593-4597, 2017. doi: 10.1109/CAC.2017.8243590

[5] M. Bindi, F. Grasso, A. Luchetta, S. Manetti, and M.C. Piccirilli "Modeling and diagnosis of joints in high voltage electrical transmission line", in 2019 Proc. of Int. Conference on Energy, Electrical and Power Engineering (CEEPE 2019), IOP Conf. Series: Journal of Physics: Conf. Series 1304 (2019) 012006, 2019. doi:10.1088/1742-6596/1304/1/012006

[6] R.K. Aggarwal, A.T. Johns, J.A.S.B. Jayasinghe, W Su, "An overview of the condition monitoring of overhead lines", Electrical Power Systems Research 53, 15-12, 2000. https://doi.org/10.1016/S0378-7796(99)00037-1

[7] C. Jinbing, W. Guogang, L. Huijun, "Mechanisms analysis and countermeasure of electrical contact overheating of overhead line," in 2014 International Conference on Power System Technology, Chengdu, 1503-1508, 2014. doi: 10.1109/POWERCON.2014.6993786

[8] G. Andersson, "Modelling and Analysis of Electric Power System", PhD Thesis, ETH Zurich, 2008.

[9] M. Davoudi, "Sensitivity Analysis of Power System State Estimation Regarding to Network Parameter Uncertainties", PhD Thesis, Politecnico di Milano, 2012.

[10] A. Cataliotti, V. Cosentino, D. D. Cara, G. Tinè, "Simulation of a power line communication system in medium and low voltage distribution networks" in 2011 IEEE International Workshop on Applied Measurements for Power Systems (AMPS), Aachen, 107-111, 2011. doi: 10.1109/AMPS.2011.6090429

[11] A. Cataliotti, D. Di Cara, R. Fiorelli, G. Tine, "Power-Line Communication in Medium-Voltage System: Simulation Model and Onfield Experimental Tests," IEEE Transactions on Power Delivery, **27**(1), 62-69, 2012. doi: 10.1109/TPWRD.2011.2171009

[12] J. S. Monteiro, W. L. A. Neves, D. Fernandes, B. A. Souza and A. B. Fernandes, "Computation of transmission line parameters for power flow studies," in 2005 Canadian Conference on Electrical and Computer Engineering, Saskatoon, 607-610, 2005. doi: 10.1109/CCECE.2005.1557004

[13] N. A. Jaffrey, S. Hettiwatte, "Corrosion detection in steel reinforced aluminium conductor cables," in 2014 Australasian Universities Power Engineering Conference (AUPEC), Perth, WA, 1-6, 2014. doi: 10.1109/AUPEC.2014.6966630

[14] G. M. Hashmi, M. Lehtonen, "Covered-Conductor Overhead Distribution Line Modeling and Experimental Verification for Determining its Line Characteristics," in 2007 IEEE Power Engineering Society Conference and Exposition in Africa - PowerAfrica, Johannesburg, 1-7, 2007. doi: 10.1109/PESAFR.2007.4498072

[15] F. Steier, "Tribologic analyses of a self-mated aluminium contact used for overhead transmission lines", in 2017 Proc. of 6th Int. Conf. on Fracture Fatigue and Wear, IOP Conf. Series: Journal of Physics: Conf. Series 843 (2017) 012057, 2017. doi :10.1088/1742-6596/843/1/012057

[16] Z. Liu, Y. Sun, "A solder joint crack — Characteristic impedance model based on transmission line theory," in 2014 10th International Conference on Reliability, Maintainability and Safety (ICRMS), Guangzhou, 237-241, 2014. doi: 10.1109/ICRMS.2014.7107178

[17] F. de Paulis, C. Olivieri, A. Orlandi, G. Giannuzzi, F. Bassi, C. Morandini, E. Fiorucci, G. Bucci, "Exploring Remote Monitoring of Degraded Compression and Bolted Joints in HV Power Transmission Lines," IEEE Transactions on Power Delivery, **31**(5), 2179-2187, 2016. doi: 10.1109/TPWRD.2016.2562579

[18] M. H. Azarian, E. Lando, M. Pecht, "An analytical model of the RF impedance change due to solder joint cracking," in 2011 IEEE 15th Workshop on Signal Propagation on Interconnects (SPI), Naples, 89-92, 2011. doi: 10.1109/SPI.2011.5898847

[19] D. Achiriloaiei, S.V. Galatanu and I. Vetres, "Stress and strain determination occurring in contact area of 450/75 ACSR conductor", in 2018 Proc. of 7th Int. Conf. on Advanced Materials and Structures AMS (Timisoara), IOP Conf. Series: Materials Science and Engineering 416 (2018) 012071, 2018. doi:10.1088/1757-899X/416/1/012071

[20] G. Fontana, A. Luchetta, S. Manetti, M. C. Piccirilli, "A Fast Algorithm for Testability Analysis of Large Linear Time-Invariant Networks," IEEE Transactions on Circuits and Systems I: Regular Papers, **64**(6), 1564-1575, 2017. doi: 10.1109/TCSI.2016.2645079

[21] G. Fontana, F. Grasso, A. Luchetta, S. Manetti, M. C. Piccirilli, A. Reatti, "A symbolic program for parameter identifiability analysis in systems modeled via equivalent linear time-invariant electrical circuits, with application to electromagnetic harvesters", Int. J. of Numerical Modelling: Electronic Networks, Devices and Fields, 32, 2019. https://doi.org/10.1002/jnm.2251

[22] G. Fontana, A. Luchetta, S. Manetti and M.C. Piccirilli, "An unconditionally sound algorithm for testability analysis in linear time invariant electrical networks" Int. J. Circ. Theor. Appl., 44, 1308-1340, 2016. doi: 10.1002/cta.2164

[23] N. Sen and R. Saeks, "Fault diagnosis for linear systems via multifrequency measurements," IEEE Transactions on Circuits and Systems, **26**(7), 457-465, 1979. doi: 10.1109/TCS.1979.1084659

[24] F. Grasso, A. Luchetta, S. Manetti, M. C. Piccirilli, "A Method for the Automatic Selection of Test Frequencies in Analog Fault Diagnosis," IEEE Transactions on Instrumentation and Measurement, **56**(6), 2322-2329, 2007. doi: 10.1109/TIM.2007.907947

[25] I. Aizenberg, Complex-Valued Neural Networks with Multi-Valued Neurons. Berlin: Springer-Verlag Publishers, 2011.

[26] I. Aizenberg, A. Luchetta, S. Manetti, "A modified learning algorithm for the multilayer neural network with multi-valued neurons based on the complex QR decomposition", Soft Computing, **16**, 563-575, 2012. doi: 10.1007/s00500-011-0755-7

[27] I. Aizenberg, "MLMVN with Soft Margins Learning", IEEE Transactions on Neural Networks and Learning Systems, 25, 1632-1644, 2014. doi: 10.1109/TNNLS.2014.2301802

[28] E. Aizenberg, I. Aizenberg, "Batch linear least squares-based learning algorithm for MLMVN with soft margins," in 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Orlando, FL, 48-55, 2014. doi: 10.1109/CIDM.2014.7008147

[29] Anjia Mao, Jing Wen and Shasha Luo, "A power system analysis model with consideration of environmental factors," in 2010 5th International Conference on Critical Infrastructure (CRIS), Beijing, 1-5, 2010. doi: 10.1109/CRIS.2010.5617557

[30] J. R. Santos, A. G. Exposito and F. P. Sanchez, "Assessment of conductor thermal models for grid studies," IET Generation, Transmission & Distribution, **1**(1), 155-161, 2007. doi: 10.1049/iet-gtd:20050472

[31] F. Grasso, A. Luchetta, S. Manetti, M. C. Piccirilli and A. Reatti, "SapWin 4.0 – a new simulation program for electrical engineering education using symbolic analysis" Computer Application to Engineering Education, 24, 44–57, 2016. doi: 10.1002/cae.21671

[32] www.sapwin.info Dept. of Information Engineering, University of Florence, Italy.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 2, 494-499 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design of Interactive Aids for Children's Teeth Cleaning Habits

Cheng Chuko, Fang-Lin Chao*, Hsin-Yu Tsai

Department of Industrial Design, Chaoyang University of Technology, Taichung, 436, Taiwan, R.O.C.

ARTICLE INFO

Article history:
Received: 09 February, 2020
Accepted: 25 March, 2020
Online: 04 April, 2020

Keywords:
Product design
Arduino
Teeth cleaning
Interactive aids

ABSTRACT

Dental plaque is considered a possible causative agent of major dental diseases. People must develop oral care skills at an early age with family support. This study aims to assist parents and to reduce children's fear while cleaning their teeth. An interactive game challenges children's brushing ability. Arduino modules and software platform were utilized to build a functional prototype. The status of the palm movements identified using a sensing module within brusher. When the cumulative number of swipes exceeds a specified target, the LED moves to the next position. Children became more attentive and adjusted to their corresponding tooth positions according to changes in the hippo's dental light. The average brushing time increased to 226 seconds, with the assistance of the partner. An interactive novelty toothbrush with Bluetooth module was built and evaluated. The contribution of the work is observing children's needs and guide them to brush their teeth from an intimate perspective.

1. Introduction

The most common dental problem of concern in children is dental caries. Dental plaque is considered a possible causative agent of major dental diseases. Oral health affects chewing, nutrition, speaking, socializing. As a result, people must develop oral care skills at an early age. The implementation of oral disease prevention includes education and the integration of family support. This design study aims to assist parents and to reduce children's fear while cleaning their teeth.

1.1. Oral Health Education

Advocacy and family coordination are vital for implementing oral health education. Children are often unfamiliar with initial skills. A survey of Taiwan's 3-6 year-olds showed that the average permanent dental caries experience index (DMFT) of school children was 3.31. This index indicates the need for promoting children's oral health programs. Taiwanese children prefer sugary drinks, which have a high correlation with severe dental caries. The frequency of sugary drinks for children in Taiwan is 4-6 times a week. Therefore, it is necessary to implement oral health education and implement correct dental cleaning habits from early childhood.

Good oral hygiene is dependent upon the effectiveness and the ease with which the procedure s carried out. The most common preventive dental program is the modified Bass approach [1] (Figure 1). Domestic academics and American public health scholars also recommend the Bass method for its high efficiency and "horizontal short-distance movements." Its main points are:

(1) Place the toothbrush at a 45-degree angle on the gum line.

(2) Brush using a gentle circular motion: downward for the upper teeth and upward for the lower teeth, and brush 2-3 teeth at a time.

(3) Use the same procedure for the lower teeth, except rotating in an upward direction toward the medial teeth.

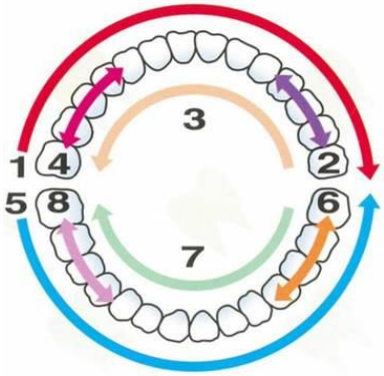


Figure 1. Brushing sequence of the Bass method.

A reduction of interdental plaque and gingival inflammation requires a suitable brushing technique. Poyato-Ferrera et al. [2] found that the modified Bass method is superior for supragingival

* Fang-Lin Chao, Email: flin@cyut.edu.tw

plaque removal. Studies also show that the modified Bass technique was useful for children [3,4]. The Scrub and Bass methods (with flossing) are effective in eliminating or reducing plaque and gingivitis in 11-year-old children using a junior toothbrush.

1.2. Analysis of Existing Products

Extant products seek to attract young children with appealing anthropomorphic images. It is difficult for purely informational products that lack substantial interaction to keep young children engaged for a long time (Figure 2). These textbooks are like a toy and cannot be interacted with in person; so it is not easy to raise the expectation of continuous exploration.



Figure 2. Interactive dental cleaning products (a) gamebook, (b) music-induced voice cleaning.

Philips Sonicare marketed by Oral Healthcare (Figure 3) is a companion product for teeth cleaning that is currently on the market. Power toothbrushes remove plaque more effectively by brush head motion. For example, Sonicare claims to remove plaque as a result of dynamic fluid activity [5]. Studies show that the use of a Sonicare Airfloss results in less removal of tooth enamel when compared to the use of a standard toothbrush [6]. A reduction in plaque in the posterior and hard-to-reach areas indicated the benefit of the design [7].



Figure 3. Sonicare children's toothbrush.

Sonicare features include timed tooth cleaning, an app music reminder to change the part of the mouth being brushed every 30 seconds, and a link with an interactive game app. Its advantage is flexible content; however, the user must be familiar with smartphone operation. Another study [8] identified several other oral health-focused apps in the Apple App Store and Google Play. Apps identified using the following search phrases: oral health, dental health, teeth health, tooth health, dental care, and oral care. However, the quality of the reviewed apps was generally poor [9]. Apps need theoretically grounded content and adherence to ethical design principles for persuasive health technologies. The Oral-B and Sonicare apps can connect with a phone via Bluetooth. However, those apps performed poorly on standard usability heuristics [10].

2. Survey and Analysis

First, we surveyed the toothbrushing habits of children to find suitable entry points for product design. Items included current brushing status, the attitudes of children and parents toward tooth brushing, and methods of parental assistance in building dental cleaning habits.

2.1. Method

A total of 204 families with schoolchildren in the age range of 3-6 years old were selected in a randomized controlled family survey. The inclusion criteria used to select the samples for the study were as follows: (i) healthy children between the age group; (ii) schoolchildren who live with their parents. The survey distributed through both paper and website.

2.2. Results and Discussion

According to the survey data, more than half of the families were double-income families, and only 40% of the carers were full-time caregivers. At present, more than 90% of the parents demonstrated an awareness of oral care. However, more than 70% of children needed to be accompanied by their parents when brushing their teeth. More than 70% of families reported purchasing education material or products for oral health care. 3.5 % families reported their child had tooth decay, and less than 20% of children demonstrated knowledge of preventative oral care.

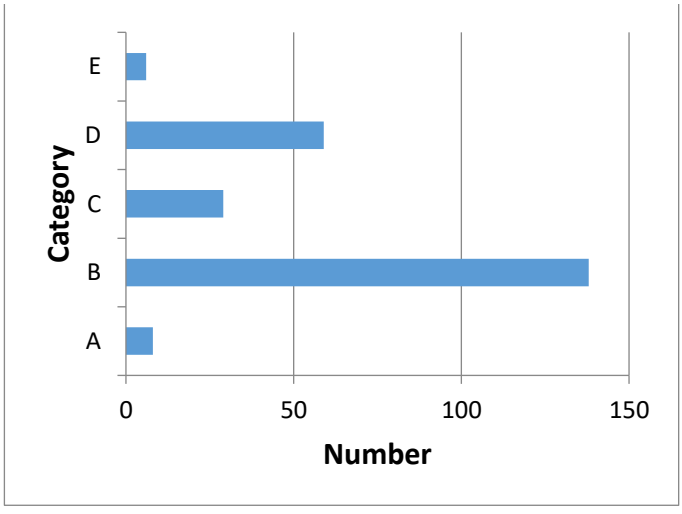


Figure 4. Children's brushing habits.

The Figure 4 shows the level of brushing reported by the families, as follows:

A: Child has not yet begun to brush their teeth.
B: The parent brushes child's teeth.
C: The child is self-motivated to brush teeth.
D: Parent motivates a child to brush teeth.
E: The child brushes teeth with parental assistance.

Items B, D, and E involved parental participation, so the children could not take the initiative to complete these tasks themselves. Therefore, an educational product that plays the role of a parent could hypothetically reduce the burden on parents. If the product is impressive, the design could improve children's motivation to brush their teeth and improve the effectiveness of tooth cleaning.

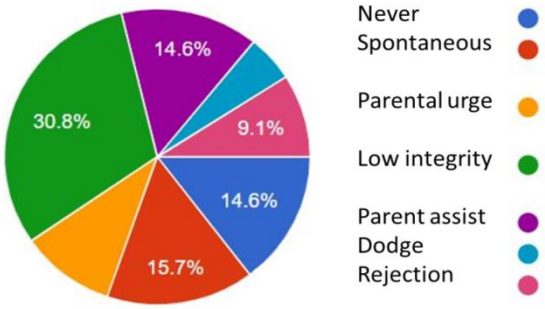


Figure 5. Children's behavior and attitude towards teeth cleaning

As shown in Figure 5, less than 20% of children can brush their teeth independently, and the rejection rate is quite high, at 15%. Low quality meant that although the child brushed their teeth, they did not do so carefully, so they remained at risk for tooth decay. The survey results revealed the Bass method needs more promotion to families.

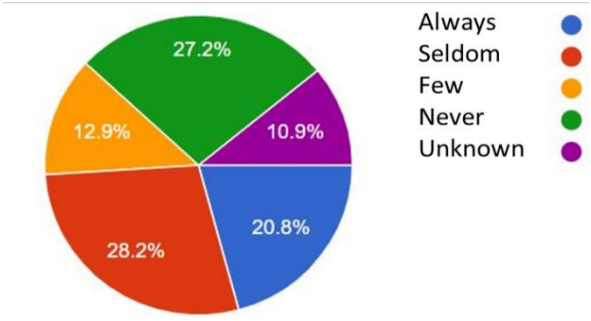


Figure 6. How often do children brush their teeth with the Bass method?

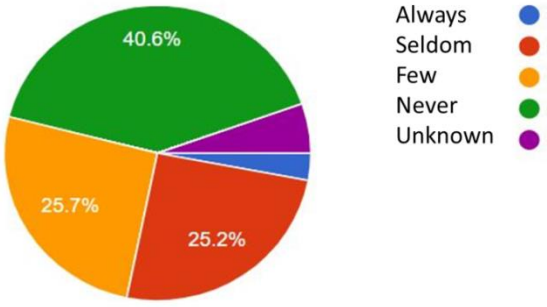


Figure 7. Completeness of Golden Teeth 333 code

As shown in Figure 6, parents report that only 21% of children can brush their teeth with the Bass method. Most children or parents are not familiar with the main points of the Bass method. The data also shows that there is a need to increase through auxiliary tools.

The "333 Code" refers a better way to brushing teeth: at least three times a day, brushing one's teeth within 3 minutes after a meal and brushing one's teeth for more than 3 minutes each time. Only less than 10% of people reported performing it correctly.

The Association of Dental Surgeons points out that the rate of tooth decay in children is high, with an average of five cases of tooth decay per child. The Department of Health appeals to all age groups, from infants, adolescents, and the elderly for independent health management.

The market survey found that most parents had an awareness of establishing healthy habits (Figure 7): 79.2% reported wanting to comply as much as possible, but also reported difficulties with implementation. Researchers hoped that assisted design could develop independent healthcare habits in children from an early age.

3. Design Approaches

Through the questionnaire survey, we understand that the current situation of children's learning is significantly different from our expectations. Through various design concepts, we proposed to increase children's attempts with interactivity, attract them to try and reduce the burden on parents. Through the sharing mechanism, excellent performance can be linked with parents' rewards to form a positive learning cycle.

3.1. Problem Definition

The target user age for this study was 3- to 6-year-old children. The goal was to remove children's fear of cleaning teeth and implement tooth cleaning habits without parental assistance. It is easier to develop a habit if the process is enjoyable. A suitable game design challenges children depending on their age and ability. Suitable game types for children include distinguishing color, number, and order, recognition, and memory. At the age of five, children have basic concepts of mathematics, association, and space. Simple vocabulary games can reduce the fear of tooth brushing in children and assist parents in encouraging children's oral health.

A bionic "dentist partner" draws children's gazes and reduces the rejection of oral cleaning appliances. Interaction and encouragement between parents and children allow children to develop habits. A visually appealing form supplemented by smart monitoring functions facilitates observation of cleaning status. A "dentist partner" game uses an LED screen to ensure that children are excited to complete the cleaning process using the Bass method.

It is not easy to cultivate oral cleaning habits at an early stage, and some children are repulsed by stimulation. Toy toothbrushes and electric toothbrushes often appeal to children with cute shapes, but parents' supervision is required to urge children to clean their teeth properly. Adding childlike connotations and introducing stories makes cleaning teeth more fun for children.

3.2. Arduino and Control Circuit

Arduino is a typical software tool for building a variety of devices with available peripherals, tools, and techniques. The

Arduino ecosystem includes LED, Servo motors and stepper motors, wireless connectivity, and the Internet of Things. The technical forum provides more prominently featured in the advanced projects. We utilized a 4-digit 7-segment display [12] with a common cathode and a digit display that turns on when the digit pin state is LOW. The lights are lit one by one in the order of the brush. The user needs to stay in this position until the light index changes. The status of the palm movements identified using IMU inertial sensor to read the accumulated number of brush. When the cumulative number of swipes exceeds a specified target, the light moved to the next position to lead children's interaction (Figure 8).

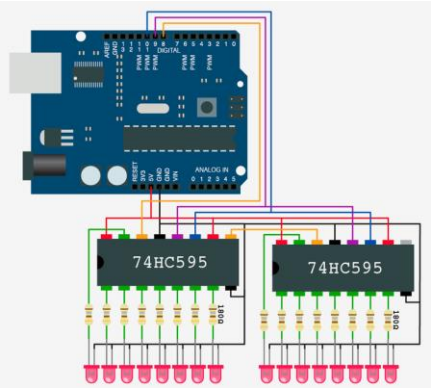


Figure 8. Hardware circuit. The device s comprised of an Arduino Nano microcontroller, MPR 121 Multiplexer, MPU 6050 IMU, and a Bluetooth [11]

3.3. Product Design

The initial concept of oral healthcare aid inspired by a whack-a-mole game. Instead of striking a mole, the idea of the game was transformed into brushing specific teeth. The device design based on an anthropomorphic hippopotamus with an open mouth and exposed teeth. A capsule machine inside the device rewarded children with a surprise gift following tooth brushing (Figure 9). Children were trained in good tooth brushing habits through daily cleaning, with additional help from a parent app.



Figure 9. Final concept sketch

The model built by vacuum forming with a polyurethane mold. The buckled structure utilized to take into account the size of the body and the internal configuration (Figure 10). The appearance of the model set to a medium-sized doll, which intended to stimulate children's alertness and curiosity.

The chassis contained the dental cleaning indicator and timing function and could transmit the cleaning record, time spent brushing, and conditions to a parent's mobile phone. The hippo-shaped teeth had indicator lights to indicate a brushing location. An Arduino program controlled the LED. Each tooth lit up sequentially according to Bass method and sensed whether there was brushing at the corresponding position. The toothbrush handle equipped with an electronic induction device, a Bluetooth module, and a battery, and the indicator light indicates the position of the brush; the inertial sensor [13] on the toothbrush handle detects whether the toothbrush is shaking to confirm that the child is brushing correctly. The counter and timer of the system can accumulate the number and time to identify the brushing situation.



Figure 10. Assembly drawing of main accessories and appearance.

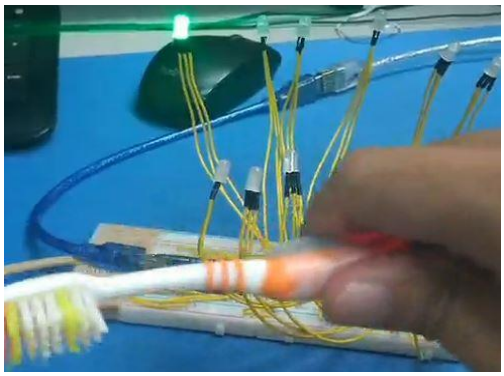
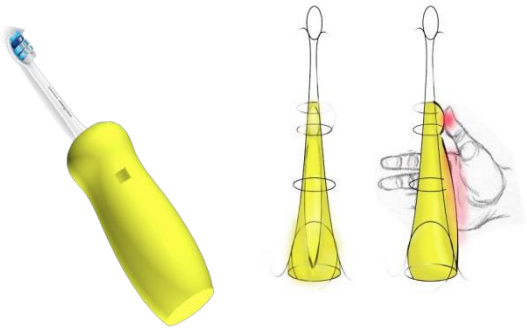


Figure 11. Sensors and LED indication configuration and the brush with an inertial sensor that wirelessly connected with the main circuit board.

Specifications of internal parts were as follows: Arduino Nano, HC-06 Bluetooth module, 74594 SIPO shift register, R.T.C. clock, 5mm LED*10, and Servo motor (Figure 11). The tooth faces indicated with different colors: green for the external surface, blue for occlusal surface, and red for the internal surface.



(a) (b)

Figure 12. The rendering of toothbrush and (a) actual installation appearance, and (b) grip shape of a possible design.



Figure 13. Operation flow and parent APP interface screen diagram.

The module comprised of a 3-watt speaker and a 3.7V lithium batteries. A brush circuit board contained the Power Module (Adio-5VCH + ADIO-USB-0506 Boost Module), Arduino IMU inertial sensor, and MPU-6050. An indicator interface placed inside each tooth, which lit up to indicate cleaning progress. The current installation appearance, and future possible grip shape are display in Figure 12.

3.4. User Interface. Design

The product placed in homes or classrooms. The intended users were children who had just begun to clean their teeth. Parents or teachers could use the game machine to demonstrate brushing to children.

Children have their feelings and tastes; designers of new technologies need to remind that children are an entirely different user. Researchers suggested "children had more direct involvement with technology developers" for technology development and invited children to test before it released [14]. The interactive experience is the basis for the toy design. Lacking researches on children's interactive experience causes mismatching of children's needs [15]. Popular toys are a dominant entertainment medium. Traditional toys have the function of supporting children's growth [16]. The user interface of a child-friendly product is one of the elements of a successful design.

The Figure 13 shows the operation process and parent's APP interface. This app can record the score and progressing of cleaning events, gift sticker reward, and parent responses. A monthly calendar is helpful to view children's oral cleaning status and progress. Designers are required to avoid using symbolic images that confuse children. The contrast formed by different color blocks is natural to arouse the user's visual interest. Deleting extra pages simplifies the process and only shows what is needed to convey important messages—the interface mainly based on natural and pleasant feelings, and brighter colors matched with cute fonts.



Figure 14. Product usage test: after the sensor has been swiped for a set number of times, the indicators will jump in the sequence. The reward gift is placed in the mouth of the villain on the lower right.

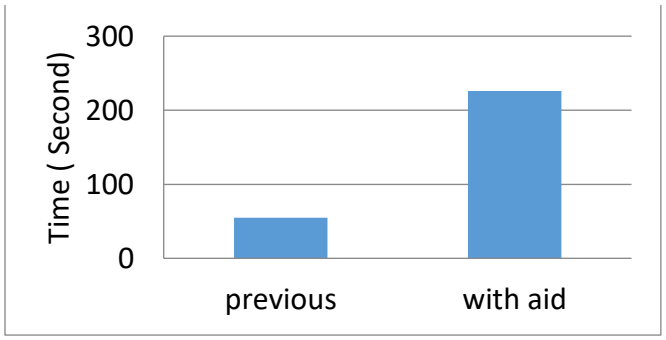


Figure 15. Product usage test: the average brushing time

3.5. User's Test

The product usage test carried on in both family house and kindergarten classrooms (Figure 14). The researcher invited five children to complete the brushing steps according to the light of assisting partners and record children's responses and the total time of brushing. From the experiment, we observed that children

became more attentive and adjusted to their corresponding tooth positions according to changes in the hippo's dental light. Figure 15 is the average brushing time result; the autonomous brushing is often completed quickly (55 seconds) with the disadvantage of less fulfilling the requirements. But the average time for brushing increased to 226 seconds with the assistance of the hippo partner. The teacher's required value is 180 seconds (3 min.) [2], and the average acting time is higher. During the boot test, the child must wait for the light to change position (time delay), so it will take more time to follow at the beginning. The classroom setting also takes advantage of the children's inclination to increase enthusiasm within a peer-collective situation [3].

Children were happy to follow the LED display of the hippo and complete the steps of brushing their teeth through interaction. When parents need to accompany them, the child will also be stressed; parents' time is limited and it is not easy to maintain good mood. Through the records left by the interaction with the companion, parents can reward and respond to the child after watching it. This approach reduces tension and wasted time on both sides.

However, the mechanism of this model encountered some configuration problems. The brush's rod is insufficiently supported, so when the child brushes his teeth, the head is biased; water may flow along the toothbrush rod into the circuit area in the handle. The added electronic induction circuit does increase the issue of safety and must be waterproof and damp-proof before entering the application. At present, the size of the product is large, and the toothbrush grip also weighs a significant amount. The component and device size can reduce in the future with the help of electrical and mechanical engineers in the cross-domain design.

4. Conclusions

Throughout the investigation, the researchers identified the current educational need for dental care. An interactive novelty toothbrush with Bluetooth module was built and evaluated. Building cross-domain design teams allowed conceptual revision and determination in the early stages of design. Through cooperation with electrical engineers and circuit implementation of Arduino, design elements were combined. The children connected with parents through the app interface. This design can reduce the burden of parental supervision and contribute to the positive cycle of child interaction. At present, the cost of firmware hardware is high, so it is more suitable for teaching aids. In the future, the miniaturization of electronic parts will make the product more suitable for children.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors acknowledge the electrical engineers in cross-domain design and programming.

References

[1] https://www.buysplat.com/pages/bass-brushing-technique

[2] M. Poyato-Ferrera, J. J. Segura-Egea, and P. Bullón-Fernández. "Comparison of modified Bass technique with normal toothbrushing practices for efficacy in supragingival plaque removal" International Journal of Dental Hygiene, 1(2), 110-114, 2003. https://doi.org/10.1034/j.1601-5037.2003.00018.x

[3] E. Robinson, "A comparative evaluation of the Scrub and Bass Methods of toothbrushing with flossing as an adjunct (in fifth and sixth graders)." American Journal of Public Health, 66(11), 1078-1081, 1976. https://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.66.11.1078

[4] S.P Patil, P.B. Patil, and M.V. Kashetty. "Effectiveness of different tooth brushing techniques on the removal of dental plaque in 6–8-year-old children of Gulbarga." Journal of International Society of Preventive & Community Dentistry, 4(2), 113, 2014. https://doi: 10.4103/2231-0762.138305

[5] P.R. Warren, M.A. Cugini, B.V. Chater, and J. Strate, "A review of the clinical efficacy of the Oral-B oscillating/rotating power toothbrush and the Philips Sonicare toothbrush in normal subject populations" International Dental Journal, 54(6), 429-437, 2004. https://doi.org/10.1111/j.1875-595X.2004.tb00300.x

[6] T. J. Holley, J. A. Ross, and T. L. Hottel, "Efficacy of Philips Sonicare AirFloss compared to manual brushing and flossing: an in vitro study" The Journal of the Tennessee Dental Association 94(1), 34-36, 2014. https://europepmc.org/article/med/25241500

[7] G.I. McCracken, P.M. Preshaw, L. Heasman, "Efficacy of plaque removal of the Sonicare Elite versus the Sonicare Advance from hard-to-reach sites" Journal of Clinical Periodontology. Nov;31(11):1007-1011, 2004. DOI: 10.1111/j.1600-051x.2004.00604.x.

[8] B. Tiffany, P. Blasi, S.L. Catz, and J.B. McClure, "Mobile apps for oral health promotion: Content review and heuristic usability analysis" JMIR mHealth and uHealth, 6(9), p.e11432., 2018.

[9] P. Nayak, S. Nayak, M. Vikneshan, S. Acharya, and D. Sathiyabalan, "Smartphone apps: A state-of-the-art approach for oral health education" Journal of Oral Research, 8(5), 386-393, 2020. https://doi.org/10.17126/10.17126/joralres.

[10] F. Zotti, D. Dalessandri, S. Salgarello, M. Piancino, S. Bonetti, L. Visconti, C. Paganelli, "Usefulness of an App in improving oral hygiene compliance in adolescent orthodontic patients" The Angle Orthodontist, (1), 101-107, 2016. https://doi.org/10.2319/010915-19.1

[11] J. Blum, Exploring Arduino: Tools and techniques for engineering wizardry, John Wiley & Sons, 2019.

[12] N. Cameron, 4-Digit 7-Segment Display, Arduino Applied. Apress, Berkeley, CA, 119-136, 2019.

[13] A. Verma, G. Culbertson, and K. Ramani, "Chiron: interpreting signals from capacitive patterns and inertial sensors for intuitive shape modeling" CHI'14 Extended Abstracts on Human Factors in Computing Systems, 1831-1836, 2014. https://doi.org/10.1145/2559206.2581161

[14] A. Druin, "The role of children in the design of new technology." Behaviour and information technology, 21(1), 1-25, 2002.

[15] X. Lv & R. Peng, "User experience research of toy design" In IEEE 11th International Conference on Computer-Aided Industrial Design and Conceptual Design, CAID and CD'2010, 2, 943--946. DOI: 10.1109/CAIDCD.2010.5681811

[16] A.A. Naufal, and A. Suzianti, "Design thinking approach for product innovation design of educational toys"In Proceedings of the 2019 5th International Conference on E-business and Mobile Commerce, 69-74, 2019. https://doi.org/10.1145/3332324.3332337

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 497-506 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

An Economic Theory Perspective for the Fight Against Poverty in the Peruvian Andes

Robert Antonio Romero-Flores*

Federico Villareal National University, Graduate University School, Lima, 15001, Peru

National University of the Altiplano, Department of Systems Engineering, Puno, 21001, Peru

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 05 November, 2020
Online: 20 November, 2020

Keywords:
Complex systems
Mathematical modelling
Economic theory of water
Systems simulation

ABSTRACT

The fight against poverty in the Peruvian Andes is a complex task in which various professionals, such as engineers, economists, anthropologists, among others, participate. The uncertainty of the decisions taken today, no matter how appropriate they may seem, such as million-dollar investments in irrigation infrastructure, can result in over-production and, therefore, in economic recessions. For this reason, a new mathematical simulation model is proposed using system dynamics to predict recession phenomena that can occur in months or after a few years of auspicious economic growth, and that can cause sales prices to be below production costs. The author has developed the conceptualization of the production system of irrigation improvement projects in several years of multidisciplinary work in the Cusco region of Peru. The primary objective of irrigation projects is to improve the socio-economic conditions of the farmer. Techniques as the fulfillment of goals have been used to quantify qualitative dimensions such as strengthening organizations and trainings that are key to guaranteeing irrigation improvement projects' sustainability in the long term. Therefore, it has been possible to identify the variables and relationships of this type of socio-economic system. To validate the model, we verified that the simulated data are consistent with the historical data collected. Likewise, if the values of the various proposed models' variables are adequately modified, these can be applied to other types of production systems under different market conditions. The dimensions addressed, such as supply, demand, sale price, land, production volume, public budget, etc., enhance the research's importance, making the simulation model formally expressed also acquire nuances from economic theory for the fight against poverty-based on water. One of the study's conclusions is to understand the production systems, it is necessary to see them in the context of their regional economy's behavior.

1. Introduction

One of the most important policies in the fight against poverty of the Peruvian government is implementing irrigation improvement projects. In this paper, we will address issues in greater depth about the experience of Plan Meriss Inka (PEPMI) to deal with the social component. We will highlight the characteristics that make this research an economic theory of water perspective that could not be exhibited in work originally submitted in the 2019 IEEE World Engineering Education Conference (EDUNINE) [1].

The primary nature of the irrigation improvement projects implemented by PEPMI is multidisciplinary, in which professionals of different natures such as engineers, economists, biologists, anthropologists, among others. One of the characteristics that stand out from this type of project is the high social component [2]. Since the first projects' implementation, it was observed that, despite providing rural communities with modern irrigation infrastructure. Farmers were reluctant to use it. A situation that has required social partners to decipher this peculiar impasse. Anthropologists' role in these cases has been essential to understanding the farmer's idiosyncrasy in terms of the intervention of new irrigation projects in their communities it refers.

From work carried out, it was concluded that the resistance to using the new irrigation infrastructure was because the farmer only uses the irrigation infrastructure that they have built and, therefore, they consider it as their own. So, not having participated in constructing this new irrigation infrastructure donated by the government, they did not consider it their own. To overcome this

*Corresponding Author: R.A. Romero-Flores, Urb. Chanu Chanu III S10, +51951885805, rromero@unap.edu.pe, 2008345041@unfv.edu.pe

problem, Community participation is incorporated as part of the project's financing, which consists mainly of the farmers' workforce's valuation in the construction of the irrigation infrastructure [3]. Similarly, there are cases of projects in which it has been possible to significantly improve water resources available without the need to build new irrigation infrastructure; it was only required to improve water management and irrigation techniques to increase agricultural productivity. This experience leaves an important lesson for professionals who are not from the social area, which is that, to achieve the objectives of the projects, they should improve their communication skills and understanding of the farmer idiosyncrasies in the area [3]. The social component is just one of the many problems. It gives an idea of the complexity of understanding and managing irrigation improvement projects in the fight against poverty in the Peruvian Andes. The ability and experience to face these peculiarities have made the PEPMI to be considered as a model project in South American by the German donors German Technical Cooperation (GTZ) and the State Development Bank of the Federal Republic of Germany (KFW).

In this understanding, the primary objective to be achieved with social intervention projects is their sustainability in strengthening the organization of farmers and profitability. In other words, farmers must manage their own resources without the accompaniment of PEPMI staff and only through its own organization and self-financed by the profit that results from the sale of their products.

In the Peruvian Andes, there is the occurrence of weather phenomena such as droughts and frosts that cause the loss of crops. The main effect of having adequate irrigation infrastructure is to ensure cultivation; firstly, supplying water in case of droughts and making it possible to recover plants from frosts. So that at least one harvest per year can be ensured. This is also known as crop safety. It is possible to increase the number of agricultural seasons per year by up to two in the best cases. Then, water administration is a strategy ancestrally used since the Incas to face unpredictable climatic changes and to guarantee food for the population [4].

To overcome these problems, millions are invested in infrastructure and training. But in the projects implemented in the Chumbivilcas-Espinar provinces, it has been shown that this does not always happen. Due to the overproduction of dairy products, there was a regional recession. A phenomenon in which supply exceeds demand and produces a drop in prices. In this case, prices fell even below production costs. So, the right short-term decisions can lead to long-term problems [5].

In this research, we are concerned with developing a production model to understand the factors that participate in it. The model simulation process has allowed the model to be validated. However, to know if the production process is profitable. It is required to evaluate it in the environment of its regional economy.

Therefore, it is also necessary to conceptualize and formulate economic factors. This gives greater importance to our research as it analyzes the Andes' economy from a perspective of fighting poverty. The methodology used to achieve our objectives is mainly system dynamics, as it includes the methods of thinking of systems and servomechanisms. The same ones that are suitable for complex systems.

As a result of the research, we have obtained the formulation of the different production sub-systems such as training, investment, available water, infrastructure, and others. And we have also managed to formulate the subsystems of the regional economy such as supply, demand, sale price, competition, etc. This makes it possible to project over time the impact of the irrigation improvement projects' anti-poverty policies. It also allows us to carry out a sensitivity analysis to know if the decisions we make in implementing the project are the correct ones, mainly to avoid regional recessions or falls in products sales prices [1].

2. Related Works

2.1. System dynamics

System dynamics is a popular simulation methodology developed by J. Forrester at the Massachusetts Institute of Technology (MIT). This simulation methodology is based on the theory of servomechanisms and feedback. It is also related to areas such as general systems theory and cybernetics. It also uses methods to study complex systems that act as an interlocutor between engineering methods and the methods for social systems [6]. One of the most important applications of System Dynamics is the world model published in 1970. A simulation of the behavior of society is shown under current "unplanned" growth conditions. The results of the simulation of the model show an over-exploitation of resources and a drop in population. The more population, the more waste is generated, and the more waste, the more diseases are generated is one of the conclusions of the model [7].

In the work of Fifth Discipline, considered to impact the business world greatly, revalues Forrester's systems dynamics. The author in [5] considers important for the understanding of modern organizations the study of dynamic systems, identifying the system's environment, and feedback mechanisms with it. He does not consider essential to arrive at the formulation of mathematical models. As a conclusion to his studies, he proposes the following disciplines: systems thinking, personal mastery, mental models, the construction of a shared vision, and team learning.

For irrigation research, dynamics can also be applied to the study of water management jobs, as demonstrated in the proposal for an optimization model for irrigation management in Australia [8]. In [9], the authors also developed a simulation model for China's water transport problem. In both works, positive results have been obtained using system dynamics.

2.2. Systems thinking

Systems thinking is the fifth discipline; the importance of systems thinking lies in being the body that unites the other disciplines of intelligent organizations. This job is incisive in pointing out that one of the factors for the failure of projects is the lack of systems thinking in their implementation [5]. Systems thinking is based mainly on the method of extension or synthesis. This becomes a premise for troubleshooting, where it is recommended to first view the system as part of a bigger system [10]. In this regard, we must comment that the classical scientific method, as we know, uses the analytical or reductionist method,

which, contrary to the synthetic method, what it does is studying the system in its components. The author in [11] categorically considers as a limitation of the scientific method. However, one of the most renowned authors on epistemology comments that systems thinking is an unfounded fashion issue and cannot generate scientific knowledge [12]. The author in [10] refuted this position, who maintains that the analytical and synthetic methods do not replace each other. On the contrary, they complement each other. In the present work, we show that it is possible to generate knowledge in the scientific method's formal language, between the system and its environment (exogenous variables) and the feedback mechanisms for adjusting the endogenous variables.

2.3. Fight against poverty

Thus, water is the engine for the economy of the Andes, although its cost is relatively low given its usefulness. To highlight the importance of water, we will mention the water-diamond paradox that tries to explain the low cost of water concerning diamonds that cannot generate life. This explains that, although the marginal utility of diamonds is much greater than the marginal utility of water, water's total utility is always greater than the total utility of diamonds [13].

To eradicate poverty, the author in [14] comments that the financial system has been saved on a global scale; for example, in Mexico, the banking system has been saved. Consequently, the cost of helping the poor is much lower and, in financial terms, more profitable if they are to be part of the aggregate demand. And, despite extreme poverty, we seem to have plenty of resources. At the same time, another study concludes that almost a billion people go to sleep hungry every day. In comparison, another billion people suffer from obesity, and 30% of food production is wasted, even from its mismanagement in harvest, sales, and post-consumption [15].

3. Methodology

The experience that served as inspiration for this work was that of the irrigation improvement projects in the Chumbivilcas-Espinar provinces that, due to their height (over 3820 m.a.s.l.) and geography, these are areas where low temperatures predominate (even below zero), all this leads to the existence of extreme poverty in the area [16]. Conditions that, in the first instance, made interventions in the so-called “high” provinces impractical. However, initially, they gave unexpected positive results thanks to the experience of the PEPMI and the adaptation of strategies that consisted, mainly, in cultivating imported pastures that resisted low temperatures and that were the main engine of the livestock industry and its derivatives in the area. Which brought the economic growth of the area. Despite this, after a few years, there was overproduction that manifested itself in the drastic decrease in sales prices. Therefore, it is necessary to manage uncertainty, the consequence of the decisions we make, and that, at first, seem the correct ones in the medium and long term, can become counterproductive. The explanation for this phenomenon was that an adequate market study was not carried out. Thus, to understand the interaction of the multiple factors, it was necessary to understand the behavior of the market and the mechanisms by which the sale price was mainly obtained. In an economic theory, appears the next components: consumers, producers, the market for goods and factors, and government intervention [13]. To formulate the simulation model of the production system, all these dimensions have been taken into account. Therefore, in this document, we will extend the original work to a perspective of economic theory whose main engine is water. In Figures 1 and 2, we can see the Sutunta lagoon and the dam built to store 4 million cubic meters to irrigate 6000 hectares.

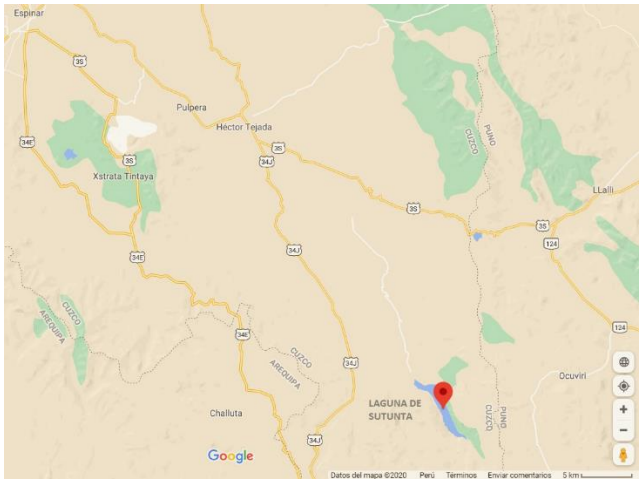


Figure 1: Location map of the Sutunta lagoon in the province of Espinar-Cusco over 4000 m.a.s.l.



Figure 2: The dam built in Sutunta lagoon

In the conceptualization and formulation of the new production model, the systems dynamics methodology was used. The same one is based on the general theory of systems proposed in [11], the author that deals mainly to provide a general framework for science. To do this, it recognizes the concept of equilibrium or homeostatic point. The same that we must reach in the interaction of supply and demand in the market model also proposed [1]. Nash’s equilibrium abstracts the theory of non-cooperative games that involve sellers' participation, sets of strategies, and profits [17].

In the conceptualization phase, approximately 285 variables have been identified. The same ones have been organized in the dimensions of investment, training, organization, water for irrigation, irrigation infrastructure, production, market, environment, public budget, and water supply. As mentioned

above, the organizational dimension is primarily qualitative, and techniques have been used to allow it to be considered and provide qualitative information. This dimension has been adequate to be considered in the proposed production model. Another characteristic is that the production model allows obtaining a production volume that is considered supply and considering the existence of competition. Therefore, it is not an empirical model for educational purposes [6]; otherwise, it can provide information in a real environment. The number of variables and the peculiarities described makes it different from the authors' production models in [18], who propose a model of positive mathematical programming for models of regional production in agricultural-environmental programs and the classic spider web model [13].

The simulation of the production model has been carried out in the VENSIM program; the simulation results have served to validate the model. For validation, the Anderson-Darling normality test was used, whose results have been superior to $\alpha = 0.05$. As a result of the optimization model, non-linear programming with restrictions has been used to develop an optimization model that allows us to find the global optimum (profit maximization, cost minimization, and equilibrium point). Likewise, an analysis was carried out on the optimization model using nonlinear programming and genetic algorithms to determine the selection operators' efficiency in restricted non-linear problems.

4. Discussion

4.1. Conceptualization of the elements of the economic system for the fight against poverty based on water

In South America, the Inca culture developed. One of the most important cultures in the world whose cultural richness was based on values such as community work, also known as the "ayni" and the "minka". The Incas built large hydraulic works with the primary purpose of supplying food to the entire population. The importance of institutions like the PEPMI lies in need to revalue the cultural richness that, in some way, was being forgotten. In the field of intervention, they can be observed as problematic situations: inadequate irrigation infrastructure (inland canals that filter water), little farmer organization, and little knowledge of irrigation techniques [4]. This situation can even be complex if the following characteristics exist: conflicts, undefined access rights to water, water scarcity, and geological problems [3].

To solve these problematic situations, PEPMI implements irrigation improvement projects whose objective is to provide the population in the project area with the possibility of improving production and, therefore, improve their quality of life through access to more and better education, health, etc. services.

4.1.1. Production systems modeling in irrigation improvement projects

Consistent with the fifth discipline [5], a work in which the author concludes that one of the causes of project failure is the lack of a systems approach. If we review the experience of the projects implemented in Chumbivilcas-Espinar, in which there were problems of regional recession, after a few years of economic growth, it can be seen that, as part of the bigger system, the "market" is an exogenous component that is always present in the environment of the system. So, it is necessary to carry out an adequate market study and, for simulation purposes, to determine its feedback mechanisms. In figure 1, it can be seen through a block diagram, the main identified subsystems. Given the complexity of these projects' implementation, approximately 285 variables and 90 feedback loops have been observed and properly grouped into subsystems, as shown in Figure 3 [1].

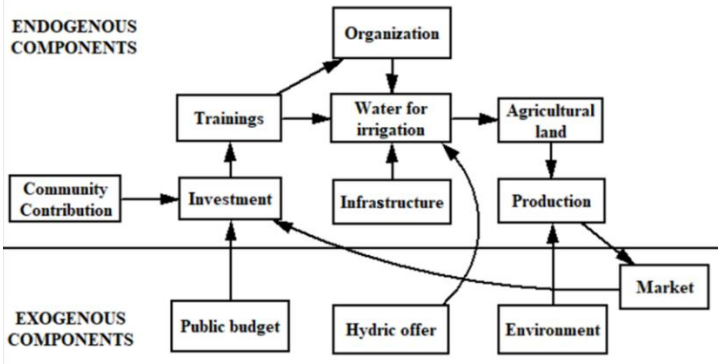


Figure 3: Block diagram of the proposed model for the agricultural production system

Figure 1 also identifies the necessary elements that participate in an economic theory (producers, consumers, goods markets, state intervention, etc.), having, in this case, water as the engine for the fight against poverty [13] or as "Theory of the economic and social value of water" states in a perfect market where there are several buyers and sellers, each commodity in the economy will be given its value [19]. As previously mentioned, the sustainability of projects is based on the strengthening of organizations and profitability. It is, then, necessary to know the profitability to identify the elements that participate under the economic theory defended by various authors [13].

4.2. Formulation and validation of the simulation model

The first objective of the work is to know the future behavior of the production system of the irrigation improvement projects implemented by the PEPMI. So, the planners will have a tool that allows them to know if they are making the right decisions or if the millionaire investments in infrastructure can cause short-medium or long-term adverse effects such as a recession. These phenomena have occurred in the provinces of Chumbivilcas-Espinar and the northern mountain range of Lima [20]. The causal diagrams made in the conceptualization phase have to be converted to mathematical logic models based on Forrester diagrams. For which the Vensim Personal Learning Edition (PLE) software has been used.

4.2.1. Subsystems formulation

Each of the subsystems shown in figure 1 are explained below:

A. Water supply

Determining the future supply of water is one of the main tasks to know. This to guarantee the water supply at the head of the plot. To achieve this, the historical information of the Pampaconga project has been considered. After analyzing the information, it has

been considered to perform a probability function for each month, using discrete probability distribution functions based on histograms. Requirements for the Monte Carlo simulation method [21]. Due to the limitations of Vensim Personal Learning Edition (PLE), the treatment of each probability function has been carried out in Excel and the results added in lookup functions in the implemented model.

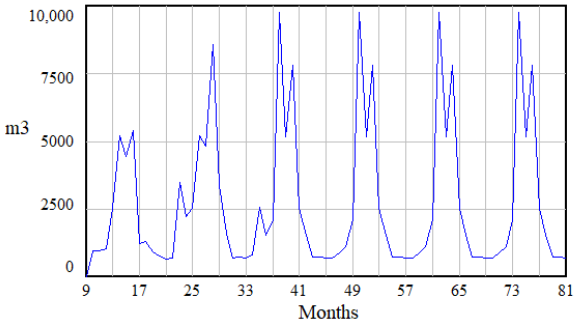


Figure 4: Projection of water availability due to rainfall

As shown in Figure 4, the hydrological year in Peru begins in September (month 9), and the high peaks represent the El Niño phenomenon. With climate change, this phenomenon will be increasingly present in the Peruvian Andes. When this occurs within the project's scope, the loss of production, and even irrigation infrastructure and all kinds of infrastructure is almost inevitable [22].

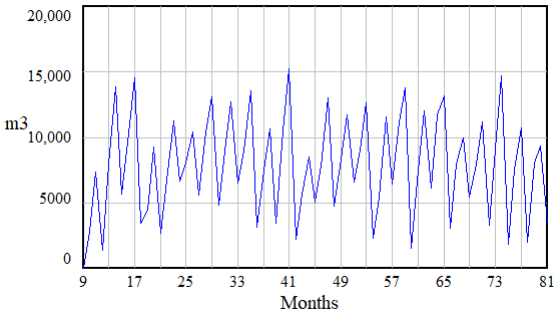


Figure 5: Projection of total water availability

The total water availability is made up of the rains and the flow of the rivers. These are exogenous variables that cannot be influenced. The projection of water availability is shown in Figure 5.

B. Investment

This subsystem is born from the public budget and external debt. For external indebtedness, there is German technical cooperation. According to the experiences of the first works implemented, the farmers showed resistance to change, to use and manage the modern irrigation infrastructure implemented by PEPMI. In addition, there is also the possibility of conflicts between communities over access to water and/or land. This demonstrates the high social component of the irrigation projects implemented [3]. Sometimes the engineer has to act as a sociologist to meet its goals [12]. To overcome these problems, the multidisciplinary team includes anthropologists, professionals who serve as important interlocutors between farmers and PEPMI.

In conclusion, the resistance to the change presented to use the new and free irrigation infrastructure was due; because farmers only use the infrastructure in whose construction they have participated. Therefore, project financing includes the community contribution, which is approximately 15 to 20% of the total project budget. This amount guarantees the participation of community members in the works to be built [3].

Another characteristic to take into account is that the Peruvian government considers irrigation improvement projects as social programs. And, given the appearance of economic problems, as the global pandemic caused by Covid-19, these programs are the first to reduce the public budget.

In the conceptualization of the production model in its interaction with the market, the investment subsystem receives feedback from the market, specifically, of the profit obtained from the sale of the products and the availability of financing that predisposes the advance of trainings and irrigation subsystems infrastructure.

C. Trainings

Training is an essential component to guarantee the sustainability of the project in its component of strengthening organizations and in terms of increasing productivity in the components of water management and the application of irrigation and cultivation techniques.

The PEPMI, being a state project, depends on the monthly budget. According to the monthly budget, it has the necessary factors to carry out the training: operational equipment, material, and workforce.

Figure 6 shows the Forrester diagram that explains the availability of resources to conduct the trainings.

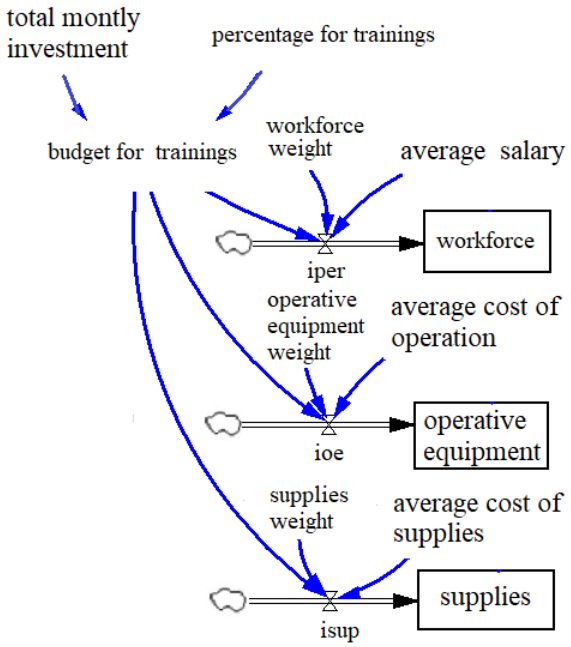


Figure 6: Forrester diagram: Availability of resources for training

The equations that govern the Forrester diagram in Figure 6 are as follows:

$$IPER = (BT*WW)/AS \quad (1)$$

where:

IPER: Personal ratio.

BT: Budget for trainings.

WW: Workforce weight = 0.5.

AS: Average salary = S/ 2000.00.

Percentage for trainings = 0.15.

$$IOE = (BT*OEW)/ACO \quad (2)$$

where:

IOE: Operative equipment ratio.

BT: Budget for trainings.

OEW: Operative equipment weight = 0.3.

ACO: Average cost of operation = S/ 1000.00.

$$ISUP = (BT*SW)/ACS \quad (3)$$

where:

ISUP: Supplies ratio.

BT: Budget for trainings.

SW: Supplies weight =0.2.

ACS: Average cost of supplies = S/ 300.00.

The trainings are carried out according to the availability of resources. The progress ratio is calculated according to the input targets programmed versus those available. And the total training per month predisposes the increase of water management, efficiency in the application of irrigation techniques, efficiency in the application of cultivation techniques, and strengthening of organizations that are governed by the following equations:

$$WM = 0.065*TPM \quad (4)$$

$$EAIT = 0.06*TPM \quad (5)$$

$$EACT = 0.07*TPM \quad (6)$$

$$SO = 0.065*TPM \quad (7)$$

where:

TPM: Trainings per month (the objective is 4 per month).

WM: Water management.

EAIT: Efficiency in the application of irrigation techniques.

EACT: Efficiency in the application of cultivation techniques.

SO: Strengthening of organizations.

As can be seen in Figure 7, in month 28, approximately the objective is achieved. This indicates that the trainings have made the farmer achieve 100% adequate water management. Similar behavior is observed in the variable's efficiency in applying irrigation techniques, efficiency in applying cultivation techniques, and strengthening organizations.

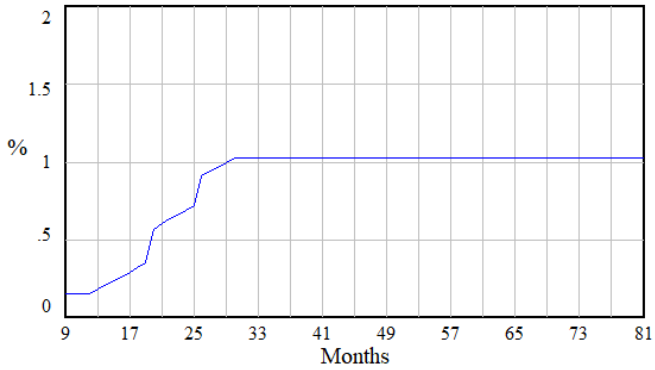


Figure 7: Advancement in water management

D. Irrigation infrastructure

Irrigation infrastructure is the component with the highest cost, depending on the size of the project. Its importance lies in the possibility of transporting water for kilometers without loss due to seepage and optimizing its storage and availability through dams and reservoirs. Advantages that the channels made on land do not have due to the high filtration rate. Irrigation infrastructure is also important to face adverse weather phenomena such as droughts, as it allows managing water scarcity, guaranteeing harvest (crop safety) and counteract adverse weather phenomena such as frost and hailstorms, since having a permanent water supply crops can recover, and, in areas where optimal conditions exist, the number of crops per year can be increased. What is also known as an increase in land use, which is the first step to fight poverty.

Then, the countries need to provide adequate irrigation infrastructure to manage the scarce water resource. This is because the imminent effects of climate change will affect primary sectors such as fisheries and agriculture, such as by 6% to the Peruvian GDP for the year 2030. One of the author's conclusions in [23] is that the emission of polluting gases and, therefore, the effects of climate change will continue. This situation evidently increases the total utility of water as part of managing scarcity and climate variability. Then, new efficient irrigation techniques such as dripping and sprinkling should also be considered. These aspects are already being considered in the making of water and irrigation laws.

Figure 8 shows the Forrester diagram that explains the availability of resources to carry out the irrigation infrastructure.

The equations that govern the Forrester diagram in Figure 8 are as follows:

Percentage for infrastructure = 0.85

$$ITS = (BI*TSW)/TSS \quad (8)$$

where:

ITS: Technical stuff ratio.

BI: Budget for infrastructure.

TSW: Technical stuff weight =0.06.

TSS: Technical stuff salary = S/ 2000.00.

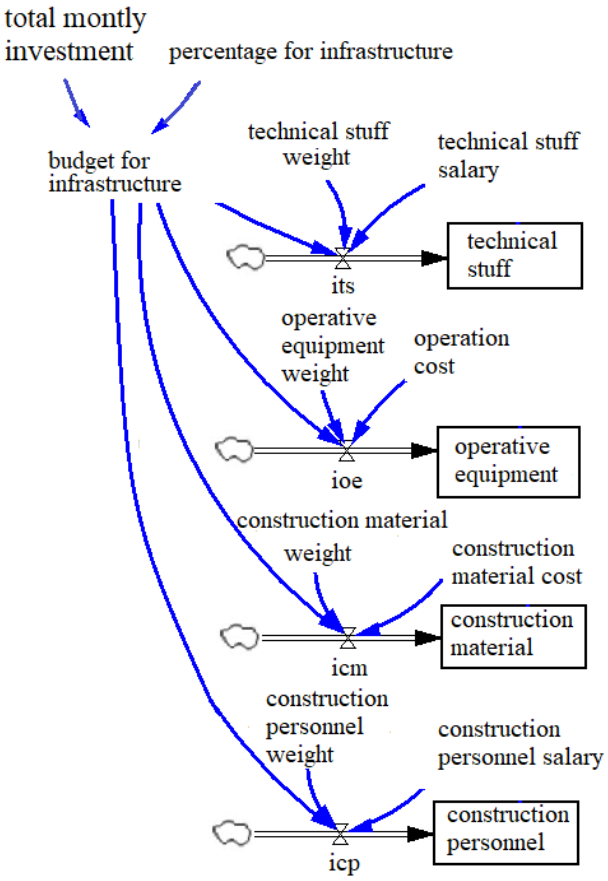


Figure 8: Forrester diagram: Availability of resources for irrigation infrastructure

$$IOE = (BI*OEW)/ACO \quad (9)$$

where:

IOE: Operative equipment ratio.

BI: Budget for infrastructure.

OEW: Operative equipment weight = 0.24.

ACO: Average cost of operation = S/ 4000.00.

$$ICM = (BI*CMW)/CMC \quad (10)$$

where:

ICM: Construction material ratio.

BI: Budget for infrastructure.

CMW: Construction material weight =0.11.

CMC: Construction material cost = S/ 15.00.

$$ICP = (BI*CPW)/CPS \quad (11)$$

where:

ICP: Construction personnel ratio.

BI: Budget for infrastructure.

CPW: Construction personnel weight =0.59.

CPS: Construction personnel salary = S/ 800.00.

Once the necessary resources are in place to carry out the irrigation infrastructure, the progress of the works depends on the availability of resources versus what is programmed and the weather conditions that allow the construction to continue. The construction progress is shown in Figure 9.

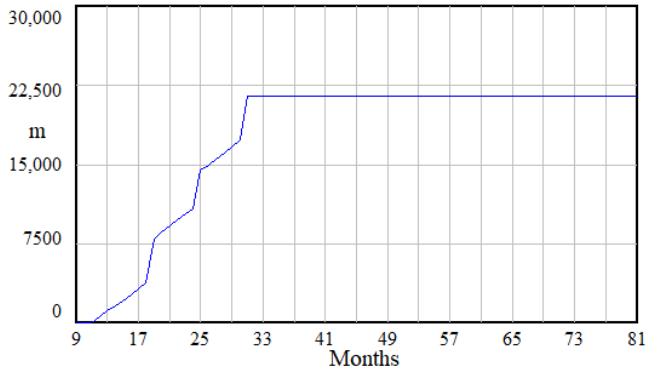


Figure 9: Progress in the construction of irrigation infrastructure

E. Water for irrigation

According to the progress of the irrigation infrastructure and the administration of water, the water for irrigation is determined according to the water demand calculated in the study phase. Figure 10 shows the Forrester diagram that explains the interaction of variables that determine the availability of water for irrigation.

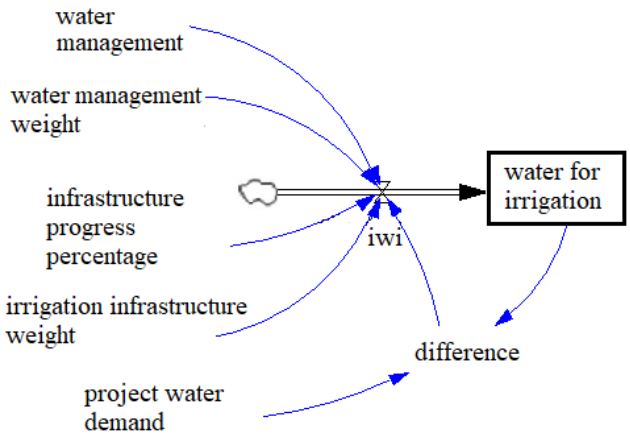


Figure 10: Forrester diagram: Water for irrigation

The equations that govern the Forrester diagram in Figure 10 are as follows:

$$IWI = [(D*WMW)*WM] + [(D*IIW)*IPP] \quad (12)$$

where:

Project water demand: 17,172.00 m3

D=PWD-WI

PWD: Project water demand.

WI: Water for irrigation.

WMW: Water management weight

WM: Water management.

IIW: Irrigation infrastructure weight

IPP: Infrastructure progress percentage

$$Water\ for\ Irrigation\ (t) = \int IWI(t)dt \quad (13)$$

where:

IWI: Water for irrigation rate

Figure 11 shows the increase in irrigation water until the project water demand is met, an important parameter that also serves to design the irrigation infrastructure (channels, reservoirs, canoes, rapids, etc.). The expansion of the agricultural frontier has a similar behavior within the project's scope as planned.

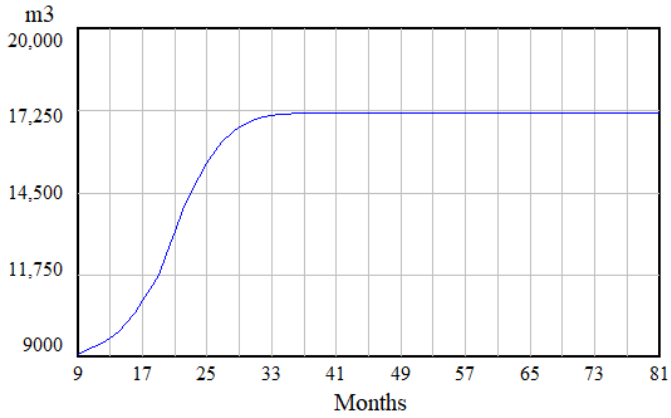


Figure 11: Increase in water for irrigation

F. Production function

The proposed simulation model for production in irrigation improvement projects includes the variables of water, supplies, training, machinery, and the human factor, whose interaction determines the quality and quantity of the cultivated land [1]. This is expressed in equation 14:

$$Adequate\ land\ for\ cultivation = \quad (14)$$

$$\begin{cases} 0 & ; WF \leq 0 \\ [WF*WWF+SF*SFW+TF*TFW+MF*MFW+HRF*HRFW]*PL & ; WF > 0 \end{cases}$$

where:

WF: Water factor.

WWF: Water factor weight (0.6).

SF: Supply factor.

SFW: Supply factor weight (0.1).

TF: Trainings factor.

TFW: Trainings factor weight (0.15)

MF: Machinery factor.

MFW: Machinery factor weight (0.1).

HRF: Human resource factor.

HRFW: human resource weight (0.05).

PL: Production land.

Note that, in extreme situations, without the water factor, production will be zero. Otherwise, it will be proportional in terms of the existence of the other production factors that have been considered. Figure 12 shows the production achieved. Note that, starting in month 33, there is a significant increase due to the already available new irrigation infrastructure, training, supplies, etc.

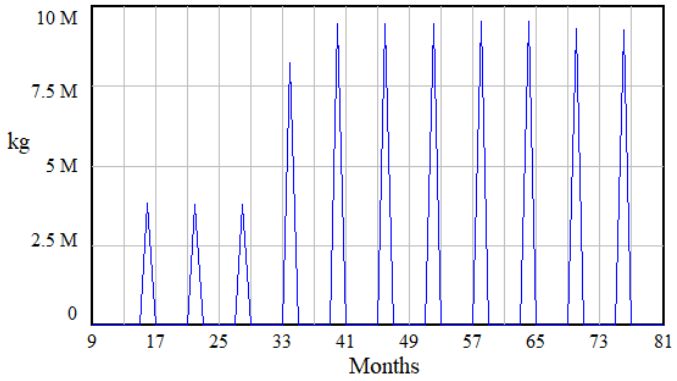


Figure 12: Production volume

G. Market

Once the production model has been obtained to determine the sustainability of the project, we must determine if there is a profit for the farmer due to sales in local markets. At this point, the proposed work requires to know the behavior of the market whose proposed model is of perfect competition [13].

Therefore, the present research becomes a broader study of the regional economy and an economic theory approach to fighting poverty, in this case, based on water as the main engine. Whose total utility in the medium term will allow farmers to access better living conditions such as health, housing, education, etc. Among these factors, we must highlight the greater and better access to education of the Andean population, as it will allow access to better economic income. About education, the winner of the last Nobel Prize in economics, in his work in Kenya,

concludes that better innovations must be made that allow more people to benefit economically from education [24].

H. Price determination

The interaction of supply and demand (which we have considered dividing into current demand and potential demand) and the seasonality of both allows us to know the product's sale price. Compared to the original work [1], in the present, we are going to present a new model described in equation 15.

$$Price = AP + [AP*(1-(SM+SMF))] \quad (15)$$

where:

AP: Average price.

SM: Satisfied market.

SMF: Stocks in the Market Factor.

$$Satisfied\ Market = Production\ volume / Current\ demand \quad (16)$$

Figure 13 shows the sale price behavior, in which it can be seen that the price reaches S/ 1.23 in times of scarcity and S/ 1.00 when there is overproduction. To carry out a sensitivity analysis, a scenario was constructed. The project's production increases by 30%, and due to market capacity (demand) occurs overproduction and, therefore, the sale price drops to S/ 0.66 per kg. These results, in a real environment, could occur in a few months or years. Note that the simulation is for 72 months, and, during the year, there are various seasons. This has been considered from the conceptualization of production and the market model. It represents an advantage and differentiates our proposal in relation to other production and market models such as the spider web.

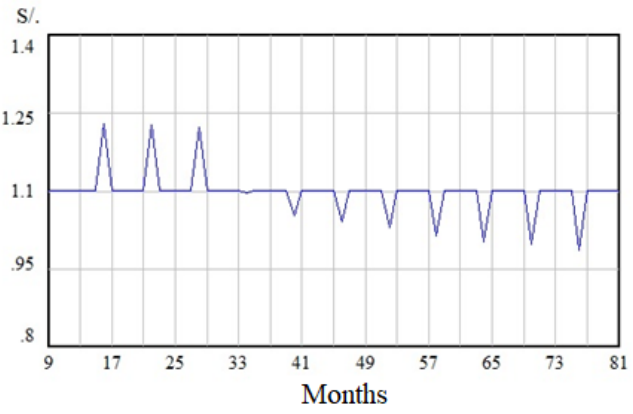


Figure 13: Sale price behavior obtained in simulation

4.3. Production optimization

Another aspect contemplated in economic theory is the maximization of producers' profits, who need to maximize the difference between total income and total cost. The break-even analysis allows us to know the minimum level of sales we must reach to avoid losses. Regarding the minimization of costs, the author in [25] recommends that this is the area we should focus on because it is difficult for producers to influence the sale price. Who governs us and defines the sale price is the market. Therefore, it is essential to reduce costs by implementing modern management techniques within organizations to obtain greater benefits. To define the optimization models, as observed through this work, we have the great restriction of the market (competition, seasonality, weather conditions, etc.). Therefore, it has been determined that the optimization of production is a restricted problem of the non-linear type. The objective function to maximize utility is shown in equation 17.

$$Max\ Z=16\varepsilon X_1\ X_2 - 11.14X_1 \quad (17)$$

Subject to:

$X_3 \leq 3600+16X_1$; *Stock on the market*

$X_3-16X_1 \leq 3600$

$X_2 \leq AP+AP*[1-16\ X_1/X_3+X_4]$; *Sale price*

Replacing and solving:

Being: AP = S/ 1.10 (Average price)

$X_2 \leq 2.2-17.6X_1/X_3+1.1X_4$

$X_1 \leq 400$; *Arable area*

$X_4 = 0.1$

Solving, it is obtained that the maximum profit is S/ 5 784,000.00 and the sale price results S/ 1.6 (X_2).

5. Conclusions

The new proposed simulation model allows for understanding the behavior of the production process of irrigation improvement projects. In this paper, the systems dynamics methodology has been used mainly for its formulation. The new model allows the projection of data in the future and, in this way, determine the impact of the decisions that are made and if they do not lead to long-term problems. In the conceptualization phase, it has been possible to identify endogenous and exogenous variables; Among the exogenous variables, the behavior of regional markets stands out, determining the sale price of products. For the validation of the model, the Anderson-Darling normality test has been used for all the identified subsystems (training, strengthening of organizations, irrigation water, farmland, etc.). We obtained from the normality tests values higher than $\alpha = 0.05$.

The production factors that we identify for irrigation improvement projects are water, inputs, training, machinery, labor, and land. The respective weights can be seen in equation 14. The irrigation infrastructure and water management determine the availability of water. All these factors in the model determine the amount of cropland and the production volume to obtain. However, to measure the profitability of production, it is necessary to know its efficiency concerning the regional economy. That's why we formulate models that allow understanding of the market's behavior and sales prices in the region. The proposed models also make it possible to carry out sensitivity analyzes and

detect the appearance of regional recession phenomena. All these models have been expressed in a formal language, a requirement of epistemology for the generation of knowledge. Then our study becomes a perspective of economic theory for the fight against poverty based on water and its correct use.

For the formulation of optimization models, the nonlinear programming techniques based on objective functions and restrictions have been considered. The dimensions evaluated were: profit maximization, cost minimization, and breakeven point. The proposed production and market models have become constraints or objective functions as appropriate in the optimization model. To validate the optimization model, it has been verified that they comply with the Kuhn Tucker conditions. The optimization problem has also been analyzed using genetic algorithms, whose application has made it possible to know that the sexual selection operator is the most efficient for achieving global and local optimums.

The difficulties we had were access to information and the models' calibration, which is a process difficult and slow in which VENSIM has been very helpful. In future works, we will apply the proposed models in the Titicaca lake basin, and we will contrast them with time series.

References

[1] R.A. Romero Flores, "Production Model for Irrigation Improvement Projects Implemented in Cusco - Peru Using System Dynamics," in 2019 IEEE World Conference on Engineering Education (EDUNINE), 1-6, 2019, doi:10.1109/EDUNINE.2019.8875796.

[2] F. Hurtado Huamán, What You Should Remember When Formulating a Rural Development Project, Editorial of the National University of San Antonio de Abad del Cusco, 2014.

[3] Plan Meriss Inka, The Water in Our Hands, Gráficos S.R. Ltda., 1995.

[4] GPER INKA, Training in Water Management and Management, Bizz Editores EIRL, 1995.

[5] P. M. Senge, The Fifth Discipline: The Art & Practice of The Learning Organization, Doubleday, 2006.

[6] J.D. Sterman, Business Dynamics: Systems Thinking and Modelling for a Complex World, McGraw-Hill, 2000.

[7] U. Bardi, The Limits to Growth Revisited, Springer-Verlag, 2011.

[8] A. Elmahdi, H. Malano, T. Etchells, K, Shahbaz, "System Dynamics Optimization Approach to Irrigation Demand Management," in 2005 MODSIM International Congress on Modelling and Simulation, 196-202, 2005, doi:10.13140/2.1.3408.4648.

[9] Z.Zhang, W.X.Lu, Y. Zhao y W. Song, "Development tendency analysis and evaluation of the water ecological carrying capacity in the Siping area of Jilin Province in China using system dynamics and analytic hierarchy process", Ecological Modelling, **275**, 9-21, 2014, doi: 10.1016/j.ecolmodel.2013.11.031.

[10] R. Ackoff, Redesigning the Future, Limusa Noriega, 1994.

[11] L. von Bertalanffy, General Systems Theory, George Braziller Inc., 2015.

[12] M. Bunge, Science, its method and philosophy, Laetoli, 2018.

[13] W.B. Zhang, The General Economy, Springer Cham, 2020.

[14] J. F. Gonzáles Bermúdez, "Income distribution and poverty in Latin America: The cases of Argentina, Brazil, Chile and Mexico. Mexico DF," PhD Thesis, National Autonomous University of Mexico, 2013.

[15] R. Tirado, Greenpeace, Ecological agriculture, Greenpeace Internacional, 2015.

[16] V. Laime Mantilla, Chumpiwillka Nation, Súper Gráfica, 2013.

[17] N.J. Forbes, "Equilibrium Points in N-person Games," Proceedings of the National Academy of Sciences of the United States of America, **36**(1), 48–49, 1950, doi:10.1073/pnas.36.1.48.

[18] O. Rohm, S. Dabbert, "Integrating Agri-Environmental Programs into Regional Production Models: An Extension of Positive Mathematical Programming," American Agricultural Economics Association, **85**(1), 254-265, 2003, www.jstor.org/stable/1244941.

[19] F.H. Arceo, "Theory of the Economic and Social Value of Water Resources", Master´s Thesis, National Autonomous University of Mexico, 2006.

[20] Ministry of Agriculture and Irrigation, Potato: Characteristics of National Production and Marketing in Metropolitan Lima, MINAGRI, 2017.

[21] V.P. Singh, System Modeling and Simulation, new age international (p) limited, 2009.

[22] A Klauer, El Niño La Niña: The ocean-atmospheric phenomenon of the South Pacific, a challenge for science and history, ISBN: 9972–817–09–1, 2000.

[23] P. Vargas, Climate Change and its Effects in Peru, Central Reserve Bank of Peru, 2009.

[24] E. Duflo, P. Dupas, M. Kremer, "School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools," Journal of Public Economics, **123**, 92-110, 2015, doi: 10.1016/j.jpubeco.2014.11.008.

[25] E. Milan, M. Stefanovic, A. Djordjevic, N. Stefanovic, M. Misic, N. Abadic, P. Popovic, "Optimization of Production Process Parameters with a New Model Based on Genetic Algorithms and ABC Classification Methods," Advances in Mechanical Engineering, **8**(8), 1-18, 2016, doi:10.1177/1687814016663477

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 499-504 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Remote Control of Garden Plantation Water Pumps using Arduino and GSM Mobile

Beza Negash Getu*, Mohamed Abdulkadir, Michael Tous

Department of Electrical, Electronics and Communications Engineering, American University of Ras Al Khaimah, Ras Al Khaimah, 10021, UAE.

ARTICLE INFO

Article history:
Received: 23 June, 2020
Accepted: 15 July, 2020
Online: 09 August, 2020

Keywords:
Soil moisture
Light intensity
Temperature
Humidity
Arduino
Greenhouse
Automatic gardening
GSM module

ABSTRACT

A remotely mobile phone controlled electronic system for supplying water to garden plantations in a greenhouse or similar environment was designed and experimentally implemented. The system monitors the garden environmental conditions such as soil moisture, temperature, humidity and sunlight and a remote user can send commands from his mobile phone such as to switch ON/OFF water pumps, supply water for the plants for certain duration and acquire the environmental status information of the plantation. The water pumps are controlled by the Arduino microcontroller that is the core part of the electronic system. The user has the ability to monitor the environmental conditions on his mobile phone. Such systems can facilitate monitoring, give flexibility of controlling, save time and human labour and increase productivity as a result of automation and remote controlling.

1. Introduction

In the agricultural sector, the traditional irrigation techniques like watering cans, drip and sprinkler irrigations require farmers to be physically present in their farm lands to regularly supply water to their crops, which is a manually controlled irrigation method that results in wastage of human power as well as time. To circumvent labour presence at the farm site, in [1] Dual Tone Multi-Frequency (DTMF) based remotely controlled electronic system was proposed to control the switching states of agricultural water pumps located in remote farm areas. On the other hand, nowadays water shortage is becoming one of the biggest problems in the world from domestic home use as well as in agricultural and industrial sectors. Many different methods are developed for conservation of water. As water is one of the scarce natural resources, it is important to properly use and manage our usage in different sectors and hence there is a need to design systems that are used to monitor and control our water usage including in homes, agriculture and industrial sectors [2], [3]. It is known that agriculture is one of the fields where water is required in tremendous quantity and there are many techniques to save or to control wastage of water from agriculture. With the development of technology in water saving irrigation and automation, automatic irrigation is going to be more popular in the farms. In [4] an automatic control system is designed and implemented for garden plantations based on soil moisture, temperature and light sensors. The system monitors and controls the soil moisture, temperature and light intensity based on the three sensors and the Arduino module that automatically controls a water pump, a fan and a lighting fixture based on the sensor information in the greenhouse. It controls the water pump automatically by the Arduino and the built electronic system.

Automation of irrigation systems based on soil moisture sensors (SMS) was also implemented in [5] to optimize water usage efficiency by maintaining soil moisture at optimum levels. It was found that a significant percentage of water savings can be achieved based on the quality of the installed sensors in the farm land. An Automatic irrigation system based on Arduino microcontroller and a sensor sensing the dampness of the soil was also proposed in [6]. An Arduino based automated irrigation system based on soil moisture sensor was also proposed in [7]. In this system, the microcontroller is interfaced with the HC-05 Bluetooth module that can be connected to Android phone for

*Corresponding Author: Beza Negash Getu, American University of Ras Al Khaimah, bgetu@aurak.ac.ae

controlling the water pumps in the farm. The drawback of this control is the limited range of the Bluetooth communication that can be from a distance of 10 m to 100 m from the transmitter [8]. In plant watering system that was proposed in [9], the soil moisture sensor checks the moisture level in the soil and if moisture level is low then the Arduino informs the user by sending message to the user cell phone then the user sends a command that switches on the water pump to provide water to the plants. However, this system is based on a PIC16F628 microcontroller that is different from Arduino and used two metal plate sensors for moisture sensing instead of standard commercially available sensors that limits the accuracy. A microcontroller based irrigation system was also proposed in [10] based on 8051 Microcontroller. However, a system based on Arduino simplifies the process of working with microcontrollers and it is inexpensive and widely used everywhere in recent days [11], [12].

The purpose of the work in this paper is to further improve the work done in [4] so that there is an ability to remotely control the water pumps depending on the farm environmental conditions received from the Arduino. The designed electronic system monitors the soil water level (moisture level) of the garden plantation and supplies water through the water pumps that are controlled by the Arduino microcontroller. The system is to be controlled by sending commands remotely from a mobile phone of the owner of the garden plantation or the farms. In addition, the system also needs to monitor the garden environmental conditions (moisture, temperature, light, humidity) and inform the user so that he can take the necessary actions. Such Global System for Mobile (GSM) based automatic irrigation system increases the productivity of farms and eliminates the cost of human labour used in manual irrigation methods. It provides convenience in accessing the system from anywhere at any time.

2. Methods and Materials

The designed system and the various components and standards used for this project are described in this section

2.1. System Design and Components

The designed system as shown in Figure 1 is implemented to provide the necessary supplies for a farm or a greenhouse. The system is designed to monitor and control the sensor elements through the Arduino microcontroller and providing the status of the land to the user through the GSM module. The sensors are connected to the Arduino to provide the microcontroller with the soil moisture, humidity and temperature and light status of the plantation. The microcontroller stores the sensors data information and then sends them to the user upon request. It also receives commands from the user through the GSM module and then supplies the land with water, light and air conditioning by switching ON/OFF the water pumps, the light and the fans respectively depending on the required needs of the plants.

The standards of the project are commercially available components brought together to form the project and make it a complete one. The design aspect is to integrate all the design components in a way to achieve our target and the target tasks are achieved through programming the Arduino microcontroller that is interacting with the various input and output components. Some of the main components of the system are listed below.

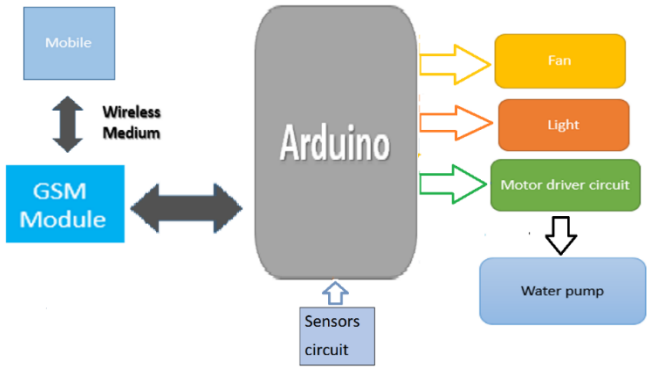


Figure 1: System Block Diagram

2.1.1. Arduino Uno

Arduino is an open source microcontroller used for hardware and software integration. Arduino can be interfaced with sensor elements and control them in the real world. Arduino has a set of digital and analogue input/output pins that can be interfaced to distinctive developed boards (shields) and different sensor circuits [12]. Serial communication interfaces including Universal Serial Bus (USB) are required for loading programs from Personal Computer (PC) to the Arduino. The Arduino board as shown in Figure 2 provides an integrated development environment (IDE) for programming the microcontroller to process various types of projects. The Arduino IDE provides a simplified integrated platform which can run on regular PCs and allows users to write programs for Arduino using C or C++ language.



Figure 2: Arduino UNO Board

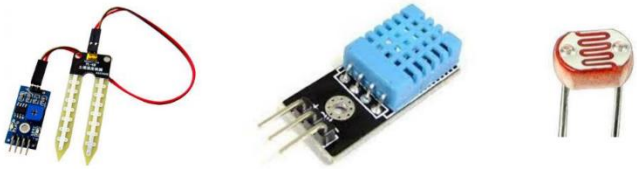
2.1.2. GSM GPRS Quad Band Board SIM900 Shield for Arduino

GSM is a second generation (2G) digital cellular networks used by mobile phones. A GSM modem is a device that modulates and demodulates signals for communications. The modem used in this project is SIMCOM SIM900 [13], which is s a quad-band GSM/General Packet Radio Service (GPRS) modem as it can operate at four frequencies (850 MHz, 900 MHz, 1800 MHz and 1900 MHz). GSM provides a wide range of things to remotely control from anywhere with your fingertips. SIM900 GSM module can be used to send and receive SMS by connecting it to a PC or Arduino when a Subscriber Identification Module (SIM) card is inserted. The GSM modem can send ATtention (AT) commands to send or receive Short Message Service (SMS) from PC via a

COM (serial or USB) port. The SIM900 GSM module used in this project as shown in Figure 3 consists of a Transistor-Transistor Logic (TTL) and an RS232 interface. The TTL interface allows you to interface directly with a microcontroller while the RS232 interface allows to communication with PC.



Figure 3: GSM/GPRS Quad Band Board



(a) (b) (c)

Figure 4: Plant Environmental Condition Sensors: (a) Soil moisture sensor, (b) The Arduino DHT11, (c) LDR

2.1.3. Plant Environmental Condition Sensors

The proposed system is intended to control and monitor four different parameters (soil moisture, temperature, humidity and light status) of a certain garden or greenhouse planation that are vital for the healthy growth of plants. Calculating the soil moisture using soil moisture sensors in the fields is important to help the farmers improve irrigation system efficiently. The quality of the crop production can be increased by management of soil moisture during critical stages of plant growth. Volumetric water content is calculated indirectly by soil moisture sensor using different parameters of the soil like dielectric constant, electrical resistance and neutrons interaction. Controlling and monitoring of the temperature around the plants is essential for maintaining their healthy growth. The system turns on the fans for cooling the plants if the temperature and/or humidity increases above a certain preset value. The threshold (preset) temperature and/or humidity is chosen depending on the requirements of the plants so that they are not affected by any excess heat which may affect their metabolism and healthy growth. Similarly, the system has a light sensor that detects the level of lighting in the plant surroundings. The sensor is a light dependent resistor (LDR) used for in light/dark sensor circuits. Normally, the resistance of an LDR is very high, can be in the order of Mega ohms (MΩ), but when it is illuminated with light, its resistance drops dramatically allowing more current to pass through it. Figure 4 shows the soil moisture sensor, the temperature/ humidity composite sensor (DHT11 sensor) and the LDR light sensor components used for this project. The system is flexible in that the preset moisture level, temperature, humidity or light levels could be modified if needed by the user depending on the requirements. By changing the Arduino code, any desired level could be programmed by the user.

2.1.4. Arduino Compatible L298N Stepper Motor Driver

This 2x2A DC Motor drive for Arduino allows driving two channel DC motors. It uses a L298N chip, which deliveries output current up to 2A for each channel. The speed control is achieved through conventional Pulse Width Modulation (PWM) which can be obtained from Arduino's PWM output Pin 5 and 6. The enable/disable function of the motor control is signaled by Arduino Digital Pins 4 and 7. The Motor shield can be powered directly from Arduino or from external power source. It is strongly encouraged to use external power supply to power the motor shield. The Arduino Motor Shield as shown in Figure 5 is based on the L298, which is a dual full-bridge driver designed to drive inductive loads such as relays, solenoids, DC and stepping motors. In this project, this component is being used because it lets to drive the water pump motors connected to the Arduino board to control the speed and direction of each one independently.

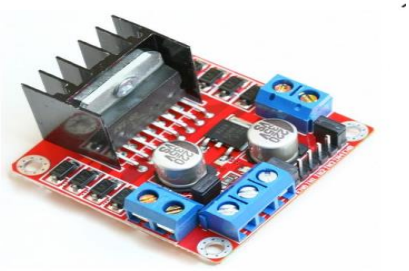


Figure 5: Arduino Compatible L298N motor drivers

2.2. Complete overall circuit

The system contains the following components: the microcontroller which is the Arduino Uno, GSM Module, the temperature and humidity sensor which is the DHT11, soil moisture sensors, a L298N motor driver, a 5V relay, water pumps, fans, LEDs, light sensor and rechargeable batteries as shown in the overall circuit in Figure 6. The micro controller used is the Arduino Uno based on Atmega328P, it has 14 electronic input/output pins, powered by a battery and connected to the GSM Module, sensors and hardware devices compatible with each sensors. All the components are tightly fixed on a wooden base and surrounded with wood and partly glass from the sides to allow light in the day time so that the plants get the amount of light it needs in the day time as seen in the prototype in [4]. When it is dark the LDR will give the readings to the Arduino Uno then the Arduino sends this information to the user through the GSM Module. Then the user will send back a message to the GSM that turns ON the lights. There are also four holes on the surrounding wood to fix the four fans. The fans will work when the user finds that the plants need cooling depending on the environmental conditions.

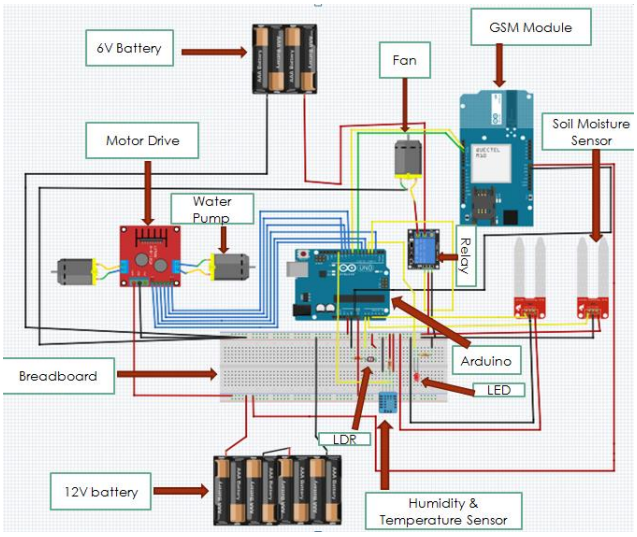


Figure 6: Complete Circuit Connections

3. Software Implementation

The Arduino Uno was programmed for preset values of moisture, temperature, humidity and light environmental conditions that can be simply changed in the program (Arduino software). For the testing purposes, when temperature is equal to or above 27 degrees and the humidity is above 70 %, the user will send a command message to turn on the fans and will stop it, if temperature is less than 27 degrees and the humidity is less than 70 %. These temperature and humidity threshold values can be easily changed in the Arduino code as per the user's interest. The DH11 composite humidity-temperature sensor will be responsible to give the reading to the system. The water pumps work with the user's permission. The speed of the water pump that supplies water to the land depends on the moisture level of the land. For the experiment, the moisture levels are divided into four levels that are **high** (65 % ≤ moisture level ≤ 100 %), **medium** (50 % ≤ moisture level < 65 %), **medium-low** (25 % ≤ moisture level < 50 %) and **very low** (0 % ≤ moisture level < 25 %). This threshold values can be easily changed in the program. When the soil moisture is high the water pump is OFF. When the soil moisture is medium the speed of the water pump is medium. When the soil moisture is medium-low the water pump is in medium-low speed. However, when the soil moisture is very low the water pump speed is high so that it pumps high quantity of water to the plants.

For simplicity, we call the text message sent by the user to be "switching states". This switching states provide information about the status of the two motor pumps, the lighting bulb (LED for the experiment) and the fan status. The set of the switching states consists of MOTOR1ON, MOTOR1OFF, MOTOR2ON, MOTOR2OFF, FANON, FANOFF, LIGHTON and LIGHTOFF. The user sends one of the mentioned switching states to turn ON/OFF, the motor pumps, the fan or the light in the farm. The other text message sent by the user is the STATUS message to get the environmental condition (moisture level, temperature, humidity and lighting conditions) of the farm or the greenhouse so that the user is aware of the status on the site and then sends the necessary commands to control the water pumps, the light and the fans accordingly. The overall working of the program that is uploaded to the Arduino can be described in a flow-chart as shown in Figure 7.

After declaration of the variables and initializations, in the main program loop, the Arduino checks if the message is sent from a specific user before executing the SMS command from the sender. This will make the system secure so that only a specific user can control the system.

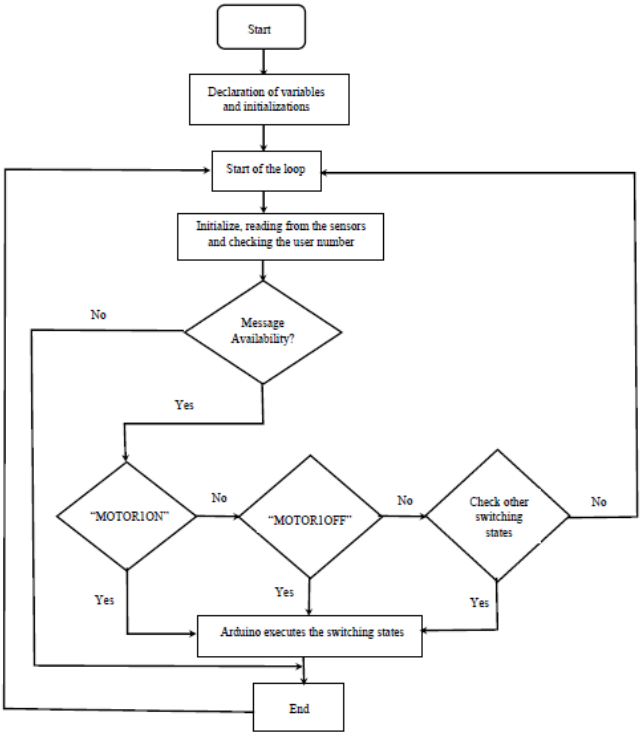


Figure 7: Flow chart of the Arduino program

A sample of the Arduino program (Algorithm 1) to switch on the motor for water pump 1 with high speed is given below. As per the implemented Arduino code, a moisture level between 0% and 25% is considered relatively dry and thus the water pump will run at high speed. A similar code is constructed to check the other three different moisture levels and the water pump will run with a speed corresponding to the proper range.

Algorithm 1: Arduino code to switch on the motor for water pump 1.

```
output_value1 = map(output_value1,1023,0,0,100); // Get the reading value of moisture sensor 1 and map the value in the range [0 100] where 100 means the maximum wet condition
if(SIM900.find("+971xxxxxxxx")>=0){  // Check whether the message is sent from a specific mobile user
if(SIM900.available()>0) // Display any text that the GSM shield sends out to the serial monitor
{
textMessage = SIM900.readString();  //Get the character from the cellular serial port
Serial.print(textMessage); //Print the incoming character to the terminal
if(textMessage.indexOf("MOTOR1ON")>=0){  // Turn on motor 1 and save current state
```

```
if(output_value1<25) //if the moisture level is less than 25%, run the water pump at high speed
  {
digitalWrite(in1, LOW);
digitalWrite(in2, HIGH);
analogWrite(enA, 255);
message = "Motor 1 is at high speed";
  }
}
sendmsg(message);   // send the message " Motor 1 is at high speed" to the user
}}
```

The message "Motor 1 is at high speed" will be sent to the farmer or the user to let him/her know that the water pump is switched on and it is running at high speed due to very low moisture level and the high need of water for the plants. The sample of the Arduino code (Algorithm 2) for this subroutine for sending the message is given below.

Algorithm 2: Arduino code for sending message to the user confirming that Motor 1 is at high speed.

```
void   sendmsg(String message)
{
SIM900.println("AT+CMGF=1");   //Set the GSM Module in Text Mode
delay(100);
SIM900.println("AT+CMGS=\"+971xxxxxxxxx\"");  // The farmer or user mobile number to whom the message to be sent
delay(100);
SIM900.println(message);   // The SMS text to be sent to the user
delay(100);
SIM900.println((char)26);  // ASCII code of CTRL+Z
delay(100);
SIM900.println();
delay(1000);
}
```

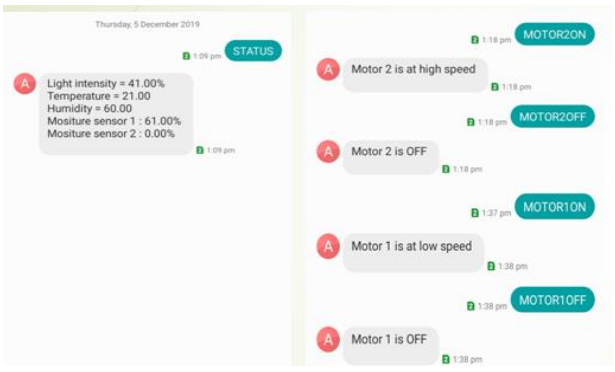
In a similar way, the subroutines corresponding to the rest of the switching states (MOTOR1OFF, MOTOR2ON, MOTOR2OFF, FANON, FANOFF, LIGHTON and LIGHTOFF) SMS command messages are incorporated in the overall program. The corresponding confirmation that provides switching status information (ON or OFF) will also be sent to the farmer or the user regarding the status of the water pumps, the fan and the light bulb. The user has also the option to obtain the environmental condition of the farm or the green house by sending the command "STATUS". When the Arduino gets this message through the GSM module, the microcontroller collects all the sensors information (light intensity, temperature, humidity and moisture levels in the two sensors) and sends this information as environmental STATUS to the user. The sample of the Arduino code (Algorithm 3) for this subroutine is given below.

Algorithm 3: Arduino code used for sending STATUS information to the user.

```
Void sendSMS(){  // AT command to set SIM900 to SMS mode
SIM900.println("AT+CMGF=1");   //Set the GSM Module in Text Mode
delay(100);
SIM900.println("AT+CMGS=\"+971xxxxxxxxx\""); //The user mobile number that gets the STATUS information
delay(100);
SIM900.print("Light intensity = "); // Send the SMS
SIM900.print(val);
SIM900.println("%");
delay(100);
SIM900.print("Temperature = ");
SIM900.println(t);
delay(100);
SIM900.print("Humidity = ");
SIM900.println(h);
delay(100);
SIM900.print("Moisture sensor 1 : ");
SIM900.print(output_value1);
SIM900.println("%");
delay(100);
SIM900.print("Moisture sensor 2 : ");
SIM900.print(output_value2);
SIM900.println("%");
delay(100);
SIM900.println((char)26);  // ASCII code of CTRL+Z
delay(100);
SIM900.println(); // Give module time to send SMS
delay(1000);
}
```

4. Experimental Test, Results & Discussions

As discussed in the above sections, an automatic Arduino and GSM based control system for garden plantations or a greenhouse was designed, built and tested in the laboratory and it was found that the system works properly. Figure 8 shows a screenshot sample results of communication between the mobile user and the designed Arduino based system to show how remote controlling is achieved. As shown Figure 8 (a), the owner of the greenhouse sent the message "STATUS" to the system and the owner received the whole status of the greenhouse. This message is helpful because it lets the owner to decide what has to be switched ON or OFF. There are two moisture sensors in this project. Since the moisture sensor 2 has a 0 % moisture, the owner of the green house has decided to switch ON Motor 2. As a result, Motor 2 pumps the water at high speed since the soil level of sensor 2 is 0 %. Similarly, this goes for soil moisture sensor 1. Since the soil moisture level is at 61%, Motor 1 pumps the water at low speed depending on the level of the soil. Depending on the status received, the user can turn ON or OFF the lights, the fans when needed as shown in Figure 8 (b).



(a) Getting the STATUS information and controlling the water pumps.



(b) Controlling the LIGHT and the FAN in the green house

Figure 8. Sample communication between the mobile user and the designed Arduino & GSM based system

5. Conclusions

An Arduino and GSM based electronic system was designed and implemented for monitoring and controlling the level of soil water, temperature, humidity and light intensity for the plants in a garden plantations or greenhouse system. These environmental factors affect the proper and healthy growth of plants. The system is designed and implemented with the Arduino being the core microcontroller interfaced with the GSM module and the sensors providing input information to the Arduino. The system was tested in Laboratory and found working properly. The system can be easily implemented with low-cost budget to make it affordable for everyone as Arduino is a cheap microcontroller with open source software availability. This system can be integrated in the agricultural sector as GSM based irrigation system since mobile phone devices are nowadays used almost everywhere in the world. It can help the farmers to easily remotely control the site and thus save human labour by eliminating the physical presence of persons on the agricultural site. Moreover, it is beneficial to properly manage, conserve water and get increased productivity in agriculture and in the end adding value to the economic growth of a country. By implementing this system not only agricultural farms but also parks, gardens, golf courses can be irrigated. The system benefits the environment positively as it saves water the scare natural resource available for human kind.

Acknowledgment

The authors appreciate the support for the research provided by the American University of Ras Al Khaimah (AURAK).

References

[1] B.N. Getu et al., "Remote Controlling of an Agricultural Pump System Based on the Dual Tone Multi-Frequency (DTMF) Technique," Journal of Engineering Science and Technology, 10 (10), 1261 – 1274, 2015. http://jestec.taylors.edu.my/Vol%2010%20Issue%2010%20October%202015/Volume%20(10)%20Issue%20(10)%201261-1274.pdf.

[2] B.N. Getu et al., "Automatic water level sensor and controller system" International Conference on Electronic Devices Systems and Applications (ICEDSA), 1-4, 2016. DOI: 10.1109/ICEDSA.2016.7818550.

[3] B.N. Getu et al., "Automatic Control of Agricultural Pumps Based on Soil Moisture Sensing. Proceedings of the IEEE AFRICON 2015 Conference, 667-671, 2015. DOI: 10.1109/AFRCON.2015.7332052.

[4] B.N. Getu et al., "Design of Automatic Control System for Garden Plantations Based on Soil Moisture, Temperature and Light Sensing" Journal of Engineering and Applied Sciences, 12 (18), 4782-4786, 2017. DOI: 10.36478/jeasci.2017.4782.4786.

[5] B.C. Lailhacar, M.D. Dukes, G.L. Miller, "Sensor-Based Control of Irrigation in Bermuda grass" ASAE Annual International Meeting, Tampa, Florida, 1-14, 2005. https://abe.ufl.edu/faculty/mdukes/pdf/publications/SMS/ASAE_05-2180.pdf.

[6] M. Nagarajapandian, "Automatic Irrigation System On Sensing Soil Moisture Content" International Journal of innovative research in electrical, electronics, instrumentation and control engineering, 3(1), 96-98, 2015. DOI 10.17148/IJIREEICE.2015.3120.

[7] S. Rakshak, R.W. Deshpande, "Automated Irrigation System Based on Arduino Controller Using Sensors" International Journal of Innovative Research in Computer and Communication Engineering, 5 (7), 13394-13400, 2017. DOI: 10.15680/IJIRCCE.2017. 0507084.

[8] K.V.S. Sairam, N. Gunasekaran S.R. Redd, "Bluetooth in wireless communication" IEEE Communications Magazine, 40 (6), 90-96, 2002. DOI: 10.1109/MCOM.2002.1007414.

[9] J. Uddin et al., "Automated Irrigation System Using Solar Power". 7th International conference on electrical and computer engineering, 228-231, 2012. DOI: 10.1109/ICECE.2012.6471527.

[10] M. Ojha et al. "Microcontroller Based Plant Irrigation" International Journal of Advanced Technology in Engineering and Science, 3 (10), 71-79, 2015. http://ijates.com/images/short_pdf/1446208511_P71-79.pdf.

[11] L. Louis, "Working Principle of Arduino and using it as a tool for Study and Research" International Journal of Control, Automation, Communication and Systems (IJCACS), 1(2), 21-29, 2016. DOI: 10.5121/ijcacs.2016.1203

[12] Arduino Board, 2019: https://www.arduino.cc/en/main/arduinoBoardUno.

[13] SIM900, 2019. https://simcom.ee/modules/gsm-gprs/sim900/.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 505-515 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Decision Making System for Improving Firewall Rule Anomaly Based on Evidence and Behavior

Suchart Khummanee*, Phatthanaphong Chomphuwiset, Potchara Pruksasri

Department of Computer Science, Faculty of Informatics, Mahasarakham University, 44150, Thailand

ARTICLE INFO

Article history:
Received: 27 June, 2020
Accepted: 23 July, 2020
Online: 09 August, 2020

Keywords:
Firewall rule anomaly
Decision support system
Firewall rule analysis
Probability of firewall rules

ABSTRACT

*Firewalls are controlled by rules which often incur anomalies. The anomalies are considered serious problems that administrators do not desire to happen over their firewalls because they cause more vulnerabilities and decrease the overall performance of the firewall. Resolving anomaly rules that have already occurred on the firewall is difficult and mainly depends on the firewall administrator's discretion. In this paper, a model is designed and developed to assist administrators to make effective decisions for optimizing anomaly rules using the probability approach (Bayesian). In this model, the firewall needs to add four property fields (Extra fields) to the firewall rules: frequency of packets matching against rules, evidence of creating rules, the expertise of rules creator and protocol priority. These fields are used to calculate the probability of each firewall rule. The probability for each rule is used while the rules conflict and administrators need to resolve them. The rule having the highest probability value indicates that it has the highest priority in consideration. Experimental results show that the proposed model allows firewall administrators to make significant decisions about solving anomaly rules. The data structure of this model is based on k-ary tree, therefore the speed of building tree, time complexity and space complexity: O(n), O($\log_m$n) and O(m*n) respectively. Besides, the confidence of the proposed firewall for resolving firewall rule anomalies of the administrator increase by 29.6% against the traditional firewall, and the reliability value between the inter-raters also increase by 13.1%.*

1. Introduction

A firewall is an indispensable system for today's computer networks. It plays an important role to prevent access to various resources on the networks, for example, networking devices, databases, web servers and etc. Besides, it can also prevent attacks and intrusions by malicious users from dark side networks. Basically, the firewall is commanded by a set of instructions, called the rule. The number of firewall rules depends on the complexity of the policies of each organization. As the number of rules increases, the number of anomaly rules is also enlarged. The anomalies arise from two or more rules overlapping but having different decisions. There are five types defined by [1]: the shadowing, correlation, generalization, redundancy, and irrelevancy anomaly. Recently, a new definition of anomalies has been advanced, which is the semantics loss of rules [2]. The rule anomalies have a great impact on the overall performance of a firewall. That is, they decrease the processing speed of rule verifying or matching. Therefore, reducing the number of anomalies improves the speed of the rule verification as well. Analyzing, managing and resolving anomalies are big problems of firewall researches and attract much interest. The first major researcher about the anomalies was Al-Shaer, who presented five types of anomaly rules and also invented an effective method for detecting anomalies, called the finite state diagram [3]. Later, several researchers contributed various methods for detecting anomalies, for example, in [4] the authors revealed a powerful algorithm to relieve the root cause of anomalies, called SDD. However, the rules are only one type either "an acceptable status" or "an unacceptable status" appearing on the firewall. Next, in [5] they proposed a method to resolve the anomaly problems effectively by using the firewall decision diagram (FDD) and this is also the prototype of much other later research. The propositional logic model was presented by [6], they claimed that their model could remove anomalies that lead to a decrease in the number of rules without changing the policy. Next, in [7] authors demonstrated the anomaly management framework which encourages systematic detection and resolution of firewall policy

*Corresponding Author: Suchart Khummanee, Department of Computer Science, Faculty of Informatics, Mahasarakham University, Thailand, Email: suchart.k@msu.ac.th

anomalies based on the average risk values, called FAME. The risk values are calculated from the Common Vulnerability Scoring System (CVSS) [8] which does not consider attacking vulnerabilities in the overview, but it is considered only the point. The next proposed solution for analyzing and managing the firewall policies was Lumeta [9] and Fang [10]. They represented the tools that are used to analyze firewall rules, but tools cannot completely verify the misconfigurations policy settings. In [11] the authors contributed an algorithm for detecting and resolving conflicts in packet filtering. Though, the algorithm can only detect certain specific conflicts. The firewall rule optimization based on Service-Grouping was proposed by [12]. The basis of this technique is resolving the conflicts segment by grouping rules from work behaviors. They claimed that the processing time and number of packet hits are better than the traditional and FIREMAN [13] firewall. FPQE [14] is an automated system to resolve rule anomalies, which does not require any admin intervention. It uses an automatic rule removal in the case of redundancy and contradiction anomaly, and uses an automatic rule permutation against shadowing and correlation. Besides, some techniques allow the firewall to automatically detect and analyze conflict rules such as [15] and [16], but they are not based on real tangible evidence. By most methods, the burden of resolving rule conflicts is often given to the administrator's discretion instead.

This paper contributes the model for optimizing firewall rule anomalies by applying the probability (Bayesian) together with evidence of each rule, i.e., frequency of packets matching against rules, evidence of creating rules, the expertise of rules and protocol priority. This model provides guidance to firewall administrators to resolve rule anomalies with probability values. In order for administrators to be confident that the decision is more accurate based on the actual evidence. This paper is organized as follows: Section 2 overviews the background and related work. Section 3 presents the key contributions. In Section 4 articulates our system design. Section 5 addresses the implementation of details and evaluations. Section 6 concludes this paper.

2. Background and Related Works

2.1. Rule Definition and Anomaly

Generally, the firewall rule consists of two parts: the condition and decision parts. Let *R* be a firewall rule, *C* as a condition part and *A* is a decision part, a firewall rule format:

$$R = C \rightarrow A \quad (1)$$

In fact, firewalls always have more than a single rule. Therefore, the first equation (1) needs to be revised to the second equation:

$$R_i = C_i \rightarrow A_i \quad (2)$$

where C_i and A_i are the condition and decision of rule R_i (Any firewall rule) by $i \in [1, n]$, and n is a non-negative integer. Given f_i representing the domain of positive integers is a finite range, denoted $D(f_i)$. For example, the domain of the source and destination address in an IP packet is [0, 2^{32} - 1] ($D(f_1)$ and $D(f_2)$), source and destination port is [0, 2^{16} - 1] ($D(f_3)$ and $D(f_4)$) and protocol is [0, 2^8 - 1] ($D(f_5)$). C_i defines a set of packet fields over the fields f_1 through f_d specified as $f_1 \in F_1 \wedge f_2 \in F_2 \wedge \ldots \wedge f_d \in F_d$ where F_i is a subset of $D(f_i)$. A_i is either *accept* or *deny* for each rule. If all conditions (f_i) in C_i are true, the decision is either *accept* or *deny* depending on the specified administrators like:

$$R_1 = (f_1 \wedge f_2 \wedge f_3 \wedge \ldots \wedge f_d)_1 \rightarrow accept_1 | deny_1$$

Given P_i as an IP packet over the d fields f_1, ..., f_d, P_i is a tuple of $d(p_1, p_2, \ldots, p_d)$ where each $p_i (1 \leq i \leq d)$ is an element of $D(f_i)$. An IP packet $(p_1, p_2, \ldots, p_d)$ matches R_i if and only if the condition $p_1 \in f_1 \wedge p_2 \in f_2 \wedge \ldots \wedge p_d \in f_d$. A set of rules $(R_1, \ldots, R_i)$ is valid when there is at least one rule in the set matching against p_i. To make sure that firewall rules are working properly, the condition of the final rule in the firewall is usually specified as $f_1 \in D(f_1) \wedge \ldots \wedge f_d \in D(f_d)$, where every packet must be matched as shown in R_3, called the implicit rule. The set of rules below shows an example of three rules over the three fields of condition $C(f_1, f_2, f_3)_i$ where $D(f_1) = D(f_2) \in [1, 100]$ and $D(f_3) \in [1, 50]$.

$R_1 = (f_1 \in [25, 50] \wedge f_2 \in [40, 60] \wedge f_3 \in [5, 25]) \rightarrow accept$
$R_2 = (f_1 \in [35, 70] \wedge f_2 \in [30, 90] \wedge f_3 \in [10, 25]) \rightarrow accept$
$R_3 = (f_1 \in [1, 100] \wedge f_2 \in [1, 100] \wedge f_3 \in [1, 50]) \rightarrow deny$

R_1 and R_2 are redundant because any packet can match both rules which have the same actions (*accept*). Furthermore, R_1 and R_2 also conflict with R_3 because both R_1 and R_2 are subsets of R_3 while they are different actions. One typical solution to resolve such conflicts, that is, firewalls choose a rule which matches with the packet p_i being considered first, called the first-match approach. The firewall rules anomalies can be classified into six [1], [2] types representing each anomaly by the theorems:

Shadow anomaly: R_x is shadowed by R_y, if and only if their intersection is equal to R_x and there are different actions illustrated in Figure 1(a).

$$R_x = C_x \rightarrow A_x, R_y = C_y \rightarrow A_y$$
$$R_x, R_y \in R_{db} \wedge \neg(A_x \leftrightarrow A_y) \wedge (C_x \cap C_y = C_x) \quad (3)$$

where R_{db} is a database of all rules, and R_y is the rule executed before R_x.

Correlation anomaly: R_x and R_y in R_{db} are correlated if their intersection is not equal to $\emptyset$, $R_x - R_y \neq \emptyset, R_y - R_x \neq \emptyset$, and they have different actions represented in Figure 1(b).

$$R_x = C_x \rightarrow A_x, R_y = C_y \rightarrow A_y$$
$$R_x, R_y \in R_{db} \wedge \neg(A_x \leftrightarrow A_y) \wedge (C_x \cap C_y \neq \emptyset) \wedge (C_x - C_y \neq \emptyset) \wedge (C_y \cap C_x \neq \emptyset) \quad (4)$$

where R_{db} is a database of all rules, and R_y is the rule executed before R_x.

Generalization anomaly: R_x is generalized by R_y, if and only if their intersection is equal to R_y, and there are different actions (Figure 1(c)), where R_y is the rule matched before R_x.

$$R_x = C_x \rightarrow A_x, R_y = C_y \rightarrow A_y$$
$$R_x, R_y \in R_{db} \wedge \neg(A_x \leftrightarrow A_y) \wedge (C_x \cap C_y \neq C_y) \quad (5)$$

Redundancy anomaly: R_x is redundant to R_y, if and only if their intersection is not equal to $\emptyset$, and they are have same actions (Figure 1(d)).

$$R_x = C_x \rightarrow A_x, R_y = C_y \rightarrow A_y$$
$$R_x, R_y \in R_{db} \wedge (A_x \leftrightarrow A_y) \wedge (C_x \cap C_y \neq \emptyset) \quad (6)$$

Irrelevance anomaly: Irrelevance anomaly occurs in the firewall if no packets can be matched against all rules in the firewall. This anomaly is caused by the administrator's misunderstanding about the network connections.

$$R_x = C_x \rightarrow A_x, R_y = C_y \rightarrow A_y, R_n = C_n \rightarrow A_n$$
$$R_1, R_2, \ldots, R_n \in R_{db} \wedge P_i \not\subset (C_i, \ldots, C_n) \quad (7)$$

where P_i is an IP packet executed by the firewall.

Semantics loss anomaly: The semantics loss represented by [2]

occurs when R_x and R_y are merged to R_z by the meaning of both old rules that have been changed or replaced by a new meaning. This anomaly is mostly caused by redundant rules as shown in Figure 1(*f*).

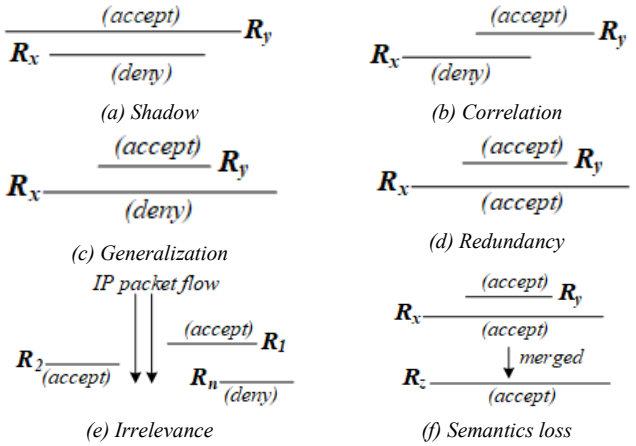


Figure 1: Firewall rule anomalies

Min-Max feature scaling [17] (also known as data normalization) is the standard method used to adjust the range of data. Since the range of data values may be very different, it is therefore a necessary step in data preprocessing before processing in the next step. It is normally used to resize any data range into the range [0, 1], called unity-based normalization. Also, it can normalize the finite range of values in the dataset between any arbitrary points t_{max} and t_{min} as the following equation.

$$m' = \frac{m - r_{min}}{r_{max} - r_{min}} \times t_{max} - t_{min} + t_{min} \quad (8)$$

Let m' denotes the value being considered has been normalized by $m \in [r_{min}, r_{max}]$. r_{min} and r_{max} denote the minimum and maximum of the measurement range. $t_{min}$and t_{max} are the minimum and maximum of the target range to be scaled.

2.3. Bayes' Theorem

Let $A_1, A_2, \ldots, A_k$ be events that partition the sample space S, i.e., $S = A_1 \cup A_2 \cup \ldots \cup A_k$ and $A_i \cap A_j \neq \emptyset$ when $i \neq j$ and let B be an event on that space for which $P_r(B) > 0$. Then Bayes' theorem is:

$$P_r(A_j|B) = \frac{P_r(A_j)P_r(B|A_j)}{\sum_{j=1}^{k} P_r(A_j)P_r(B|A_j)} \quad (9)$$

2.4. Moving Average (MA)

$$SMA = \frac{A_1 + A_2 + A_3 + \ldots + A_n}{n} \quad (10)$$

where A_i is an average in period n, and n is the number of periods. EMA is a weighted average of the last n data, where the weighting decreases exponentially with each previous data per period. In other words, the formula gives greater weight to more recent data. The formula for the exponential moving average is the following.

$$EMA_t = [V_t \times \frac{s}{1+d}] + EMA_y \times [1 - \frac{s}{1+d}] \quad (11)$$

where EMA_t = EMA today, V_t = Value today, EMA_y = EMA yesterday, s = smoothing, d is the number of day.

2.5. Converting an IP Address to a non-negative Integer

The Internet Protocol Address (as known as IP Address) is a unique address that networking devices such as routers, switching, and computers use to identify themselves and communicate over other devices in computer networks. An IPv4 address (IP version 4) is equal to 32 bits, ranging from 0 to 2^{32} – 1 address space. It is usually divided into four parts, each part (8 bits = an octet) separated by a dot, e.g., $A_1.A_2.A_3.A_4$ where $A_{(1-4)} \in [0, 255]$. IPv4 address can be converted to any non-negative integer with the following equation.

$$IPv4' = A_1 \times 2^{24} + A_2 \times 2^{16} + A_3 \times 2^8 + A_4 \times 2^0 \quad (12)$$

where $IPv4'$ is a new IP address to be converted, for example, 1.2.3.4 will be convert to:

$$IPv4' = 1 \times 2^{24} + 2 \times 2^{16} + 3 \times 2^8 + 4 \times 2^0 = 16,909,060$$

2.6. Arithmetic Mean and Kappa Statistics

We use the average method ($\bar{x}$) to evaluate the administrator's satisfaction with the proposed firewall and use the Cohen's kappa coefficient ($\widehat{K_F}$) [19] to measure the interrater reliability as the following equations.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (13)$$

where $\bar{x}$ is an average (or arithmetic mean), n is the number of terms (e.g., the number of items or numbers being averaged), and x_i is the value of each individual item in the list of numbers being averaged.

$$\widehat{K_F} = \frac{\bar{P}_a - \bar{P}_e}{1 - \bar{P}_e} \quad (14)$$

where: $\bar{P}_a$ denotes the relative observed agreement among raters, $\bar{P}_e$ denotes the hypothetical probability of chance agreement as in the equation 15 and 16.

$$\bar{P}_a = \frac{1}{r}\sum_{i=1}^{r} Z_j, \bar{P}_e = \sum_{k=1}^{q} P_j^2 \quad (15)$$

$$Z_j = \frac{1}{m(m-1)}\left(\sum_{k=1}^{q} n_{ij}^2 - \sum_{k=1}^{q} n_{ij}\right), P_j = \frac{1}{rm}\sum_{i=1}^{r} n_{ij} \quad (16)$$

where m is the number of raters, the objective of the assessment is r, q is the type of information that needs to be evaluated (e.g., most satisfied, very satisfied, ..., least satisfied), and n_{ij} denotes the observed cell frequencies.

3. Key Contributions

As rule anomalies occur over firewalls, the decision-making power to resolve the anomalies mainly depends on the administrator's discretion. However, the decisions made often result in errors or loopholes over the existing rules, if admins

cannot entirely understand the relationship between conflict rules. Therefore, it is necessary to develop a decision support system for admins to assist decision-making during real-time anomaly detection. The system consists of four procedures:

1) Firstly, preparing various information to be ready before processing,
2) Analyzing and detecting the rule abnormalities by the Path Selection Tree (PST),
3) Calculating the probability (Bayesian) of each rule based on the frequency of packets matched against rules, evidence of creating rules, expertise on creating rules, and protocol priority to help admins decide before optimizing the rules,
4) Lastly, optimizing anomaly or conflict rules based on the probability.

4. The System Design

There are four steps in the system design as shown in Figure 2.

4.1. Preparing Rule-Based Firewall (Step 1)

The conditions (C_i) and decision (A_i) of each rule: Referring to R_i in the equation (2), in general, the members of C_i have five fields ($f_1 \wedge \ldots \wedge f_5$), where f_1 = source IP address (*SIP*), f_2 = destination IP address (*DIP*), f_3 = source port (*SP*), f_4 = destination port (*DP*) and f_5 = protocol (*PRO*) respectively as shown in Table 1. According to R_1 of Table 1, the preparation process of firewall rules begins with converting the IP addresses of f_1 and f_2 into a range of positive integers by equation (12). Hence, f_1 and f_2 are then converted into the following numbers: $f_1 \in [16909056, 16909066]$ and $f_2 \in [1, 256]$. The fields converted in the next order are f_3 and f_4 which contain the numbers ranging from 0 to $2^{16} - 1$: $f_3 \in [0, 65535]$ and $f_4 \in [80, 80]$, where $*$ means all numbers in such domain. The field f_5 is both TCP and UDP protocol, thus they are translated to: $f_5 \in \{6, 17\}$, where TCP = 6 and UDP = 17. In case of decision field (A_i), it is changed to a positive integer either 0 or 1, such as $A_i \in \{0, 1\}$, where *accept* = 1 and *deny* = 0. As a result of all these calculations, R_1 is converted to:

$$R_1: (f_1 \in [16909056, 16909066] \wedge f_2 \in [1, 256] \wedge f_3 \in [0, 65535] \wedge f_4 \in [80,80] \wedge f_5 \in \{6, 17\}) \rightarrow 1$$

Table 1: The basic member fields of C_i and A_i

R_i	$C_i(f_1 \wedge f_2 \wedge f_3 \wedge f_4 \wedge f_5)$					A_i
R_1	f_1(SIP)	f_2(DIP)	f_3(SP)	f_4(DP)	f_5(PRO)	decision
R_1'	1.2.3.0-1.2.3.10	0.0.0.1-0.0.1.0	*	80	TCP, UDP	*accept*

Calculating Probability of Extra Fields of Each rule: To determine the probability of each rule in this model, there are four additional fields added including the frequency of packets matching against rules (*FPM*), evidence of creating rules (*ECR*), the expertise of rules creator (*ERC*) and protocol priority (*PRI*). Let $P(e_1)$, $P(e_2)$, $P(e_3)$ and $P(e_4)$ are the probability of *FPM*, *ECR*, *ERC*, and *PRI* respectively. Therefore, the sum of the probability of rule R_i is equal to the equation (17).

$$P(E_i) = \sum_{j=1}^{4} P(e_j) \quad (17)$$

Where $P(E_i)$ is the probability of R_i. For example, the information of extra fields of R_i as shown in Table 2. From R_1 in Table 2., matching rate ($FPM : e_1$) between packets and R_1 is equal to 2,125 times. e_2(*ECR*), e_3(*ERC*) and e_4(*PRI*) are 3, 2 and 4 respectively, explained more details in the next section. These four extra fields are calculated to be the probability pattern in which data is in the range from 0.0 to 1.0 ($R_1'$by the Min-Max Feature Scaling in the equation (8).

In the e_1 case: it is the frequency of the packets matching against any rules over the firewall, the counting process starts from the time at which a rule had been created and continues until the present time. For example, if the maximum and minimum number of matching of any rules in the firewall are 5,000 and 1,200 times respectively, then e_1' here is equal to:

$$e_1' = \frac{m - r_{min}}{r_{max} - r_{min}} \times t_{max} - t_{min} + t_{min}$$
$$= \frac{2{,}125 - 1{,}200}{5{,}000 - 1{,}200} \times (1.0 - 0.0) + 0.0 = 0.24$$

where m = 2,125, r_{min} = 1,200, r_{max} = 5,000, t_{min} = 0.0 and t_{max} is equal to 1.0. However, recording e_1 in the firewall requires the equation (10 and 11) to make the data smoother since the recorded data may be a swinging data caused by network attacks, user behaviors, or network usage during rush hours, etc. The period for calculating data with EMA method will depend on the suitability of each organization. For this research, e_1 is recorded every hour per day as in the following example:

Given e_1 of R_1 for each hour per a day: 1300, 1500, 1200, 1300, 1400, 1500, 1800, 4500, 6000, 6300, 5500, 1000, 2400, 2800, 2600, 2600, 2400, 1900, 1500, 1200, 1000, 800, 700, 600 times, then it can calculate the EMA of e_1 using five hours in the past as follows.

$$\text{SMA of 5th hour} = \frac{(1{,}300+1{,}500+1{,}200+1{,}300+1{,}400)}{5} = 1{,}340.0$$
$$\text{EMA of 6}^{\text{th}}\text{ hour} = 1{,}340.0 + \frac{2}{(5+1)} \times (1{,}500 - 1{,}340.0) = 1{,}393.3$$
$$\text{EMA of 7}^{\text{th}}\text{ hour} = 1{,}393.3 + \frac{2}{(5+1)} \times (1{,}800 - 1{,}393.3) = 1{,}528.8$$
$$\text{EMA of 8}^{\text{th}}\text{ hour} = 1528.8 + \frac{2}{(5+1)} \times (4{,}500 - 1528.8) = 2{,}519.2$$
$$\text{EMA of 9}^{\text{th}}\text{ hour} = 2{,}519.2 + \frac{2}{(5+1)} \times (6{,}000 - 2{,}519.2) = 3{,}679.5$$
$$\vdots$$
$$\text{EMA of 24}^{\text{th}}\text{ hour} = 1{,}094.2 + \frac{2}{(5+1)} \times (600 - 1{,}094.2) = 929.5$$

Calculating the EMA for twenty-four hours as shown in Figure 3, the results will be calculated with SMA again. In order to find the average value of each day, which the calculated value will be recorded as e_1' in Table 2. To calculate the average value of e_1 of each day as follows.

$$e_1 = \frac{\sum_{i=5}^{n} EMA_i}{n - (hours\ used\ in\ the\ past)}$$

$$= \frac{1{,}340.0 + 1{,}393.3 + \ldots + 929.5}{24 - 5} = 2{,}463.0$$

where n is the number of hours per day subtracts by the number of hours used in the past.

Table 2: Four extra fields of R_i

R_i	$C_i(f_1 \wedge f_2 \wedge f_3 \wedge f_4 \wedge f_5)$			
R_1	e_1(FPM)	e_2(ECR)	e_3(ERC)	e_4(PRI)
R_1	2,125	3	2	4
R_1'	e_1'=0.24	e_2'=1.0	e_3'=0.66	e_4'=0.62

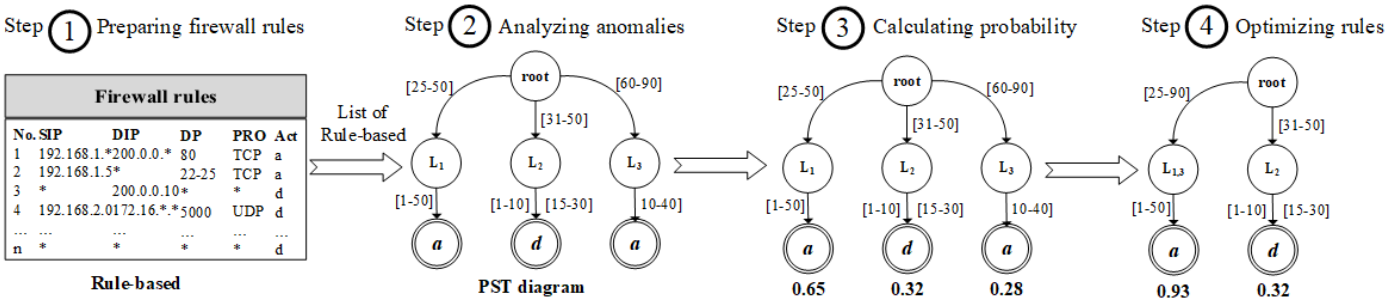


Figure 2: Overview of the system design

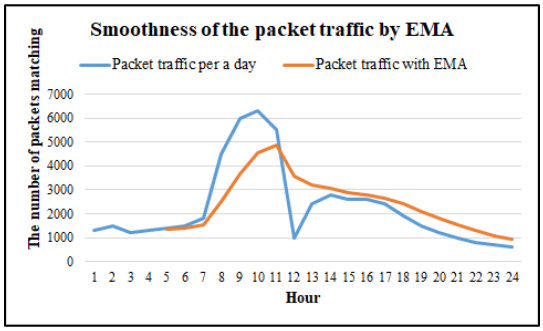


Figure 3: Adjusting raw packets traffic to be more smooth by EMA

***In the e_2 case*:** it refers to documents or pieces of paper used to confirm that such rules are approved to create them. In this paper, for example, the evidence for creating rules is divided into four levels: there is no evidence for approval, a firewall administrator is an approver, the head of the department is the approver, and approval is made by the owner of the organization. By dividing the weight of evidence according to the priority of the document approver: no evidence = 0, an administrator = 1, a head of the department = 2 and an owner of the organization = 3 respectively. If the weight of the document obtained is calculated by the Min-Max equation (8), the result will be the $e_2{}'$. Let the owner of the organization be approved to create the rule R_1, the result of the calculation is equal to:

$$e_2{}' = \frac{m - r_{min}}{r_{max} - r_{min}} \times t_{max} - t_{min} + t_{min}$$
$$= \frac{3-0}{3-0} \times (1.0 - 0.0) + 0.0 = 1.0$$

where $m = 3$, $r_{min} = 0$, $r_{max} = 3$, $t_{min} = 0.0$ and t_{max} is equal to 1.0.

In the e_3 case: similar in the case of evidence for creating rules, the expertise in creating rules is also divided into four levels: newbie admins, normal admins with sufficient expertise, professional admins and very expert administrators. The newbie admins mean those who have recently been assigned to configure the firewall system and have the least experience. When they have configured the firewall for a while, they will be more proficient, which should have at least 3 - 5 years of working hours, called normal admins. For those who have a lot of experience and training or firewall customization, with working hours of 5 – 10 years, they will be professional administrators. Finally, those who have received a lot of training and certificates about firewalls will be considered expert admins. From the statistics, those who are very skilled will be able to analyze and design firewall rules to minimize errors as well. In this paper, therefore, determines the weight of the following expertise e_3: newbie = 0, normal = 1, professional = 2 and very expertise administrator = 3. Let professional admins create the rule R_1, the result is calculated as.

$$e_3{}' = \frac{2-0}{3-0} \times (1.0 - 0.0) + 0.0 = 0.66$$

***In the e_4 case*:** the protocols communicating on the computer networks are always prioritized, such as video conferencing, and must be smooth throughout the meeting. On the other hand, sending electronic mail does not have an urgent need to be sent or received immediately. The protocol prioritization can be done depending on the policies of each organization. In this research, prioritization of the protocol is based on priorities from 3 GPP QoS Class Identification QCI categories [20] by IP Multimedia singling having the highest priority (1 = highest); Chat, FTP and P2P having the lowest priority (9 = lowest). From e_4 in Table 2, it is a teleconference application with a priority of 4.When processing in the form of probability using the Min-Max Scaling, the result is equal to:

$$e_4{}' = \frac{6-1}{9-1} \times (1.0 - 0.0) + 0.0 = 0.625$$

Where $m = 6$ (Teleconference = 4), $r_{min} = 1$, $r_{max} = 9$, $t_{min} = 0.0$ and t_{max} is equal to 1.0. Notice that the priority of the protocols calculated must always reverse priorities, such as from 9 to 1 and from 1 to 9. For example, the priority of 4 is reversed to 6. Last, in Table 3, 4. represents examples of firewall rules consisting of all anomalies as previously mentioned, and these rules will be processed in the next step. Extra fields of each rule, when passing the data preparation process, it will produce the following results $R_i \rightarrow R_i{}'$.

4.2. Analyzing and Detecting Anomalies (Step 2)

In this phase, the rules from the 1st step are used to build a tree structure, called the Path Selection Tree (PST), to analyze the anomalies. The algorithm begins with the creation of the root node of PST. After that, field f_1 of the first rule is created as the first node on the tree, namely SIP_1 as shown in Figure 4(a). In this node records source IP addresses of R_1 to be <R_1:[1, 100]>, where $f_1 \in$ [1, 100]. The next node (DIP_1) stores the range of destination IP addresses (f_2) of R_1 ranging from 1 to 100. Next, it is the node that records source port ranges from 0 to 65535, called SP_1. The next node as DP_1, this node contains a group of destination ports f_4 between 80 and 85 (<R_1 : [80, 85]>). The final field f_5 of R_1 as PRO_1 which keeps the range of TCP and UDP protocols. In the decision field, a bottom rectangular box in the tree, contains an acceptance decision (1) of R_1. At the end of the tree, it records what rules members of this path are like <R_1>.

Table 3: Examples of firewall rules (R_i)

R_1: $f_1 \in [1,100] \wedge f_2 \in [1,100] \wedge f_3 \in [0,65535] \wedge f_4 \in [80,85] \wedge f_5 \in \{6,17\} \rightarrow 1$
R_2: $f_1 \in [10,50] \wedge f_2 \in [20,60] \wedge f_3 \in [0,65535] \wedge f_4 \in [80,80] \wedge f_5 \in \{6,17\} \rightarrow 0$
R_3: $f_1 \in [20,40] \wedge f_2 \in [30,70] \wedge f_3 \in [0,65535] \wedge f_4 \in [80,90] \wedge f_5 \in \{6,17\} \rightarrow 0$
R_4: $f_1 \in [20,30] \wedge f_2 \in [20,30] \wedge f_3 \in [0,65535] \wedge f_4 \in [80,82] \wedge f_5 \in \{6,17\} \rightarrow 0$
R_5: $f_1 \in [1,100] \wedge f_2 \in [1,100] \wedge f_3 \in [0,65535] \wedge f_4 \in [30,90] \wedge f_5 \in \{6,17\} \rightarrow 1$
R_6: $f_1 \in [1,100] \wedge f_2 \in [1,100] \wedge f_3 \in [0,65535] \wedge f_4 \in [0,65535] \wedge f_5 \in \{6,17\} \rightarrow 0$

Table 4: Examples of extra fields (R_i')

R_1: $e_1 = 2{,}500, e_2 = 1, e_3 = 2, e_4 = 6 \rightarrow R_1'$: $e_1' = 0.26, e_2' = 0.33, e_3' = 0.67, e_4' = 0.67$
R_2: $e_1 = 1{,}500, e_2 = 3, e_3 = 2, e_4 = 3 \rightarrow R_2'$: $e_1' = 0.06, e_2' = 1.0, e_3' = 1.0, e_4' = 0.33$
R_3: $e_1 = 2{,}000, e_2 = 2, e_3 = 1, e_4 = 8 \rightarrow R_3'$: $e_1' = 0.16, e_2' = 0.67, e_3' = 0.33, e_4' = 0.89$
R_4: $e_1 = 3{,}200, e_2 = 1, e_3 = 2, e_4 = 5 \rightarrow R_4'$: $e_1' = 0.40, e_2' = 0.33, e_3' = 0.67, e_4' = 0.56$
R_5: $e_1 = 1{,}200, e_2 = 3, e_3 = 0, e_4 = 2 \rightarrow R_5'$: $e_1' = 0.0, e_2' = 1.0, e_3' = 0.0, e_4' = 0.22$
R_6: $e_1 = 500, e_2 = 0, e_3 = 3, e_4 = 9 \rightarrow R_6'$: $e_1' = 0.76, e_2' = 0.0, e_3' = 1.0, e_4' = 1.0$

In the next order, the second rule R_2 is imported into PST as illustrated in Figure 4(b). Firstly, f_1 of $R_2 \subset f_1$ of R_1, thus R_2 (f_1) uses the same route as R_1 (f_1) and also records <R_2:[10, 50]> into the SIP_1 node. Likewise, R_2 (f_2) $\subset R_1$ (f_2), it is also recorded to the same node (DIP_1 = <R_1:[1, 100], R_2:[20, 60]>), and travels over the same route as R_1. Similar to R_2 (f_3), it is equal to R_1 (f_3), hence R_2 (f_3) is appended in the SP_1 node to be <R_1, R_2:[0, 65536]>. In case of R_2 (f_4) and R_1 (f_4), R_2 (f_4) is a subset of R_1 (f_4), so the data of DP_1 is updated as <R_1:[80, 85], R_2:[80, 80]> as well as PRO_1 updated to <R_1, R_2:{6, 17}>. On the other hand, the decision of R_1 and R_2 are not the same, so the decision path must be separated from each other, where <R_1> = 1, <R_2> = 0. For inserting R_3 (Figure 4(c)) into PST, there is not much difference from inserting R_2, it is slightly different in the position of the protocol level in the tree. Since $R_3(f_4)$ is a superset of $R_1(f_4)$ and $R_2(f_4)$, some destination ports of $R_3(f_4)$ have to be separated into another node of the tree, namely DP_2, which stores the destination ports ranging from 86 to 90 ($R_3(f_4)$ - $R_1(f_4)$) like <R_3: [86, 90] >. Remaining destination ports are combined with DIP_1 in the first path together with R_1 and R_2 as <R_1, R_3:[80, 85], R_2:[80, 80]>. The decision of R_3 is not allowed in both paths, where <R_3> = 0. Remaining firewall rules (R_4, R_5, R_6) will be executed like the previous rules (R_1, R_2, R_3). If all rules have been implemented successfully over PST, the results are shown in Figure 5.

In the process of checking the rule anomalies, the algorithm uses the information recorded on each node to detect anomalies by using the Cartesian product of all nodes separated from the protocol layer (PRO_i) and looking back from the bottom to the root as follows.

***Group 1*:** path number 1 and 2 under the node PRO_1

$$CP(< R_1, R_5 >) = (R_1, R_5) \quad (18)$$

$$CP(< R_2, R_3, R_4, R_6 >) = (R_2, R_3), (R_2, R_4), (R_2, R_6), (R_3, R_4), (R_3, R_6), (R_4, R_6) \quad (19)$$

$$CP(< R_1, R_5 >, < R_2, R_3, R_4, R_6 >) = (R_1, R_2), (R_1, R_3), (R_1, R_4), (R_1, R_6), (R_2, R_5), (R_3, R_5), (R_4, R_5), (R_5, R_6) \quad (20)$$

***Group 2*:** path number 3 and 4 under the node PRO_2

$$CP(< R_3, R_6 >) = (R_3, R_6) \quad (21)$$

$$CP(< R_3, R_6 >, < R_5 >) = (R_3, R_5), (R_5, R_6) \quad (22)$$

***Group 3*:** path number 5 and 6 under the node PRO_3

$$CP(< R_5, R_6 >) = (R_5, R_6) \quad (23)$$

where CP is the Cartesian product. The results of each pair of the Cartesian product are calculated by equations (3) to (7) to find out what kind of anomalies they are. For example, in the equation (18) of group 1, (R_1, R_5) has the same decision (decision = 1), so it is executed by the equation (6). The result of the execution is a redundant anomaly. Next example, in equation (19), they consist of $(R_2, R_3), (R_2, R_4), (R_2, R_6)$, $(R_3, R_4), (R_3, R_6)$ and (R_4, R_6) by every pair of rules has the same decisions, thus all is executed by the equation (6) as same as the equation (18). Results of Cartesian product in the equation (20): $(R_1, R_2), (R_1, R_3), (R_1, R_4), (R_1, R_6)$, (R_2, R_5), $(R_3, R_5), (R_4, R_5)$ and (R_5, R_6) , and every pair has different decisions, so they are executed by equations (3), (4) and (5) respectively. The calculated results: (R_2, R_3) = Redundancy and Semantics loss (Executed by equation (6)), (R_2, R_4) = Redundancy and Semantics loss (6), (R_2, R_6) = Redundancy and Semantics loss (6), (R_3, R_4) = Redundancy and Semantics loss (6), (R_3, R_6) = Redundancy and Semantics loss (6), (R_4, R_6) = Redundancy and Semantics loss (6), (R_1, R_2) = Shadowing (3), (R_1, R_3) = Correlation (4), (R_1, R_4) = Shadowing (3), (R_1, R_6) = Generalization (5), (R_2, R_5) = Generalization (5), (R_3, R_5) = Generalization (5), (R_4, R_5) = Generalization (5), (R_5, R_6) = Generalization (5). The results obtained from the calculations of group number 2 and 3 in equation 21 to 23: (R_3, R_6) = Redundancy and Semantics loss (Executed by equation (6)), (R_3, R_5) = Generalization (5), (R_5, R_6) = Generalization (5).

Losing the meaning of rules always occurs by redundant rules, thus all members in equations 18, 19 and 21 are possible to be the semantics loss as well.

4.3. Calculating Probability of Each Path of PST (Step 3)

The PST obtained from the previous steps is used to calculate the probability of each path in order to advise administrators to make decisions about optimizing firewall rules effectively, which has the following steps. Let R as a firewall rule, e as an attribute field of a rule, and S is a sample space, then the conditional probability of R given e is the equation (24) and shown in Figure 6.

$$P(R|e) = \frac{P(R|e)}{P(e)} \quad (24)$$

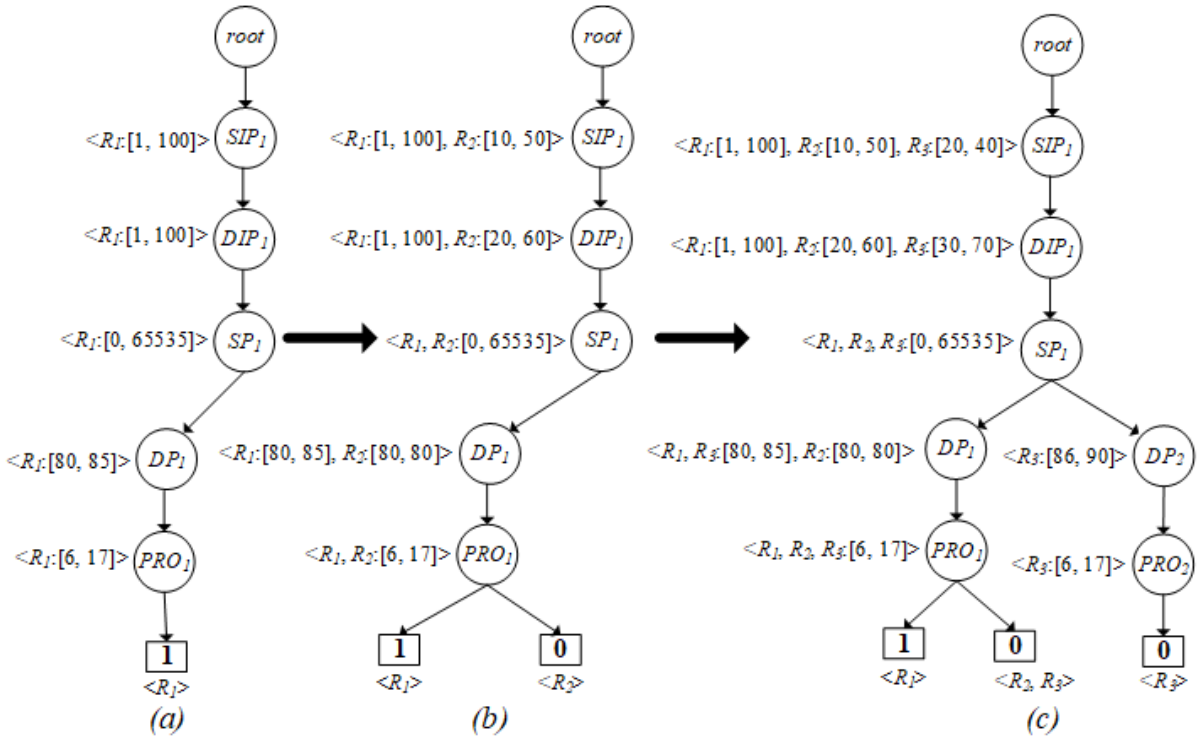


Figure 4: Creating rule $R_1(a)$, $R_2(b)$ and $R_3(c)$ into PST

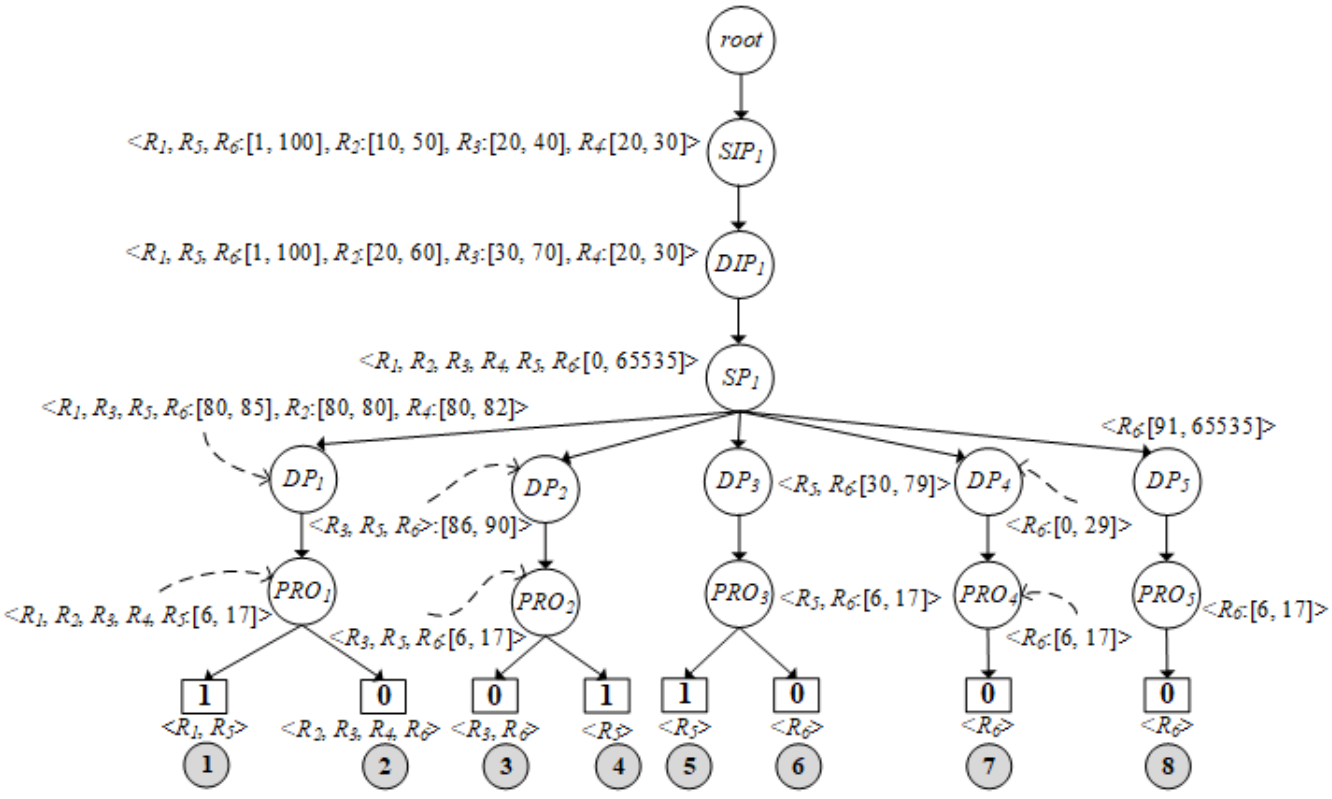


Figure 5: Complete PST structure after compiling all rules

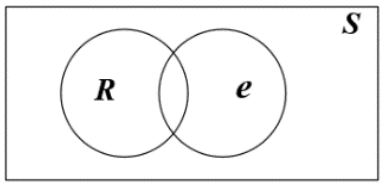


Figure 6: Conditional probability of R given e in Venn diagram

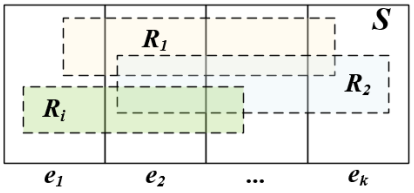


Figure 7: Conditional probability of R_i given e_k

According to Figure 7., given R_i as any rule, e_k is any attribute (Extra field) of R_i:

$$e_i \cap e_k = \emptyset;\ \forall_i, k; i \neq k$$

$$e_1 \cup e_2 \cup e_3 \cup \ldots \cup e_k = S$$

$$P(R_i) = P(R_i \cap e_1) \cup P(R_i \cap e_2) \cup \ldots \cup P(R_i \cap e_k) \quad (25)$$

From the equation (24), $P(R|e) = \frac{P(R|e)}{P(e)}$, so $P(R|e)$: $P(e)P(R|e)$ or $P(R)P(e|R)$. Since we already know the value of the $P(e)$, then we choose $P(R \cap e) = P(e)P(R|e)$ and represent i and k into the equation as follows.

$$P(R_i \cap e_k) = P(e_k)P(R_i|e_k) \quad (26)$$

Applying the equation (26) instead of the equation (25):

$$P(R_i) = P(e_1)P(R_i|e_1) \cup (e_2)P(R_i|e_2) \cup \ldots \cup (e_k)P(R_i|e_k) \quad (27)$$

$$P(R_i) = \sum_{i,k=1}^{n}(e_k)P(R_i|e_k) \quad (28)$$

Given e_k is any property considering when giving $P(R_i)$ on the firewall. Finally, we can substitute this into Bayes' rule (Equation 9) from above to obtain an alternative version of Bayes' rule, which is used heavily in Bayesian inference:

$$P(e_k|R_i) = \frac{P(R_i \cap e_k)}{P(R_i)} = \frac{P(e_k)P(R_i|e_k)}{\sum_{i,k=1}^{n}(e_k)P(R_i|e_k)} \quad (29)$$

From examples of the property fields (Extra fields) in Table 4., there are four fields (e_1', e_2', e_3', e_4'), where e_1' = the frequency of packets matching against rules (*FPM*), e_2' = evidence of creating rules (*ECR*), e_3' = expertise of rules creator (*ERC*) and e_4' = protocol priority (*PRI*). Assuming the weight of each feature is equal, so e_1', e_2', e_3' and e_4' are equal to 25% (0.25). Substituting various values in equations (28) and (29), the calculated results:

$$P(e_1') = 0.25, P(e_2') = 0.25, P(e_3') = 0.25, P(e_4') = 0.25$$

In case of $\boldsymbol{P(R_1')}$:

$$P(R_1|e_1') = 0.26, P(R_1|e_2') = 0.33, P(R_1|e_3') = 0.67, P(R_1|e_4') = 0.67$$

$$\begin{aligned} P(R_i) = {} & P(e_1')P(R_1|e_1') + P(e_2')P(R_1|e_2') + P(e_3')P(R_1|e_3') \\ & + P(e_4')P(R_1|e_4') + P(e_1')P(R_2|e_1') + P(e_2')P(R_2|e_2') \\ & + P(e_3')P(R_2|e_3') + P(e_4')P(R_2|e_4') + \cdots \\ & + P(e_1')P(R_6|e_1') + P(e_2')P(R_6|e_2') + P(e_3')P(R_6|e_3') \\ & + P(e_4')P(R_6|e_4') \end{aligned}$$

$$\begin{aligned} \boldsymbol{P(R_i)} = {} & \big((0.25 \times 0.26) + (0.25 \times 0.33) + (0.25 \times 0.67) \\ & + (0.25 \times 0.67)\big) \\ & + \big((0.25 \times 0.06) + (0.25 \times 1) + (0.25 \times 1) \\ & + (0.25 \times 0.33)\big) + \ldots + \big((0.25 \times 0.76) \\ & + (0.25 \times 0) + (0.25 \times 1) + (0.25 \times 1)\big) = \boldsymbol{3.07} \end{aligned}$$

$$P(e_1'|R_1) = \frac{P(e_1')P(R_1|e_1')}{P(R_i)} = \frac{0.25 \times 0.26}{3.07} = 0.0211$$

$$P(e_2'|R_1) = \frac{P(e_2')P(R_1|e_2')}{P(R_i)} = \frac{0.25 \times 0.33}{3.07} = 0.0268$$

$$P(e_3'|R_1) = \frac{P(e_3')P(R_1|e_3')}{P(R_i)} = \frac{0.25 \times 0.67}{3.07} = 0.0545$$

$$P(e_4'|R_1) = \frac{P(e_4')P(R_1|e_4')}{P(R_i)} = \frac{0.25 \times 0.67}{3.07} = 0.0545$$

$$\therefore \boldsymbol{P(R_1')} = P(e_1'|R_1) + P(e_2'|R_1) + P(e_3'|R_1) + P(e_4'|R_1) = \boldsymbol{0.157}$$

In case of $\boldsymbol{P(R_2')}$:

$$P(R_2|e_1') = 0.06, P(R_2|e_2') = 1.0, P(R_2|e_3') = 1.0, P(R_2|e_4') = 0.33$$

$$P(e_1'|R_2) = 0.0488, P(e_2'|R_2) = 0.0814, P(e_3'|R_2) = 0.0814, P(e_4'|R_2) = 0.0268$$

$$\therefore \boldsymbol{P(R_2')} = P(e_1'|R_2) + P(e_2'|R_2) + P(e_3'|R_2) + P(e_4'|R_2) = \boldsymbol{0.194}$$

In case of $\boldsymbol{P(R_3')}$:

$$P(R_3|e_1') = 0.16, P(R_3|e_2') = 0.67, P(R_3|e_3') = 0.33, P(R_3|e_4') = 0.89$$

$$P(e_1'|R_3) = 0.0130, P(e_2'|R_3) = 0.0545, P(e_3'|R_3) = 0.0268, P(e_4'|R_3) = 0.0724$$

$$\therefore \boldsymbol{P(R_3')} = P(e_1'|R_3) + P(e_2'|R_3) + P(e_3'|R_3) + P(e_4'|R_3) = \boldsymbol{0.166}$$

In case of $\boldsymbol{P(R_4')}$:

$$P(R_4|e_1') = 0.40, P(R_4|e_2') = 0.33, P(R_4|e_3') = 0.67, P(R_4|e_4') = 0.56$$

$$P(e_1'|R_4) = 0.0325, P(e_2'|R_4) = 0.0268, P(e_3'|R_4) = 0.0545, P(e_4'|R_4) = 0.0456$$

$$\therefore \boldsymbol{P(R_4')} = P(e_1'|R_4) + P(e_2'|R_4) + P(e_3'|R_4) + P(e_4'|R_4) = \boldsymbol{0.159}$$

In case of $\boldsymbol{P(R_5')}$:

$$P(R_5|e_1') = 0.0, P(R_5|e_2') = 1.0, P(R_5|e_3') = 0.0, P(R_5|e_4') = 0.22$$

$$P(e_1'|R_5) = 0.0, P(e_2'|R_5) = 0.0814, P(e_3'|R_5) = 0.0, P(e_4'|R_5) = 0.0179$$

$$\therefore \boldsymbol{P(R_5')} = P(e_1'|R_5) + P(e_2'|R_5) + P(e_3'|R_5) + P(e_4'|R_5) = \boldsymbol{0.099}$$

In case of $\boldsymbol{P(R_6')}$:

$$P(R_6|e_1') = 0.76, P(R_6|e_2') = 0.0, P(R_6|e_3') = 1.0, P(R_6|e_4') = 1.0$$

$$P(e_1'|R_6) = 0.0618, P(e_2'|R_6) = 0.0, P(e_3'|R_6) = 0.0814, P(e_4'|R_6) = 0.0814$$

$$\therefore \boldsymbol{P(R_6')} = P(e_1'|R_6) + P(e_2'|R_6) + P(e_3'|R_6) + P(e_4'|R_6) = \boldsymbol{0.224}$$

According to the weight of each rule property, admins can adjust the weight of each property as needed, such as $P(e_1') = 35\%\ (0.35), P(e_2') = 15\%\ (0.15), P(e_3') = 25\%\ (0.25)$ and $P(e_4') = 25\%\ (0.25)$ depending on each organization to give weight to their properties. After calculating all the probability values successfully, the algorithm inserts these probabilities to each path of PST as shown in Figure 8.

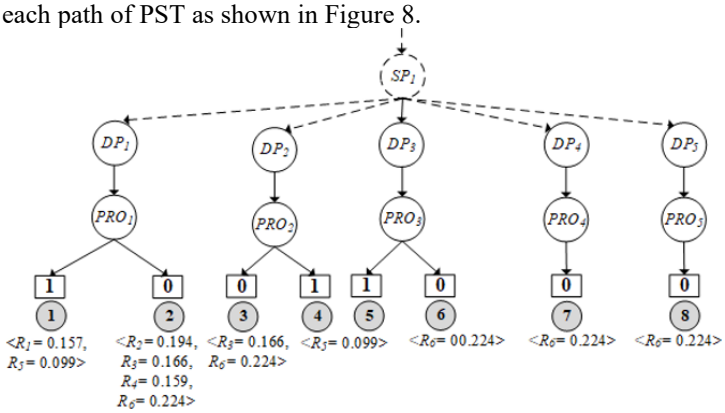


Figure 8: Inserting probability of each R_i into PST

4.4. Optimizing Rule Anomalies (Last step)

Anomalies occurred over firewall rules have different solutions, for example, the redundant anomaly is solved by merging the rules together. However, this method may result in semantics loss instead. Other anomalies such as shadowing, correlation, and generalization, should not use the merging method because their decisions are different. Sometimes, administrators choose to resolve problems by switching rules, but they are not sure what will happen in the future. Therefore, this research uses the calculated probability in each rule to help administrators decide how to proceed with anomalies to achieve maximum efficiency and reasonableness. For example, the path number 1 of Figure 5., R_1 and R_5 are redundancy. If admins decide to combine the two rules together, the result is.

$$R_1 \Phi\, R_5 = R_1(f_1) \cup R_5(f_1) \wedge R_1(f_2) \cup R_5(f_2) \wedge \ldots \wedge R_1(f_5) \cup R_5(f_5)$$

$$R_1 \Phi\, R_5 = ([1,100] \cup [1,100]) \wedge ([1,100] \cup [1,100]) \wedge ([0,65535] \cup [0,65535]) \wedge ([80,85] \cup [80,85]) \wedge (\{6,117\} \cup \{6,17\})$$

$$R_{new} = f_1 \in [1,100] \wedge f_2 \in [1,100] \wedge f_3 \in [0,65535] \wedge f_4 \in [80,85] \wedge f_5 \in \{6,17\} \rightarrow 1$$

The property fields of R_1: $e_1 = 2500$, $e_2 = 1$, $e_3 = 2$, $e_4 = 6$; and R_5 are $e_1 = 1200$, $e_2 = 3$, $e_3 = 0$, $e_4 = 2$. Thus, $R_1(e_i)\ \Phi\ R_5(e_i)$:

$$R_1(e_i)\ \Phi\ R_5(e_i) = (R_1(e_1) + R_5(e_1)), Max(R_1(e_2), R_5(e_2)), Max(R_1(e_3), R_5(e_3)), Max(R_1(e_4), R_5(e_4))$$

$$R_1(e_i)\ \Phi\ R_5(e_i) = (R_1(2500) + R_5(1200)), Max(R_1(1), R_5(3)), Max(R_1(2), R_5(0)), Max(R_1(6), R_5(2))$$

$$R_{1_{new(e)}} \rightarrow e_1 = 3700, e_2 = 3, e_3 = 2, e_4 = 6$$

where $R_{1_{new}}$ as a new merged rule, Φ as a merging operator for the same decisions, Max is a function calculated the maximum value. In the same way as (R_2, R_3), (R_2, R_4),..., and (R_3, R_6), which are a redundant conflict, so they can solve the problem by combining rules like (R_1, R_5). The methods of resolving the remaining anomalies (Shadowing, Correlation, and Generalization) can be done in three ways: merging, swapping and removing rules. Nevertheless, admins must be highly skilled and aware of the consequences, almost all researchers do not recommend using these methods and often push the burden to the discretion of administrators instead. If the admins choose one of three methods, they can do by checking the probabilities of each rule. If the probability of any rule is the highest, it means that admins have the opportunity to resolve anomalies to be more effective. For example, (R_1, R_2) is the shadowing anomaly. If admins need to delete, merge or swap rules, they should give priority to R_2 rather than R_1 because R_2 is a higher probability (R_1 = 0.157, R_2 = 0.194) as shown in Figure 9.

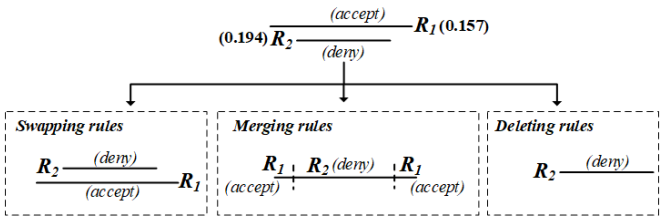


Figure 9: Resolving shadowing by swapping, merging and deleting rules

Updating property of R_1 and R_2 is not necessary in case of swapping and deleting rules, but in the case of merging, there are the following details.

$$R_2 \Theta\, R_1 = R_2(f_1) - R_1(f_1) \wedge R_2(f_2) - R_1(f_2) \wedge \ldots \wedge R_2(f_5) - R_1(f_5)$$

$$R_2{}' = f_1 \in [10,50] \wedge f_2 \in [20,60] \wedge f_3 \in [0,65535] \wedge f_4 \in [80,80] \wedge f_5 \in \{6,17\} \rightarrow 0$$

$$R_1{}' = f_1 \in [1,9], [51,100] \wedge f_2 \in [1,19], [61,100] \wedge f_3 \in [0,65535] \wedge f_4 \in [81,85] \wedge f_5 \in \{6,17\} \rightarrow 1$$

where Θ is a merging operator for different decisions. Although, $R_2(f_3) - R_1(f_3)$ and $R_2(f_5) - R_1(f_5)$ are equal to $\emptyset$. However, for this model, both are not equal to $\emptyset$ because they share the same path in the tree. The property fields of R_1: $e_1 = 2500$, $e_2 = 1$, $e_3 = 2$, $e_4 = 6$; and R_2 are $e_1 = 1500$, $e_2 = 3$, $e_3 = 2$, $e_4 = 3$. Thus, $R_1(e_i)\ \Theta\ R_2(e_i)$:

$$R_2(e_i)\ \Theta\ R_1(e_i) = (R_2(1500) + R_1(2500)), Max(R_2(3), R_1(1)), Max(R_2(2), R_1(2)), Max(R_2(3), R_1(6))$$

$$R_2{}'(e_i) \rightarrow e_1 = 4000, e_2 = 3, e_3 = 2, e_4 = 6$$

$$R_1{}'(e_i) \rightarrow e_1 = 2500, e_2 = 1, e_3 = 2, e_4 = 6$$

Note that while updating each conflict rule each time, the PST structure will be changed, which means that the probability has to be recalculated whenever when resolving conflicts.

5. Implementation and Performance Evaluation

PST uses the k-ary tree structure (also known as m-ary or k-way tree) to develop, so the processing speed to build the tree is $O(n)$, where n is the number of nodes of the given k-ary tree. The number of levels of the existing k-ary tree is L, the depth of the k-ary in the worst case is $N - 1$, where N is the number of nodes in a tree. The k-ary tree can also be stored in breadth-first order as an implicit data structure in pointer-based, each node would have an internal array for storing pointers to each of its m children. So, the space complexity of k-ary tree structure is $O(m \times n)$. Traversing the k-ary tree is very similar to binary tree traversal. Besides, the worst-case time complexity is $O(log_m n)$. In practically, PST is developed on Intel Core i7 2.3 GHz, 8 GB RAM. The software developed includes Python language version 3.7 (64-bit), Graphviz [21] and NetworkX version 1.11 [22] running on Linux kernel (Version 4.4). The proposed model is illustrated in Figure 10.

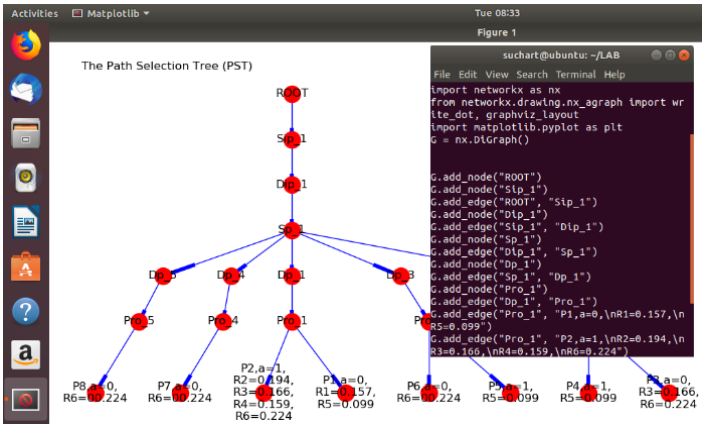


Figure 10: The developed PST structure over Linux operating system

In this paper, we used $\bar{x}$ (Equation 13) to evaluate the satisfaction of the firewall administrators for resolving firewall rule anomalies of both traditional (No recommendation system) and our proposed firewall (Recommendation system with probability). In which of the confidence test consists of ten scenarios to resolve anomalies and the total number of testers (Firewall expert) is five as shown in Table 5.

Referring to Table 5, the average ($\bar{x}$) of the five administrators' confidence for resolving ten scenarios of rule anomalies based on their skills for the traditional firewall is equal to 2.68; however, the average confidence of our proposed firewall which is a recommendation system based on probability is 4.16, which the confidence rate of the proposed firewall increased by 29.6% from the conventional firewall. In the case of evaluating reliability between raters, we apply Kappa statistics [19] in the equation (14) with the data from Table 5. The reliability value between the inter-raters of the conventional firewall is equal to 0.379, which means that the reliability is at a fair agreement as shown in Table 6. The reliability value between the assessors of the proposed firewall is equal to 0.510 (Moderate agreement), which means that the reliability increased from 37.9% to 51% significantly.

Table 5: Confidence tests to resolve anomalies of firewall rules

Firewall rule anomaly and resolving methods	**The confidence of five admins for resolving rule anomalies**									
	Traditional firewalls (No the recommendation system) Admins judged by their experience					Our proposed firewall (Use the recommendation system) Admins judged by probability				
	5	4	3	2	1	5	4	3	2	1
Shadowing										
- Merging rules	1	1	2	1	0	0	0	1	2	2
- Swapping rules	1	1	2	0	1	0	0	1	3	1
- Deleting rules	0	2	2	1	0	0	0	2	2	1
Correlation										
- Merging rules	1	2	2	0	0	0	0	1	2	2
- Swapping rules	0	2	2	1	0	0	0	1	2	2
Generalization										
- Merging rules	1	2	2	0	0	0	0	2	1	2
- Swapping rules	1	1	2	1	0	0	0	1	2	2
- Deleting rules	0	1	2	1	1	0	0	1	2	2
Redundancy										
- Merging rules	0	2	2	1	0	0	0	0	2	3
Semantics loss										
- Merging rules	1	1	2	1	0	0	0	1	2	2

Remarks: the level of satisfaction is divided in to five levels: 5, 4, 3, 2, and 1; 5 is the highest confidence.

Table 6: Interpretation of the reliability between Inter-raters [19]

Kappa statistic value	**Inter-rater reliability description**
0	Agreement equivalent to chance
0.1 - 0.20	Alight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

6. Conclusion

Practically, fixing anomalies of firewall rules is quite complex, depending on the administrator's perspective and experience. Correcting mistakes may lead to other anomalies. For example, when resolving the redundant anomaly, it may become the semantics loss of rules. In order to reduce the impact of errors and to resolve anomalies of administrators, this paper has designed and developed a system to assist in the decision-making of administrators by using probability together with four additional features of rules: frequency of matching between packets and rules, evidence of creating rules, expertise of rules creator and protocol priority. For each rule, the probability is calculated based on their features. If the probability of any rule is high, it indicates that the rule has a high priority. While rules in the firewall are conflicts, the rule that has a high probability value is always considered first. As a result of system testing, administrators can make more accurate decisions about conflict rules in the firewall. For the overall efficiency of the system, the time complexity of creating a system (PST) is equal to $O(n)$, searching time over PST is $O(log_m n)$ and the space complexity is $O(m \times n)$. However, the system still has a limitation against the establishment of the tree structure. As resolving any anomaly of rules in each period, it needs to reconstruct the whole PST tree structure. For the evaluation of confidence for resolving firewall rule anomalies, the firewall that we have proposed on the basis of probability obtains a confidence value more than the traditional firewall by 29.6%, and the reliability between Inter-raters of proposed firewall is in the moderate agreement (0.51), which increased by 13.1% from the traditional firewall.

Acknowledgment

This research project was financially supported by Faculty of Informatics, Mahasarakham University 2020.

References

[1] E. S. Al-Shaer and H. H. Hamed, "Modeling and management of firewall policies," IEEE Transactions on Network and Service Management, 1(1), 2–10, April 2004. https://doi.org/10.1109/TNSM.2004.4623689

[2] S. Khummanee, "The semantics loss tracker of firewall rules," in IC2IT 2018 Recent Advances in Information and Communication Technology, 769, Jun. 2018, 220–231, 2018. https://doi.org/10.1007/978-3-319-93692-5_22

[3] E. Al-Shaer, H. Hamed, R. Boutaba, and M. Hasan, "Conflict classification and analysis of distributed firewall policies," IEEE Journal on Selected Areas in Communications, 23(10), 2069–2084, Oct 2005. https://doi.org/10.1109/JSAC.2005.854119

[4] S. Khummanee, A. Khumseela, and S. Puangpronpitag, "Towards a new design of firewall: Anomaly elimination and fast verifying of firewall rules," in The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE), May 2013, 93–98, 2013. https://doi.org/10.1109/JCSSE.2013.6567326

[5] A. X. Liu, "Formal verification of firewall policies," in 2008 IEEE International Conference on Communications, May 2008, 1494 – 1498, 2008. https://doi.org/10.1109/ICC.2008.289

[6] M. Rezvani and R. Aryan, "Analyzing and resolving anomalies in firewall security policies based on propositional logic," in 2009 IEEE 13th International Multitopic Conference, Dec 2009, 1–7, 2009. https://doi.org/10.1109/INMIC.2009.5383125

[7] H. Hu, G. Ahn, and K. Kulkarni, "Detecting and resolving firewall policy anomalies," IEEE Transactions on Dependable and Secure Computing, 9(3), 318–331, May 2012. https://doi.org/10.1109/TDSC.2012.20

[8] S. R. Peter M. Mell, Karen A. Scarfone. (015) A complete guide to the common vulnerability scoring system version 2.0. [Online]. Available: https://www.nist.gov/publications/complete-guidecommon-vulnerability-scoring-system-version-20

[9] A. Wool, "Architecting the lumeta firewall analyzer," in SSYM'01 Proceedings of the 10th conference on USENIX Security Symposium, 10, Aug 2001, p. 7.

[10] A. Mayer, A. Wool, and E. Ziskind, "Fang: a firewall analysis engine," in Proceeding 2000 IEEE Symposium on Security and Privacy. S P 2000, May 2000, 177–187, 2000. https://doi.org/10.1109/SECPRI.2000.848455

[11] A. Hari, S. Suri, and G. Parulkar, "Detecting and resolving packet filter conflicts," in Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer

and Communications Societies (Cat. 00CH37064), 3, March 2000, 1203–1212 3, 2000. https://doi.org/10.1109/INFCOM.2000.832496

[12] L. Zhang and M. Huang, "A firewall rules optimized model based on service-grouping," in 2015 12th Web Information System and Application Conference (WISA), Sep. 2015, 142–146, 2015. https://doi.org/10.1109/WISA.2015.47

[13] Lihua Yuan, Hao Chen, Jianning Mai, Chen-Nee Chuah, Zhendong Su, and P. Mohapatra, "Fireman: a toolkit for firewall modelling and analysis," in 2006 IEEE Symposium on Security and Privacy (S P'06), May 2006, 15 – 213, 2006. https://doi.org/10.1109/SP.2006.16

[14] A. Saˆadaoui, N. B. Y. B. Souayeh, and A. Bouhoula, "Formal approach for managing firewall misconfigurations," in 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), May 2014, 1–10, 2014. https://doi.org/10.1109/RCIS.2014.6861044

[15] W. Krombi, M. Erradi, and A. Khoumsi, "Automata-based approach to design and analyze security policies," in 2014 Twelfth Annual International Conference on Privacy, Security and Trust, July 2014, 306–313, 2014. https://doi.org/10.1109/PST.2014.6890953

[16] A. Saˆadaoui, N. B. Y. B. Souayeh, and A. Bouhoula, "Automated and optimized fdd-based method to fix firewall misconfigurations," in 2015 IEEE 14th International Symposium on Network Computing and Applications, Sep. 2015, 63–67, 2015. https://doi.org/10.1109/NCA.2015.31

[17] C. C. Aggarwal, Neural Networks and Deep Learning. Boca Raton, FL, USA: Springer, 2018.

[18] D. Morris and M. Koning, Bayes' Theorem Examples: A Visual Introduction for Beginners. 80 Strand, London, WC2R 0RL UK: Blue Windmill Media, 2016.

[19] P. W. M. J. Kenneth J. Berry, Janis E. Johnston, The Measurement of Association: A Permutation Statistical Approach. Boca Raton, FL, USA: Springer, 2018.

[20] CelPlan Technologies. Gpp qos class identification qci categories 2019. [Online]. Available: http://www.celplan.com

[21] Graphviz. Graph visualization software 2019. [Online]. Available: https://www.graphviz.org

[22] NetworkX. Software for complex networks, 2019. [Online]. Available: https://networkx.github.io

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 08-23 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

A Novel Representative *k*-NN Sampling-based Clustering Approach for an Effective Dimensionality Reduction-based Visualization of Dynamic Data

Dharamsotu Bheekya*, Kanakapodi Swarupa Rani, Salman Abdul Moiz, Chillarige Raghavendra Rao

School of Computer and Information Sciences, University of Hyderabad, Hyderabad, Telangana, 500046, India.

ARTICLE INFO

Article history:
Received: 13 May, 2020
Accepted: 19 June, 2020
Online: 06 July, 2020

Keywords:
Dimensionality Reduction
Exploratory Analysis
Visualization
Clustering
Sampling
t-Distributed Stochastic Neighbor Embedding
Cluster Validation Index
Interpolation

ABSTRACT

Visualization plays a crucial role in the exploratory analysis of Big Data. The direct visualization of Big Data is a challenging task and difficult to analyze. Dimensionality Reduction techniques extract the features in the context of visualization. Due to the unsupervised and non-parametric nature, most of the dimensionality reduction techniques are not evaluated quantitatively and not allowed to extend for dynamic data. The proposed representative k-NN sampling-based clustering, determines the underlying structure of the data by using well-known clustering techniques. The external cluster validation index determines the order sequence of clustering techniques from which the appropriate cluster techniques are recommended for the given datasets. From the recommended set, the samples of the best clustering technique are considered as representative samples which can be used for generating the visual representation. The t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm is applied to generate a low-dimensional embedding model of representative samples, which is more suitable for visualization. The new data samples are added to the generated model by using the interpolation technique. The low-dimensional embedding results are quantitatively evaluated by k-NN accuracy and trustworthiness. The performance analysis of representative k-NN sampling-based clustering results and embedding results accomplished by seven differently characterized datasets.

1 Introduction

Exploratory analysis of Big Data is ubiquitous in an increasing number of fields and vital to their progress. Visualization plays a paramount role in an exploratory study. Data visualization is applicable for the limited number of dimensions, which depends on the perceptual capability of the analyst. For exploratory analysis, traditional visualization techniques may not provide useful visual insights of high-dimensional data and they are restricted for a limited number of dimensions. The conventional feature selection methods may not provide helpful visual insights for exploratory analysis, which happens due to the inappropriate feature selection. There is a requirement of the feature extraction technique, which shows the correlation between the original features of the data. The Dimensionality Reduction (DR) technique transforms the data and extract new features, which makes data analysis tractable. From the last few decades, researchers have proposed various linear as well as non-linear DR techniques in the context of visualization. The linear DR techniques like Principal Component Analysis (PCA) [1], Multidimensional Scaling (MDS) [2] and Factor Analysis (FA) [3] deals with the simple structured data. It is challenging to extract valuable information from complex structured data using linear DR. In contradiction to linear DR, the non-linear DR techniques like Isomap [4], Local Linear Embedding (LLE) [5], Laplacian Eigenmap (LE) [6] and Stochastic Neighbor Embedding (SNE) [7] deals with non-linear data. The paper [8] provides a complete comparative study of DR techniques. The methodology of the DR technique depends on its feature extraction criteria. The feature extraction depends on the characteristics of interest in the data such as inter-point distances, reconstruction weights, variation, linear subspace, geodesic distances, linear tangent space, neighborhood graph and conditional probability distribution.

Among all the DR techniques, t-Distributed Stochastic Neighbor Embedding (t-SNE) [9], is an improved version of SNE, introduced by Laurens Van Der Maaten and Geoffrey Hinton in 2008. The t-SNE has gained impressive attention and enormous popularity in several fields [10–12]. The empirical study states that the low-dimensional visual representation of t-SNE is more robust than any other DR technique. The t-SNE algorithm most commonly used to preserve the original structure of the high-dimensional data in

*Corresponding Author: Dharamsotu Bheekya, +919573466965, & 15mcpc21@uohyd.ac.in

very low-dimensional embedding (i.e., either 2D or 3D). From past several years the t-SNE has been explored in various aspects such as optimization [13, 14], scalability [14–17], dealing with non-numeric data [18], outliers separation [17] and many more.

The t-SNE is a non-parametric technique that provides flexibility in learning and reduces computational complexity. The non-parametric nature limits t-SNE applicability to the out-off-sample extension, which means the addition of new data samples into the existing t-SNE environment is not possible. If we want to add a new sample, then we should re-run the entire t-SNE model by including a new sample. When the addition of new data points increases, the computational cost of t-SNE also increases monotonically. Therefore, it does not apply to time-series and streaming data. The LION-tSNE [17] of Boytsov et al. addressed the problem of adding new data into the existing t-SNE environment using Local Inverse Distance Weighting Interpolation (LIDWI) without re-running. The outlier's handing is also addressed by LION-tSNE using outlier placement heuristic, which assumes that some percentile of outliers present in the designed t-SNE and determine the outliers from the newly added data points. In LION-tSNE, the sample t-SNE model is designed based on the random sample selection, which may cause the non representativeness of the data. The representative samples are selected by our earlier approach called k-NN sampling [19] and the resutlts are statistically significant which is measured by statistical method pairwise t-test [20].

This paper is an extension of our earlier paper presented in High Performance Computing, Data and Analytics (HiPC) [19], which deals with the preservation of the underlying cluster structure of high-dimensional data in low-dimensional t-SNE embedding with a representative sample. The underlying cluster structure preservation is measured in terms of a quantitative metric. In the existing methods, the low-dimensional embedding of t-SNE describes the quality of the structure preservation. Still, it is an open problem for giving the quantitative proof for the number of clusters that exist in the original data. The proposed novel representative k-NN sampling-based clustering approach for effective dimensionality reduction-based visualization finds the solution. In the first step, the proposed approach determines several distinct data samples using our earlier proposed k-NN sampling algorithm. The number of samples depends on the range of k (i.e., $1 \leq k \leq m$), which gives the neighborhood representation.

In the second step, the effective number of clusters existing in the original dataset is determined by the sampled data using clustering techniques such as k-Means [21], Agglomerative Hierarchical Clustering (AHC) [22], Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [23], Fuzzy c-means (FCM) [24], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [25], Ordering Points To Identify the Clustering Structure (OPTICS) [26], Mean Shift (MS) [27], Spectral Clustering (SC) [28], Expectation-Maximization Gaussian Mixture Model (EMGMM) [29], Affinity Propagation (AP) [30] and Mini Batch k-Means (MBKM) [31]. Each clustering technique generates cluster labels for sampled data. The cluster labels of remaining data (i.e., data points other than selected samples) are labeled by the k-Nearest Neighbor (k-NN) algorithm. The k-NN of each remaining data sample is subject to the selected samples. After assigning the labels to the remaining data samples, the representative sample of each clustering technique is determined by their external cluster validation index called Fowlkes-Mallows Index (FMI) [32]. In our contribution, we are also recommending the order sequence of clustering techniques among the selected techniques for a given dataset. We are also presenting the Compactness (CP) [33], Calinski-Harabaz Index (CHI) [33] and Contingency Matrix (CM) [34] of clustering techniques for gaining a more detailed analysis about clustering. The first technique in the order sequence denotes the best clustering technique and the representative samples of it considered as the samples for t-SNE model design for a given dataset. The cluster validation index provides the comparison between representative k-NN sampling-based clustering and the aggregate clustering (i.e., clustering on the whole dataset). The proposed representative k-NN sampling-based clustering is scalable to all clustering techniques which are suitable for numerical datasets. Also, we can apply any clustering technique to the large scale dynamic data with representative k-NN sampling-based clustering. Due to the paper limitation, we have chosen the most popular clustering techniques from different groups.

In the third step, the sample t-SNE model is designed on a representative sample of best clustering techniques, which transforms the high-dimensional data into low-dimensional embedding. The remaining data samples are added to the sample t-SNE environment with the help of LIDWI. The outliers from the remaining samples are identified and controlled by proposed heuristic and the identified outliers are placed into the t-SNE environment using outlier placement heuristic of Boystov et.al [17]. In the fourth step, the t-SNE embedding of input data is quantitatively evaluated by k-NN accuracy in the context of clustering and trustworthiness. The quantitative evaluation answers the question, how much structure of high-dimensional data is preserved by the low-dimensional t-SNE embedding. The k-NN accuracy of t-SNE embedding is measured in two ways: baseline accuracy and sampling accuracy. For quantitative performance evaluation, the k-NN accuracy of t-SNE embedding of representative k-NN sampling-based clustering and aggregate clustering results are compared with the ground truth class labels, which is measured in our earlier paper. In our earlier approach, we used ground-truth class labels for obtaining the k-NN accuracy of the t-SNE embedding but here we are using for checking the derived cluster purity of representative k-NN sampling-based clustering. The k-NN sampling-based clustering results and t-SNE embedding of it are analyzed by seven differently characterized toy and real-world datasets. In summary, our contribution consists of

- The order sequence of applicable clustering methods with representative k-NN sampling-based clustering using cluster validation index. The set of recommended techniques for the given dataset using a threshold. The comparison between representative k-NN sampling-based clustering and aggregate clustering results.
- The outliers from the remaining samples are identified and controlled by the proposed heuristic before adding into the t-SNE embedding. The embedding structure quantitatively evaluated by k-NN accuracy in the context of clustering and trustworthiness.

The organization of the paper is as follows. In section 2, we present the background knowledge that gives the basic mathemati-

cal intuition required for understanding the clustering techniques, t-SNE and other supplements. Section 3 presents the related work. In section 4, we present a detailed description of the proposed *k*-NN sampling-based clustering for the visualization of dynamic data. In section 5, we are providing the details of datasets, experimental setup and analysis of results. Finally, section 6 gives directions for future work and conclusions.

2 Background Knowledge

The following section describes the few techniques for understanding the background formulation of our work. Section 2.1 gives a detailed description of the selected clustering techniques for representative *k*-NN sampling-based clustering. Section 2.2 describes the mathematical intuition of the t-SNE algorithm. The intuition behind the addition of the inlier data point into an existing t-SNE environment is explained in section 2.3. In section 2.4, we are giving a detailed description of metrics that are used for the quantitative evaluation of embedding.

2.1 Clustering and their performance measures

Section 2.1.1 describes the selected clustering techniques which are used for representative *k*-NN sampling-based clustering. 2.1.2 provides detailed mathematical intuition of cluster validation indexes such as Fowlkes-Mallows Index (FMI), Compactness (CP), Calinski-Harabaz Index (CHI) and Contingency Matrix (CM).

2.1.1 Clustering Techniques

Clustering is unsupervised learning, where the similarly characterized data objects are grouped. The clusters of data objects can be represented as a set C of subsets $C_1, C_2, ..., C_k$ such that $\cap_{i=1}^{k} C_i = \phi$. The different clustering algorithms are proposed based on their measures of similarity: partitional, hierarchical, fuzzy theory-based, distribution-based, density-based, graph partition-based, grid-based, model-based and many more. The user decides the number of clusters present in the dataset by using a heuristic, trail-and-error and evolutionary approaches such as density and probability density. From the groups mentioned above, we have chosen the most frequently and popularly used algorithms for experimental evaluation. But proposed representative *k*-NN sampling-based clustering approach is scalable to all the clustering techniques which are suitable for the numerical datasets. We explored traditional clustering algorithm [35–37] for our work.

In partitional clustering, data points assigned to any one of the k-clusters using distance similarity measures such as Euclidean distance. k-means [21] clustering is one of the simplest, best-known and benchmarked partition-based clustering. The k-means clustering classifies the given data points through a user-defined number of clusters. The main goal of the k-means clustering is the initialization of an appropriate k-centroids, one for each cluster. The objective of k-means is the minimization of the sum of square distance which can be defined as follows

$$SSD = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \quad (1)$$

where $\| x_i^{(j)} - c_j \|^2$ is a L2 norm between a data point $x_i^{(j)}$ and the cluster center c_j. In AP [30], at the initial stage, all data points are considered as centroids and nodes of the network. The clusters and their centroids are measured by transmitting the similarity message recursively. The Mini Batch k-Means (MBKM) [31] is an improved version of k-means, which performs clustering on batches instead of considering each point. Therefore, MBKM requires less computational time and is also applicable to large datasets.

In hierarchical clustering, groups are formed by iteratively dividing the data objects either in the bottom-up or top-down approach. The bottom-up approach is also known as Agglomerative Hierarchical Clustering (AHC) [22], in which initially each data object is considered as a separate group, then merging these small groups into larger and larger groups until a single group formation or certain threshold. The top-down approach is also known as Divisive Hierarchical Clustering (DHC), and it works in reverse order of the bottom-up approach. Both hierarchical approaches depend on the linkage criteria such as single, complete and average linkage. The linkage criteria determines the metric for merging two similar characterized small clusters. For example, single linkage criteria determines the minimum distance pair from the neighbor clusters $min\left\{d(x,y) : x \in C_i, y \in C_j\right\}$. The BIRCH [23] method has proposed to deal with large datasets, outliers in robust and also to reduce the computational complexity. The BIRCH method works on the idea of Cluster Features (CF) which is a height-balanced tree.

The basic idea of fuzzy theory-based clustering is that the discrete labeling is converted to continuous intervals, to describe the belonging relationship among objects more reasonably. The Fuzzy c-means (FCM) [24] is an extension of *k*-means where each data point can be a member of multiple clusters with membership value. The main advantage of FCM is that the formed groups are more realistic.

The density-based clustering finds the clusters based on the density of data points in a region. The principal objective of density-based clustering is that for each instance of a group, there should be at least a minimum number of neighbor instances within the given radius. The DBSCAN [25] is the most well known density-based clustering. In DBSCAN clustering, the data objects fall into three groups: core-object, border-object and noise-object. The data points of the core-object group have enough number of neighbors in the given radius, and these data points are from the higher density region. The data points of the border-object group have fewer neighbors than the required number of neighbors in the given radius and they are in the neighborhood of the core object group. The data points of the noise-object group are not in either core or border-object group. The advantage of DBSCAN clustering is that the generated clusters are in arbitrary shape based on the given parameters such as radius and number of minimum instances. The improved version of DBSCAN is OPTICS [26], which overcomes the limitations of DBSCAN. Mean-shift [27] algorithm determines the mean of offset of the current data point, the next point is identified based on the current point and the offset. The process will continue until some criteria is satisfied.

The spectral clustering [28] is a graph partition-based technique in which clustering is obtained by similarity graph partition. The paper [28] describes the different ways of constructing the similarity graph. While building the similarity graph, the data objects are con-

sidered as vertices and the similarity (i.e., the similarity distance) between the data objects is a weighted edge. The critical idea of similarity measure is to find a method of graph partition by making edges between different groups have low edge weight and the edges of the same group have more edge weights. In similarity graph partitioning, the cluster analysis is carried out by feature space obtained by the eigenvectors corresponding to k largest eigenvalues of the laplacian matrix [38] of similarity graph. The spectral clustering is suitable for the datasets with arbitrary shape, high-dimensions and outliers.

The EMGMM [29] is a model as well as a distribution-based clustering technique in which cluster data obey the same independent Gaussian distribution which exists in the original data distribution. The generated clusters are more realistic because a probability distribution obtains them.

The performance and applicability of clustering depends on the used dataset characteristics. Therefore, the researchers empirically state that none of the clustering technique is generalized for any datasets.

2.1.2 Cluster validation Measures

In our experimentation, we have used two internal and one external cluster validation measure to evaluate the performance of clustering techniques. The internal measure computes the performance of clustering without using the ground-truth class labels. The compactness (CP) [33] is an internal cluster validation measure which can be computed as follows

$$CP = \frac{1}{n}\sum_{l=1}^{k} n_l \left(\frac{\sum_{x_i, x_j \in C_l} d(x_i, x_j)}{n_l(n_l-1)/2} \right) \tag{2}$$

where $d(x_i, x_j)$ is Euclidean distance between two objects in cluster C_l and n_l is the number of objects in C_l. The smaller CP value of clustering is more compact and gives better clustering results. The Calinski-Harabaz Index (CHI) [33] or Variance Ratio Criterion is defined as the ratio of between-clusters dispersion mean and the within-cluster dispersion mean. The CHI is an internal cluster validation measure which is given by

$$CHI = \frac{\sum_{l=1}^{k} n_l d^2(C_l, C)/(NC-1)}{\sum_{l=1}^{k}\sum_{x \in C_l} d^2(x, C_l)/(n-NC)} \tag{3}$$

Where n is the dataset size, C is the dataset center, n_l is the size of lth cluster, C_l is the lth cluster centroid and NC is the number of clusters.

The external validation measure knows the ground truth class labels. The primary purpose of the external validation index is to choose an optimal clustering algorithm for a given dataset. The external validation measure also checks the cluster purity. In our experiment, we used Fowlkes-Mallows Index (FMI) [32] is the geometric mean between precision and recall. The FMI defined as follows

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{4}$$

The FMI value is bounded between 0 and 1, 1 denotes that the obtained clusters and the given ground truth classes are the same. The large FMI indicates that the obtained clustering is purer.

We also measured the contingency matrix that reports the intersection cardinality for every true/predicted cluster pair. The contingency matrix gives sufficient statistics for all clustering metrics. But it is hard to interpret the contingency matrix of extensive clustered data.

2.2 Student t-Distributed Stochastic Neighbor Embedding (t-SNE)

The t-SNE [9] algorithm introduced by Laurens Van Der Maaten and Geoffrey Hinton in 2008, based on the SNE algorithm. The principal objective of SNE is to preserve the underlying structure of high-dimensional data in low-dimensional embedding space. Lets assume the given input data $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ where each $x_i \in \mathbb{R}^D$ is a D-dimensional vector. The t-SNE computes the embedding $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ of $\mathcal{X}$ where each $y_i \in \mathbb{R}^d$ is a d-dimensional vector, where $d \ll D$ and most commonly d =2 or 3. The similarity between x_i and x_j of $\mathcal{X}$ is calculated by conditional probability p_{ij} is given by

$$p_{j/i} = \frac{exp\left(\frac{-d(x_i,x_j)^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}^{N} exp\left(\frac{-d(x_i,x_k)^2}{2\sigma_i^2}\right)},\ p_{i/i} = 0,\ p_{ij} = \frac{p_{j/i} + p_{i/j}}{2N} \tag{5}$$

where the bandwidth σ_i of Gaussian kernel, is obtained by binary search by matching the perplexity of P_i and the user-defined perplexity (μ) which is given as a parameter. The perplexity is a smooth measure of an adequate number of neighbors for each data point. The equality and perplexity of P_i is as follows

$$\mu = 2^{H(P_i)} \quad where \quad H(P_i) = -\sum_{j}^{N} p_{j/i} \log_2 p_{j/i} \tag{6}$$

where $H(P_i)$ denotes the entropy and P_i is the conditional probability distribution across all data points for the given x_i. The y_i and y_j are the corresponding low-dimensional values of x_i and x_j (i.e., the values of $\mathcal{Y}$ are initialized by Gaussian or uniform distribution). The similarity between y_i and y_j is defined as

$$q_{ij} = \frac{(1 + d(y_i, y_j)^2)^{-1}}{\sum_{l}^{N}\sum_{k\neq l}^{N}(1 + d(y_l, y_k)^2)^{-1}} \tag{7}$$

where $d(y_i, y_j)$ is defined as distance similarity measure such as Euclidean distance. In low-dimensional embedding, the similarity q_{ij} is obtained by student t-distribution with one degree of freedom. But, the similarity p_{ij} of high-dimensional data uses the Gaussian distribution. The cumulative function curvature of Gaussian distribution is flatter than the cumulative function curvature of student t-distribution with one degree of freedom. The principal idea of using student t-distribution in low-dimensional embedding is to overcome the crowding problem [7].

If $p_{ij} \sim q_{ij}, \forall_{i\neq j} \in N$, then the given data is perfectly embedded into the low-dimensional space. Otherwise, compute the KL-divergence (i.e., error) between p_{ij} and q_{ij} that is equal to the cross-entropy in Information Retrieval System (IRS). The cost function (**C**) or objective of t-SNE is defined as follows

$$\mathbf{C} = KL(P \parallel Q) = \sum_{j}^{N} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{8}$$

The simple gradient descent applied on cost function **C** for obtaining the optimization or minimization of it. The simple gradient descent of cost function **C** given by

$$\frac{\partial \mathbf{C}}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j) \tag{9}$$

where $Z = (1 + d(y_i, y_j)^2)^{-1}$ is a normalization term of student t-distribution. The *Equation 9* produce either positive or negative values depending on the p_{ij} value. If an *Equation 9* gives positive value then there is an attractive force among the y_i and y_j of embedding space. Otherwise, there is a repulsive force among the y_i and y_j of embedding space. The degree of repulsion depends solely on the closeness of points in the embedding space. In the optimization process, the early exaggeration coefficient $\alpha > 1$ plays a paramount role in forming groups of similar objects of high-dimensional data in low-dimensional embedding space. In the early exaggeration process, the elements of similarity matrix (P) (i.e., p_{ij}'s) multiplied by the early exaggeration coefficient, which is measured by the intuition given by George C. Linderman and Stefan Steinerberger [39]. Therefore, similar data objects bring near to each other in low-dimensional embedding space. This process can happen at the early stage of optimization. The gradient descent of **C** after early exaggeration is

$$\frac{\partial \mathbf{C}}{\partial y_i} = 4 \sum_{j \neq i} \alpha p_{ij} q_{ij} Z (y_i - y_j) - 4 \sum_{j \neq i} q_{ij}^2 Z (y_i - y_j) \tag{10}$$

The momentum and learning rate parameters improve the optimization process of the cost function. The momentum parameter reduces the number of iterations of the cost function optimization. At the initial stage of iteration, the momentum value is small until the map points have become moderately well organized. The optimization is improved by input approximation and tree-based algorithms [13], which reduce the memory and computational complexity. The updated values of $\mathcal{Y}$ at iteration t is obtained by

$$\mathcal{Y}^{(t+1)} = \mathcal{Y}^{(t)} + \eta \frac{\partial \mathbf{C}}{\partial y_i} + \alpha(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}) \tag{11}$$

where $\alpha(t)$ is momentum at t^{th} iteration, η is learning rate. The work-flow of t-SNE is shown in the *Figure 1*.

The well-separated clusters of input data are well preserved in the low-dimensional embedding by setting the early exaggeration coefficient α and learning rate η. The intuition for setting the above parameter is derived by George C. Linderman and Stefan Steinerberger in [39]. According to the George C. Linderman and Stefan Steinerberger observations the early exaggeration coefficient α, learning rate η, and minimum probability p_{ij}'s of same cluster objects(i.e., x_i and x_j belong into the cluster C_l for all $i \neq j$) is defined as follows.

$$\alpha \sim \frac{n}{10} \quad , \quad \eta \sim 1 \quad and \quad p_{ij} = \frac{1}{10n \mid \pi^{-1}(\pi(i)) \mid} \tag{12}$$

where $\pi : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., k\}$ assigns each data point to one of the k clusters.

2.3 Local Inverse Distance Weighting Interpolation (LIDWI)

The LIDWI [40] maps new data point $x \in \mathcal{R}^D$ into the existing embedding $y_i \in \mathcal{R}^d$, where $\{i = 1, 2, ...m\}$. The LIDWI determines the value of x as a weighted sum of values y_i, where weight is proportional to inverse distances. The LIDWI of x is

$$LIDWI(x) = \sum_{\|x - x_i\| \leq r_x} w_i(x).y_i, \; w_i(x) = \frac{\| x - x_i \|^{-p}}{\sum_{\|x - x_i\| \leq r_x} \| x - x_i \|^{-p}} \tag{13}$$

for instance, when the data point $x \rightarrow x_i$, the inverse distance $\| x - x_i \|^{-1} \rightarrow \infty$, the corresponding weight $w_i(x) \rightarrow 1$ (i.e., $\forall_{j \neq i} w_j(x) \rightarrow 0$ due to the normalization) and $LIDWI(x) \rightarrow y_i$. The neighbor points selection is obtained by a radius r_x parameter. The parameter r_x value is calculated by the heuristic proposed by Andrey Boytsov et.al. [17]. In LIDWI, the power parameter p plays an important role. For instance, very small value of p predicts the value of x around the center: $y \approx mean(y_i)$ (unless $x = x_i$) even the distance $\| x - x_i \|$ is low because the weight distribution is close to uniform. When the power parameter is high and the distance $\| x - x_i \|$ is low, the weight $w_i(x)$ of very first nearest neighbor is dominating all other neighbors, therefore $y \approx y_i$ where $i = argmin \| x - x_j \|$. The overfitting suffers from either too small or too large values of power parameter p. In LOIN-tSNE, the authors proposed a generalization for obtaining power parameter by using leave-one-out cross-validation of the training sample. The computation of the generalized power parameter is obtained by applying the LIDWI for each training sample that produces the estimation of each y_i. Then the mean square distance between the estimated y_i's and real y_i's is computed. The optimal power parameter is obtained by optimizing the mean square error (i.e., the mean square distance is minimum). The obtained power parameter is considered as a metric. However, this metric is heuristic, not an exact criterion.

2.4 Performance evaluation metrics

2.4.1 k-NN accuracy

The existence of the cluster structure of high-dimensional data in low-dimensional embedding is quantitatively measured by k-NN accuracy of the t-SNE embedding in the context of clustering. The k-NN accuracy is defined as the percentage of the neighbors having the cluster label equivalent to the observational point cluster label.

2.4.2 Trustworthiness

Trustworthiness [41] is one of the measure to evaluate the quality of the t-SNE embedding. Trustworthiness is defined as any unexpected nearest neighbors in the output space are penalized in proportion to their rank in the input space.

$$T(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{n} \sum_{x_j \in U_k(x_i)} (r(x_i, x_j) - k) \tag{14}$$

where $U_k(x_i)$ is the k-NN of x_i in embedding space, $r(x_i, x_j), i \neq j$ the rank of x_j when the data vectors are ordered based on their

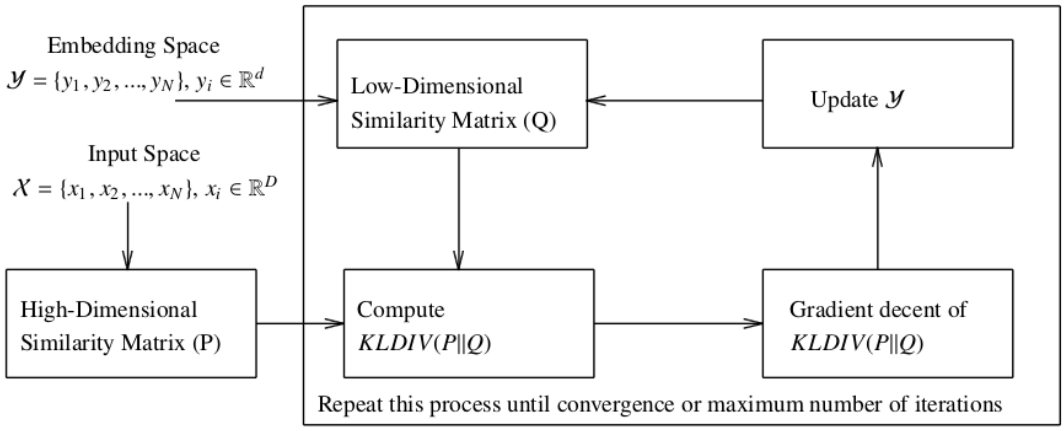


Figure 1: Work-flow of t-SNE algorithm [19]

Euclidean distance from the data vector x_i in the original data space. It is bounded between 0 and 1, where 1 represents the complete structure of the data preserved in the embedding space.

3 Related Work

Our earlier proposed k-NN sampling based tSNE mainly concentrates on the preservation of the underlying structure of high-dimensional data in low-dimensional embedding space with representative samples. The obtained low-dimensional embedding structure describes the quality of the structure using ground-truth class labels. But the low-dimensional embedding does not give the quantitative proof for the number of clusters that exist in the original data. In [39] authors gave the theoretical observations for well-separated clusters of high-dimensional data in low dimensional embedding space.

The researchers have proposed various methods to incorporate new data samples into the existing t-SNE environment. Most of the existing techniques designed a mapping function $f : \mathcal{X} \rightarrow \mathcal{Y}$, which accepts multi-dimensional data and returns its low-dimensional embedding. The designed mapping functions are used for incorporating the new data samples into the existing t-SNE environment. In [14–16] authors have proposed different approaches for adding a new data sample or scaling up the t-SNE algorithm.

Andrey Boytsov et.al. [17] proposed the LION-tSNE algorithm based on local IDWI for adding new data sample into an existing t-SNE environment. It also addresses the outlier handling approaches. Our earlier work extended the idea of the LION-tSNE algorithm by proposing a k-NN sampling method for designing a representative sampling based t-SNE model. It allows the selection of the sample concerning their k-nearest neighbors instead of random sampling. In this paper, we are proposing the novel representative k-NN sampling-based cluster approach for effective dimensionality reduction-based visualization of dynamic data, which determines the underlying cluster by using the most popular clustering techniques. The obtained cluster structure is quantitatively evaluated by k-NN accuracy in the context of clustering and trustworthiness.

4 Proposed representative k-NN sampling-based clustering for effective visualization Framework

The proposed representative k-NN sampling-based clustering for effective dimensionality reduction-based visualization of dynamic data framework is shown in the *Figure 2*. It has four phases: at phase 1, data preprocessing is done for removing the redundant data points and filling the empty variables with appropriate values. At phase 2, the proposed k-NN sampling-based clustering is applied to determine: (i) The generation of k distinct samples using mutual k-NN sampling with static graph updation algorithm. (ii) The list of clustering techniques applicable for the given dataset. (iii) The optimal sample size produces the best grouping for each concerned clustering technique. (iv) The ordered sequence of recommended clustering techniques for the given dataset from which the best clustering technique is selected by using FMI scores. The detailed representative k-NN sampling-based is clustering presented in section 4.1. At phase 3, the dimensionality reduction-based visualization is obtained by t-SNE algorithm, which derives the low-dimensional embedding of data and the LIDWI algorithm is used to interpolate the new data points into the learned t-SNE model. The outliers from the remaining data points are identified and controlled by the proposed heuristic. The dimensionality reduction-based visualization is described in section 4.2. In the final phase, the quantitative measure of the t-SNE embedding is computed by k-NN accuracy and trustworthiness. The process of quantitative metric derivation is described in section 4.3.

4.1 Representative k-NN sampling-based clustering

Initially, the k distinct samples are generated by the modified algorithm of our earlier proposed method. The sample generation depends on the parameter k of k-NN sampling. Each k generates one distinct sample; the boundary of k is denoted as $1 \leq k \leq m$, where m is the maximum number of neighbors required for preserving the behaviour of any data point of the given dataset. Our earlier approach has given two different k-NN sampling strategies

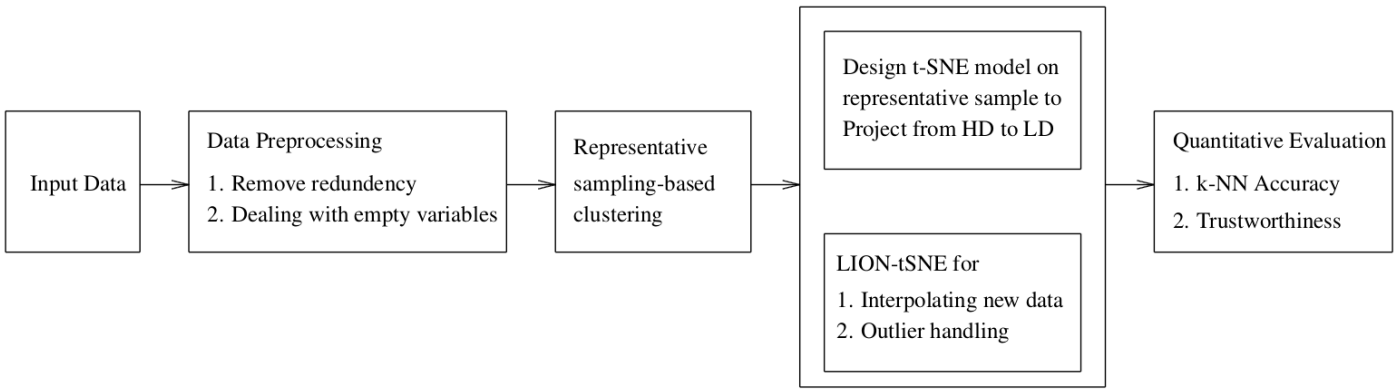


Figure 2: Proposed framework of representative k-NN sampling-based clustering for an effective dimensionality reduction-based visualization of dynamic data using LION-tSNE with quantitative measure of cluster accuracy

such as static and dynamic k-NN sampling based on k-NN graph modification. In this paper, we are considering only static k-NN sampling in which the k-NN graph is modified statically that means no new edge is added among any two vertices of the k-NN graph. The main reason for selecting static k-NN sampling is that the t-SNE embedding results of it are more statistically significant than the dynamic one. The Nearest Neighbor score (NN_score) and Mutual Nearest Neighbor score (MNN_score) of each node of the k-NN graph are used for selecting the samples. Lets assume k-NN graph is a directed graph $G = (V, E)$, the edge $E(v_1, v_2)$ gives v_2 as a neighbor of v_1 and neighborhood of v_1 is denoted by N_{v1}. The *out-degree* of each vertex is equal to k, and the *in-degree* of a vertex depends on the neighborhood property of other vertices (i.e., the data point x_i belong into the neighborhood of other datapoints). In our method, each data point is a vertex of the k-NN graph and k is a parameter for deriving the edges between neighboring vertices. The NN_score of data point x_i is equal to the *in-degree* of x_i which is defined as follows

$$NN\_score(x_i) = \mid \{x_j \mid x_i \in N_{x_j}\} \mid, \forall_{j \neq i} x_j \in \mathcal{X} \tag{15}$$

where $\mathcal{X}$ denotes whole data set, N_{x_j} denotes the neighborhood of x_j. The MNN_score of data point x_i is at most k which is given by

$$MNN\_score(x_i) = \mid \{x_j \mid x_i \in N_{x_j}\} \mid \quad where \quad x_j \in N_{x_i}, \forall_{j \neq i} x_j \in \mathcal{X} \tag{16}$$

From the NN_score and MNN_score matrix the Representative Sample (RS) x_i selection and the sample x_j Represented by Representative Sample (RRS) x_i is given by

$$RS(x_i) = first\_index\{argmax_{x_i \in \mathcal{X}}\{NN\_score(x_i)\} \cap argmax_{x_i \in \mathcal{X}}\{MNN\_score(x_i)\}\} \tag{17}$$

$$RRS(x_j) = \{x_j \mid x_j \in N_{x_i}\} \quad where \quad x_i \in N_{x_j}, \forall_{j \neq i} x_j, x_i \in \mathcal{X} \tag{18}$$

The updated mutual k-NN sampling with a static graph updation algorithm is shown in *Algorithm 1*. Initially, the Train_samples (i.e., representative samples set) and Rep_samples (i.e., the sample represented by the selected training sample that is the first nearest mutual neighbor of train sample) are null sets. For each iteration, the data point x which has maximum NN and maximum MNN score appended to the Train_samples set. The data point y is appended to the Rep_samples if $y \in N_x$ and $x \in N_y$ that denotes the point x and y are more similar (i.e., the distance $d(x, y) < d(x, z)$ where $\forall z \in N_x$) to each other. The data point x and y are deleted from $\mathcal{X}$ and their corresponding vertices are deleted from the k-NN graph. After deleting the vertices of x and y, the corresponding in-edges and out-edges are removed from the k-NN graph and the graph is updated accordingly. The elements of Train_samples and Rep_samples are obtained iteratively, the iteration repeats until the NN_score of the remaining $\mathcal{X}$ is equal to zero.

Algorithm 1: Mutual k-NN Sampling with a static graph updation

Data: data set $\mathcal{X} = \{x_1, x_2, ...x_N\}$, parameter k for minimal training sample selection
Result: Return Train_samples, Rep_samples
$Train\_samples = \emptyset$
$Rep\_samples = \emptyset$
begin
  Compute k-NN graph of $\mathcal{X}$
  repeat
    Compute $NN\_Score(\mathcal{X})$
    Compute $MNN\_Score(\mathcal{X})$
    $index = [NN\_Score(\mathcal{X}) == argmax\{NN\_Score(\mathcal{X})\}]$ /* gives all index which are having same NN_Score */
    if $len(index) > 1$ **then**
      $train\_index = argmax\{MNN\_Score(x_i)\}$ where $i \in index$
    end
    else
      $train\_index = index$
    end
    $x = \mathcal{X}[train\_index]$
    $Train\_samples = Train\_samples \cup x$
    Determine the mutual neighbors of $train\_index$
    $y = Most\_NMN(x)$ /* return first mutual nearest neighbor of x */
    if $y \in N_x \& x \in N_y$ **then**
      $Rep\_samples = Rep\_samples \cup y$
    end
    Delete x and y from $\mathcal{X}$ and update the k-NN graph
  until $(NN - Score(\mathcal{X}) = 0))$;
end

In our earlier approach of static graph updation, we have considered the whole mutual neighborhood set of x as the samples represented by x. The whole neighborhood set selection causes

loss of information due to that reason the earlier approach select Train_samples set at an early stage of k value. The data points of the remaining $\mathcal{X}$ (i.e., the data point do not belong into either Train_samples set or Rep_samples set) is handled in two different ways. Handling of the remaining samples is described in subsection 4.4. The mutual k-NN sampling with a static graph updation algorithm generates a various set of Train_samples and Rep_samples with different k-values.

The clustering techniques which are described in subsection 2.1.1 are applied on each Train_samples set of mutual k-NN sampling with a static graph updation algorithm that is called a sampling-based clustering. The cluster labels for the data points other than the Train_samples are assigned by using k-NN algorithm. For example, the dataset size is N, the Train_samples size is n and then the unmarked sample size is $N - n$(i.e., the samples are not assigned with any labels). The k-NN of each unmarked sample is derived from the associated Train_samples set. The label assignment of each unmarked sample depends on the labels of its k-NN, a label with maximum occurrence in k-NN that is assigned to a sample. From each clustering technique, the optimal Train_samples set is obtained by the FMI score. The FMI is an external validation index that uses the ground truth class labels. Therefore, the number of clusters is defined as a constant that is equal to the number of ground truth classes. The FMI score of optimal Train_samples of clustering techniques generates an order sequence of the clustering techniques. From the order sequence, we can recommend the most desirable techniques for a given dataset from the selected set of techniques. From this recommendation, the best suitable technique is selected and its optimal Train_samples set is considered as the best representative sample. The threshold parameter is used to derive the recommended techniques, which is defined as the FMI score difference between two adjacent techniques of order sequence. The three cluster validation index such as FMI score, CHI score and CP of representative k-NN sampling-based clustering are compared with their aggregate clustering (i.e., clustering on whole dataset) validation index of chosen clustering techniques. The result comparison is discussed in subsection 5.3. The algorithm of representative k-NN sampling-based clustering is shown in *Algorithm 2*. The embedding of selected optimal Train_samples and addition of other data samples into an existing t-SNE environment is described in the following section.

4.2 Dimensionality reduction-based visualization

The subsection 4.2.1 describes the low-dimensional embedding of a representative sample with t-SNE algorithm. Subsection 4.2.2 describes the addition of new data samples into an existing t-SNE environment that is called out-off-sample extension.

4.2.1 Low-dimensional embedding of a representative sample with t-SNE algorithm

Barnes-Hut t-SNE (BH-tSNE) [13] algorithm is an optimized version of the t-SNE algorithm. BH-tSNE optimizes the t-SNE objective function by input similarity approximation and gradient descent approximation. Therefore, it generates low-dimensional embedding of data with minimum computational and memory complexity than original t-SNE. In our approach, the BH-tSNE algorithm is used in two different ways for calculating the low-dimensional embedding space: 1. Baseline t-SNE embedding, 2. Sampled t-SNE embedding. The baseline t-SNE embedding is obtained by applying BH-tSNE on the whole dataset. In contrast, the sampled t-SNE embedding is obtained by applying BH-tSNE on the best representative sample, which is selected from the representative k-NN sampling-based clustering. The Baseline t-SNE embedding results analyze the overall structure of the data in low-dimensional embedding. The sampled t-SNE embedding results analyze the data structure with sampled data and allows the addition of new data samples into an existing t-SNE environment, which solves the scalability issue of the t-SNE. For obtaining a well-separated cluster in low-dimensional t-SNE embedding, the value of early exaggeration coefficient α, learning rate η and input similarity probability p_{ij}'s are adjusted according to the George C.linderman and Stefan Steinerberger intuition. In our experimentation, the initial solutions of t-SNE is assigned in three different ways, such as random, PCA based and MDS based initial solutions. The random initial solution takes many iterations for convergence. The PCA and MDS based initial solutions overcome the problem of random initialization and they produce better accuracy results, but their initial solution is cost-effective. Adding new data points into a designed t-SNE model is discussed in the below section.

4.2.2 Out-off-sample extension: Interpolation and Outlier handling

The addition of new data point to t-SNE embedding depends on the parameter r_x, r_y and r_{close} values. The value of parameter r_x, r_y and r_{close} is obtained from the best representative sample and the t-SNE embedding of it. The parameter r_x is defined as the percentile of the 1-NN distance of a representative sample that decides whether the given new data point is either inlier or outlier. In our proposal, we came to know that the objects of the representative sample set are representing at least one sample of the dataset. Therefore, the representative sample set does not contain any outlier object and we have considered the parameter r_x as the maximum 1NN distance of it. If the new data point x has at least one data point within the r_x from the representative sample set, then x is an inlier, otherwise outlier. The dilation factor (df) is used to derive a heuristic $r_x = (1 + df) * r_x$ which controls the consideration of outliers percentage. The LIDWI interpolation technique is used for adding an inlier data point to the t-SNE embedding of the representative sample set. The outlier placement depends on the parameter r_y and r_{close}. The outliers placed into the t-SNE embedding of the representative sample set according to the heuristic of Boytsov et.al. [17]. The k-NN accuracy and trustworthiness of t-SNE embedding quantitatively evaluate the existence of a high-dimensional cluster in the low-dimensional embedding. The quantitative evaluation described in the following section.

4.3 Quantitative metric derivation: k-NN accuracy

In our experimentation, the k-NN accuracy is calculated in two different ways: baseline accuracy and sampling accuracy. The k-NN accuracy of t-SNE embedding of aggregate data is known as

Algorithm 2: Representative k-NN Sampling-based clustering

Data: data set $\mathcal{X} = \{x_1, x_2, ...x_N\}$, parameter k for minimal training sample selection
Result: Order sequence of recommended techniques, Representative samples, samples represented by representative samples and best cluster technique
$Recommended\_tech = \emptyset$
begin
  for $k \leftarrow 1$ **to** Max_kval **do**
    train_sample, rep_sample = k-NN Sample(k)
    for $i \leftarrow 0$ **to** $NumberOfClusterTechniques$ **do**
      train_labels = Clustering(train_sample, i) /* apply i^{th} clustering technique on train_sample set */
      $remaining\_samples = \mathcal{X} - train\_sample$
      Compute labels of remaining_samples using k-NN algorithm (i.e., remain_labels)
      cluster_labels=train_labels ∪ remain_labels
      Compute cluster validation indexes such as FMI, CHI and CP using cluster labels, ground truth class labels
      if $FMI > optimal\_FMI[i]$ **then**
        /* $optimal\_FMI[i] = 0$, where i = 0,1,..., NumberOfClusterTechniques */
        optimal_FMI[i] = FMI
        RS[i] = train_sample
        RRS[i] = rep_sample
      end
    end
  end
  Sort optimal_FMI list: $optimal\_FMI[1] > optimal\_FMI[2] > .... > optimal\_FMI[l]$
  Swap the RS and RRS lists values and maintain the tech_index list of clustering techniques according to the sorted optimal_FMI
  for $i \leftarrow 1$ **to** $size(tech\_index)$ **do**
    if $optimal\_FMI[i] - optimal\_FMI[i+1] \leq threshold$ **then**
      $Recommended\_tech = Recommended\_tech \cup tech\_index[i]$
    end
  end
  The First element of Recommended_tech represents the best clustering technique among the selected clustering techniques and corresponding RS is chosen as sample for t-SNE embedding.
end

baseline accuracy. The sampling accuracy is computed in three different forms such as training accuracy, test accuracy and overall accuracy. The k-NN accuracy of t-SNE embedding of the representative sample is known as training accuracy. The k-NN accuracy of interpolated samples (i.e., the samples other than the representative sample) is known as test accuracy. The k-NN accuracy of combined low-dimensional space (i.e., the integration of t-SNE embedding of sampled data and interpolated data) is known as overall accuracy. The k-NN accuracy typically depends on the parameter k, which is considered as fixed in our experimentation which is discussed in subsection 5.2. For instance, the smaller k will give good performance accuracy and while increasing the k, performance accuracy will decrease. The selection of parameter k also plays a paramount role in k-NN accuracy measure. The relationship between the parameter k and accuracy is shown in the *Figure 3*. For quantitative performance evaluation, the k-NN accuracy of representative k-NN sampling-based clustering and aggregate clustering are compared with the k-NN accuracy of ground truth class labels, which is discussed in subsection 5.3.

4.4 t-SNE sample selection criteria

In our experimentation, we have considered two different sample selection criteria for designing the t-SNE model. In the first criteria, the samples are mostly representative of at least one or more other samples. The samples are obtained by a mutual k-NN sampling with a static graph updation algorithm and the best representative sample is selected by the representative k-NN sampling-based clustering. We observe that most of these samples were chosen from the dense region because they produce good NN_score and MNN_score. Therefore, these samples do not contain any outliers. In this criterion, the outliers from the remaining sample addition are not sufficiently identified by the adequate 1-NN distance. The outliers consideration is controlled by the proposed heuristic, which is discussed in subsection 4.2.2. In the second criteria, the data points added to the representative sample from the remaining data points using (t, m, s)-Nets sampling [42]. For the addition of (t, m, s)-Nets samples, we used the same procedure of our earlier work. The (t, m, s)-Nets select samples randomly in a uniform distribution. These samples may not represent any other samples and the samples may change among the executions due to the randomness. Therefore, the samples of this criteria may contain outliers. The outliers of this criteria are handled in similar ways as they are handled in our earlier approach.

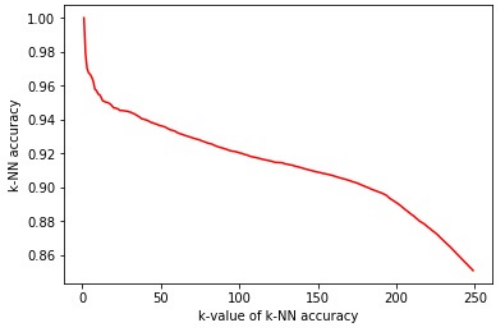


Figure 3: Relationship between k-value of k-NN accuracy and k-NN accuracy which is obtained from the baseline accuracy of k-NN sampling-based clustering.

5 Experimental result analysis

Here we are briefing the datasets, experimental setup and the result analysis. Section 5.1 describes the numerical datasets of different characterizations. The details about the experimental setup is given in section 5.2. The results analysis of each dataset is described in section 5.3.

5.1 Datasets

In our experimental evaluation, we have considered differently characterized datasets such as IRIS, Breast-Cancer, Leukemia, Wine, MNIST, Olivetti-Faces, and COIL-20. The *Table 1* provides a detailed description of all the datasets which are downloaded from the UCI Machine Learning repository [43]. For our experimentation, the features of IRIS, Breast-Cancer and Wine datasets are normalized between zero and one, which improves the computational complexity. Initially, the KernelPCA [44] is applied to reduce the dimensionality of the high-dimensional datasets such as Leukemia, MNIST, Olivetti-Faces and COIL-20. The Leukemia dataset is a micro-array of gene expression. The MNIST, Olivetti-Faces and COIL-20 are image datasets that are represented in pixel orientations. The MNIST is a handwritten digits dataset. The Olivetti-faces dataset consists ten face images of 40 individuals with small variation in viewpoint, the addition of glass and large variation in expression. The COIL-20 is an image of 20 group objects such as animals, furniture and etc.

Table 1: Overview of datasets along with their size, dimensions, and classes

Dataset Name	Size	Dimensions	# Classes
IRIS	150	4	3
Breast Cancer	569	30	2
Leukemia-ALL-AML	72	7129	2
Wine	178	13	3
MNSIT	70K	784	10
Olivetti faces	400	10304	40
COIL-20	1440	1024	20

5.2 Experimental configuration

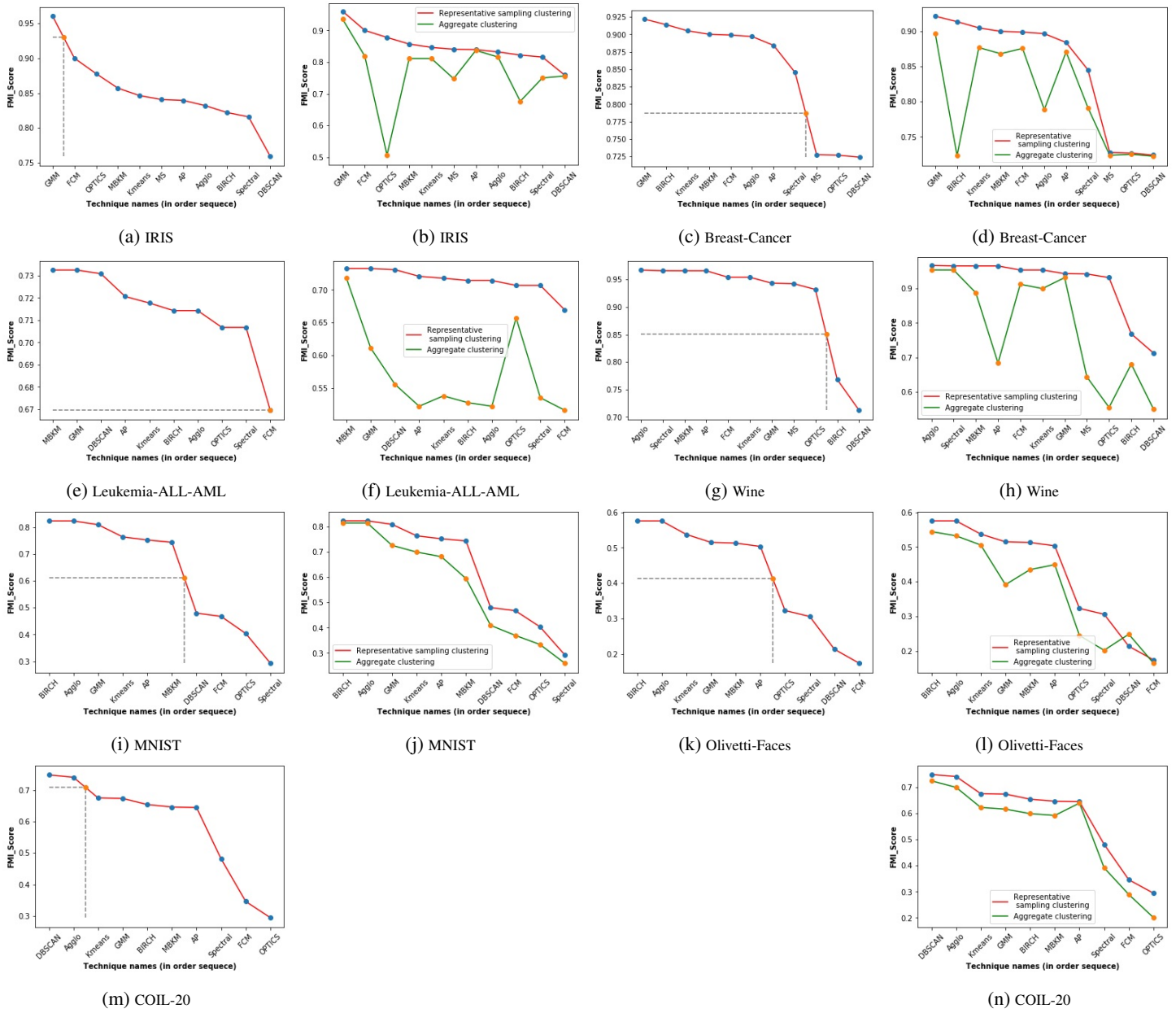
In our experimentation, The parameter k of *Algorithm 1* is initialy considered as $1 \leq k \leq 50$. The upperbound of k is equal to largest perplexity value from the literature study [9]. The perplexity is set between 5 and 50 for a fairly good visual representation of any real-world data. In our proposal, the samples of any clustering technique depends on the parameter k. Also, we observed that when there is an increment in parameter k then there is an increment or no change in FMI score of clustering until certain k value. The FMI score becomes stable afterwards. In our experimentation at most of the cases, the selected clustering technique generates representative sample with k value less than or equal to 20. In other cases, the cluster technique generates representative sample with k value greater than 20. But there is an small increment in FMI score comparatively FMI score of clustering with k value less than or equal to 20. Therefore, we have generalized the k-value as less than or equal to 20. The sensitivity of parameter k needs to be investigated more in future. The number of clusters is defined as a constant that is equal to the number of ground truth classes. The original class labels are not used anywhere in the experimental evaluation. The ground truth class labels are used only for measuring the FMI score that determines the cluster purity. The threshold parameter is set to 0.05 that provides the intuition for selecting the recommended clustering techniques from the order sequence. The threshold parameter derived from the statistical method where the maximum allowable difference between two consecutive values of either increasing or decreasing order sequence is 5%. The recommended set size and threshold parameters are inversly proposional to each other. The sensitivity of threshold parameter needs to be investigated further. From the recommended set, the best technique is chosen and the optimal sample of it being considered as the best representative sample for designing the t-SNE model. The representative sample is embedded in a 2D space using the BH-tSNE algorithm. The parameters of BH-tSNE are set up according to the paper [13] experimental setup. In addition to that, we initialized the embedding space $\mathcal{Y}$ by sampling the point y_i from a uniform distribution with $[-0.02, 0.02]^2$ for obtaining well-separated clusters in embedding space. The datasets with more than 50 dimensions, their dimensionality is reduced to 50 by kernel PCA. The dimensionality reduction speeds up the computation of the probability distribution of the input similarity and suppresses some noise. The results of the BH-tSNE algorithm are shown in 2D scatter-plot representation. The minimum value of p_{ij} of clustered data, the early exaggeration factor α and the learning rate η values are assigned similar to the George C.linderman and Stefan Steinerberger paper.

The data points other than the representative samples are interpolated into the BH-tSNE of a representative sample using LIDWI. The parameter r_x is obtained by either proposed heuristic or intuition of LION-tSNE. In the proposed heuristic dilation factor is bounded between 0 and 1. The parameter r_y, r_{yclose} and power are measured similar to LION-tSNE algorithm.

Table 2: Parameters setting for the experimental setup

Parameter	Value
k of k-NN sampling	$1 \leq k \leq 20$
Threshold	0.05
Perplexity	5 - 50
Early exaggeration coefficient	$\sim \frac{N}{10}$
Adaptive learning rate	$1 - 200$
Dilation Factor	$0 \leq df \leq 1$
r_x at *dist_perc*	95 - 100
r_y at *dist_perc*	100
r_{yclose} at *dist_perc*	10
Fixed k-value for k-NN accuracy	3 - 10

Table 2 provides the parameter settings of our experimental evaluation. The parameter *perplexity* represents the effective number of neighbors for each data point. For instance, the small value of *perplexity* creates subgroups within the same cluster of t-SNE results. In contrast to small, the large value of it does not maintain the clear separation between two clusters of t-SNE results. Both cases suffer from either under-fitting or over-fitting problem that causes a lack of visual clarity. The empirical studies state that the *perplexity* value between 5 to 70 gives a good visual representation of t-SNE results. The parameter *dist_per* represents the overall percentile of the representative sample that needs to be considered as inliers. For



(a) IRIS (b) IRIS (c) Breast-Cancer (d) Breast-Cancer

(e) Leukemia-ALL-AML (f) Leukemia-ALL-AML (g) Wine (h) Wine

(i) MNIST (j) MNIST (k) Olivetti-Faces (l) Olivetti-Faces

(m) COIL-20 (n) COIL-20

Figure 4: The column one and column three (i.e., sub-figure (*(a), (c), (e), (g), (i), (k)* and *(m)*) shows the threshold derivation (i.e., the integration of doted horizontal and vertical lines) of all seven datasets from the order sequence curve. The column two and column four (i.e., sub-figure *(b), (d), (f), (h), (j), (l)* and *(n)*) shows the relationship curve of all datasets from the order sequence, which provides the relationship between the FMI scores of proposed representative *k*-NN sampling-based clustering and aggregate clustering.

example, if we take *dist_per* as 95th percentile, that means out of 100 points, 95 points are considered as inliers and the remaining 5 points are outliers. The parameter *k* plays an important role in the computation of *k*-NN accuracy of the data. The effect of parameter *k* is shown in the *Figure 3*. It is clear that when there is an increment in *k* value, accuracy decreases monotonically. The result evaluation of representative *k*-NN sampling-based clustering is discussed in the next section.

5.3 Result analysis

In our experimentation, we have considered most frequently and popularly used clustering techniques. The *Figure 4* shows the curves of all seven datasets of subsection 5.1. The *Figure 4* gives the threshold derivation for the recommended order sequence and relationship between the FMI scores of proposed representative *k*-NN sampling-based clustering and aggregate clustering. In the *Figure 4*, the x-axis represent the clustering technique names in the order sequence of FMI_scores (i.e., $FMI(t_1) > FMI(t_2) > .... > FMI(t_l)$ where t_i represents i^{th} best clustering technique) and y-axis represents the corresponding FMI_Scores. Column one and column three of *Figure 4* (i.e., sub-figure *(a), (c), (e), (g),(i), (k)* and *(m)*) shows the threshold derivation for the recommendation of clustering techniques for all the given seven datasets. The column two and column four of *Figure 4* (i.e., sub-figure *(b), (d), (f), (h), (j), (l)* and *(n)*)

Table 3: The applicable clustering and their recommended order sequence for the given seven datasets.

Dataset	Applicable cluster methods	Recommended order sequence of cluster methods
IRIS	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS, MS	EMGMM
Breast-cancer	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS, MS	EMGMM, BIRCH, k-Means, MBKM, FCM, Agglomerative, AP, Spectral
Leukemia	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS	MBKM, EMGMM, DBSCAN, AP, k-Means, BIRCH, Agglomerative, OPTICS, Spectral, FCM
Wine	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS, MS	Agglomerative, Spectral, MBKM, AP, FCM, k-Means, EMGMM, MS, OPTICS
MNIST	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS	BIRCH, Agglomerative, EMGMM, k-Means, AP, MBKM
Olivetti-faces	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS	BIRCH, Agglomerative, k-Means, EMGMM, MBKM, AP
COIL-20	k-Means, Agglomerative, BIRCH, Spectral, EMGMM, FCM, DBSCAN, AP, MBKM, OPTICS	DBSCAN, Agglomerative

shows the relationship curve that provides the relationship between the FMI_Score of representative *k*-NN sampling-based clustering and aggregate clustering for all the given seven datasets. The *Figure 4* clearly shows that the proposed representative *k*-NN sampling-based clustering results are superior to the aggregate clustering results.

The sets of applicable and recommended order sequence of clustering techniques for the given seven datasets is listed in the *Table 3*. The CP and HCI of representative *k*-NN sampling-based clustering and aggregate clustering of all clustering techniques are listed in the *Table 4* and *5*. The *Table 4* and *5* clearly states that the CP of the proposed representative *k*-NN sampling-based clustering is smaller than the aggregate clustering for some techniques. In contrary techniques, the compactness is in reverse order, but the compactness difference is very minute in both situations. The CHI is also deriving the relationship between both clustering as same as CP where the CP is small, and then there is a large CHI score.

The contingency matrices of representative *k*-NN sampling-based clustering and aggregate clustering of IRIS, Wine, Breast-cancer and Leukemia datasets are shown in the *Table 6, 7, 8* and *9*. If the number of classes of the given dataset is larger, then it is difficult to analyze the data with the contingency matrix. Therefore, the contingency matrix of MNIST, Olivetti-Faces and COIL-20 datasets are not addressed. The *Table 6, 7, 8* and *9* clearly states that the proposed representative *k*-NN sampling-based clustering classifies the labels much similar to the ground truth class labels compared to aggregate clustering. The *Tabel 9* represents the contingency matrix of the Leukemia dataset, which indicates that none of the selected technique gives the best clustering results. Still, representative *k*-NN sampling-based clustering produces better results than the aggregate clustering.

For Quantitative evaluation, the baseline *k*-NN accuracy of proposed representative *k*-NN sampling-based clustering is compared with baseline *k*-NN accuracy of aggregate clustering and baseline *k*-NN accuracy of ground truth class labels. The baseline *k*-NN accuracy and trustworthiness of representative *k*-NN sampling-based clustering, aggregate clustering and ground truth class labels of seven datasets are listed in *Table 10*. The *Table 10* clearly indicates that the proposed method produces more robust results than others.

Table 4: Compactness and CHI score of representative *k*-NN sampling-based clustering, overall sampling clustering (i.e., assigning the cluster labels based on the KNN of remaining samples in the context of representative samples) and the aggregate clustering of IRIS, Breast-Cancer, Leukemia and Wine datasets for the selected clustering techniques.

IRIS

Method	Samp. size	Compactness			CHI Score		
		Samp. cltr	Overall cltr	Aggr. cltr	Samp. cltr	Overall cltr	Aggr. cltr
k-Means	73	0.4945	0.5601	0.5557	221.54	347.56	358.56
Agglomerative	63	0.5143	0.5579	0.5606	178.78	353.21	348.03
Birch	73	0.5291	0.5646	0.8507	183.85	335.22	193.97
Spectral	60	0.5234	0.5606	0.5709	167.457	348.03	322.48
EMGMM	63	0.5316	0.5754	0.5789	161.44	317.59	307.76
FCM	72	0.5107	0.5633	0.5576	197.71	340.79	355.71
DBSCAN	73	0.6379	0.7047	0.6723	104.84	175.96	124.78
AP	73	0.52	0.557	0.5588	196.99	354.54	353.22
MBKM	74	0.4834	0.5613	0.5562	234.32	343.95	356.28
OPTICS	73	0.5894	0.6071	1.0919	120.53	259.45	15.99
MS	70	0.527	0.557	0.6002	188.2	354.45	289.52
Breast-Cancer							
k-Means	291	1.5608	1.6064	1.6013	203.91	355.42	364.09
Agglomerative	290	1.5735	1.6073	1.6322	199.73	356.80	319.01
Birch	291	1.5698	1.6083	1.9509	193.53	352.67	53.96
Spectral	289	1.5694	1.6177	1.6449	197.34	351.45	328.74
EMGMM	276	1.5744	1.6096	1.6168	178.38	350.08	337.00
FCM	276	1.5645	1.6048	1.6001	186.30	358.22	363.04
DBSCAN	291	1.9493	2.0004	1.9735	21.27	19.66	38.11
AP	213	1.5146	1.6088	1.6088	156.95	356.82	358.62
MBKM	268	0.8824	1.6047	1.6007	180.883	359.63	363.83
OPTICS	268	1.9635	2.0179	2.0069	21.72	24.65	38.41
MS	213	1.9138	2.0054	1.9412	19.79	17.65	68.97
Leukemia-ALL-AML							
k-Means	22	3.4413	3.9367	3.9279	1.567	1.173	1.685
Agglomerative	19	3.3494	3.9315	3.9222	1.874	1.269	1.756
Birch	19	3.3494	3.9315	3.919	1.874	1.269	1.921
Spectral	19	3.3076	3.915	3.92	1.707	1.279	1.845
EMGMM	20	3.4012	3.9415	3.9237	1.337	0.953	1.591
FCM	20	3.3743	3.939	3.9271	1.622	1.117	1.619
DBSCAN	17	3.2654	3.9436	3.9225	1.351	0.959	1.326
AP	14	2.8993	3.895	3.9355	1.548	0.882	1.292
MBKM	22	3.3088	3.8957	3.9366	1.102	0.859	1.122
OPTICS	19	3.315	3.922	3.9071	0.859	0.717	1.317
MS	...	....	....	....	....	....	....
Wine							
k-Means	82	1.3275	1.4298	1.4289	47.255	82.828	83.373
Agglomerative	87	1.3207	1.4356	1.4328	45.964	80.465	81.327
Birch	82	1.555	1.6567	1.629	19.859	34.395	42.564
Spectral	87	1.3532	1.4348	1.4298	42.372	81.014	82.828
EMGMM	71	1.3481	1.4331	1.4348	39.846	81.796	81.698
FCM	85	1.3473	1.4317	1.4305	45.029	82.346	83.135
DBSCAN	64	1.6181	1.6656	1.9606	15.719	37.502	3.102
AP	82	1.5534	1.3368	1.4458	45.362	80.714	80.828
MBKM	87	1.3562	1.4848	1.4278	42.472	81.784	81.523
OPTICS	85	1.4484	1.4531	1.4248	38.866	82.796	81.698
MS	71	1.3489	1.4371	1.4748	49.846	81.797	80.598

Table 5: Compactness and CHI score of representative k-NN sampling-based clustering, overall sampling clustering (i.e., assigning the cluster labels based on the KNN of remaining samples in the context of representative samples) and the aggregate clustering of MNIST Handwritten digits, Olivetti-Faces and COIL-20 datasets for the selected clustering techniques.

MNIST							
Method	**Samp. size**	**Compactness**			**CHI Score**		
		Samp. cltr	**Overall cltr**	**Aggr. cltr**	**Samp. cltr**	**Overall cltr**	**Aggr. cltr**
k-Means	4481	3816.4	3897.8	3877.8	233.97	477.45	492.48
Agglomerative	4900	3936.2	3974.2	3974.2	206.16	409.22	395.43
Birch	4900	3936.2	3974.2	3974.2	206.16	409.22	395.43
Spectral	4735	4662.2	4694.5	4695.0	1.048	1.0308	1.0459
EMGMM	4206	3862.2	3930.6	3977	200.8	443.59	400.32
FCM	4835	3962.2	4374.5	4555.	1.248	1.308	1.59
DBSCAN	4735	4662.2	4694.5	4695.	1.048	1.0308	1.0459
AP	4496	3832.5	3978.6	3897.3	213.97	457.78	472.34
MBKM	4783	4216.4	4597.1	4577.3	224.17	467.65	452.18
OPTICS	4625	3962.5	4094.3	4195.7	1.258	1.13	1.045
MS	...	....	....	....	....	....	....
Olivetti-Faces							
k-Means	200	1568	1784.5	1565.8	10.384	14.009	21.210
Agglomerative	178	1495.9	1798.8	1546.8	10.718	13.634	22.008
Birch	178	1495.9	1798.8	1546.8	10.718	13.634	22.008
Spectral	188	2235	2545.9	2608.4	0.6858	0.6839	0.9924
EMGMM	165	1528.9	1812.8	1591.5	9.211	13.168	20.272
FCM	178	2479.3	2659.4	2732.6	4.321	4.874	4.178
DBSCAN	194	2679.4	2750.5	2745.1	4.021	3.804	5.782
AP	194	1628.9	1852.8	1691.5	10.321	14.168	20.872
MBKM	178	1668.6	1852.8	1671.5	9.711	12.168	21.275
OPTICS	165	1598.4	1932.8	1891.5	8.217	13.468	22.728
MS	...	....	....	....	....	....	....
COIL-20							
k-Means	679	10.333	10.362	10.180	87.911	183.571	188.454
Agglomerative	701	10.239	10.237	10.294	90.220	184.703	181.461
Birch	701	10.577	10.562	11.1165	87.377	180.283	162.538
Spectral	676	14.852	14.944	17.447	28.585	57.964	27.4288
EMGMM	679	10.273	10.243	10.356	87.832	186.73	187.686
FCM	689	12.872	13.645	15.745	34.784	56.768	34.58
DBSCAN	679	18.843	18.962	19.769	19.532	38.902	38.548
AP	701	10.573	10.253	10.856	83.83	188.63	186.656
MBKM	679	10.253	10.143	10.366	80.832	188.73	187.656
OPTICS	679	10.243	10.233	10.326	87.432	185.738	184.286
MS	...	...	...	...	...	...	...

Table 6: Contingency matrix of k-NN sampling-based and original clustering on IRIS dataset with best k-NN sampling-based clustering technique (i.e., EMGMM).

		Class labels generated by					
		***k*-NN sampling-based clustering**			**Aggregate clustering**		
		C1	**C2**	**C3**	**C1**	**C2**	**C3**
Original	**C1**	50	0	0	50	0	0
	C2	0	49	1	0	45	5
	C3	0	2	48	0	0	50

Table 7: Contingency matrix of k-NN sampling-based and original clustering on Wine dataset with best k-NN sampling-based clustering technique (i.e., Agglomerative).

		Class labels generated by					
		***k*-NN sampling-based clustering**			**Aggregate clustering**		
		C1	**C2**	**C3**	**C1**	**C2**	**C3**
Original	**C1**	0	0	59	2	0	57
	C2	69	2	0	69	2	0
	C3	1	47	0	0	48	0

The sampling k-NN accuracies such as train, test and overall of t-SNE embedding of a representative sample and interpolation of other samples are listed in *Table 11* and the overall trustworthiness is also covered. The best representative sample of representative k-NN sampling-based clustering and random sampling of IRIS dataset is shown in *Figure 5*. The 2D scatter-plots shown in *Figure 5* are the first two coordinates of IRIS dataset. The *Figure 5(a)* shows the best representative sample of k-NN sampling and the *Figure 5(c)* shows the representative sample of k-NN sampling along with samples of (t,m,s)-Nets of remaining data points. The *Figure 5(b)* and *5(d)* shows the random sampling of size equal to the sample size of *Figure 5(a)* and *5(c)*. The samples of *Figure 5(a)* are constant and consistent comparatively other sampling.

Table 8: Contingency matrix of k-NN sampling-based and original clustering on Breast-Cancer dataset with best k-NN sampling-based clustering technique (i.e., EMGMM).

		Class labels generated by			
		***k*-NN sampling-based clustering**		**Aggregate clustering**	
		C1	**C2**	**C1**	**C2**
Original	**C1**	21	191	16	196
	C2	352	5	340	17

Table 9: Contingency matrix of k-NN sampling-based and original clustering on Leukemia dataset with best k-NN sampling-based clustering technique (i.e., MBKM).

		Class labels generated by			
		***k*-NN sampling-based clustering**		**Aggregate clustering**	
		C1	**C2**	**C1**	**C2**
Original	**C1**	0	47	9	38
	C2	2	23	6	19

Table 10: Baseline k-NN accuracy of k-NN sampling based clustering, original clustering and ground truth class labels

Dataset Name	**Trust**	**Baseline *k*-NN Accuracy**		
		Sampling Clustering	**Aggregate Clustering**	**Ground-truth class labels**
IRIS	0.9861	0.9735	0.9666	0.96
Breast-Cancer	0.958	0.9876	0.9862	0.9577
Leukemia	0.6799	0.9875	0.6643	0.7485
Wine	0.9552	0.9774	0.9828	0.9717
MNIST Digits	0.9895	0.9259	0.9483	0.9512
Olivetti face	0.9494	0.6775	0.8287	0.88
COIL-20	0.9972	0.9554	0.9709	0.9743

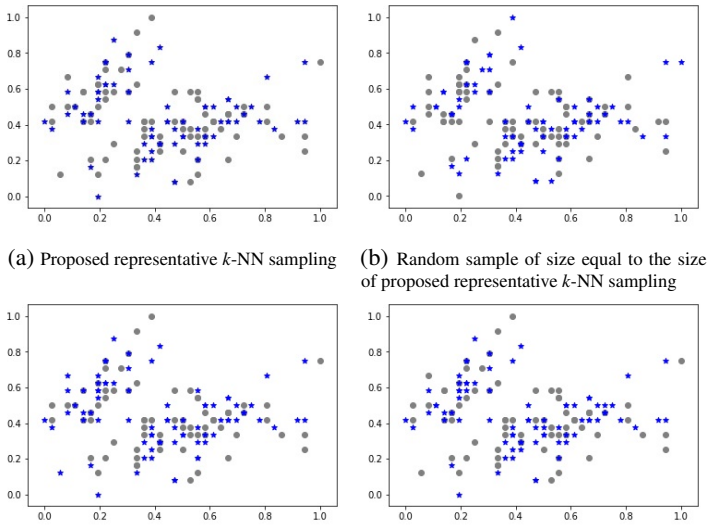
The *Figure 6* shows the 2D scatter-plot visualization of IRIS data with outliers projection. In *Figure 6*, we are also showing the outliers (i.e. denoted by large grey color circle) of original data after finding the outliers from the addition of remaining data points with the radius r_x which is obtained by the proposed heuristic $r_x = (1 + df) * r_x$ with $df = 0.2$. It clearly states that sampled data obtain the outliers of original data.

The 2D scatter-plot visualization of embedding space of representative k-NN sampling-based clustering of the MNIST dataset is shown in the *Figure 7*. The *Figure 7* shows the baseline 2D scatter-plot of t-SNE embedding of whole MNIST data of size 10K

Table 11: Quantitative measure using k-NN accuracy of k-NN sampling based clustering, original clustering and ground truth class labels

Optimal Sample											
Dataset Name	**Sample Size**	**Trust**	**Sampling k-NN Accuracy**								
			Sampling-based clustering			**aggregate clustering**			**Ground truth class labels**		
			Train	**Test**	**Overall**	**Train**	**Test**	**Overall**	**Train**	**Test**	**Overall**
IRIS	73	0.9594	0.9726	0.9594	0.9661	0.9753	0.9405	0.9652	0.9589	0.9324	0.9585
Breast Cancer	291	0.9239	0.9896	0.9856	0.9820	0.9690	0.9856	0.9876	0.9452	0.9676	0.9546
Leukemia	33	0.5665	0.9416	1.0	0.9875	0.6714	0.4102	0.6266	0.6363	0.5897	0.6247
Wine	87	0.9299	0.9377	0.956	0.9379	0.966	0.967	0.9604	0.9511	0.945	0.9548
MNIST	4900	0.9635	0.9346	0.8956	0.9120	0.9201	0.91	0.9206	0.9335	0.8972	0.9170
Olivetti-Faces	200	0.968	0.6076	0.5128	0.6045	0.6356	0.7948	0.8270	0.6334	0.7794	0.8578
COIL-20	701	0.9941	0.9841	0.9417	0.9509	0.9529	0.9634	0.9749	0.9426	0.9688	0.9727
Random Sample of size equal to optimal sample											
IRIS	73	0.9654	0.9726	0.9859	0.98	0.9808	0.9802	0.977	0.9616	0.9661	0.9728
Breast Cancer	291	0.9337	0.9759	0.9712	0.9753	0.9828	0.9892	0.9929	0.9553	0.9496	0.9472
Leukemia	33	0.5495	1.0	0.9722	0.9866	5654	0.6111	0.6133	0.5771	0.4444	0.6087
Wine	87	0.9316	0.9639	0.9662	0.9517	0.9431	0.9662	0.949	0.9671	0.9775	0.9547
MNIST	4900	0.9599	0.9122	0.8918	0.9001	0.9157	0.8977	0.9081	0.9249	0.8956	0.9048
Olivetti-Faces	200	0.9578	0.5080	0.5050	0.5860	0.7235	0.7121	0.7869	0.6967	0.6767	0.7842
COIL-20	701	0.9935	0.9543	0.9362	0.9449	0.9699	0.9606	0.9652	0.9273	0.9375	0.9515
Optimal Sample plus Lhs Sample											
IRIS	79	0.965	0.9746	1.0	0.9797	0.9733	0.9852	0.9862	0.9594	0.9852	0.9632
Breast Cancer	305	0.9348	0.9736	0.9772	0.9806	0.9934	0.9848	0.9912	0.9608	0.9810	0.9701
Leukemia	41	0.5979	0.9777	1.0	0.9723	0.5361	0.6551	0.5994	0.5964	0.6551	0.6285
Wine	95	0.9356	0.956	0.9512	0.9488	0.9461	0.9634	0.9717	0.9671	0.9512	0.9545
MNIST	5327	0.9684	0.9289	0.9092	0.9144	0.9195	0.9158	0.9210	0.9276	0.9073	0.9180
Olivetti-Faces	211	0.9641	0.6591	0.5271	0.6175	0.6418	0.7826	0.8492	0.6888	0.7989	0.886
COIL-20	729	0.9942	0.9705	0.9464	0.9536	0.9425	0.9690	0.9743	0.9263	0.9760	0.9785
Random Sample of size equal to Optimal Sample plus (t,m,s)-Nets Sample											
IRIS	79	0.9626	1.0	0.9705	0.9661	0.9772	0.9558	0.9657	0.9866	0.9705	0.9728
Breast Cancer	305	0.9339	0.9802	0.9734	0.9806	0.9902	0.9924	0.9894	0.9542	0.9772	0.9683
Leukemia	41	0.5557	0.9492	1.0	0.9732	0.6333	0.6451	0.6365	0.7289	0.4516	0.679
Wine	95	0.9434	0.9444	0.9518	0.949	0.9456	0.9638	0.9493	0.9666	0.9518	0.9487
MNIST	5327	0.9616	0.9326	0.8977	0.9138	0.9316	0.8977	0.9172	0.9433	0.8959	0.9177
Olivetti-Faces	211	0.9708	0.6642	0.4867	0.6011	0.6938	0.7671	0.8108	0.6673	0.7724	0.85
COIL-20	729	0.9949	0.9838	0.9381	0.9558	0.9381	0.9620	0.9721	0.9330	9648	0.9666

with representative k-NN sampling-based clustering labels as the colors of scatter point groups.



(a) Proposed representative k-NN sampling

(b) Random sample of size equal to the size of proposed representative k-NN sampling

(c) Proposed representative k-NN sampling along with the samples from remain samples using (t,m,s)-Nets

(d) Random sample of size equal to the size of proposed representative k-NN sampling plus (t,m,s)-Nets samples

Figure 5: Training sample selection (i.e., represented by blue star scatter-plot point) from IRIS dataset with four different strategies.

The *Figure 8* shows the four different 2D scatter-plot representation of MNIST dataset. In *Figure 8(a)*, we are representing the 2D scatter-plot of sampled t-SNE along with the interpolation of inliers and placement of outliers. The sampled t-SNE is designed based on the best representative sample of the representative k-NN sampling-based clustering concerning the best clustering technique. The inliers of remaining samples (i.e, other than representative samples) are interpolated to t-SNE with the parameter r_x that is obtained from the proposed heuristic. The outlier of remaining samples are placed into an existing t-SNE environment based on the Boystov et.al heuristic.

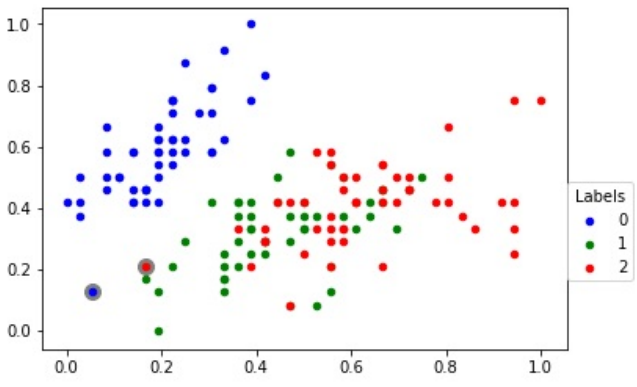


Figure 6: The 2D scatter-plot of IRIS data which denotes the outliers from the newly added sample with proposed heuristic.

In *Figure 8(a)*, the outliers are placed separately from other embedding points. The *Figure 8(b)* shows the 2D scatter-plot of

(t,m,s)-Nets based t-SNE along with interpolation of other samples. The (t,m,s)-Nets based t-SNE is designed with a combination of representative k-NN sampling-based clustering and (t,m,s)-Nets of remaining samples. The (t,m,s)-Nets samples are selected from the data samples, which are having NN-score as zero after selecting the samples using a mutual k-NN sampling with a static graph updation algorithm. The inliers of remaining samples (i.e, other than representative samples plus (t,m,s)-Nets samples) are interpolated to t-SNE with the parameter r_x that is obtained from the Boystov et.al heuristic. The *Figure 8(c)* and *8(d)* shows the 2D scatter-plots of a random sampled t-SNE along with the addition of new data samples. The random sample size is equal to the sample size of *Figure 8(a)* and *8(b)*. The inliers and outliers of remaining samples are placed according to the intuitions of *Figure 8(a)* and *8(b)*. The representative k-NN sampling-based results are statistically significant than the random sampling based results which is derived in our earlier paper [19]. The following section describe the conclusion and future work.

6 Conclusion and Future Work

In this paper, we have proposed a representative k-NN sampling-based clustering approach, which generates cluster results on a sampling basis. The most frequently used clustering techniques are applied to obtain the sampling-based cluster results. Initially, we determine the applicable set of techniques for the given dataset. From the applicable set, sampling-based clustering results of each technique are evaluated by an external cluster validation index called FMI-score. The applicable techniques are arranged in an orderly sequence of their FMI scores. Some threshold parameter derives the recommendation of clustering techniques for the given dataset. From the recommended set, the first technique is selected as the most desirable clustering for the given dataset.

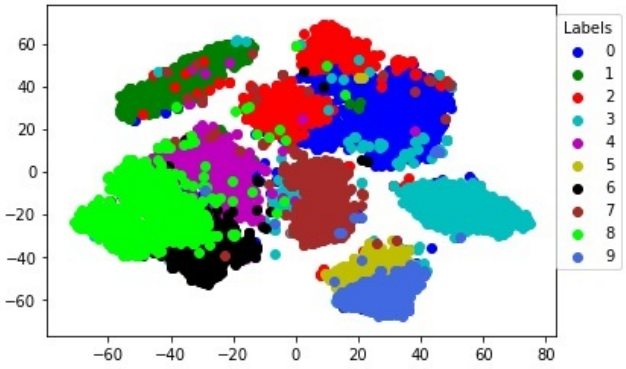
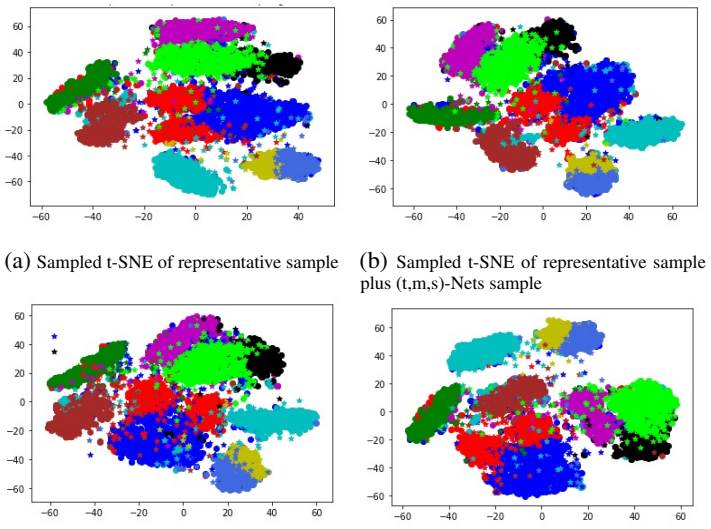


Figure 7: The baseline 2D visual representation of Representative k-NN sampling-based clustering t-SNE embedding where t-SNE applied on whole data

The samples of this technique are used for generating the low-dimensional embedding of input data. The embedding results are visualized and quantitatively evaluated in the context of sampling-based cluster results. The proposed approach is expanded to identify all the applicable set of clustering techniques for the given dataset, which can be done as future work. There is scope for implementing the proposed algorithm in a distributed environment that can be applied to Big Data.



(a) Sampled t-SNE of representative sample

(b) Sampled t-SNE of representative sample plus (t,m,s)-Nets sample

(c) Sampled t-SNE of random sample of size equal to representative sample size

(d) Sampled t-SNE of random sample of size equal to representative sample plus (t,m,s)-Nets sample size

Figure 8: 2D representation of four different t-SNE embedding of MNIST dataset, (*a*) t-SNE embedding of representative train sample and the interpolation of inliers as well as placement of outliers with proposed heuristic, (*b*) t-SNE embedding of representative train sample which includes the samples from (t,m,s)-Nets and the interpolation of inliers as well as placement of outliers with A Boystov et.al heuristic, (*c*) t-SNE embedding of random sample of size equal to the size of sub-figure (*a*) and the interpolation of inliers as well as placement of outliers with proposed heuristic, (*d*) t-SNE embedding of random sample of size equal to the size of sub-figure (*b*) and the interpolation of inliers as well as placement of outliers with A Boystov et.al heuristic.

Acknowledgment I would like to express my sincere gratitude to Visveswaraya PhD Scheme for providing financial support to pursue my PHD degree.

References

[1] M. Partridge, R.A. Calvo, "Fast dimensionality reduction and simple PCA" Intelligent data analysis, **2**(3), 203-214, 1998.

[2] T.F. Cox, M.A. Cox, Multidimensional scaling, Chapman and hall/CRC, 2000.

[3] J. Shi, J. Malik, "Normalized cuts and image segmentation" IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**(8), 888-905, 2000, doi: 10.1109/34.868688.

[4] J.B. Tenenbaum, V.D. Silva, J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction" Science, **290**(5500), 2319–2323, 2000. DOI: 10.1126/science.290.5500.2319

[5] S.T. Roweis, L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding" Science, **290**(5500), 2323-2326, 2000. DOI: 10.1126/science.290.5500.2323

[6] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation" Neural computation, **15**(6), 1373-1396, 2003. https://doi.org/10.1162/089976603321780317

[7] G.E. Hinton, S.T. Roweis, "Stochastic neighbor embedding" in Advances in neural information processing systems, pp. 857-864, 2003.

[8] L.V.D. Maaten, E. Postma, J.V.d. Herik, "Dimensionality Reduction: A Comparative" J Mach Learn Res, **10**(13), 66-71, 2009.

[9] L.V.D. Maaten, G. Hinton, "Visualizing data using t-sne" J Mach Learn Res, **9** 2579-2605, 2008.

[10] W. Li, J.E. Cerise, Y. Yang, H. Han, "Application of t-sne to human genetic data" Journal of Bioinformatics and Computational Biology, **15**(4) 1750017, 2017. https://doi.org/10.1142/S0219720017500172

[11] M. Nguyen, S. Purushotham, H. To, C. Shahabi, "m-tsne: A framework for visualizing high-dimensional multivariate time series" arXiv preprint arXiv:1708.07942, 2017.arXiv:1708.07942

[12] W.M. Abdelmoula, B. Balluff, S. Englert, J. Dijkstra, M.J.T. Reinders, A. Walch, L.A. McDonnell, B.P.F. Lelieveldt, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data" in Proceedings of the National Academy of Sciences, **113**(43) 12244-12249, 2016. https://doi.org/10.1073/pnas.1510227113

[13] L.V.D. Maaten, "Accelerating t-sne using tree-based algorithms" J Mach Learn Res, 15, 3221-3245, 2014.

[14] N. Pezzotti, B.P.F. Lelieveldt, L.V.D. Maaten, T. Höllt, E. Eisemann, A. Vilanova, "Approximated and user steerable tsne for progressive visual analytics" IEEE transactions on visualization and computer graphics, **23**(7) 1739-1752, 2016. DOI: 10.1109/TVCG.2016.2570755

[15] L.V.D. Maaten, "Learning a parametric embedding by preserving local structure" in Artificial Intelligence and Statistics, pp 384-391, 2009.

[16] A. Gisbrecht, A. Schulz, B. Hammer, "Parametric nonlinear dimensionality reduction using kernel t-sne" Neurocomputing, 147, 71-82, 2015. https://doi.org/10.1016/j.neucom.2013.11.045

[17] A. Boytsov, F. Fouquet, T. Hartmann, Y. LeTraon, "Visualizing and exploring dynamic high-dimensional datasets with lion-tsne" arXiv preprint arXiv: 1708.04983, 2017. arXiv:1708.04983

[18] S. Ingram, T. Munzner, "Dimensionality reduction for documents with nearest neighbor queries" Neurocomputing, 150 557-569, 2015. https://doi.org/10.1016/j.neucom.2014.07.073

[19] B. Dharamsotu, K.S. Rani, S.A. Moiz, C.R. Rao, "k-NN sampling for visualization of dynamic data using LOIN-tSNE" in 2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC), pages 63–72, 2019. DOI: 10.1109/HiPC.2019.00019

[20] L.J. Williams, H. Abdi, "Fishers least significant difference (lsd) test," Encyclopedia of research design, 218, 840-853, 2010.

[21] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation" IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**(7) 881-892, 2002. DOI: 10.1109/TPAMI.2002.1017616

[22] W. Zhang, D. Zhao, X. Wang, "Agglomerative clustering via maximum incremental path integral" Pattern Recognition, **46**(11) 3056-3065, 2013. https://doi.org/10.1016/j.patcog.2013.04.013

[23] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: an efficient data clustering method for very large databases" ACM Sigmod Record, **25**(2) 103-114, 1996. https://doi.org/10.1145/235968.233324

[24] J¿ C. Bezdek, R. Ehrlich, W. Full, "FCM: The fuzzy c-means clustering algorithm" Computers and Geosciences, **10**(2) 191-203, 1984.

[25] E. Schubert, J. Sander, M. Ester, H.P. Kriegel, X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN" ACM Transactions on Database Systems (TODS), **42**(3) 1-21, 2017. https://doi.org/10.1145/3068335

[26] J.C. Dassun, A. Reyes, H. Yokoyama, P.B.E. Jireh, M. Dolendo, "Ordering points to identify the clustering structure algorithm in fingerprint-based age classification" Virtutis Incunabula, **2**(1) 17-27, 2015.

[27] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis" IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**(5) 603-619, 2002. DOI: 10.1109/34.1000236

[28] U.V. Luxburg, "A tutorial on spectral clustering" Statistics and Computing, **17**(4) 395-416, 2007.

[29] D.A. Reynolds, "Gaussian mixture models" Encyclopedia of biometrics, 741, 2009.

[30] B.J. Frey, D. Dueck, "Clustering by passing messages between data points" Science, **315**(5814) 972-976, 2007. DOI: 10.1126/science.1136800

[31] David Sculley, "Web-scale k-means clustering" in Proceedings of the 19th international conference on World Wide Web, pages 1177-1178, 2010. https://doi.org/10.1145/1772690.1772862

[32] E.B. Fowlkes, C.L. Mallows, "A method for comparing two hierarchical clusterings" Journal of the American Statistical Association, **78**(383) 553–569, 1983.

[33] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, S. Wu, "Understanding and enhancement of internal clustering validation measures" IEEE Transactions on Cybernetics, **43**(3) 982-994, 2013. DOI: 10.1109/TSMCB.2012.2220543

[34] S. Tsumoto, "Contingency matrix theory: Statistical dependence in a contingency table" Information Sciences, **179**(11) 1615-1627, 2009. https://doi.org/10.1016/j.ins.2008.11.023

[35] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, M.J. Er, W. Ding, C. Lin, "A review of clustering techniques and developments" Neurocomputing, 267 664–681, 2017. https://doi.org/10.1016/j.neucom.2017.06.053

[36] A.K. Jain, M.N. Murty, P.J. Flynn, "Data clustering: a review" ACM computing surveys (CSUR), **31**(3) 264-323, 1999. https://doi.org/10.1145/331499.331504

[37] P. Berkhin, "A survey of clustering data mining techniques" in Grouping multidimensional data, pp 25-71. Springer, 2006.

[38] A.Y. Ng, M.I. Jordan, Y. Weiss, "On spectral clustering: Analysis and an algorithm" in Advances in neural information processing systems, pp 849-856, 2002.

[39] G.C. Linderman, S. Steinerberger, "Clustering with t-sne, provably" SIAM Journal on Mathematics of Data Science, **1**(2) 313–332, 2019. https://doi.org/10.1137/18M1216134

[40] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data" in Proceedings of the 1968 23rd ACM national conference, pp 517-524, 1968. https://doi.org/10.1145/800186.810616

[41] J. Venna, S. Kaski, "Neighborhood preservation in nonlinear projection methods: An experimental study" in International Conference on Artificial Neural Networks, pp 485-491, Springer, 2001.

[42] T. Kollig, A. Keller, "Efficient multidimensional sampling" Computer Graphics Forum, **21**(3), 2002.

[43] D, Dua, C. Graff, UCI machine learning repository, 2019.

[44] B. Schölkopf, A. Smola, K. Müller, "Kernel principal component analysis" in International conference on artificial neural networks, pages 583–588, Springer, 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 506-511 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Growth Models and Age Estimation of Rice using Multitemporal Vegetation Index on Landsat 8 Imagery

Abdi Sukmono*, Arief Laila Nugraha, Arsyad Nur Ariwahid, Nida Shabrina

Department of Geodetic Engineering, Faculty of Engineering, Diponegoro University, Semarang, 50275, Indonesia

ARTICLE INFO

Article history:
Received: 16 July, 2020
Accepted: 29 August, 2020
Online: 24 September, 2020

Keywords:
Age of Rice
Growth Model of Rice
Multitemporal EVI

ABSTRACT

Age and growth are two essential rice biophysics parameters used to determine the health parameters and production rate. The spatial data of both parameters can utilize remote sensing technology, which in turn makes use of several vegetation indices to achieve accurate estimation. However, due to the rapid changes in rice plants' characteristics, it is essential to study vegetation index utilization using a multitemporal method to improve its accuracy. Therefore, this research uses a multitemporal Enhanced Vegetation Index (EVI) to estimate rice's age and growth model. The multitemporal EVI patterns were observed to estimate the Time Early Planting (TEP) and the maximum EVI value of rice in an area. The results showed that the maximum EVI value in the rice fields of Demak Regency has a class range of 0.4 to more than 0.9. The highest value is in the class of 0.80 - 0.85 covering 12023.28 ha, followed by 0.75 - 0.80 at 11834.19 ha. Furthermore, the multitemporal EVI method on Landsat 8 images was used to estimate the rice age with accuracy or RMSE of 7.7 days. The result also showed that this value is good enough because the RMSE is still in the same range of paddy growth phases.

1. Introduction

Rice is the essensial food for about half of the world's population. It is supply 20 % of the calories consumed worldwide [1]. Almost 90 % the global rice is produced in Asia, Africa and America Latin [2]. Most of country in thats region used rice as the staple food, especiaally in Indonesia. According to Saliem [3], close to 100% of the Indonesian population consume this product, thereby making it a staple food, with significant economic importance. The continuous increase in population leads to a rise in rice demand, therefore, an essential management strategy, such as the rice intensification system, to boost production [4].

The intensification and management of rice cultivation are inseparable from high-tech approaches due to its ability to monitor biophysical parameters such as the growth phase closely. Detailed information on the growing phase is needed to evaluate rice development [5],[6]. Furthermore, the biophysical monitoring of these parameters is valuable for the growth model and prediction of rice production [7], [8]. This growth phase model is also closely related to rice's age and used to obtain predictable information related to future harvest time.

*Corresponding Author: Abdi Sukmono, Diponegoro University, +62 85733065477, sukmono35@gmail.com

Spatial data is needed to monitor rice growth and age models for proper distribution and analysis using remote sensing technology. Several approaches are used to estimate the rice-growing phase, such as the Vegetation Index value approach, as stated in a research carried out by [9] and [10]. However, this study utilizes the Normalized Difference Vegetation Index (NDVI) to estimate rice growth. NDVI is a method occasionally used to monitor vegetation growth from space [11], [12]. Furthermore, various vegetation index techniques were developed to improve their accuracy and efficiency. Therefore, rice plants with a fast-changing level of green sensitivity utilized various combinations of techniques. The multitemporal data technique is one of the methods that can provide a solution [13]. Generally, indices such as NDVI show the growth status of green vegetation, therefore plant monitoring can be realized using remote sensing with time series or multitemporal data [14].

The utilization of this technique was carried out by [15] to detect rice biological parameters with NDWI and multitemporal EVI on Modis imagery, which is good for growth monitoring because of its daily temporal resolution. However, this method has a low spatial resolution, therefore it is only intended for regional plant growth studies. NDVI also affects soil and atmospheric reflection [16]. This tends to affect rice plants that have background soil types and diverse vegetation densities. [17] carried out a research to modify modified NDVI and ensure the

canopy background and atmospheric noise are minimized to Enhanced vegetation Index (EVI). Therefore, this study aims to utilize multitemporal Enhanced vegetation Index (EVI) techniques on Landsat 8, which has a moderate spatial resolution that is sufficiently resolved to natural change from local to global scale [18]. Landsat 8 imagery needs to be assessed to determine its use in monitoring growth models and estimating the rice age due to its efficiency, medium spatial resolution, and freedom of usage. Therefore, this study's results can be used to estimate the growth model and age of rice in a cheap, efficient, and accurate method.

2. Method

2.1. Description of Study Area

This study was conducted in all rice fields in Demak Regency, Central Java province, Indonesia. This region is the second-largest rice producing region in Central Java Province. In 2019, rice fields were 106 629.56 hectares (ha) with the production of 666 141, 30 Metric Ton (MT) of dry grain [19]. Demak is located adjacent to the Capital of Central Java Province, making it the major food supplier to Semarang City. The rice fields in this province have similar characteristics in terms of shape, function, and varieties. The paddy has a large plot above 100 m x 100 m, thereby making it suitable for analysis using pixel-based satellite image data, where the pixel size of the Landsat image is 30 m x 30 m. The majority of the Paddy fields in Demak are irrigated and dominated by Ciherang and Mekongga rice types, with some comprising of IR 64 during the dry season. When the same rice varieties varieties are planted, it allows for a fairly balanced analysis to estimate its growth model. The study areas in this study are shown in Figure 1.

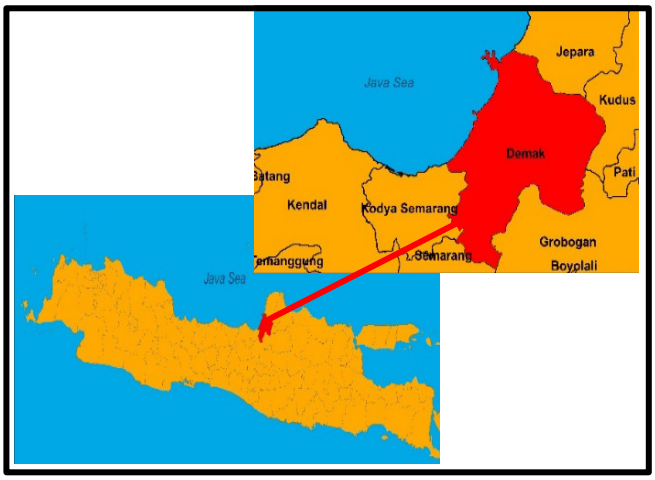


Figure 1: Location of Study Area

Paddy in Demak Regency is planted twice a year, namely season 1 (March - June) and season 2 (September-December). This study was conducted from March –June 2019, also known as planting season 1.

2.2. Preparation and Preprocessing Data

Data were obtained from Landsat 8 OLI multi-temporal acquisition images on March 4, 2019, March 20, 2019, April 5, 2019, April 21, 2019, May 8, 2019, May 24, 2019, June 9, 2019, and June 25, 2019. Therefore, the acquisition images were obtained for 4 months with a range of 16 days. Furthermore, supporting data was also prepared, using the Indonesian Topographic Map in Demak Regency for geometric tests and rice field map for cropping images at a scale of 1: 25,000.

Image preprocessing was carried out to determine the quality of images from geometric, radiometric, and atmospheric errors. In this study, Landsat 8 OLI imagery was tested for geometric accuracy using the image to map method. This test used 20 test points (Independent Control Points) extracted from Topographic Map data on a scale of 1: 25,000 and distributed in the study area. This study's geometric accuracy-test requirements need to meet Circular Error 90 (CE 90) using a map accuracy of 1: 25,000. This was followed by determining the geometrical quality and radiometric calibration process. The result of accuracy test of geometric has the RMSE 24.9 m. So it meets the requairement of the map accuracy 1 : 25.000. This research also converted the pixel value of DN (Digital Number) and image data to the BoA (Bottom of Atmospheric) reflected value. This was carried out using the second simulation of a satellite signal in the solar spectrum-vector (6SV) algorithm. Furthermore, it also utilized the sun's azimuth angle at the time of image recording. The final image preprocessing stage was used to carry out the cloud removal by the BQA (Band Quality Assessment) method before the image is cropped with the Rice Field Area of Interest in Demak Regency for effective and efficient processing.

2.3. Growth of Rice Modelling Method

Rice growth models were obtained using multitemporal vegetation index values at 30 sampling points to obtain a chart value. The rice growth pattern in the graph is polynomial and shows the level of the greenness of plants from one time to another. Healthy plants have a high maximum vegetation index value on the graph.

This study utilized the EVI (Enhanced Vegetation Index), which is an improvement from the previous algorithm, known as NDVI (Normalized Difference Vegetation Index). This Algorithm is sensitive to the "red" band channel that absorbs the canopy, with lower optical penetration depth, thereby, enabling it to absorb more quickly in high biomass areas. EVI is becoming increasingly sensitive to NIR bands on moderate to high amounts of vegetation with greater optical depth penetration into the canopy. Therefore, it is better to describe variations in biophysical canopies' structure and fails to absorb in areas with high biomass [20], [21].

EVI is derived from the soil-adjusted vegetation index (SAVI) and atmosphere-resistant vegetation index. It is an optimized combination of blue, red, and NIR channels, based on Beer's law of canopy radiation transfer, designed to extract its greenness, regardless of soil background and atmospheric aerosol variations [21] as shown in equation 1.

$$EVI = \mathrm{G} * \frac{NIR - RED}{(\mathrm{L} + NIR + C1\ RED + C2\ BLUE)} \quad (1)$$

Where L value is a calibration factor of the canopy and soil effect, G is a scale factor, with the EVI value in the range between -1 to 1. Furthermore, C1 and C2 are the weight of aerosol resistance. The variable and values of the equation coefficients above are L = 1, C1 = 6, C2 = 7.5, and G = 2.5 [22].

2.4. Age of Rice Estimation Method

The rice plants' growth profile is determined using the Landsat 8 imagery, which has a multi-temporal 16-day recording time using the EVI algorithm. The beginning of planting was marked by an EVI value lower than the harvest time and vacant land. Based on vegetation index changes, it is used to determine the spatial distribution of age (HST) and rice growth at a certain time or period [15].

The early growth phase of rice had EVI values lower than at harvest time (100-120 days) with vacant land. The Time Early Planting (TEP) can be spatially estimated when the maximum EVI is obtained. It can be calculated using the modified equation from [15], as shown in equation (2).

$$TEP = T_{-EVI}Max - \frac{1}{2} * (\frac{LP}{Period}) \quad (2)$$

Where $T_{-EVI}Max$ is the time when the maximum EVI value is obtained with the unit in Julian date, LP is the planting time for rice plants, and the period is the temporal resolution of Landsat 8. When the TEP value is obtained, the age of the rice at the time of image acquisition is determined by calculating the difference in the acquisition date with TEP. So the equation modified become equation 3.

$$TEP = T_{-EVI}Max - \frac{1}{2} * (\frac{110}{16}) \quad (3)$$

Equation 3 was modified the band in EVI and Landsat 8 period or temporal resolution with the rice planting time in study area is 110 days.

2.5. Limitation of The Study

This research was conducted in a research area characterized by irrigated rice fields during the dry season. rice varieties in the study area were homogeneous, most of which planted medium time planting rice. So that the rice planting time (LP) is assumed to be homogen, that is 110 days.

3. Result and Discussion

3.1. Growth Model of Rice Result

The development of rice growth models is obtained by arranging a multitemporal EVI pattern (March 4 2019, March 20 2019, April 5 2019, April 21 2019, May 8 2019, May 24 2019, June 9 2019, and June 25 2019) to produce its graphical representation. In this study, 25 sampling points were chosen to determine the multitemporal EVI pattern in the study area with the graph shown in figure 2.

Figure 2 shows more than one whole wave (one peak and one trough) on the multitemporal EVI graph. This is used to illustrate the model of rice growth patterns in the study area. It also means that there is more than one planting season in the study area. The minimum EVI value (trough) indicates that the land was vacant at that time, and the maximum value indicates that the rice was at the peak greening phase.

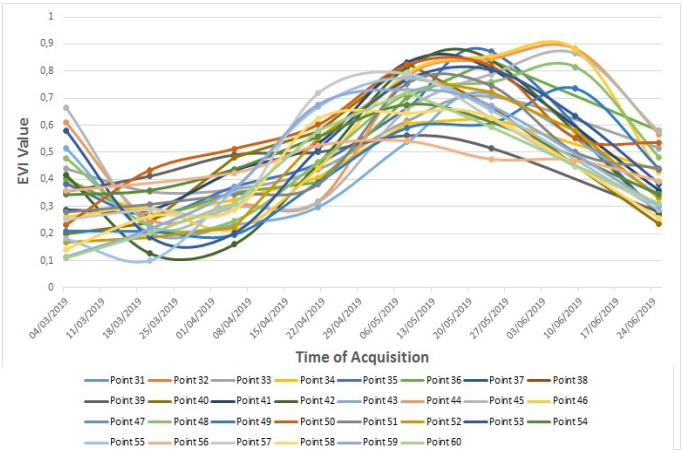


Figure 2: Graph of multitemporal EVI patterns at sample points

In general, the rice growth pattern shows the EVI value moves up towards the peak and decreases with the plant turning yellow when the rice has passed the flowering phase. This is influenced by rice flowers' effects and continues to decline because the seeds start maturing while drying the stems. A maximum EVI image is needed to determine the image with the highest value at each pixel. It is necessary because the highest pixel values among all multitemporal images represent the phase with the highest greenish level at 60-64 days after planting (HST), also known as the flowering phase. The maximum EVI that has been obtained is classified into 11 classes with an interval of 0.05, as shown in Figure 3.

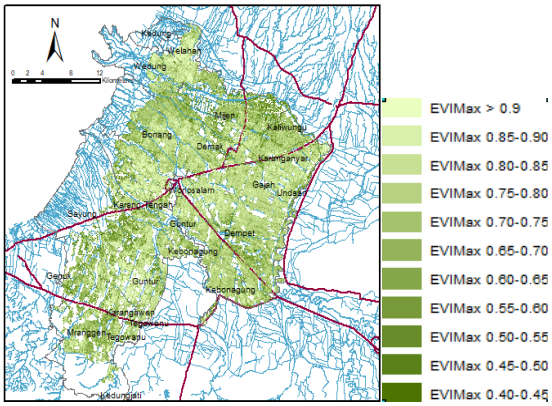


Figure 3: Maximum EVI classification results in Demak Regency

Figure 3 shows the spatial distribution of EVI max for each rice field area. This EVI class shows that the higher the value, the healthier the plants in the rice fields. This classification can be recapitulated in each class area shown in Table 1 and the graph in Figure 4.

Table 1 and Figure 4 show each EVI Max class area, which indicates that the EVI Max rice in Demak Regency in the first Planting Season of 2019 is the widest at 0.80-0.85 covering 12023.28 ha, followed by 0.75-0.80 ha at 11834 ha. Figure 4 shows that more than 70% of rice fields in Demak Regency in the planting season 1 of 2019 have an EVI Max above 0.70. Based on [16], the high EVI and NDVI represent the healthy vegetation. Therefore, the pattern of rice growth in the first planting season is healthy,

with the greenness of rice (chlorophyll), which is closely related to crop health.

Table 1: Area of each EVI Max class in Demak Regency rice field

Class of EVI Max	Area (ha)
EVIMax 0.4-0.45	961.02
EVIMax 0.45-0.50	1299.42
EVIMax 0.50-0.55	1800.99
EVIMax 0.55-0.60	2617.92
EVIMax 0.60-0.65	3781.71
EVIMax 0.65-0.70	5665.05
EVIMax 0.70-0.75	8542.08
EVIMax 0.75-0.80	11834.19
EVIMax 0.80-0.85	12023.28
EVIMax 0.85-0.90	7320.6
EVIMax > 0.9	2731.32

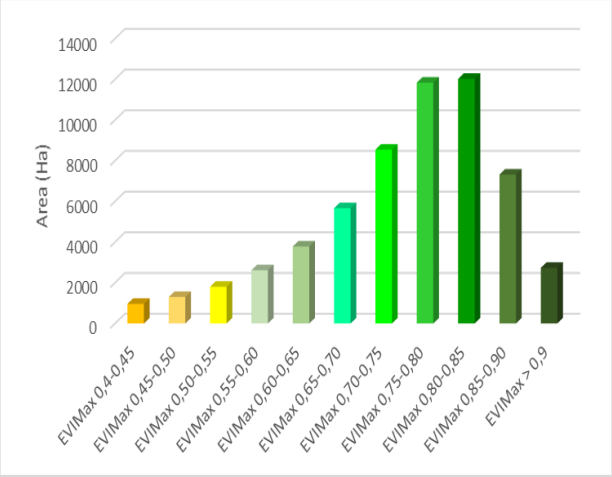


Figure 4: Graphic of Area in each EVI Max class of Rice field in Demak Regency from March – June 2019

3.2. Age Estimation of Rice Result

In this study, the age of rice was calculated using the Time Early Planting (TEP) estimation approach. This was obtained by running equation 1 on the EVI Max image, which was previously extracted from the time when EVI Max occurred on the Julian Date unit. Age estimation results are obtained from the difference between the time of acquisition or observation with TEP. The results of processing Time Early Planting and Age Estimation of rice in Demak Regency as of July 26, 2019 are shown in Figure 5 and Table 2.

Figure 5 and Table 2 show the results of Time Early Planting (TEP) and Age Estimation of Rice in Demak Regency on July 26, 2019. The widest TEP and Age Estimation was obtained on March 20, 2019 (Age of 99 Days) covering 8482, 245 Ha, followed by March 4, 2019 (Age of 115 Days) at 7088, 355 Ha and March 21, 2019 (Age of 67 Days) at 6846, 277 Ha. It shows that the rice field in Demak Regency on July 26, 2019, was dominated by Ripening's growing phase. The harvesting time estimation is obtained by adding the Time Early planting (TEP) with the Long Period (LP). It is the planting time for rice in the study area. Most of the rice varieties in Demak in planting season II were medium rice types (aged after planting 110 days). So that the harvesting time can be obtained by adding 110 days to TEP.

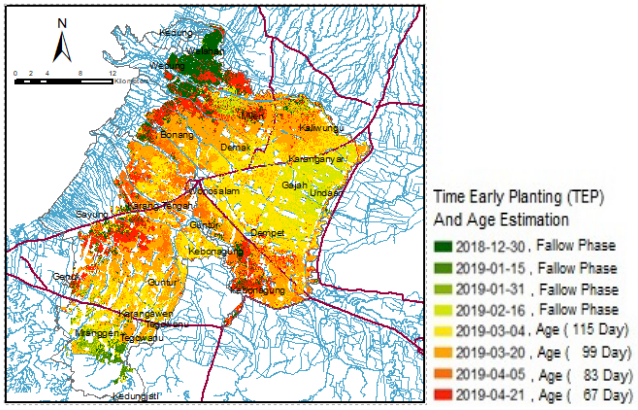


Figure 5: Time Early Planting and Age Estimation as of July 26, 2019

Table 2: Time Early Planting estimation result

Time Early Planting Estimation	Time Harvesting Estimation	Area (ha)
2018-12-30	2031-11-30	4718.604
2019-01-15	2021-04-21	827.056
2019-01-31	2021-11-16	1020.586
2019-02-16	2030-09-13	4227.501
2019-03-04	2038-07-30	7088.355
2019-03-20	2042-06-09	8482.245
2019-04-05	2029-01-31	3589.963
2019-04-21	2038-01-17	6846.277

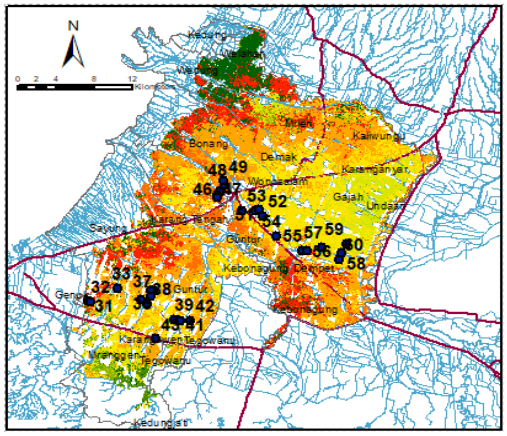


Figure 6: Spatial distribution of the validation sample point

3.3. Validation

Validation is used to determine the accuracy of rice's estimated age with the Time Early Planting (TEP) approach of EVI

Multitemporal Landsat 8. In this study, this process was carried out at 25 points distributed in the study area, as shown in Figure 6. This was carried out by field observations and interviews with farmers in order to obtain the actual age of rice. The validation was carried out on July 26, 2019.

The distributed validation test sample points also considered the representation of each rice age estimation results using the proportional random sampling technique. It aims to ensure that the validation is sufficient to meet the existing population. The results of this validation are shown in Table 3.

Table 3: Validation result of rice age estimation

Sample Point	Age of Rice In Situ	Age of rice Estimation	Error (Days)
31	48	67	19
32	59	67	8
33	95	99	4
34	96	99	3
35	104	99	5
36	104	99	5
37	97	99	2
38	106	99	7
39	104	115	11
40	117	115	2
41	109	115	6
42	98	99	1
43	104	99	5
44	97	83	14
45	97	83	14
46	95	83	12
47	95	83	12
48	82	83	1
49	87	83	4
50	107	115	8
51	107	115	8
52	107	115	8
53	92	99	7
54	112	115	3
55	112	115	3
56	114	115	1
57	117	115	2
58	112	115	3
59	117	115	2
60	107	115	8
		RMSE	7.698

The Root Mean Square Error (RMSE) or the accuracy of the multitemporal EVI method is obtained to estimate the rice age by 7.698 days. This is sufficient when compared to the Landsat 8 temporal resolution of 16 days. Furthermore, this method is quite effective for identifying vegetative, reproductive, and ripening phases. This value is good enough because the RMSE is in the same range of growth phases of rice. Another study by [23] using time series MODIS data with the method wavelet and fourier transform in EVI graphic to estimate phenological age of rice. The result of this method has the RMSE of 12.1 day. Then [24] also used the MODIS time series data to estimate phenological stage. They used savitky golay transform in EVI multitemporal graphic. It has RMSE of 10 days. The last study propose by [25] used the machine learning in one time Landsat 8 imagery to clasification the growth phases of rice. The result of this study show that SVM radial data has the overal accuracy 70,5 % in classification of growth phase of rice. Therefore multitemporal EVI using the TEP method of Landsat 8 that has a RMSE of 7.698 days, was appropriate to estimate the age of rice..

4. Conclusion

The following conclusions were made based on the research analysis:

In conclusion, the multitemporal EVI from Landsat 8 imagery can describe rice growth models with the ability to provide information on the number of growing seasons in the study area and the health level of rice from the maximum EVI value. Furthermore, more than 70% of rice fields in the Demak Regency in the planting season 1 of 2019 have an EVI Max above 0.70, therefore it is in the quite healthy category.

The study also showed that the multitemporal EVI method of Landsat 8 Imagery can estimate rice age using the Time Early Planting (TEP) approach. It has an RMSE accuracy of 7.7 days and sufficient for monitoring the rice age because the RMSE value is still in the same growth phase range.

Conflict of Interest

The authors declared no conflict of interest.

Acknowledgment

The authors are grateful to the Faculty of Engineering – Diponegoro University for funding this research through the 2020 Strategic Research grant. This article publication is suported by World Class University (WCU) Program from Diponegoro Univeristy. We are also grateful to all those that supported this research.

References

[1] M. Kubo, M. Purevdorj. The Future of Rice Production and Consumption, Journal of Food Distribution Research **35**(1), 129-142, 2004.

[2] Bandumula, N. Rice Production in Asia: Key to Global Food Security. Proc. Natl. Acad. Sci., India, Sect. B Biol. Sci. **88**, 1323–1328, 2018. https://doi.org/10.1007/s40011-017-0867-7

[3] H. Saliem, N. Hermanto, E. Suryani, R. Suhaeti, & M. Arian. The Dynamics of Indonesian Consumption Patterns of Rice and Rice-Based Food Eaten Away From Home. Analisis Kebijakan Pertanian, [S.l.], **17**(2), 95-110, dec. 2019. ISSN 2549-7278. doi: http://dx.doi.org/10.21082/akp.v17n2.2019.95-110

[4] H. Maridjo, Y. Mudayen, & A. Prihantoro. Increased productivity and technical efficiency of rice farming with the System of Rice Intensification (SRI) method in Purworejo District, Central Java. Journal of Economics, Business, & Accountancy Ventura, [S.l.], **19**(1), 49-58, july 2016. doi: http://dx.doi.org/10.14414/jebav.v19i1.535

[5] J. R. Porter, M. A. Semenov . Crop responses to climatic variation. Phil. Trans. R. Soc. B **360**: 2021–2035. 2005. doi: http://doi.org/10.1098/rstb.2005.1752

[6] S. FUKUI, Y. ISHIGOOKA, T. KUWAGATA, T. HASEGAWA. A methodology for estimating phenological parameters of rice cultivars utilizing data from common variety trials. Journal of Agricultural Meteorology, **71**(2), 77-89. 2015. https://doi.org/10.2480/agrmet.D-14-00042

[7] W. Koppe, M. L. Gnyp, C. Hut, Y. Yao, Y. Miao, X. Chen, G. Bareth. Rice monitoring with multi-temporal and dual-polarimetric TerraSAR-X data. Int

J Appl Earth Obs Geoinformation **21**, 568–576. 2012. https://doi.org/10.1016/j.jag.2012.07.016

[8] R. Nurtyawan, A. Saepuloh, A. G. Harto, K. Wikantika, A Kondoh. Satellite Imagery for Classification of Rice Growth Phase Using Freeman Decomposition in Indramayu, West Java, Indonesia. HAYATI Journal of Biosciences, **25**(3), 126-137, 2014. https://doi.org/10.4308/hjb.25.3.126

[9] K. L. Liu, Y. Z. Li, H. W. Hu. Predicting ratoon rice growth rhythm based on NDVI at key growth stages of main rice. Chilean J. Agric. Res. **75(**4), 410-417 Chillán dic. 2015. http://dx.doi.org/10.4067/S0718-58392015000500005.

[10] R Rosle, N. N. Che'Ya, N. A. Roslin, R. M. Halip, M. R. Ismail. Monitoring Early Stage of Rice Crops Growth using Normalized Difference Vegetation Index generated from UAV. IOP Conf. Series: Earth and Environmental Science **355**, 2019. 012066. doi:10.1088/1755-1315/355/1/012066

[11] P. Santos, A. J. Negri. A Comparison of the Normalized Difference Vegetation Index and Rainfall for the Amazon and Northeastern Brazil. J. Appl. Meteor. (1997) **36** (7): 958–965. 1997. https://doi.org/10.1175/1520-0450(1997)036<0958:ACOTND>2.0.CO;2

[12] C. J. Trucker, P. J. Sellers. Satellite remote sensing of primary production. Int. J. Remote Sens., **7**, 1395–1416. 1986. https://doi.org/10.1080/01431168608948944

[13] C. Yonezawa, M. Negishi, K. Azuma, M. Watanabe, N. Ishitsuka, S. Ogawa, G. Saito. Growth monitoring and classification of rice fields using multitemporal RADARSAT-2 full-polarimetric data. Int. J. Remote Sens., **33**(18), 2012. https://doi.org/10.1080/01431161.2012.665194

[14] Jiang, D., Wang, N., Yang, X. et al. Study on the interaction between NDVI profile and the growing status of crops. Chin. Geograph.Sc. **13**, 62–65. 2003. https://doi.org/10.1007/s11769-003-0086-4

[15] D. D. Domiri. The method for detecting biological parameter of rice growth and early planting of paddy crop by using multi temporal remote sensing data. IOP Conf. Series: Earth and Environmental Science **54**(2017) 012002. 2018. doi:10.1088/1755-1315/54/1/012002

[16] J. Xue, B. Su. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. Journal of Sensors **2017**, Article ID 1353691, 17 pages. https://doi.org/10.1155/2017/1353691

[17] H. Q. Liu and A. Huete, "Feedback based modification of the NDVI to minimize canopy background and atmospheric noise," IEEE Transactions on Geoscience and Remote Sensing, **33**(2), 457–465, 1995.

[18] D. P. Roy, M. A. Wulder, T. R. Loveland, C. E. Woodcock, R. G. Allene, M. C. Anderson. Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment, **145**, 154-172. 2014. https://doi.org/10.1016/j.rse.2014.02.001

[19] BPS. 2020. Central Java In Figure. BPS-Statistics of Central Java Province, 2020.

[20] X. Gao, A. Huete, W. Ni, T. Miura. Optical-biophysical relationships of vegetation spectra without background contamination. Remote Sens Environ **74**, 609–620. 2000. https://doi.org/10.1016/S0034-4257(00)00150-4

[21] A. Huete, K. Didan, W. V. Leeuwen, T. Miura, E. Glenn. MODIS Vegetation Indices. In: Ramachandran B., Justice C., Abrams M. (eds) Land Remote Sensing and Global Environmental Change. Remote Sensing and Digital Image Processing, **11,** 2010 Springer, New York, NY. https://doi.org/10.1007/978-1-4419-6749-7_26

[22] B. D. Wardlow, S. L. Egbert. Analysis of time series MODIS 250m Vegetation index Data for Crop Calssification in the U.S Central Great Plains. Remote Sensisng of Environment, **83**, 97-111. https://doi.org/10.1080/01431160902897858

[23] T. Sakamoto, M. Yokozawa, H. Toritani, M. Shibayama, N. Ishitsuka, H. Ohno. A crop phenology detection method using time-series MODIS data. 2004. Remote Sensing of Environment **96**(3–4), 366-374, 2005. https://doi.org/10.1016/j.rse.2005.03.008

[24] L. Shihua, X. Jiangtao, N. Ping, Z Jing, W Hongshu, W. Jingxian. Monitoring paddy rice phenology using time series MODIS data over Jiangxi Province, China. Int J Agric & Biol Eng, 2014; **7**(6): 2 — 36. DOI: 10.3965/j.ijabe.20140706.005

[25] F. Ramadhani , R. Pullanagari , G. Kereszturi, J. Procter, Mapping of rice growth phases and bare land using Landsat-8 OLI with machine learning, International Journal of Remote Sensing, **41**(21), 8428-8452, 2020. DOI: 10.1080/01431161.2020.1779378

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 512-525 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Genetic Organization and Evolution of Electromechanical Objects with Adaptive Geometry of Active Zone

Vasyl Shynkarenko[1], Ali Makki[2], Viktoriia Kotliarova[1,*], Anna Shymanska[1], Pavlo Krasovskyi[1]

[1]*Department of Electromechanics, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 03056, Ukraine*

[2]*KEEP'MOTION Company, Luzinay, 38200, France*

ARTICLE INFO

Article history:
Received: 12 August, 2020
Accepted: 01 September, 2020
Online: 24 September, 2020

Keywords:
Electromechanical object
Active zone
Variable geometry
Electromagnetic chromosome
Genetic code
Macrogenetic program
Synthesis
"Elastic" electromechanics
Functional evolution

ABSTRACT

The paper is devoted to the presentation of a new methodological approach (new philosophy) to the formulation and solution of directed search and synthesis of electromechanical objects for a given function problems. The object under the study is a class of electrical machines and electromechanical devices which operation is carried out with a variable structure or geometry of the active zone. The research is relevant due to the characteristic trend in the evolution of modern technology which is associated with the creation of complex technical systems with the ability to change the structure and spatial geometry of the executive body in accordance with changing of external factors. The novelty of the synthesis methodology is determined by genetic nature of the technical evolution of electromechanical objects and by genetic programs of structure formation using. According to the results of research the genetic principles and macrogenetic programs of electromechanical objects with variable spatial geometry of active parts have been determined for the very first time. The area of existence and results of genetic synthesis of "elastic" electromechanics objects are presented. The reliability of genetic models, genetic programs and the results of the electromechanical objects with adaptive spatial geometry of the active zone synthesis is confirmed by the results of evolutionary experiments.

1. Introduction

One of the current trends in the development of mobile technical systems is associated with the transition from classical rigid forms and monostructures to spatially distributed structures with adaptive spatial geometry which functioning is as close as possible to natural analogues. This trend is relevant due to the growing demands of improving the dynamic characteristics and mass and size dimensions, increasing energy efficiency and reliability of technical systems based on the use of the latest materials and the widespread introduction of digital technologies. In the process of interaction between the object and external environment, the energy characteristics of the whole system change which is accompanied by a changing of the parameters and physical state of its elements. Factors of external influence often determine the necessity of the spatial geometry and structure of the system adaptation in the area of their direct interaction. Spatially adaptive systems and technologies are already used in modern aviation, space and military technics, electronics, automotive, robotics, machine-building and other science-intensive technical fields [1]-[5]. The latest direction of technology development has not yet received an unambiguous definition and occurs under various terminological names: "elastic mechanics", "machines with adaptive morphology", "soft robotics", "flexible electronics", "reconfigured systems", "objects-transformers", etc.

Analysis of the functional and technical evolution of electromechanical converters of energy (EMCE) shows a great variety of electrical machines and electromechanical devices, which operation of which is directly related to changes in the structure and geometry of their active zone in space and time. In the concept of structural-system approach, these electromechanical objects (EM-objects) and devices belong to the class of genetically organized systems (GOS) which structural diversity is determined by the relevant genetic programs and their development is regulated by systemic laws of genetic evolution [6].

The object of research is the functional classes of modular EMCE which operating routines are realized with a change in the spatial geometry and structure of the active parts. An

*Corresponding Author: Viktoriia Kotliarova, sharik_2004@ukr.net

informational sources review shows the absence of a systematic analysis of the structural organization and technical evolution of these objects and is currently limited to describing only their known technical implementations. The tasks of determination the principles of structure formation of EM-objects with variable geometry of the active zone and their genetic programs and of analysis the directions of their structural and functional evolution.

2. Genetic Organization of Electromagnetic Systems

The genetic approach in modern science acquires a key role in understanding of the fundamental principles of structural organization and laws of complex systems developing which opens the possibility of implementing scenarios for predicting and managing heredity in the generalized language of genetic information not only in biological systems but also in systems of anthropogenic origin [7], [8]. General system principles of hierarchical structural organization of complex systems, interdisciplinary structural isomorphisms and homologies indicate that the class of GOS includes not only systems of natural origin (biological, chemical, crystallographic) but also technical, musical, linguistic, mathematical and other anthropogenic systems. Evolutionary alive and anthropogenic systems regardless of their level of complexity have two alternative properties: heredity and variability. Heredity is manifested in the fact that an arbitrary object, population or species as a whole, retain in a number of generations their inherent genetic characteristic and properties. The property of variability predetermines a variety of options for the implementation and ways of development of hereditary structures by adapting them to changing environmental conditions.

The theoretical basis of GOS are interconnected system principles of structural organization and information heredity which predetermine the diversity and development of systems of different physical nature. The principles of structure formation and the genetically acceptable diversity of GOS objects are determined through the genetic information of the elemental basis, which is ordered by genetic classifications (GC) of the natural type. Periodic genetic classifications in the concept of GOS perform the function of generative systems. The category of generating systems includes the periodic classification of chemical elements, the classification of 230 groups of crystals symmetry, the classification of geographical zoning, the periodic classification of natural numbers and others. Genetic organization and evolution are determined by a system of interrelated principles and laws, the form of presentation of which is the generative system – the Periodic Genetic Classification (Table 1). The absence of certain invariant principles in GOS of a certain physical nature should be considered as a temporary phenomenon which is explained by the unevenness of structural and systemic studies or the lack of genetic classifications (generative systems) of the corresponding elemental basis. The biological science is particularly at the level where the theoretical basis operates on almost all system-wide principles of genetic evolution except its own generative periodic table the search for which is continued today [9].

As for the electromechanics principles are shown in Table 1 were absent in the classical theory of electric machines, and became known only after the discovery and analysis of the invariant properties of GC (Figure 1). Analysis of intersystem isomorphisms, after some time, revealed the principles of invariance of GOS also in the system of natural numbers [10].

Table 1: System-wide Principles of Invariance in Genetically Organized Systems of Different Physical Nature (Fragment)

Invariants of GOS	Biology	Music	Linguistics	Chemistry	Electro-mechanics	Natural numbers
Own elements base	+	+	+	+	+	+
Generative periodic system of elements	?	+	?	+	+	+
Universal genetic code	+	+	?	+	+	+
The principle of symmetry saving	+	+	+	+	+	+
The principle of topological invariance	+	+	+	+	+	+
The principle of genetic information saving	+	+	+	+	+	+
Structures genetic memory	+	+	+	+	+	+
The principle of parity	+	+	+	+	+	+
Homology	+	?	+	+	+	+
Isotopy	+	+	+	+	+	+
Isomerism	+	+	+	+	+	+

+ discovered or have been investigated;
? currently missing or need to be clarificated.

Electromagnetic field is a fundamental physical property of material and natural phenomena. The magnetic field was formed in the early stages of the evolution of the universe and is primary comparatively with chemical and biological evolution.

The information which accompanies magnetic fields has exceptional importance for studying the evolution of different physical nature systems. The discovery of the Genetic Classification of primary sources of the electromagnetic field have established that EMCE as systems of both natural and anthropogenic origin belong to the category of genetically organized systems [6]. The elemental basis of the GC is represented by the primary sources of the electromagnetic field (electromagnetic chromosomes) which perform the role of the original elementary structures in the tasks of combinatorial and topological synthesis of structures in arbitrary level of complexity. The position and invariant properties of an arbitrary primary source of field in the structure of the GC are determined by its genetic information which is represented by a universal genetic code. Electromagnetic, topological and geometric properties of the primary structures of the GC are determined by the system-wide principles of electric charge, symmetry and topology stability.

Elemental and information basis of the first large period of GC is represented by 36 oriented field sources (electromagnetic chromosomes) of the basic level, the location and genetic information of which unambiguously determine their belonging to the corresponding geometric class of active surfaces (Genus taxonomic category), their connectivity and orientation (Homological series), genetic code (Species category) and the principle of parity (Twin-species category). There is a deterministic connection between the elemental basis of GC and EM-objects-descendants which is determined by the principles of genetic structure and information stability. Therefore, an arbitrary EM-object is a carrier of genetic information which recognizes its location in the subject area of the generative system as well as determines the taxonomic status and the corresponding genetic programs of structure and its relationship with other elements of the system.

The structure of the GC is a form of presenting the fundamental principles the structure and information stability as well as the integral periodic law.

According to the results of genetic analysis and numerous evolutionary experiments it is established that the structural diversity and properties of electromagnetic and electromechanical objects were created by many generations of specialists are endowed with highly ordered system connections recognized by universal genetic codes of elemental GC basis.

The set of these provisions is a system basis for setting and solving fundamentally new problems including problems of determining and decoding of genetic programs of arbitrary functional classes of electromechanical systems (EM-systems), implementation of structural prediction and genetic synthesis technologies, development of genetic data banks and knowledge bases.

3. Spatial Geometry and Topology of Electromagnetic Structure

Spatial geometry and topology in the evolution of structures are fundamental problems in the theory of genetically organized systems of arbitrary physical nature. “The problem of the inheritance of forms is one of the most important problems of the human mind. Any form, any state of a certain process is a product of the union or intersection of elementary forms (atoms) that have similarity, indivisibility and immutability... Any evolution comes down to changes in the relative position, organization of these elementary forms” [11]. “…In modern physics it is assumption that the set of elementary forms and functions of matter is limited, and new forms and functions appear only as a result of combinations and transformations of basic forms and functions” [12]. These provisions are necessary conditions for the development of GOS.

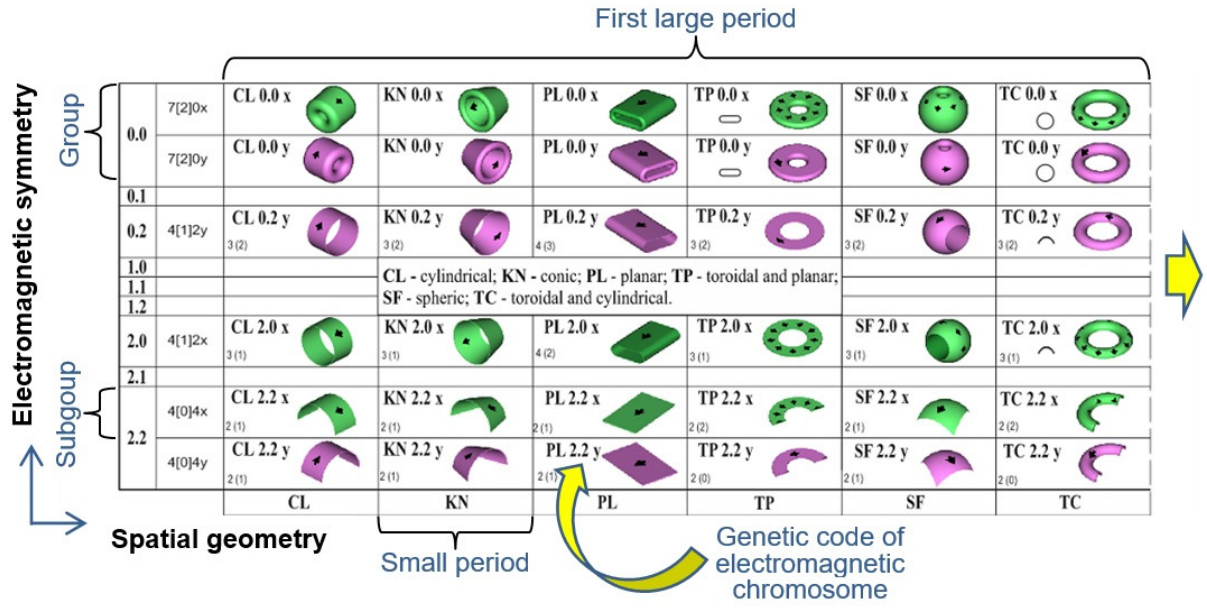


Figure 1: Generative System − Genetic Classification of Primary Sources of Electromagnetic Field (the First Large Period)

Spatial geometry, electromagnetic symmetry and topology of primary sources of electromagnetic field (electromagnetic chromosomes) define the concept of genetic information and play the role of invariants in the structure of arbitrary Species of EMCE. In the structure of the universal genetic code, spatial geometry is encoded by the first component of genetic information which determines the Genus affiliation of objects-descendants. The structural equivalent of the spatial geometry of the primary field source is the active surface of the corresponding EM-object. The geometry of the active zone, electromagnetic symmetry and topological properties of the electromagnetic structure determine the principles of structure formation of EM-objects at all levels of their structural organization, starting from the level of electromagnetic chromosomes and ending with complex systems created in the process of evolution (Figure 2).

Changing of the spatial geometry and parameters of the secondary environment is a necessary condition for the functioning of the vast majority of EMCEs designed for direct implementation of technologies (magnetic and electrodynamic separators, electromechanical disintegrators, magnetohydrodynamic converters, etc.). So their analysis is a separate task. In this study structures with only variable spatial orientation and geometry of the primary active part (stator, armature, inductor, poles, etc.) are analyzed.

4. Macrogenetic Programs

The limits of arbitrary functional classes of EMCE existence and the quantitative composition of genetically acceptable Species are determined by their macrogenetic programs [13]. Macrogenetic programs are carriers of systematized information on the Species diversity of structures, including structural representatives of known Species and Species that are not involved yet in the technical evolution of the functional class. The problem of determining the genetic program will be considered on the example of a class of EM-objects the which functioning is carried out with a variable spatial orientation (OZ-rotation) of the active parts (Figure 3).

The integrated search function F_{OZ} is determined by a set of the following partial requirements:

- the presence of a solid secondary part (C_2);

- the possibility of OZ-rotation ($\beta = 0 \div \pi$) of the primary part relatively to the *OZ*-normal;

- invariability of the air gap ($\delta \approx const$);

- ensuring the maximum area of overlap of the active zone ($S_a = max$).

Then the vector of the integral function F_{OZ} in the multidimensional search space R^n takes the following form:

$$F_{OZ} = \{C_2, (\beta = 0 \div \pi), (\delta \approx const), (S_a = max)\} \subset R^n. \quad (1)$$

The search space is limited by the elemental basis of the first large period of GC (excluding sources-isotopes). This set of requirements is met by electromagnetic chromosomes of three Genuses: Spherical (SF), Flat (PL) and Toroidal Flat (TP) primary sources of the electromagnetic field. The results of genetic analysis show that the electromagnetic chromosomes of the Spherical and the Flat Genuses are endowed with the greatest tendency to rotational symmetry of the *OZ*-type

$$Q_{SF} = (SF0.0y;\ SF0.0x;\ SF0.2y;\ SF2.0x;\ SF2.2y;\ SF2.2x); \quad (2)$$

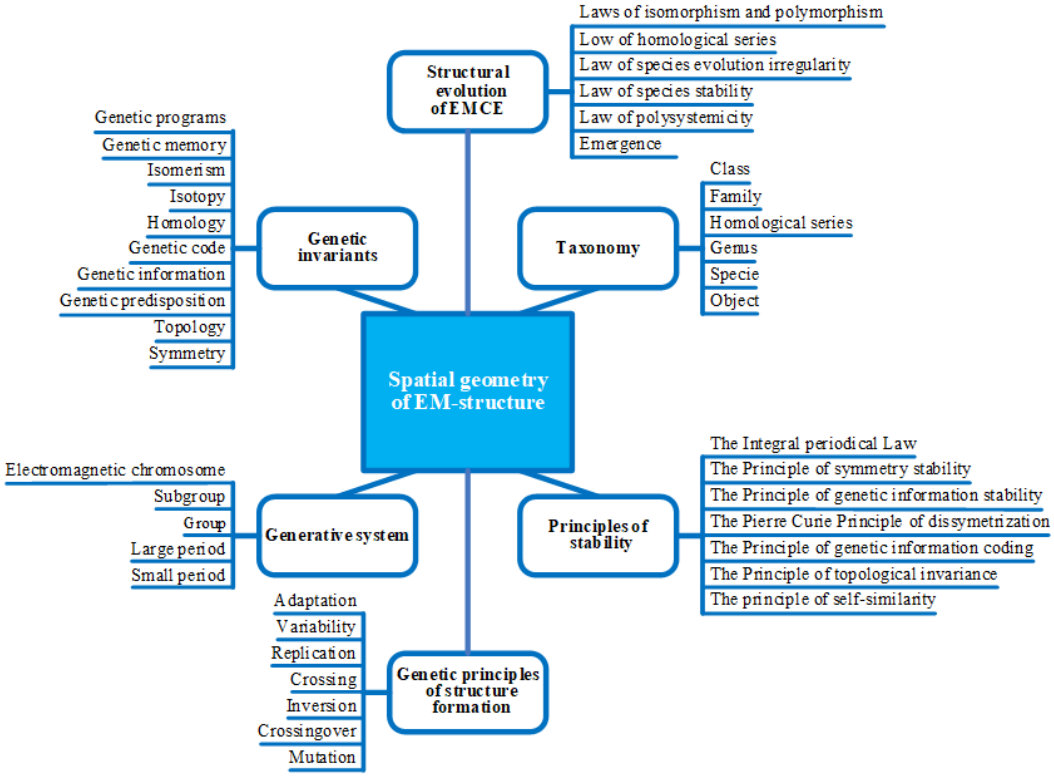


Figure 2: Ontology of the Category "Spatial Geometry of Electromagnetic Structure" in the Theory of Genetic Evolution of Electromechanical Converters of Energy

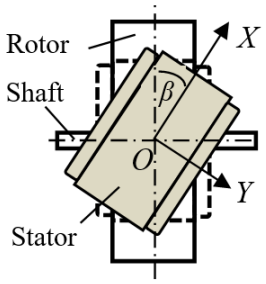


Figure 3: The Model of Electric Machine with OZ-rotary Stator

$$Q_{PL} = (PL0.0y;\ PL0.0x;\ PL0.2y;\ PL2.0x;\ PL2.2y;\ PL2.2x), \quad (3)$$

as well as two chromosomes of the Toroidal Flat Genus

$$Q_{TP} = (TP2.2y;\ TP2.2x). \quad (4)$$

Spatial OZ-rotation is also allowed by hybrid electromagnetic chromosomes of the intraspecific level:

$$Q_H = \{(PL2.2x,y)_1 \times (CL)_2;\ (PL2.2x,y)_1 \times (TP)_2;\ (PL2.2x,y)_1 \times (SF)_2;\ (TP2.2x,y)_1 \times (PL)_2\}. \quad (5)$$

The finite set of electromagnetic chromosomes (2–5) is generalized by the concept of macrogenetic program of structure formation of a functional class of objects with *OZ*-rotating active parts. The macrogenetic program contains information about the genetically acceptable elemental basis, which determines the Species diversity of both known and potentially possible EM-objects with a rotary inductor of the *OZ*-type.

Analysis of the spatial geometry and topology of paired electromagnetic chromosomes of the first large GC period shows that asymmetric field sources of group 2.2 with x-y topological invariance are endowed with the maximum genetic predisposition to change the spatial orientation. Such topological properties (provided that the value of the air gap $\delta \approx const$ is unchangeable) occur only in two primary sources of the electromagnetic field with genetic codes:

$$PL\ 2.2y = PL\ 2.2x;\ SF\ 2.2y = SF\ 2.2x, \quad (6)$$

as well as at the level of their electromechanical pairs:

$$(PL\ 2.2y)_1 \times (PL)_2 = (PL\ 2.2x)_1 \times (PL)_2; \quad (7)$$

$$(SF\ 2.2y)_1 \times (SF)_2 = (SF\ 2.2x)_1 \times (SF)_2. \quad (8)$$

The spatial geometry of the active zone of the descendant structures which are synthesized on the basis of chromosomes (6) is not "sensitive" to changes in the orientation of the active surfaces relative to the *OZ*-normal. But the change of the orientation of the active parts significantly affects the electromagnetic, electromechanical and functional properties of EM-structures-descendants. Genetic principles of mutation and structural replication determine the principles of structural synthesis of multi-element electric machines and electromechanical devices of modular design [14], [15], [16].

The relationship between genetic codes and the tendency to relative rotation at the level of paired electromagnetic chromosomes is evidence that electromechanical structures (EM-structures) with variable active zone geometry will occur at all other levels of structural complexity of a genetically organized EM-system: chromosomal (combinatorial space of paired electromagnetic chromosomes), elemental (active sections of windings, poles, moving active elements), object (inductors, armatures, rotors) and systemic (electromechanical systems with spatially distributed and active parts, etc.) levels.

5. Principles of Structure Formation of Objects with Adaptive Geometry of the Active Surface

In the theory of genetic synthesis electromagnetic structures with a violation of proportions, geometric relationships, or a violation of the electromagnetic symmetry of the active parts belong to the class of genetically mutated [6]. The system-wide principle of mutation is one of the most productive and least studied mechanisms of structure formation in GOS both of natural and of anthropogenic origin. In the methodology of genetic synthesis the corresponding group of homeomorphic transformations is matched to the mutation operator *M*. The genetic operator of the mutation changes the genetic information of an arbitrary structure, generating a sequence of homologous structures within a given topological space R^T:

$$M\,(S) \rightarrow (S_{M1}, S_{M2}, \ldots, S_{Mn}) \subset R^T, \overline{1, \infty}. \quad (9)$$

The topological equivalent of mutational transformations is volume or surface deformations of active elements or surfaces as well as changes in the spatial geometry and orientation of their relative position. The most well-known representatives of genetically mutated EM-objects are electric machines with beveled grooves, functional classes of electric motors with a rolling rotor, EM-objects with rotary stators and variable geometry of the active surface and so on. Topological model (9) indicates the unique role of mutations in understanding the mechanisms of intraspecific, interspecific and intergeneric structural relationships and transformations. Analysis of the mechanisms of mutagenesis in technical objects opens the possibility of structural prediction and synthesis of fundamentally new structural varieties.

In applied tasks of genetic synthesis, the operator *M* is matched with a group of continuous spatial deformations, compression, tension, shear, rotation, torsion and others. In procedures for the synthesis of multi-element EM-structures the mutation operator forms a stable combination with the replication operator and isomeric compositions the structural equivalents of which are spatially distributed modular structures.

In the general case for the replication coefficient $k_r \geq 2$ the mutated structure S_M allows *N* variants of spatial isomeric compositions in R^3 the number of which will increase with increasing value of k_r.

$$M(S_M) \rightarrow k_r\ s \in K_i \subset R^3,\ i = \overline{1 \div N}, \quad (10)$$

where: S_M – electromagnetic mutant chromosome; k_r – replication rate; R^3 – three-dimensional space of existence of isomeric compositions K_i.

The basis of the structure formation of EM-structures classes which allow spatial elastic deformation of the active surface is the fundamental principle of topological invariance which determines the structure and properties of primary elements within an arbitrary subgroup of periodic GC. From the point of genetic information topologically equivalent chromosomes form the corresponding horizontal series of homologous EM-structures which are subject to the realization of elastic deformation of the active zone provided the electromagnetic structure and topology stability. In the structure of the first large period of the GC there are six such series

$$T = (H_{00y}, H_{00x}, H_{02y}, H_{20x}, H_{22y}, H_{22x}). \quad (11)$$

According to the criterion of maximum adaptation to changes in the spatial geometry of the active surface of EM-objects the highest level of genetic predisposition is endowed with electromagnetic chromosomes of group 2.2. Within the first large period of GC the genetic information of the sources of group 2.2 is represented by two horizontal homologous series

$$T_{22y} = (CL\ 2.2y;\ KN\ 2.2y;PL\ 2.2y;\ TP\ 2.2y;\ SF\ 2.2y;\ TC\ 2.2y); \quad (12)$$

$$T_{22x} = (CL\ 2.2x;\ KN\ 2.2x;PL\ 2.2x;\ TP\ 2.2x;\ SF\ 2.2x;TC\ 2.2x). \quad (13)$$

For electric machines with solid secondary parts which implement rotational or translational spatial motion the range of generating structures is limited by the chromosome set of the subgroup T_{22y}. It si known that the changing in the spatial geometry of homologous electromagnetic chromosomes is related to the homeomorphism relation:

$$H_{22y} = (\ldots \leftrightarrow CL \leftrightarrow KN \leftrightarrow PL \leftrightarrow TP \leftrightarrow SF \leftrightarrow \leftrightarrow TC \leftrightarrow \ldots)2.2y \subset R^3. \quad (14)$$

In synthesis algorithms the homeomorphism of chromosomes (14) is realized by procedures of intergeneric chromosomal mutations which leads in changing of the Generic geometry of the corresponding chromosome within a given topological space. The genetically admissible set of intergeneric chromosomal mutations is theoretically determined by the combinatorial space of chromosomes of group 2.2 and is practically limited by their possibility of technical realization of structures-descendants. Paired chromosomal mutations ($N_1 = 2$) will be of the greatest practical value within subgroup 2.2y whose genetic predisposition is sensitive both to changes in the spatial orientation of active parts and to changes in the geometry of the active surface. These chromosomal mutations have a deterministic relationship with the spatial geometry of the active surface and the type of spatial motion of objects-descendants (Table 2).

Systematized information which is presented in Table 2 can be considered as a system basis for the structure formation of objects of "elastic electromechanics" of subgroup 2.2y for the combinatorial space $N_1 = 2$.

The invariant component of the genetic code (*2.2y*) indicates a common topological space within homeomorphic transformations are allowed. Changing the geometry of an arbitrary chromosome of a series (14) is carried out using a genetic mutation operator (9). Homeomorphism of primary sources of the electromagnetic field indicates the practical possibility of "elastic" formation at the level of electromagnetic twin structures which elemental basis is determined in the periodic structure of GC by isotope sources [17].

Table 2: Relationship between Intergeneric Chromosomal Mutations, Active Surface Geometry and the Type of Spatial Motion of EM-objects-descendant (Subgroup T_{22y})

Intergeneric chromosomal mutations	Type of spatial deformation	Structural formula of mutated chromosome	Type of spatial motion of objects-descendant
$PL \leftrightarrow CL$	M_{OY}	$2.2y(PL \leftrightarrow CL)_1$	Progressive (V_{OX}) ↔ rotational (ω_{OY})
$PL \leftrightarrow TP$	M_{OZ}	$2.2y(PL \leftrightarrow TP)_1$	Progressive (V_{OX}) ↔ rotational (ω_{OZ})
$PL \leftrightarrow SF$	$(M_{OY} \leftrightarrow M_{OX})_R$	$2.2y(PL \leftrightarrow SF)_1$	Progressive (V_{OX}) ↔ rotational (ω_{OY})
$CL \leftrightarrow SF$	M_{OX}	$2.2y(CL \leftrightarrow SF)_1$	Rotational (ω_{OY})
$PL \leftrightarrow KN$	$M_{OZ} \leftrightarrow M_{OY}$	$2.2y(PL \leftrightarrow KN)_1$	Progressive (V_{OX}) ↔ rotational (ω_{OY})
$TP \leftrightarrow KN$	M_{OY}	$2.2y(TP \leftrightarrow KN)_1$	Rotational (ω_{OZ}) ↔ (ω_{OY})
$CL \leftrightarrow KN$	M_{OZ}	$2.2y(CL \leftrightarrow KN)_1$	Rotational (ω_{OY})
$KN \leftrightarrow SF$	$(M_{OX} \leftrightarrow M_{OY})_R$	$2.2y(KN \leftrightarrow SF)_1$	Rotational (ω_{OY})
$TP \leftrightarrow SF$	$(M_{OX} \leftrightarrow M_{OY})_R$	$2.2y(TP \leftrightarrow SF)_1$	Rotational (ω_{OZ}) ↔ (ω_{OY})
$CL \leftrightarrow TP$	$M_{OY} \leftrightarrow M_{OZ}$	$2.2y(CL \leftrightarrow TP)_1$	Rotational (ω_{OY}) ↔ (ω_{OZ})
$CL \leftrightarrow TC$	$[M_{OY} \leftrightarrow (M_{OX})_{R1} \leftrightarrow (M_{OZ})_{R2}]$, where $(R_2 > R_1)$	$2.2y(CL \leftrightarrow TC)_1$	Rotational (ω_{OY}) ↔ (ω_{OZ})
$PL \leftrightarrow TC$	$[(M_{OX})_{R1} \leftrightarrow (M_{OZ})_{R2}]$, where $(R_2 > R_1)$	$2.2y(PL \leftrightarrow TC)_1$	Progressive (V_{OX}) ↔ rotational (ω_{OZ})
$KN \leftrightarrow TC$	$[(M_{OZ})_{R2} \leftrightarrow (M_{OX})_{R1}]$, where $(R_2 > R_1)$	$2.2y(KN \leftrightarrow TC)_1$	Rotational (ω_{OY}) ↔ (ω_{OZ})
$SF \leftrightarrow TC$	$[(M_{OX})_{R1} \leftrightarrow (M_{OZ})_{R2}]$, where $(R_2 > R_1)$	$2.2y(SF \leftrightarrow TC)_1$	Rotational (ω_{OY}) ↔ (ω_{OZ})

This pattern of interspecific topological formation is also characteristic of some biological species which was first noticed by the Scottish biologist and mathematician D'Arcy Thompson in 1917 [18]. The use of a group of topologically equivalent transformations (continuous deformations) allows the transition from the spatial geometry of the structure of one Genus to the spatial forms of other Homologous Genus. Therefore, the procedures for the synthesis of homologous EM-structures according to model (14) acquire the status of intergeneric mutations which are directly related to the change of the first component of the genetic information of the genetic code. This once again confirms the conclusion about the structural evolution of electromechanical objects occurs not only on the principle of heredity (i.e. within vertical Homologous Series) but also largely through the horizontal exchange of information within horizontal Homologous Series.

6. Genetic Synthesis

The homeomorphism of the primary structures of subgroups opens the possibility of the synthesis of EM-objects which are endowed with the property of changing their own spatial geometry within a genetically determined topological space. For example genetic information for a structure whose active surface can change geometry within three topologically equivalent geometric shapes ($N_G = 3$) for subgroup H_{22} can be represented as the result of a local intergeneric mutation

$$S_M = (H_{CL} \leftrightarrow H_{PL} \leftrightarrow H_{TP})2.2y \subset H_{22}. \quad (15)$$

Formula (15) corresponds to the model of a mutated EM-structure the geometry of the active surface of which can alternately exist in one of three generic forms flat, cylindrical or toroidal flat (Figure 4).

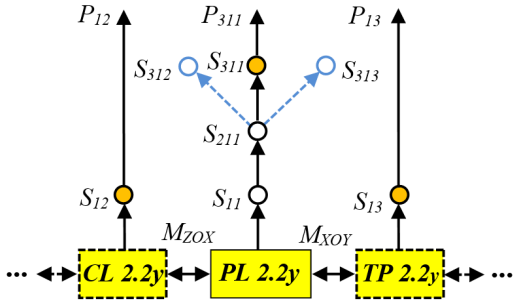


Figure 4: Genetic Model of Intergeneric Mutated EM-structure ($N_G = 3$) with Variable Geometry of Active Surface Synthesis

The results of the microgenetic program decoding confirm the dominant role of the mutation operator (M_i) in the procedures of formation of EM-objects with variable geometry of the active surface (Table 3).

The results of genetic analysis of the model allow to obtain the structural formula of the integrated mutated chromosome S_{M3}

$$S_{M3} \rightarrow (PL2.2_y)_1 \rightarrow \begin{Bmatrix} M_{ZOX} \rightarrow (CL2.2_y)_1 \times (CL2.2_y)_2 \\ M_{OZ} \rightarrow [(PL2.2y): M_{XOY}]_2 \\ M_{XOY} \rightarrow (TP2.2_y)_1 \times (TP2.2_y)_2 \end{Bmatrix} \quad (16)$$

Technical implementation of the synthesized on the basis of the S_M chromosome structure is possible if two opposite requirements are provided: providing of elastic deformation and maintaining the required mechanical stiffness of the active surface of the inductor. Modern polymers and composites allow to implement these requirements. One of the possible technical solutions involves the implementation of a hybrid design of the inductor, in which U-shaped magnetic conductors with concentrated windings, mechanically combined into a monolithic structure using a polymer-reinforced filler with a high coefficient of elasticity (Figure 5).

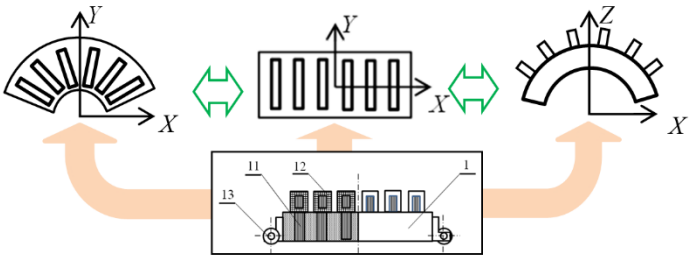


Figure 5: Multifunctional Inductor Module with Spatially Adaptive Active Surface (Chromosomal Mutation $CL2.2y \leftrightarrow PL2.2y \leftrightarrow TP2.2y$) Was Synthesized on the Basis of the Electromagnetic Chromosome S_{M3}

Genetically synthesized inductor is endowed with a unique ability to function with the ability to sequentially change the geometry of the core (cylindrical *CL* ↔ flat *PL* ↔ toroidal flat *TP*), which opens the possibility of implementing both translational and rotational motion of the moving part. Electromechanical objects with such properties can be generalized to the concept of objects of "elastic" electromechanics.

Table 3: The Results of Decoding of the Electromagnetic Chromosomes Local Intergeneric Mutation Microgenetic Program ($N_G = 3$)

Chromosomal code	Structural formula of synthesized chromosome	Chromosomal status	Weighting factor
PL 2.2y	*PL 2.2y*	Parental	–
	$[(PL\ 2.2y):M_{ZOX}]_1 \rightarrow (CL\ 2.2y)_1$	Parental mutated	–
	$[(PL\ 2.2y):M_{XOY}]_1 \rightarrow (TP\ 2.2y)_1$	Parental mutated	–
S_{11}	$(PL2.2y)_1 \times (PL2.2y)_2$	Electromagnetic double, informational	–
S_{12}	$(CL2.2y)_1 \times (CL2.2y)_2$	Electromagnetic double, generative	0.3
S_{13}	$(TP2.2y)_1 \times (TP2.2y)_2$	Electromagnetic double, generative	0.6
S_{211}	$[(PL2.2y):M_{OZ}]_1 \times (PL2.2y)_2$	Mutated (*OZ*-rotative), informational	–
S_{311}	$[(PL2.2y):M_{OZ}]_1 \times [(PL2.2y):M_{XOY}]_2$	Mutated $(M_{XOY})_2$, generative	0.8
S_Σ	$(S_{12} \leftrightarrow S_{13} \leftrightarrow S_{311})$	Integral, generative	1.0

7. Practical Implementation

One of the urgent tasks is to create modern technological complexes designed for technological preparation and processing of rolled steel and large diameter steel pipes. The problem of synthesis will be considered on the example of creating a traction spatially adaptive EM-system of a robotic technological complex for processing large diameter steel pipes. The initial information for the synthesis of a genetic model is a set of requirements for the desired structure of the EM-system and the results of the analysis of its macrogenetic program.

The integral function of F_S synthesis must satisfy the following set of partial requirements:

1) To provide the possibility of functioning with steel products of both cylindrical and flat spatial shape (G_{2CL}, G_{2PL}) which simultaneously perform the function of the secondary part of the EM system;
2) To implement progressive, rotational and complex spatial motion of the technological object ($\omega \times V$);
3) To minimize the number of functional modules of the EM system ($M_I \rightarrow min$);
4) To provide ability of quick changing of the spatial layout and number of active inductor modules in the structure of the manipulator ($N_I \rightarrow var$).

Taking into account these partial requirements, the vector of the integral synthesis function in the search space R^n takes the form:

$$F_S = [(G_{2CL}, G_{2PL}); (\omega \times V); (M_I \rightarrow min); (N_I \rightarrow var)] \subset R^n. \quad (17)$$

Requirement 3 is provided by changing the spatial orientation of the elementary inductor module relative to the surface of the process object (*OZ*-rotation) and changing the spatial geometry of the active surface of the inductor module under the action of magnetic gravity whith condition to provide constantally of the air ($\delta_{OZ} \rightarrow const$).

The multilevel genetic model of the divergent type is corresponds to given search function (17) (Figure 6).

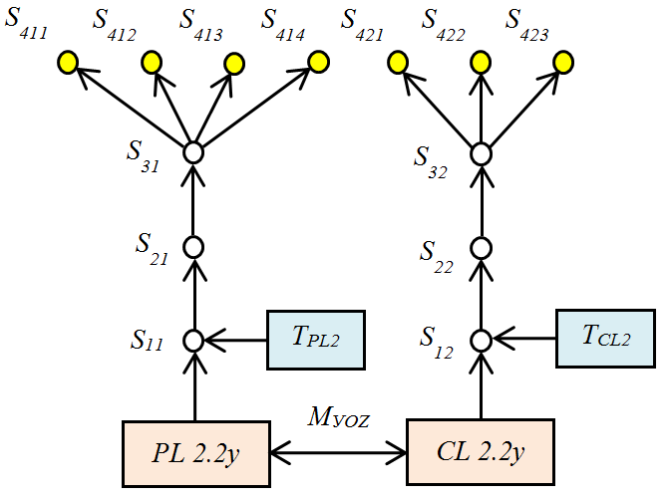


Figure 6: Genetic Model of Spatially Distributed Modular EM-system with Adaptive Geometry of Active Zone

The structure of the model contains five levels of genetic complexity which are represented by: the parent electromagnetic chromosome *PL 2.2y* of the dominant type of EM-structures (zero level); parallel sequences of genetically modified electromagnetic chromosomes of information type (S_{11}–S_{32}) which determine the trajectory of synthesis of the desired EM-structure (levels 2-4); isomeric spatial compositions of generating electromagnetic chromosomes (S_{411}–S_{414}; S_{421}–S_{423}) which satisfy the given function of synthesis and model the spatially distributed structure of the modular EM-system (level 5).

A characteristic feature of the genetic model (Figure 6) is determined by the homeomorphism of the parent chromosomes *PL 2.2y* and *CL 2.2y* which can coexist in two geometric forms of the active surface through the operator of the intergeneric mutation (M_{YOZ})

$$M_{YOZ}\,[2.2y\,(PL \leftrightarrow CL)_{ZOX}]_1 \subset T_{22Y}. \quad (18)$$

Another type of mutation is represented by chromosomes S_{21} and S_{22} which simulate their ability to implement the relative *OZ*-rotation of the active parts. These chromosomal mutations determine the procedure of synthesis of the EM-structure which is able to alternately implement the function of a flat inductor of progressive motion or an arc cylindrical inductor of rotational motion which depends on the type of secondary chromosome which simulates the spatial geometry of the object. Variants of the spatial geometry of the active surface of the technological object (requirement 1) in the model are represented by secondary transit chromosomes T_{PL2} and T_{CL2}. The results of decoding the microgenetic program are given in Table 4.

The results of the analysis of the microgenetic program (Table 4) confirm the dominant role of mutation (M_i) and replication (R_i) operators in the formation procedures of isomeric compositions which determine the genetic nature of the structure of modular EM-objects with variable structure and active zone geometry. The structural formulas of isomeric chromosomal compositions S_{411}–S_{414} (for functioning with flat objects) and S_{421}–S_{423} (for functioning with cylindrical objects) which have the status of generative satisfy the given F_S function. Each of these isomeric compositions is associated with a specific structural equivalent of the generic geometry, spatial location and orientation of the inductor modules.

Genetic synthesis procedures reproduce the deterministic relationship between genetic synthesis operators and isomeric compositions of electromagnetic chromosomes on the one hand and the pattern of inductor modules and the type of spatial motion on the other. Structural isomerism is a systemic property of genetically organized systems of arbitrary physical nature with a multi-element $(N \geq 2)$ structure, to form N_i variants of spatial compositions. The phenomenon of isomerism of primary sources of electromagnetic fields was first discovered and described in the analysis of invariant properties of the system model − Genetic Classification of primary sources of electromagnetic fields [6]. The synthesized isomeric compositions allow to determine the spatial layout schemes of inductor modules and the corresponding types of spatial motion of the technological object depending on its type and spatial geometry (Table 4).

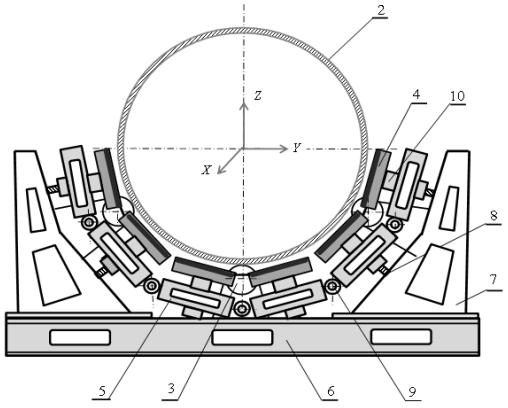
Table 4: The Results of a Modular EM-system with Variable Spatial Structure and Active Zone Geometry Structure Formation Microgenetic Program Decoding

Chromo-somal number	Structural formula of synthesized chromosome	Chromosomal status
PL2.2y	*PL 2.2y*	Parental
	$[(PL\ 2.2y){:}M_{ZOY}]_1 \rightarrow (CL\ 2.2y)_1$	Parental, intergeneric mutant
S_{11}	$(PL2.2y)_1 \times (T_{PL})_2$	Electromagnetic double, informational
S_{21}	$[(PL2.2y){:}M]_1 \times (T_{PL})_2$	Mutated, informational
S_{31}	$[(PL2.2y){:}M{:}R]_1 \times (T_{PL})_2$	Replicated, informative
S_{411}	$[(PL2.2y){:}M{:}R_{XOY}{:}V_{OX}]_1 \times (T_{PL})_2$	Isomer (*OX*-oriented), generative
S_{412}	$[(PL2.2y){:}M{:}R_{XOY}{:}V_{OY}]_1 \times (T_{PL})_2$	Isomer (*OY*-oriented), generative
S_{412}	$[(PL2.2y){:}M{:}R_{XOY}{:}\omega_{OZ}]_1 \times (T_{PL})_2$	Isomer (*OZ*-rotative), generative
S_{413}	$[(PL2.2y){:}M{:}R_{XOY}{:}V_{OX}:V_{OY}:\omega_{OZ}]_1 \times (T_{PL})_2$	Isomer $(V, \omega)_{XOY}$, generative
S_{12}	$(CL2.2y)_1 \times (T_{CL})_2$	Electromagnetic double, informational
S_{22}	$[(CL2.2y){:}M]_1 \times (T_{CL})_2$	Mutated, informational
S_{32}	$[(CL2.2y){:}M{:}R]_1 \times (T_{CL})_2$	Replicated, informative
S_{421}	$[(PL2.2y){:}M{:}R_{ZOY}{:}V_{OX}]_1 \times (T_{CL})_2$	Isomer (*OX*-oriented), generative
S_{422}	$[(CL2.2y){:}M{:}R_{ZOY}{:}\omega_{OX}]_1 \times (T_{CL})_2$	Isomer (*OX*-rotative), generative
S_{423}	$[(PL2.2y){:}M{:}R_{ZOY}{:}V_{OX}) \leftrightarrow (CL2.2y){:}M{:}R_{ZOY}:\omega_{OX}]_1 \times (T_{CL})_2$	Isomer $(V, \omega)_{OX}$, generative

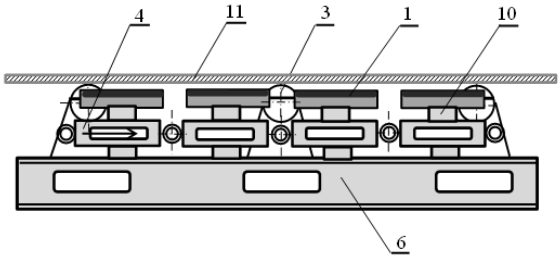
Technical implementation of a multifunctional EM-system with spatially distributed inductors carried out on a unified module with variable spatial geometry which is synthesized by a genetic formula

$$S_{M2} \rightarrow \left(PL2.2_y\right)_1 \rightarrow \begin{Bmatrix} M_{OZ} \rightarrow [(PL2.2y){:}\, M_{XOY}]_2 \\ M_{ZOX} \rightarrow \left(CL2.2_y\right)_1 \times \left(CLP2.2_y\right)_2 \end{Bmatrix} \quad (19)$$

Elastic deformation of the active surface of the modules is carried out under the action of electromagnetic gravity which allows to use one module to operate both steel pipes of different diameters and rolled steel sheets (Figure 7). The function of *OZ*-rotation of modules relative to the surface of the object of operation provides the implementation of various types of spatial motion of the technological object (progressive, rotating, helical, plane-parallel or complex programmable motion).



a)



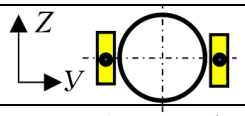
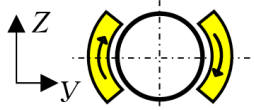
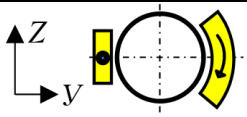
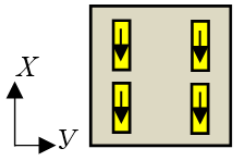
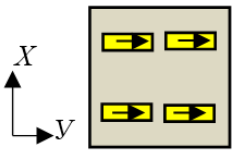
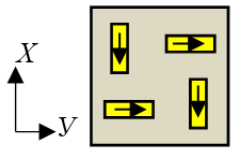
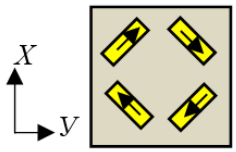
b)

Figure 7: Multifunctional Electromechanical System Based on Unified Adaptive Inductor Modules (S_{M2}) for Technological Manipulator of Steel Pipes (a) and Rolled Steel (b)

Systematized variants of spatial configurations of inductor modules (Table 5) perform the function of a genetic catalog to select the modes of operation of the spatially distributed EM-system of the technological manipulator. The results of the analysis are also the source of information for the development of algorithms for automated process control.

The speed of the object is regulated by the changing the spatial orientation of the inductor modules using frequency or amplitude-phase control. The modular structure of inductors and unification of units of the technological manipulator at the minimum expenses of time provides adaptation of spatial structure and geometry of EM-system under the given technological process. The spatial structure of the system also provides the ability to operate the number of active modules which allows to implement energy-saving algorithms for controlling the modes of operation of objects with different dimensions and weight.

Table 5: Relationship between Genetic Information of Isomeric Compositions, the Spatial Arrangement of Active Inductor Modules and the Type of Spatial Motion of the Technological Object

Type of spatial motion	Structural formula of isomeric compositions	Geometry and spatial arrangement of inductor modules*
The technological object is a steel pipe		
Progressive: $\pm (V_{OX})$;	$4(PL\ 2.2x)_{XOY}$	
Rotative: $\pm (\omega_{OX})$;	$4[(CL\ 2.2y):M_{OZ}(\pi/4):M_{ZOY}]_{ZOY}$	
Rotative-progressive (helical): $\pm (V_{OX} \times \omega_{OX})$;	$2(PL\ 2.2x)_{XOY} \times 2[(CL\ 2.2y):M_{OZ}(\pi/4)]_{ZOY}$	
The technological object is a steel sheet		
Progressive: $\pm (V_{OX})$;	$4(PL\ 2.2x)_{OX}$	
Progressive: $\pm (V_{OY})$;	$4[(PL\ 2.2y):M_{OZ}(\pi/4)]_{OY}$	
Flat-parallel: $\pm (V_{OX} \times V_{OY})$;	$2(PL\ 2.2x)_{OX} \times 2[(PL\ 2.2y): M_{OZ}(\pi/4)]_{OY}$	
Rotative: $\pm (\omega_{OZ})$;	$(PL\ 2.2y): M_{OZ}\ [(-\pi/4) \times (\pi/4) \times (3\pi/4) \times (5\pi/4)]_{XOY}$	

*For $N_I = (2\times2)$

8. Genetic Evolution

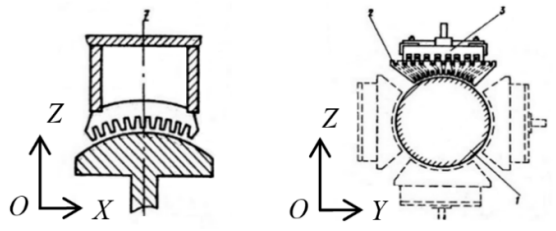
The results of genetic programs decoding were verified by carrying out of evolutionary experiments. The main task of the evolutionary experiment is to determine the deterministic correspondence between the genetic information of the elemental basis of genetic programs and the genetic codes of the real diversity of EM-objects-descendants that have historically emerged in the process of technical evolution [19].

The first structural representatives of objects with a change in the orientation of the active parts were induction electric machines with a rotating stator. The term "rotary stator" became known in the technical literature in the late eighteenth century after the publication of the French engineer P. Boucher [20]. But the concept of "rotary stator" turned out to be broader and ambiguous as its structural identification is associated with the specification of the spatial orientation of the axis of rotation relative to the axis of rotation or direction of movement of the moving part which in general case could be unconformitive. According to this feature electrical machines and electromechanical devices with the function of the relative orientation of their active parts can be divided into two subclasses:

- EMCE with the possibility of rotation of the active parts relative to the axis of rotation (*OY*-rotation);

- EMCE with the possibility of rotation of the active parts relative to normal (*OZ*-rotation).

In electromechanical objects the normal vector is perpendicular to the active surface of the converter. The rotation of the stator relative to the axis *OZ* allowed to implement one of the ways of non-contact speed control of the induction motor with the possibility of reversing. Such engines were patented in the late 40's of last century (Figure 8) [21], [22].



a) b)

Figure 8: The First Electric Machines with Rotary OZ-symmetry: a) Asynchronous Motor of the Spherical Genus (SF 2.2y); b) Induction Electric Machine of the Cylindrical Genus with a Hybrid Design of Rotary Stators (4PL 2.2y).

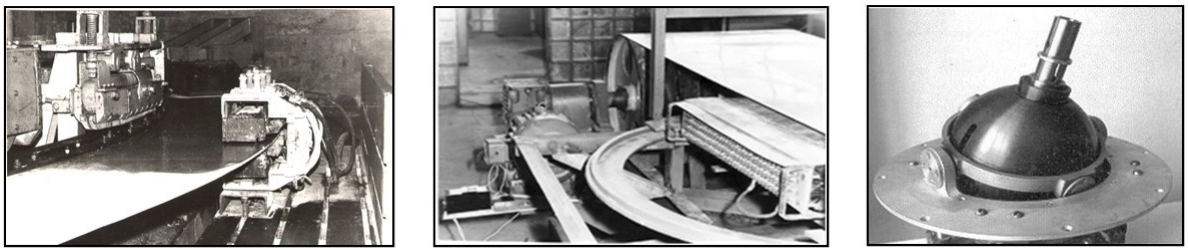
The historical stage of the genetic evolution of EMCE with variable spatial geometry of active parts is represented by a wide variety of electrical machines and electromechanical devices for different functional purposes. In 1966–1969 the first traction linear asynchronous motors (LAM) with a capacity of 5, 10 and 40 kW for driving experimental and full-scale cars of passenger overpass electric vehicles were designed, created and tested at the Special Design Bureau of Linear Electric Motors in Kyiv [23]. One of the projects involved the development of a motor with a could change the geometry of a flat active surface into a curved (arc) which significantly reduced the turning radii. In fact it was the first structural representative of traction linear motors in the design of which the principle of spatial deformation of the active surface of a moving inductor is implemented which is matched by an intergeneric mutation

$$M_2(S_M) \rightarrow 2[(PL2.2y) \leftrightarrow (CL2.2y)]_1 \subset H_{22}. \quad (20)$$

For dynamic tests of full-scale traction linear asynchronous motors a ring-type stand with an active part of 9.74 m diameter was put into operation at the Special Design Bureau of Linear Electric Motors (Kyiv). The design of the stand was implemented according to the formula of an intraspecific hybrid $(PL2.2x)_1 \times (TP0.2y)_2$. Dynamic tests of traction induction motors were carried out on the stand with the possibility of changing both the parameters of the secondary part and changing the geometry and power of their inductors (Figure 9, a). The development of space technology and autonomous robotic complexes have had significant affect on the intensity of research and development of spatially adaptive multi-coordinate electromechanical systems based on spherical electric machines (Figure 9, c) [24], [25], [26].

The current stage in the evolution of modular objects of spatially adaptive electromechanics is characterized by the use of new polymer and composite electromagnetic materials, printing technologies for magnetic circuits and windings as well as the capabilities of additive technologies of 3D- and 4D-printing. These technologies and the widespread use of magnetoelectric excitation systems have stimulated the development of technical solutions related to the implementation of the modular principle at all levels of the EM-object organization (Figure 10).

Thus the development of electromechanics is in accordance with the systemic laws of genetic evolution. The species diversity of EM-objects which determines the structural potential of a certain functional class is gradually expanding which is clearly illustrated by the dynamics of macroevolution of EM-objects with adaptive spatial geometry (Figure 11). The trajectory of macroevolutionary events reproduces the relationship of the elemental basis of Genetic Classification (Generative System) with real historical events of the appearance of structural representatives with the corresponding genetic codes which confirms the validity of the theory and practice of genetic evolution.



a) b) c)

Figure 9: Examples of EM-objects with Variable Spatial Orientation and Geometry of Active Parts: a) Universal Stand (Diameter 9.74 m) for Dynamic Tests with the Possibility of Changing the Geometry of Traction Linear Induction Motors; b) Electrodynamic Separator with Flat OZ-rotary Inductor (Specie PL 2.2y) [22]; c) Three-coordinate Synchronous Motor with a Rotary Stator for Orientation Systems (Representative of the Spherical Genus) [26].

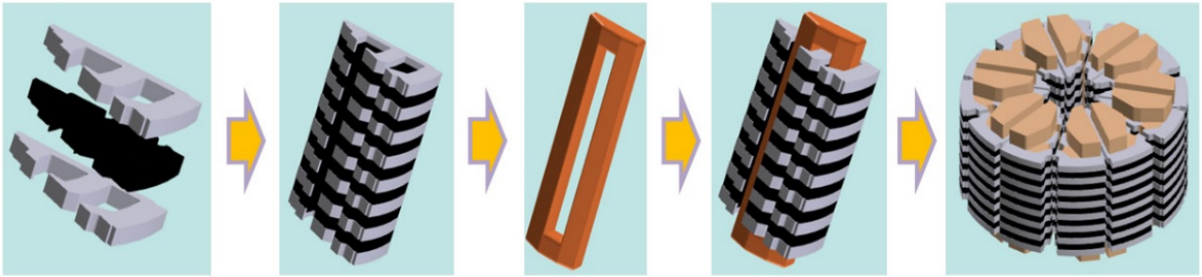


Figure 10: Multilevel Implementation of the Modular Principle in the Genetic Design of Linear Electric Motors with Transverse Magnetic Flux (Representative of the Specie CL 2.0x) by the Technology of KEEP'MOTION Company

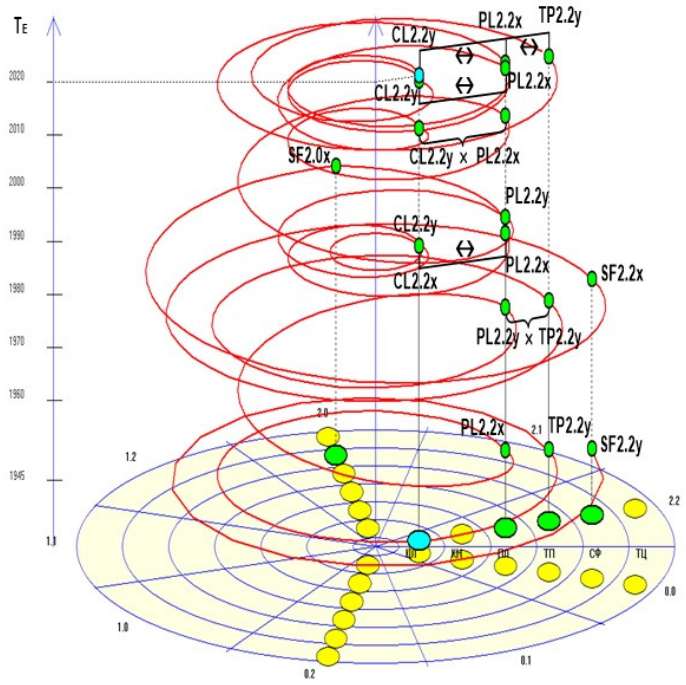


Figure 11: Macrogenetic Evolution of Species Diversity of Electric Machines with Adaptive Spatial Geometry of the Active Zone (T_E = 75 years)

The current stage of functional evolution of spatially adapted modular EM-systems is represented by a wide variety of EM-objects for different purposes (Figure 12). Modern EM-systems combine the functions of executive bodies with elements of spatial self-organization and artificial intelligence. Examples of such systems are robotic complexes with reconfigured structures [27], spatially distributed electromechanical systems of cableless passenger elevators of the MULTI series, implemented by MAGLEV technology [28], electromechanical systems of branched 3D networks-conveyors, grips and manipulators of anthropomorphic robotic systems, machining centers [5] and others. A promising area of practical use of EM-objects of elastic electromechanics are inductor systems of magnetic therapy equipment and implanted artificial life support systems in medicine [29].

The results of genetic analysis of the directions of functional evolution confirm the reliability of genetic programs and form the basis for the tasks of structural prediction and innovative synthesis of their new structural varieties. The technical evolution of spatially adaptive electromechanics in the nearest future will be largely determined by advances in materials science, genetic electromechanics, additive technologies and artificial intelligence.

9. The New Philosophy of Synthesis

The presence of the Generative Periodic System of primary sources of the electromagnetic field and the methodology of deciphering the structural information potential of genetic programs open the possibility of transition to the fundamentally new philosophy of creating complex technical systems with the ability to predict and discover new types and functional classes of EMCE. Such tasks are directly related to the theory of genetic Specie formation of EM-systems and the formulation of the corresponding prognostic-evolutionary experiment. The concept of Species is fundamental in the evolution of GOS as only through the category of Species it became possible to explain the mechanisms of micro- and macroevolution of EM-systems, determine genetic programs of structure formation and develop principles of genetic taxonomy of EMCE.

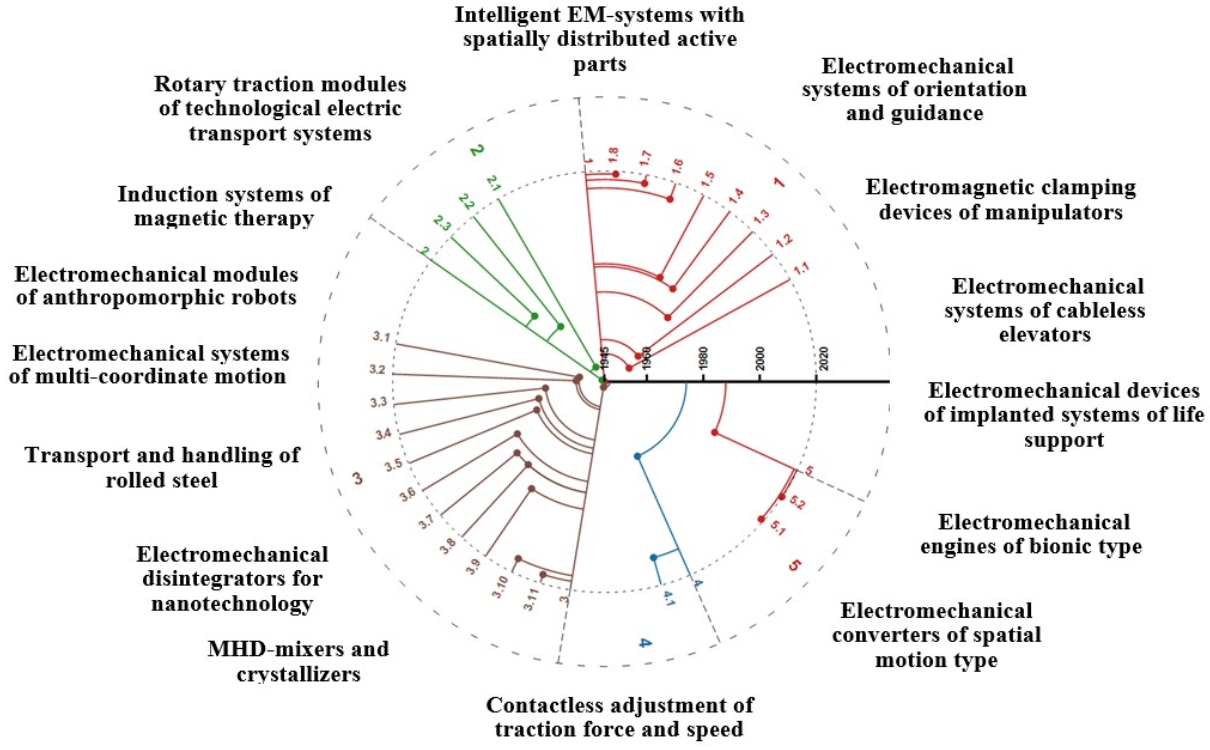


Figure 12: The Main Directions of Functional Evolution of EMCE with Variable Structure and Spatial Geometry of Active Parts

Until recently the discovery of new Species was considered the prerogative of biological science the systemic foundation of which is biological taxonomy. The discovery of the Biological Species is a landmark scientific event of international level. In the concept of GOS the concept of Species is a natural and necessary component in the hierarchy of interrelated levels of genetic complexity of electromagnetic systems: "Electromagnetic gene" → "Electromagnetic chromosome" → "Electromagnetic structure" → "Population of structures" → "Species" → "Genus" → "Functional class of EM-objects". But the establishment of the relationship and logical sequence of these categories became possible only after the discovery of the Periodic Table of primary sources of the electromagnetic field in 1991 (Figure 1) [6].

The reliability of the theory of genetic Specie formation is confirmed by evolutionary experiments the results of deciphering dozens of macrogenetic programs [30] and the creation of technology for genetic prediction of new types of EM-objects [31].

The genetic concept of Specie formation establishes a deterministic relationship between the information of the elemental basis of GC and the Species diversity of EM-objects, between the processes of macro- and microevolution of electromechanical objects and between the history of electromechanics and its future.

Opposite to the problem of biological Species which has been remainding controversial since Darwin due to the lack of their generating system the diversity of electromagnetic Species is strictly regulated by the periodic structure of genetic classification where information on both species involved in technical evolution and currently absent stored in the form of appropriate genetic codes. This sets a precedent for setting fundamentally new intellectual tasks – the prediction, discovery and introduction into technical evolution of new Species of EMCE.

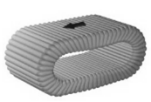
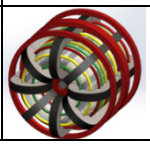
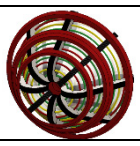
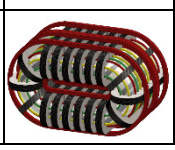
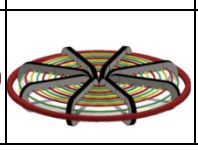
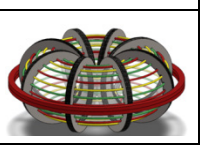
The task of new Species discovering is directly related to the synthesis and technical implementation of their first structural representatives. As an example, consider the problem of finding EM-objects with the maximum use of active volume (V_A) of an electric machine or electromechanical device. According to the results of the analysis of electromagnetic symmetry and topology of electromagnetic chromosomes of the first large period of GC (Figure 1) it is established that only electromagnetically symmetric chromosomes of group 0.0 are endowed with the maximum genetic predisposition to the criterion ($V_A \rightarrow max$). The geometry and topology of the active surfaces of the zero group is limited by the elemental basis of two topologically equivalent subgroups (y- and x-orientation) with Betty number β (S) = 2. At the level of multiphase windings such topology and symmetry can be provided only by magnetic systems with ring-distributed windings and the internal location of the active zone.

The reliability of the results of genetic prediction is confirmed by the results of group synthesis of representatives of a new structural class of electromechanical disintegrators (EMD) which designed for direct implementation of nanotechnologies (fine and ultrafine grinding, intensive mixing, preparation of homogeneous powder mixtures and compositions, enrichment technologies, etc.). Implementation of contradictory requirements (V_A = 100 %) while providing technical access to the internal working chamber during operation, was ensured by implementing the modular principle with variable spatial geometry of the external electromagnetic system of EMD (Table 6).

The development of technical solutions for EMD and their subsequent group patenting simultaneously proved the discovery and introduction into technical evolution of 6 new Species of EMCE which structural representatives were absent in the history of structural electromechanics. It should be noted that in the real evolution of technology the introduction into the evolution a new Specie occurs once every 30–40 years.

Actually, genetic electromechanics has accumulated experience in solving a number of unique problems which formulation were impossible a few years ago. It becomes obvious that the new understanding of synthesis is based on a rethinking and widespread use of systemic principles of structural and informational heredity. In each system it is possible to allocate as a building brick the elementary structures underlying technology of genetic synthesis. It is an atom in nanotechnology, a gene in biotechnology, a bit in information technology, a neuron in cognitive technology. Such system of interconnected initial structures is invariant to the time of evolution, the level of complexity of objects-descendants and their functional purpose. The transition to genetic technologies is a natural way to unity and harmony with nature, discovery and rational use of the hidden structural potential of genetic programs, creation of multifunctional integrated systems of natural-anthropogenic type, synthesized by genetic codes and models of living nature.

Table 6: Synthesized Structures of New Species of Electromechanical Disintegrators with Variable Geometry of Electromagnetic System and with 100% Using of the Active Volume for Direct Implementation of Nanotechnologies

Genus geometry	Cylindrical	Conical	Flat type	Toroidal flat	Spherical	Toroidal cylindrical
Genetic code	*CL 0.0x*	*KN 0.0x*	*PL 0.0x*	*TP 0.0x*	*SF 0.0x*	*TC 0.0x*
Electromagnetic chromosome						
3D-model of EMD structure						

10. Conclusions

The main results of the research can be summarized as follows:

- for the first time it is established that the principles of genetic structure formation of spatially distributed modular EM-systems with adaptive geometry of active zone are determined by electromagnetic symmetry and topological properties of generative sources of the electromagnetic field;

- for the first time macrogenetic programs of genetically admissible basic Species of EM-structures with variable orientation of active parts (18 Species) and objects of «elastic» electromechanics (12 Subspecies) which allow changing of spatial geometry of active surface are defined;
- it is established that in the algorithms of spatially adaptive systems synthesis the dominant role belongs to the operators of replication (modularity) and mutation ("elasticity") which determine the structure of modularity and "elasticity" of the synthesized structures;
- for the first time genetic programs are defined and algorithms of intergeneric mutation are developed, which determine the principles of structure formation of "elastic" electromechanics modular objects, which functioning is carried out with variable geometry of the active surface;
- the presence of a deterministic relationship between the elemental basis of the Periodic system of primary sources of the electromagnetic field (genetic program), isomeric chromosomal compositions, the spatial arrangement of active inductor modules and the type of spatial motion of the technological object is established;
- the reliability of genetic models, programs and synthesized structures is confirmed by the results of evolutionary experiments;
- practical implementation of research results is carried out on the example of project development of a competitive modular EM-system with adaptive spatial structure which is the part of a robotic technological complex designed for processing steel pipes and rolled products.

The results of research form the basis for the problems solving of innovative synthesis of new structural varieties of modular EM-systems with adaptive spatial structure of active elements and using of "elastic" electromechanics innovative technology.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A.N. Knaian, K.C. Cheung, M.B. Lobovsky, A.J. Oines, P. Schmidt-Neilsen, N.A. Gershenfeld, "The Milli-Motein: A self-folding chain of programmable matter with a one centimeter module pitch," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 1447–1453, 2012, doi:10.1109/IROS.2012.6385904.

[2] I.-C. Cheng, S. Wagner, Overview of Flexible Electronics Technology, Springer, Boston, MA: 1–28, 2009, doi:10.1007/978-0-387-74363-9_1.

[3] E. Tsihosh, Supersonic aircrafts. Reference manual. Translation from Polish, Mir, Moscow, 1983.

[4] V. Shikhirin, "Elastic machines and mechanisms of the future," The Summary of Technologies, **2**(6), 37–42, 2001.

[5] Y.N. Kuznetsov, G.J.A. Hamuyela, T.O. Hamuyela, Morphological synthesis tools and mechanisms: Monograph, Ltd "Gnosis," Kiev, 2012.

[6] V.F. Shynkarenko, Fundamentals of the Theory of Evolution of Electromechanical Systems, Naukova Dumka, Kyiv, 2002.

[7] V. Shynkarenko, V. Kotliarova, A.N. Al-Husban, "Genetical Modeling, and Experimental Analysis of the Functional Evolution of Electromechanical Energy Converters," Journal of the Technical University of Gabrovo, **49**, 15–20, 2015.

[8] M. Zagirnyak, V. Prus, V. Shynkarenko, "The assessment of the processes of aging of the electric machines with structural unit defects using the genetic approach," Przegląd Elektrotechniczny, **95**(1), 145–148, 2019, doi:10.15199/48.2019.01.37.

[9] S.G. Inge-Vechtomov, "The search for a periodic system… in evolution," Science First Hand, **2**(3), 20–25, 2004.

[10] V.F. Shynkarenko, "Isomorphisms of generative systems (on the example of electromagnetic and numeric)," Electromechanical and Energy Saving Systems, **1**(45), 46–55, 2019, doi:10.30929/2072-2052.2019.1.45.46-55.

[11] R. Thom, Structural stability and morphogenesis, Logos, Moscow, 2002.

[12] A. Lima-de-Faria, Evolution without Selection: Form and Function by Autoevolution, Mir, Moscow, 1991.

[13] V. Shynkarenko, Iu. Gaidaienko, A.N. Al-Husban, "Decoding and functional analysis of genetic programs of hybrid electromechanical structures," Modern Applied Science, **8**(2), 36–48, 2014, doi:10.5539/mas.v8n2p36.

[14] V. Shynkarenko, A. Makki, V. Kotliarova, A. Shymanska, "Modular Principle in the Structural organization and Evolution of Electromechanical Objects," in 2019 IEEE International Conference on Modern Electrical and Energy Systems (MEES), 162–165, 2019, doi:10.1109/MEES.2019.8896446.

[15] V. Shynkarenko, A. Makki, A. Shymanska, V. Kotliarova, "Genetic Synthesis of Electromechanical Objects of the Modular Type," in 2019 IEEE International Conference on Modern Electrical and Energy Systems (MEES), 166–169, 2019, doi:10.1109/MEES.2019.8896596.

[16] V. Shynkarenko, V. Kotliarova, P. Krasovskyi, N. Misan, "Principles of structural construction of spatially adaptive electromechanical systems with variable structure and geometry of the active surface," Bulletin of NTU "KhPI", Series: "Electric Machines and Electromechanical Energy Conversion," **3**(1357), 62–70, 2020, doi:10.20998/2409-9295.2020.3.11. (In Ukrainian)

[17] V. Shynkarenko, V. Kotliarova, "Twin-objects and double-objects in the structural evolution of electromechanical energy converters," in 2017 International Conference on Modern Electrical and Energy Systems (MEES), IEEE: 116–119, 2017, doi:10.1109/MEES.2017.8248865.

[18] N. Witkovski, The Sentimental History of Science. Translation from French, CoLibri, Moscow, 2007.

[19] V.F. Shinkarenko, I.A. Shvedchikova, V.V. Kotlyarova, "Evolutionary Experiments in Genetic Electromechanics," in Proc. 13th Anniversary International scientific Conference "Unitech'13," 289–294, 2013.

[20] K.I. Shenfer, Asynchronous machines, Gosizdat, Moskow–Leningrad, 1929.

[21] P.A. Fridkin, "A new type of arc stator for an electric drive with low and adjustable rotation speeds," Electricity, **1**–**2**, 22–23, 1945.

[22] O.N. Veselovsky, M.N. Godkin, Induction motors with open magnetic circuit, Informelectro, Moskow, 1974.

[23] V.F. Shynkarenko, "Developments of Special Design Bureau of Linear Electric Motors and their place in evolution of new types of electric transport," Collection of Scientific Works "Research on the History of Technology," **3**, 8–31, 2003.

[24] D. Martínez Muñoz, Spherical machines: a literature review, 2005.

[25] O. Saßnick, Application of the Spherical Induction Motor to Dynamically Stable Robots. Marshall Plan Scholarship Research Report, 2015.

[26] A.E. Antonov, Electric machines of magnetoelectric type. Fundamentals of the theory and synthesis, IED NASU, Kiev, 2011.

[27] I.M. Makarov, V.M. Lokhin, S.V. Manko, M.P. Romanov, M.V. Kadochnikov, "Knowledge processing technologies in control problems of autonomous mechatronic-modular reconfigurable robots," Information Technologies, **8**, 2–29, 2010.

[28] J. Condliffe, World's First Cable-Free Elevator Zooms Horizontally and Vertically Using Maglev Tech, MIT Technology Review, 2017.

[29] N.I. Kulikov, A.D. Kupriyanov, "Implantable auxiliary blood circulation system based on a controlled valve motor integrated with a diaphragm-type blood pump," Biotekhnosfera, **4**(16), 9–14, 2011.

[30] A.A. Avgustynovych, Theoretical bases of electric machines Species genetic systematics creation, PhD Thesis, IED NASU, 2008. (In Ukrainian)

[31] V.F. Shinkarenko, "Genetic Foresight as System Basis is in Strategy of Management Innovative Development Technical Systems," Proceedings of the Tavria State Agrotechnological University, **4**(11), 3–9, 2011.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 516-522 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

IoT Based Human Activity Recognition System Using Smart Sensors

Deepti Sehrawat*, Nasib Singh Gill

Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, 124001, India

ARTICLE INFO

Article history:
Received: 24 May, 2020
Accepted: 19 July, 2020
Online: 09 August, 2020

Keywords:
Types of IoT Sensors
HAR Systems
Sensor Applications
IoT and Sensors

ABSTRACT

Internet of Things provides a virtual view of real-life things by guiding challenges faced by persons in daily life. It is reforming our world with trillions of sensors and other IoT enabled devices by creating a smart environment. Effective use of IoT sensors in various smart IoT applications is very important. After analyzing different sensor applications, this paper presents various types of wearable sensors used for monitoring of human activities along with different locations optimal for their placement. This paper enlightens sensors suitable for any particular application. IoT has opened up a new avenue of research in the field of human activity recognition using wearable sensors. It provides an individual's valuable information regarding functional ability and lifestyle. Human activities are monitored automatically to provide personalized support to different individuals. Recently, various researchers presented different human activity recognition systems, a few are cumulated in this paper. Furthermore, a Human Activity Recognition (HAR) system is also proposed in this paper for a smart IoT environment that would be secure enough to use.

1. Introduction

The advent of Internet of Things (IoT) has opened a plethora of opportunities. In recent years, human activity recognition has received growing attention. This paper is an extension of work originally presented in the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) [1]. IoT not only connects animate/living things but inanimate/non-living things are also connected in a smart environment. This brings revolutionary changes in the world [2]. In any IoT based applications, sensors are very significant. Any physical/chemical change can be detected by these sensors which are then collected and processed to automate the application/devices to make it smarter in the form of smart application. IoT integrates several sensors of different types attached to the devices and nodes that can communicate without any human intervention [3].

Wearable sensors and the advances of wireless networks has opened the door to a world of novel interactive IoT applications. Different types of inexpensive and unobtrusive body-worn sensors like accelerometers, gyroscopes, and microphones have been often used in daily life activities [4]. Various researchers worked on detecting human activities by observing different parameters. It can be used for the monitoring of people of any age group like for infants, young people, disabled people, and elderly people in a solitary life. These monitoring systems are usually smaller and are wearable [5].

The mental or physical health of a user can be monitored remotely and then the results can be sent to their remote family. Most of the earlier researchers in this area made use of multiple body-worn sensors worn/ placed at different body parts like chest, thigh, waist, ankle, knee, wrist, ear, eyes, and so on [6]. The purpose of those research work was to categorize a definite subset of activities for particular applications. For instance, for upper body movements, wrist and arm sensors were used whereas for simple lifestyle monitoring a single sensor would be enough [7]. Most of the human activity recognition systems use accelerometers for the classification and assessment of day-to-day activities. Some used single accelerometers while some used multiple accelerometers attached to different sites on one's body. Wearing multiple accelerometers gives a good level of accuracy for certain activities but wearing multiple sensors all the time is not feasible. Wearable sensors are mostly used to detect physical movements but these can also be used for monitoring academic performance, analysing sleep class/patterns, stress level, and mental fitness conditions. Researchers in [8] gathered data from 66 individuals and analysed behavioural and physiological patterns. They also identified factors affecting performance in academics, sleep quality, stress level, and mental health. In the process of human activity recognition first step is to collect the data and prepare a

*Corresponding Author: Deepti Sehrawat, dips.scorpio@gmail.com

dataset which is then pre-processed and then the classification of data is carried out. This paper also presents various steps and techniques used in the process of recognition of human activities.

For the recognition of human activities from wearable, there are different classification techniques supervised and unsupervised (both) which are widely used by different researchers. Supervised techniques are considered more accurate when compared to unsupervised techniques [9].

IoT has numerous applications where different types of sensors are used in different types of applications. This paper studies different IoT applications and enlightens which smart application demands which type of sensor. Besides, various types of wearable sensors which are used for monitoring of human activities and application area where these are used are also presented in this paper.

This paper presents a detailed review of different locations where we can place wearable sensors and explains which one is ideal for a given group. It is also considered whether the sensor is comfortably wearable or not. For this, it should be small, lightweight, and able to provide the information content for which it is meant [10]. The positioning of wearable sensors along with the selection of relevant features for different activities pose challenges in this area. One of the limitations of wearable sensors in the prediction of optimal location on the body of the subject which can provide the most relevant features regarding activity classification [10].

Furthermore, various open areas are also explored in this paper so that a possible future research direction could be recognized and in the near future, we will come up with a good solution in this direction.

Section II of this paper presents various application areas of IoT where different types of sensors are used. Then, in section III, an analysis is made to identify the type of sensors that are used in different IoT applications. This helps various researchers in this area to identify which sensor is suited best for any specific type of application. Sensors are also used for HAR systems. Section IV presents several HAR systems and then categorizes various types of activities that are recognized in different HAR systems. Finally, a new HAR system is proposed in section V that helps the users to identify different activities. And soon, we come up with the detailed design of this HAR system that would be beneficial for the society in general.

2. Sensor Applications for a Smarter World

There are various issues and challenges in implementing and operating smart environment applications. Advanced smart applications need sustainability, reliability, and efficiency. A smart environment commonly has various kinds of Sensor nodes that control the current environment condition and energy usage. Sensors give their measurements and observations to the control system and then intelligent algorithms are implemented by the control system in the network. For surveillance and monitoring of smart environment applications, sensors are implemented in wide areas. Various examples are Highway Data Collection, Electric Toll Collection System, Traffic Signal Monitoring, Structural Health Management System, Traffic Management System, etc. Figure 1 presents the major application areas of IoT where IoT is marking its presence to create a smart environment.

In this section, we have presented some of the applications of IoT sensors.

2.1. Intelligent Transportation:

It is one of the prominent domains of Smart City application. All aspects of VANET safety involve communication, decision making, observation, and action. Currently, self-driving cars are also high attention to intelligent transportation. And this is possible due to sensors embedded in them [11].

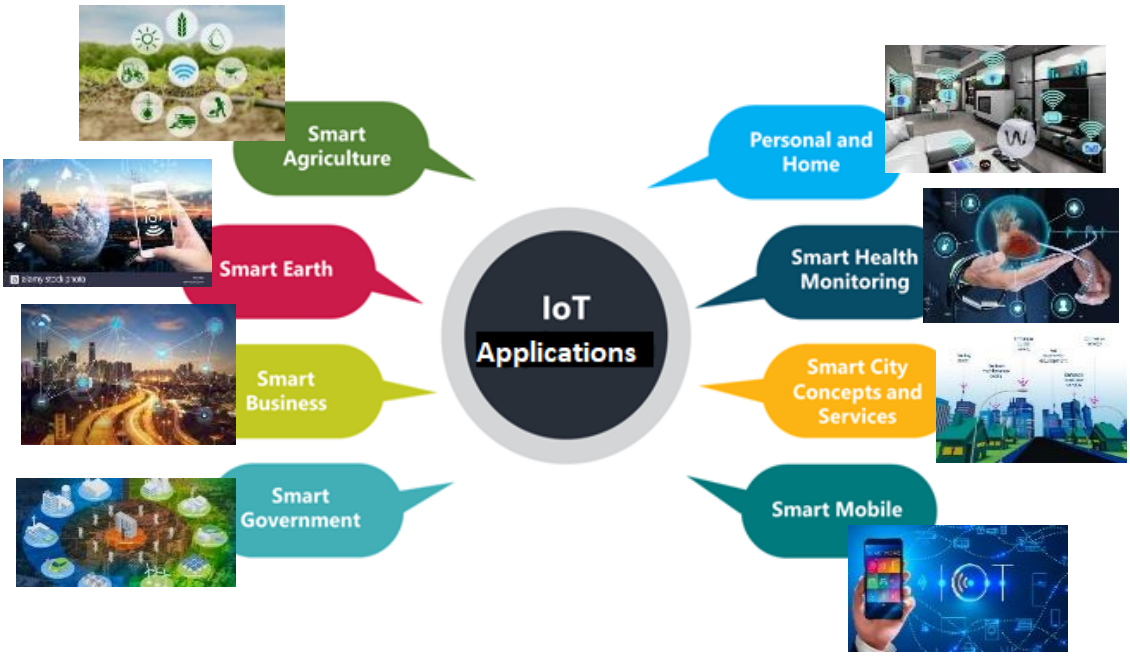


Figure 1: Major IoT Applications

2.2. Military Applications

In Defense Operation, Information sharing between soldiers is an important issue for effective combat. But, in the hostile sector, there is a very difficult process for building infrastructure, most of which, communication infrastructure has always become the first target of the enemy. Hence creation and maintaining communication infrastructure is the essential issue of a military operation. In defense Operations, Information sharing between soldiers is an important issue for minimal casualties and effective combats. Sensors are used in various areas in military like Reconnaissance of opposing forces and terrain, Battlefield surveillance, Battle damage assessment, Equipment, and ammunition, monitoring friendly forces.

2.3. Medical and Healthcare Applications

In the medical & Healthcare applications, the Sensors promises to solve various issues that are related to medical discoveries on incurable diseases. Most of the Healthcare devices are linked wirelessly through sensors [12]. In medical, IoT is used in identification management (IDM). In IDM each thing (device or user) needed to have a Unique Identification and it ensures identity and information safety. IDM also relays on sensors. Authors in [13], proposed a NIGHTCare platform that overnight detects the abnormal activities and events of patients in a health center. In this application, sensors are attached to the clothes of the patients. Another researcher in the same area presented a low-cost rural e-health monitoring model [14]. A similar smart health application by authors in [15] monitors the physiological parameters of the patients. Sensors that detect the temperature and humidity of a patient's body are used by the researchers in [16]. These sensors are placed inside the shoe insole.

2.4. Vehicular Network

VANET usually linked to the internet using APs using WLAN (wireless local area network) Technology such as Sensors, Bluetooth, WiMAX, WI-Fi. Hence sensors play an important role in this area. VANET application: Prevention of Disaster scenarios, Evacuation emergency, Proximity, and correlation analysis, Traffic flows, Congestion, Mobile sensor networks through vehicle communications, Road congestion avoidance, etc.

2.5. Smart Cities

Sensors aim to make things convenient and dynamic. Things can exchange data using wireless technology and provides many convenient services. We need new techniques to address various challenges in Smart City. We need to come up with a new approach, where the internet of thing is used to minimize the communication problem of the sensors. IoT deployed in a smart environment comprises IoT devices to transmit the packets to the sink. The Internet of Things encapsulates sensors as an essential component. For surveillance and monitoring of smart environment applications, sensors are implemented in wide areas. Various examples are Highway Data Collection, Electric Toll Collection System, Traffic Signal Monitoring, Structural Health Management System, Traffic Management System, etc [11] [17].

2.6. Heterogeneity of Network

Intelligent transportation and smart building involve Sensors to add the different components within the network. In these systems, sensors need to be capable to co-exist. The network requires accurate mapping of different control information within the Network Stack.

2.7. Wired and Wireless Connectivity

Almost all the applications in the smart environment require wireless communication and wireless communication need sensors. The Internet of Things has a discrete topology that ties heterogeneous things/ devices together. Since in the Smart City, the data is inexhaustible and it has an expiration time associated with it means the information must reach the destination within a limited mention time and with good connectivity. Sensors play a vital role to provide good connectivity.

2.8. Smart Environment/ Smart Earth

Sensors are also used for monitoring the environmental conditions of the earth to provide a smart environment. It helps in the development of the country. This is not so simple and easy as it seems to be. Monitoring earth is a challenging area because of various constraints like harsh operating conditions and difficulty in maintenance. Also, highly developed and latest advancements make it more challenging [18].

2.9. Smart Water

Water is an essential part of life for every living being. It is necessary to monitor the quality of water in terms of pollution, minerals, its level from the ground, and many more. Sensors are widely used to monitor water quality. It's one widely used application is detecting leakages and water overflow in tanks. This not only saves water but also assist in water management [19]. Some of the smart applications of water are detecting water leakages, monitoring seawater pollution level, smart pools, monitoring river floods, dams and reservoirs levels, and many more. All this is possible due to the sensors that are attached to various sites.

2.10. Smart Security:

IoT is marking its presence in almost all areas. In this scenario, our day-to-day very personal data is over network all the time. So, it becomes more important to secure this data. Also, sensors are attached to the walls of the containers in various applications for security purposes like intrusion detection systems in homes and offices, monitoring gas levels in sensitive applications, smart monitoring radiation levels, and early detection of leakages. This not only saves a life but also helps in timely management. Sensors are also used for implementing smart perimeters around specific areas that restrict the entry of intruders at that place. A similar application is designed by the authors in [19] where an intelligent intrusion system attached at door, named ADXL 345 accelerometer is given. A similar embedded security framework for IoT is proposed by the researchers in [20].

2.11. Smart Agriculture

Sensors are extensively used in several agriculture applications. When attached to the soil, it detects moisture in the soil and by monitoring weather conditions like humidity and temperature it sprinkles water and pesticides in the fields according to the grown crops. An application model is proposed by the

researchers in [21] which detects the humidity, moisture in the soil, rainfall, pH value of soil, temperature, and seed recognition. A similar study in [22] used a drone for smart agriculture. Water sprinkling in IoT agriculture is presented in [23], it is a case of vertical farming.

2.12. Smart Homes

Sensors are widely used in creating a smart home. Sensors are attached to various household devices like refrigerator, doors, lights, fan, AC, washing machine, grass, water tank, kids' room, school bags, curtains, tv, and almost to every device or appliance. Its application for smart home includes intruder detection, switching on/off electric appliances, opening/closing doors, sending meeting information, smart cameras for guests, detecting the amount of ingredient in cooking and so many. This all needs sending alerts to provide a better lifestyle. Various researchers worked in this area. Researchers in [24] proposed an intrusion detection system that intimates the owner via messages whenever senses motion in the door.

3. Analysis of IoT Applications and Sensors

Different types of sensors are widely used in different IoT applications. After analyzing different applications of IoT, this section recommends which sensors are more suitable and efficient for a particular application. For example, smart city monitors movement on highways which requires position sensor, velocity sensor, accelerometer sensor; toll collection system requires proximity sensor, position sensor, and accelerometer sensor; traffic management in a smart city requires infrared sensor, GPS, and position sensor; and to monitor building structural health, it requires humidity, pressure and temperature sensors.

When talking about a smart environment, it requires monitoring of weather conditions for which a light sensor, temperature sensor, humidity sensor, and gyroscope are required. Whereas, to monitor the emission of harmful gases in the environment, there is a need for chemical sensors, biosensors, and also optical sensors.

Similarly, smart agriculture requires monitoring of soil via humidity and chemical sensors. For monitoring weather conditions for sprinkling water, it requires temperature, humidity sensors. When there is a need for monitoring of intruders in the field it requires proximity or position sensor. A water quality sensor helps in detecting the water level of the soil.

Smart health monitors a patient, so it requires different bio-related sensors that detect abnormal behavior in a patient like biosensors, chemical sensors, magneto sensors, and pressure sensors. To detect the physical location of patients or health providers there is a need of a position sensor.

Different types of sensors are used and required in different IoT applications. Figure 2 shows which type of sensor is used for a particular IoT application.

The analysis of various IoT applications and sensors shows that some applications require more than one type of sensor. Besides, selecting any sensor for an IoT application requires analyzing and identifying different types of activities for which we need sensors.



Figure 2: Types of sensors used in major IoT applications

4. Human Activity Recognition (HAR) System

HAR System detects human activities by observing various parameters through different sensors attached to the human body via different wearables. Various researchers worked in this field and proposed numerous HAR systems.

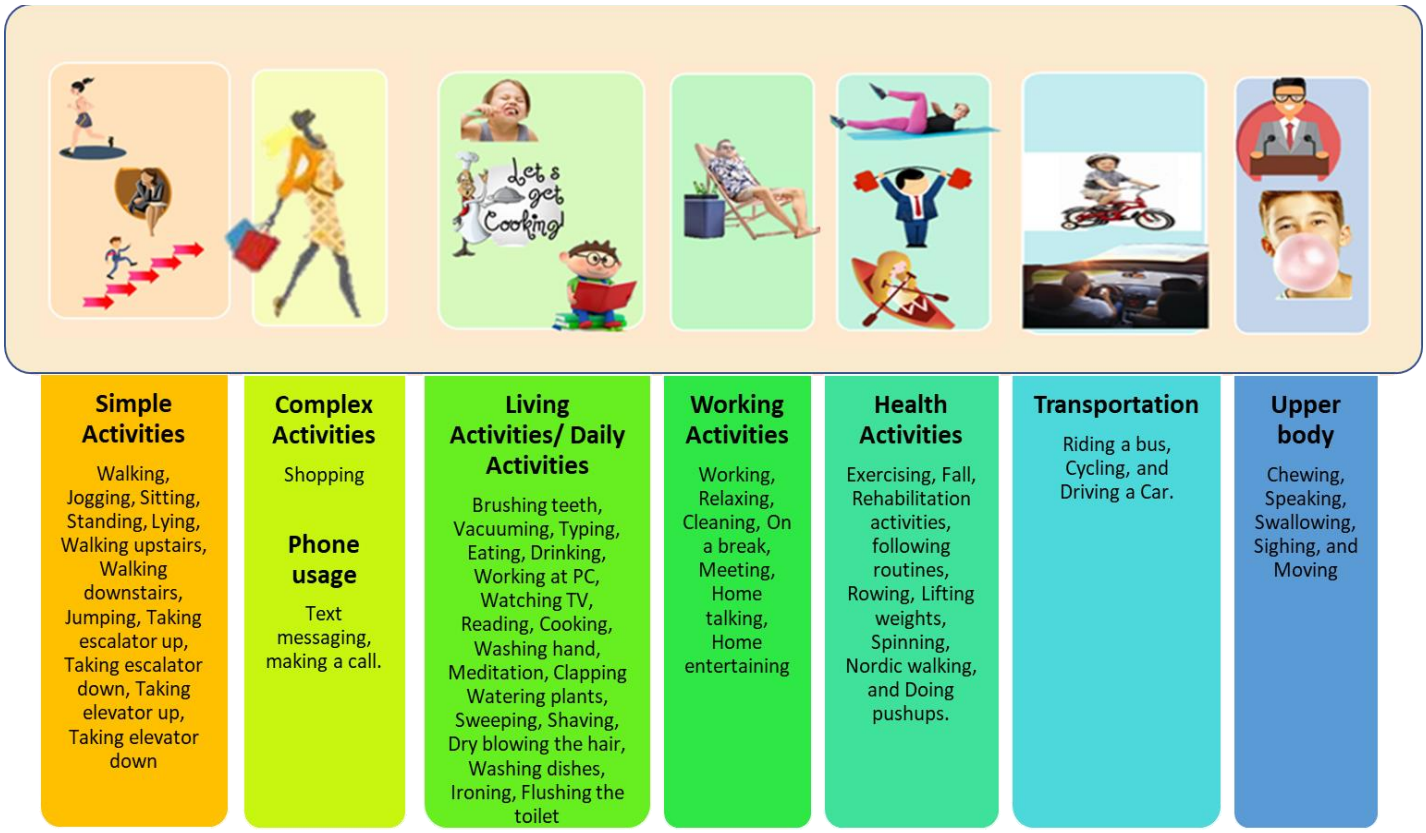


Figure 3: Various types of activities according to category

These monitoring systems are usually smaller and are wearable. Various researchers worked on detecting human activities by observing different parameters. It can be used for the monitoring of people of any age group like for infants, young people, disabled people, and elderly people in a solitary life. Researchers in [25] presented a wearable monitoring system for a patient that monitors various physiological parameters of the patient in healthcare. Notifications are sent to the concerned (doctor/ relative) in any abnormal event via a twitter account. Another similar study in [26] proposed a personalized health monitoring system. This system attaches a sensor inside the shoe that monitors the pressure, temperature, and humidity and sends the health data to the user via a hand-held device. A smart university campus is proposed by the authors in [27] to make smart rooms, smart parking, and smart education a reality. Wearable sensors are mostly used to detect physical movements but these can also be used for monitoring academic performance, analyzing sleep quality/patterns, stress level, and mental health conditions. Researchers in [8] gathered data from 66 individuals and analyzed behavioral and physiological patterns. They also identified factors affecting performance in academics, sleep quality, stress level, and mental health. HAR systems recognize various activities these are generally classified in different groups. Figure 3 enlists these types of activities according to the category.

5. Security Mechanism in IoT

As we know that in IoT enabled smart environments, everything is connected via the internet and data is communicated with other similar devices. In IoT based environment, there are approximately 6 times IoT enabled devices than people which are connected in IoT. Every minute, every second our very personal and sensitive data is over the network. This is serious and couldn't be overlooked. So, this smart environment requires an even more strong security mechanism than ever before. To fulfil the needs of IoT security, strong cryptographic solutions are required. Numerous security algorithms are present now-a-days like AES, DES, RSA, and so many. But using these traditional security algorithms in the IoT environment is not possible because of various constraints. As sensors are the main elements of IoT which are attached to even very small devices. So, implementing traditional security mechanisms in IoT enabled devices is not possible as these are having high memory requirements that are not available in small sensors. Also, there is a different operating environment, low available power, and frequent maintenance in IoT enabled devices is not possible. So, there is a need for such a security mechanism which fulfils the needs of constraint devices. In this direction, there comes a new area of cryptography, which is known as lightweight cryptography. Numerous researchers in the area of IoT security has given different lightweight security algorithms. Moving in the same direction, we have also proposed three different lightweight security algorithms, BRIGHT [28], [29], [30], and UBRIGHT [31]. A total of 9 different security algorithms are proposed in these three security algorithms.

6. Proposed Work

In this section, a HAR System is proposed. This system enables users to identify various types of activities. Security is an essential part of any HAR system, so in the proposed system for enhancing the security, one of the newly designed lightweight cipher out of BRIGHT, UBRIGHT, and Expanded-BRIGHT is proposed to be used.

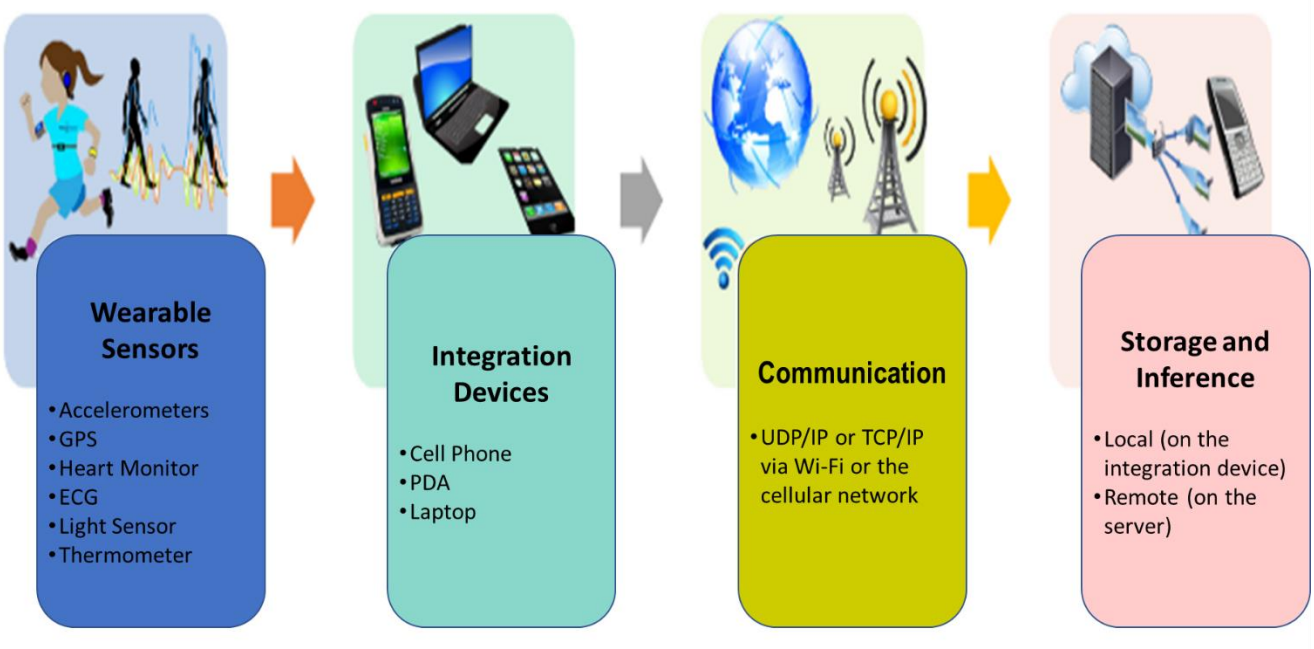


Figure 4: Generic data acquisition architecture for Proposed HAR systems

The first step in any HAR system is collecting raw data using sensors which are then passed through feature extraction and learning processes that recognize the activities. This is achieved through machine learning techniques that may be supervised or unsupervised or semi-supervised.

Figure 4 represents generic data acquisition architecture for the proposed HAR systems. First using different wearable sensors or accelerometer, data are gathered which is sent to some integrated devices where processing is carried out.

Feature Extraction is the next step where the sensors sampling rate is faster as compared to the time taken by human activities. A single sample at any specific instant of time is not sufficient to provide useful information. So, to successfully recognize any activity sampling must be considered for time windows rather than at any specific instant of time. The main underlying problem is that it is not possible to make a comparison between two-time windows. It's almost impossible for signals from the same subject performing the same activity to be identical. So, feature extraction comes into the picture and plays an important role in any HAR systems. There are two types of techniques for feature extraction from time-series data, these are statistical and structural. To extract features from the raw data, statistical methods make use of quantitative characteristics of the data. Fourier transform and the wavelet transform are examples of statistical methods. The methods in which interrelationship among data are considered into account are known as structural approaches.

Learning is a necessary step to build useful patterns that describe and analyze the data. Machine learning is carried out from examples or observations called a training set. Machine learning may be supervised or unsupervised. In unsupervised learning, it is quite not possible to distinguish the activities that are to be recognized. In supervised approaches, the labeling of different activities along with supervising collection process is carried out by some person whereas in an unsupervised learning technique no person is involved. Different researchers in this area are working on both supervised and unsupervised learning techniques. Then this data is communicated to the concerned and finally stored on some integration device or remote server.

7. Conclusion

IoT has numerous applications where different types of sensors are used in different types of applications. This paper discusses different IoT applications and enlightens which smart application demands which type of sensor. Besides, various types of wearable sensors which are used for monitoring of human activities and application area where these are used are also presented in this paper. The positioning of wearable sensors along with the selection of relevant features for different activities pose challenges in this area. Our very personal and sensitive data is over the network in this smart environment. This is serious and couldn't be ignored. So, it requires a strong security mechanism in the form of lightweight cryptography. In this paper, a HAR system is also proposed that enables users to identify various types of activities. To provide high enough security to the proposed system, very efficient and secure crypto solutions are also used.

References

[1] D. Sehrawat, N. S. Gill, "Smart Sensors: Analysis of Different Types of IoT Sensors" in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (523-528), 2019. IEEE. https://doi.org/10.1109/ICOEI.2019.8862778

[2] D. Sehrawat, N. S. Gill, "Security Requirements of IoT Applications in Smart Environment", in: Proc. of 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018), 324-329, 2018. https://doi.org/10.1109/ICOEI.2018.8553681

[3] M. Ambrosin, A. Anzanpour, M. Conti, T. Dargahi, S. R. Moosavi, A. M. Rahmani, P. Liljeberg, "On the Feasibility of Attribute-Based Encryption on Internet of Things Devices", IEEE Micro, **36**(6), 25–35, 2016. http://doi.org/10.1109/MM.2016.101

[4] D. Sehrawat, N. S. Gill, "Deployment of IoT based Smart Environment: Key Issues and Challenges", International Journal of Engineering & Technology. **7**(2), 544-550, 2018. http://dx.doi.org/10.14419/ijet.v7i2.9504

[5] H. Lee, K. Kim, J. Kwon, "A pervasive interconnection technique for efficient information sharing in social iot environment," Int. J. Smart Home, **10**(1), 9–22, 2016.

[6] S. Kale, M. Satendra, P. Patil, "IOT based Wearable Biomedical Monitoring System," in International Conference on Trends in Electronics and

Informatics ICEI 2017 IOT, 2017, 971–976. https://doi.org/10.1109/ICOEI.2017.8300852

[7] J. Wilden, A. Chandrakar, A. Ashok, N. Prasad, "IoT Based Wearable Smart Insole," in Wireless Summit (GWS), 2017 Global IEEE., 2017, 186–192. https://doi.org/10.1109/GWS.2017.8300466

[8] A. Sano, A. J. Phillips, Z. Y. Amy, A. W. McHill, S. Taylor, N. Jaques, N., ... & R. W. Picard, "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones", in 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (1-6) , 2015. IEEE. https://doi.org/10.1109/BSN.2015.7299420

[9] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat, "Physical human activity recognition using wearable sensors" Sensors, 15(12), 31314-31338, 2015. https://doi.org/10.3390/s151229858

[10] L. Atallah, B. Lo, R. King, G. Z. Yang, "Sensor placement for activity detection using wearable accelerometers" in 2010 International conference on body sensor networks, 24-29, 2010. IEEE. https://doi.org/10.1109/BSN.2010.23

[11] A. Zanella, N. Bui, a Castellani, L. Vangelista, M. Zorzi, "Internet of Things for Smart Cities," IEEE Internet Things, **1**(1), 22–32, 2014. 10.1109/JIOT.2014.2306328

[12] D. Zissis, D. Lekkas, "Addressing cloud computing security issues", Futur. Gener. Comput. Syst., 28(3), 583–592, 2012. Available Online: https://doi.org/10.1016/j.future.2010.12.006

[13] C. Occhiuzzi, C. Vallese, S. Amendola, S. Manzari, G. Marrocco, "NIGHT-care: A passive RFID system for remote monitoring and control of overnight living environment", Procedia Comput. Sci., **32**, 190–197, 2014. https://doi.org/10.1016/j.procs.2014.05.414

[14] R. Chanchal, J. Chaman, W. Arif, "HEMAN : Health Monitoring and Nous", in IEEE WiSPNET 2017 conference., 2017, 2115–2119. 10.1109/WiSPNET.2017.8300134

[15] S. Kale, M. Satendra, P. Patil, "IOT based Wearable Biomedical Monitoring System", in International Conference on Trends in Electronics and Informatics ICEI 2017 IOT, 2017, 971–976. 10.1109/ICOEI.2017.8300852

[16] J. Wilden, A. Chandrakar, A. Ashok, N. Prasad, "IoT Based Wearable Smart Insole", in: Wireless Summit (GWS), 2017 Global IEEE., 2017, 186–192. 10.1109/GWS.2017.8300466

[17] P. Patil, "Smart IoT Based System For Vehicle Noise And Pollution Monitoring Piyush," in Proc. of International Conference on Trends in Electronics and Informatics, 2017, 322–326. 1109/ICOEI.2017.8300941

[18] M. T. Lazarescu, "Design of a WSN Platform for Long-Term Environmental Monitoring for IoT Applications," IEEE J. Emerg. Sel. Top. CIRCUITS Syst., 3, 45–54, 2013. 10.1109/JETCAS.2013.2243032

[19] R. Khan, S. U. Khan, R. Zaheer, S. Khan, "Future internet: The internet of things architecture, possible applications and key challenges," in: Proc. of - 10th Int. Conf. Front. Inf. Technol. FIT 2012, April 2017, 257–260. 10.1109/FIT.2012.53

[20] A. Stango, N. R. Prasad, J. Sen, "Proposed Embedded Security Framework for Internet of Things (IoT)," in: Proc. of 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), IEEE, 2011, 1–5. 10.1109/WIRELESSVITAE.2011.5940923

[21] P. Patil, V. Sachapara, "Providing Smart Agricultural Solutions/Techniques By Using Iot Based Toolkit," in: Proc. of International Conference on Trends in Electronics and Informatics ICEI 2017 Providing, 2017, 327–331. 10.1109/ICOEI.2017.8300942

[22] A. K. Saha *et al.*, "IOT-Based Drone for Improvement of Crop Quality in Agricultural Field," IEEE Internet Things J., 612–615, 2018. 10.1109/CCWC.2018.8301662

[23] M. Ikhwan, N. M. Thamrin, "IoT Implementation for Indoor Vertical Farming Watering System", IEEE Internet Things J., 89– 94, 2017. 10.1109/ICEESE.2017.8298388

[24] S. Nazeem Basha, S. Jilani, M. Arun, "An Intelligent Door System using Raspberry Pi and Amazon Web Services IoT", Int. J. Eng. Trends Technol., **33**(2), 84-89, 2016.

[25] M. Ermes, J. Pärkkä, J. Mäntyjärvi, L. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions", IEEE transactions on information technology in biomedicine, **12**(1), 20-26, 2008. https://doi.org/10.1109/TITB.2007.899496

[26] L. Bao, S. S. Intille, "Activity recognition from user-annotated acceleration data", in *International conference on pervasive computing*, 1-17, 2004. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24646-6_1

[27] K. Kunze, P. Lukowicz, H. Junker, G. Tröster, "Where am i: Recognizing on-body positions of wearable sensors", in International Symposium on Location-and Context-Awareness, 264-275, 2005. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11426646_25

[28] D. Sehrawat, N. S. Gill, "BRIGHT - Proposed Family of Lightweight Block Ciphers for IoT-Enabled Smart Environment", International Journal of Innovative Technology and Exploring Engineering (IJITEE). **8**(9), 584-592, 2019.https://www.ijitee.org/wp-content/uploads/papers/v8i9/I7778078919.pdf

[29] D. Sehrawat, N. S. Gill, "BRIGHT: A Small and Fast Lightweight Block Cipher for 32-bit Processor", International Journal of Engineering and Advanced Technology, **8**(5), 1549-556, 2019. https://www.ijeat.org/wp-content/uploads/papers/v8i5/E7302068519.pdf

[30] D. Sehrawat, N. S. Gill, "Performance Evaluation of Newly Proposed Lightweight Cipher, BRIGHT", International Journal of Intelligent Engineering & Systems, **12**(4), 71-80, 2019. http://www.inass.org/2019/2019083108.pdf

[31] D. Sehrawat, N. S. Gill, "UBRIGHT: Ultra BRIGHT, a Tiny and Fast Ultra Lightweight Block Cipher for IoT", International Journal of Scientific & Technology Research, **9**(2), 1063-1068, 2020. http://www.ijstr.org/final-print/feb2020/Ultra-Bright-A-Tiny-And-Fast-Ultra-Lightweight-Block-Cipher-For-Iot.pdf

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 2, 545-550 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Capacity Factor of Renewable Energy Power Plants During Electric Power Peak Times in Jeju Island

Zulmandakh Otgongerel, Gaemyoung Lee*, Ankhzaya Baatarbileg

Department of Electrical Engineering Jeju National University, Jeju, 63243, South Korea

ARTICLE INFO

Article history:
Received: 29 January, 2020
Accepted: 27 March, 2020
Online: 08 April, 2020

Keywords:
Renewable energy
Solar power plant
Wind power plant
Capacity Factor
Peak period

ABSTRACT

Jeju Island is the largest island of South Korea and has an independent electric supply system. The island has a special energy plan that realizes itself a carbon-free region by 2030. Wind and solar energy resources are considered as an important means. The introduction of weather-variable renewable energy generation sources into the electric power system makes it unstable. The main content of this study was to investigate the capacity factor of large scale renewable energy (solar, wind) power plants operated in Jeju Island during the electric power peak time. We can see their utilization rate or contribution rate during the power peak time by the value of their capacity factors. The largest 18 solar power plants (SPP) and 8 wind power plants (WPP) on the island were chosen and the 4, 12, 25, 50 and 100% electric power peak times were set for this study. There are two electric power peak seasons, summer and winter, in Jeju Island when the electric power demand arrives at peak. Generation of renewable energy plants depends on weather and the climates of the 5 regions of Jeju Island are a little different from one another. So, we investigated the difference between the average capacity factors of the 5 region's renewable plants. WPPs showed a high contribution to the power grid only during winter peak time and SPPs did high contribution to the power grid only during summer peak time, while WPPs showed low contribution to the power grid during summer peak time and SPPs did low contribution to power grid during winter peak time. WPPs and SPPs have inter-compensation relation in the aspect to contribute to the power grid.

1. Introduction

Population increase and industrialization have resulted in rapid increases in energy consumption and greenhouse gas (GHG) emissions. Therefore, many countries have been focusing on the wider application of renewable energy. South Korea also announced the "Renewable Energy 3020 Plan" in 2017 [1]. The global utilization of renewable energy increases annually and renewable plants are one of the largest factors in the world energy system. The capacity of the renewable power plants installed in 2018 was 181 GW and the annual electric energy generated by renewable sources grew by an average of 3.8 percent from 1990 until 2018, which is higher than the average increase rate of 2.9 percent of total electricity generation [2]. In addition, renewable power capacity totaled 2,378 GW (including hydropower) and 1,246 GW (not including hydropower). By the end of 2018, 230 countries proposed 100% renewables in at least one sector, and 169 countries adopted renewable energy targets. [3]. The international energy agency reported that wind power systems are the second fastest growing source of renewable electric energy after solar power systems. Furthermore, between 1990 and 2018, growth of electric energy generated by wind power plants from 3.8 TWh to 745.2 TWh showed an average annual increase rate of 20.7% [4]. Between 2005 and 2017, the world photovoltaic electricity production increased from 4TWh to 444TWh [5]. In 85 countries, states and provinces with renewable portfolio standards have the targets to produce more than 50% of their electric energy using renewable energy plants [6]. These documents showed that solar and wind resources are the lead renewable energies worldwide and this power plant technology has developed rapidly over the last few years.

Jeju's regional government announced the "Carbon Free Island by 2030 Plan (CFI2030 Plan)" in 2012 [7], which is to realize itself a carbon-free region by 2030. The capacity of renewable (solar and wind) electric power plants were expected to reach 4.15GW by 2030, which consists of on-shore wind power plants 0.45 GW, off-

*Gaemyoung Lee, Jeju National University, myounglk@jejunu.ac.kr

shore wind power plants 1.9 GW and photovoltaic power plants 1.8 GW.

Also, about 377,000 electric vehicles will run on the island by 2030 [8]. In the case of the Jeju Island, WPPs and SPPs were estimated about 260.6MW and 160.8MW in 2018 [9]. In the aspect of supply and consumption of electric energy, demand and supply must be always equal to each other. Namely, the electric power of the supply side must be always equal to one of the demand side. Also, the supply side always has more than electric power capacity than the demand side needs and then has sufficient electric generation plants to supply timely the demand side with electric energy. The problem is that renewable electric power plants such as WPPs and SPPs can't always control the amount of generation according to the request of the demand side because their generation depends on the weather. Unlike the existing other power generation resources, renewable power generation sources cannot be rated. If extra generation facilities that cover the stop of the renewable electric power plants were reserved, the problem will be solved, but it needs big cost. Therefore, the estimation of the average contribution of renewable electric power plants during the electric power peak time when most generation facilities will be operated is very important.

There are a few case studies on how to approximate the supply capacity of renewable power resources [10, 11]. Some studies used the capacity factor as a measure of how much energy is generated by power plants compared to their maximum output. Ref [12] explained and compared the peak-period capacity factor method and effective load-carrying capability method, and Ref [13] analyzed the capacity value of locations in the south-western United States. Similarly, ref [14] estimated the capacity factor of SPPs in Korea.

The electric energy supply side must have sufficient electric power during even electric power peak times in summer and winter seasons. It is very important in designing a stable energy mix for a region to know the past statistical contribution of renewable power plants (SPPs and WPPs) during the peak times in the peak seasons because the renewable power plants cannot be always operated when the demand increases. This study is going to examine the contribution of SPPs and WPPs to the power grid by evaluating their capacity factors during the electric power peak times in Jeju Island.

2. Methodology

2.1. Regional grouping of renewable power plants

Jeju Island is the largest island in South Korea and locates in North-East Asia. The island has four distinct seasons. Winter is cool and dry while summer is hot, sometimes rainy and humid [15]. In the case of renewable energy plants, unlike conventional thermoelectric power plants, their power capacity possible to supply is determined by climate or weather. To suggest the regional generation characteristics of renewable power plants, grouping them needs according to the climates of the regions at which they are installed. In this study, 8 wind power plants and 18 solar power plants were selected and grouped into the three and the four regions respectively. Furthermore, we chose the WPPs and SPPs with the power capacity of more than 0.5MW and 15MW respectively for this study.

The Korea Meteorological Administration divided Jeju Island into the 5 regions according to weather; western, eastern, southern, northern and mountain regions.

Figure 1 shows the regional grouping of solar power plants (SPPs). The mountain region is not selected because it has a few SPPs. In this paper, we selected the SPPs that have generation data for more than two years.

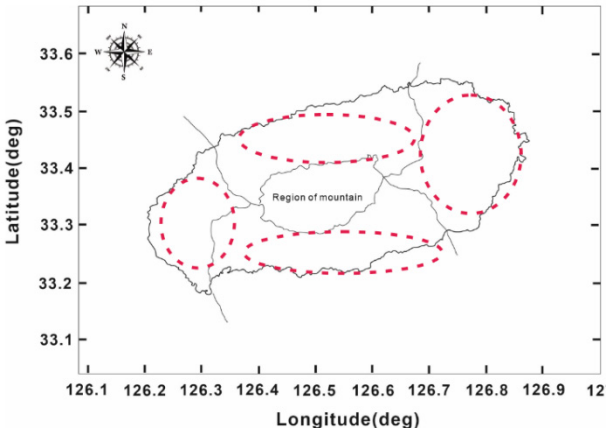


Figure 1: Regional grouping of solar power plants

Figure 2 shows the regional grouping of wind power plants. Wind farms were usually constructed in the west and east parts because both regions have a strong windy climate. Let call the 3 regions western, eastern and southern regions respectively.

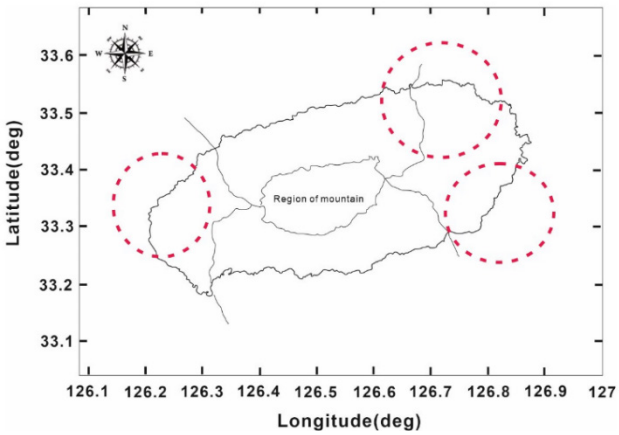


Figure 2: Regional grouping of wind power plants

2.2. Data sources

The data on the hourly average electric power demand of Jeju Island for three years (2015-2017) was used to define its electric power peak time and was obtained from the Korean Power Exchange Company (KPX). The data on the hourly electric energy generated by wind and solar powers over three years (2015-2017) are also obtained from KPX.

2.3. Choosing the electric power peak seasons

Figure 3 shows the average electric power demand of Jeju Island for three years (2015-2017). The electric power peak occurred in July and August in the summer season and December and January in the winter season.

Winter electric power peak season was chosen from December to January and the summer electric power peak season was done from June to August in this study.

2.4. Defining the electric power peak times

The electric power peak time can be selected from the data on the electric demand during the electric power peak day within the specific period, or can be done from the data on the average electric demand during the specific electric power peak season. We choose the latter because, usually, the renewable power plants always operate in the case that weather matches. In this study, the data on the average electric demand during the winter and the summer electric power peak seasons was used in determining the electric power peak time.

This study suggested an electric power peak time system with the five types of electric power peak times: the 4, 12, 25, 50 and 100% electric power peak times. They mean to have the 1, 3, 6, 12 and 24 hours intervals, respectively, in the order of the volume of electric power demand. Let's call them as the % electric power peak times. The electric power demand curve in Figure 4 shows the July-August average power demand for the 3 years (2015-2017) in the respective time interval. In Figure 5 does the December-January's one. We selected the % electric power peak times during the summer and the winter electric power peak seasons from Figure 4 and 5.

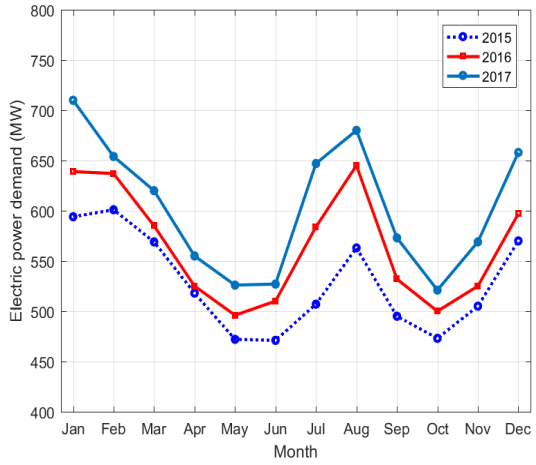


Figure 3: Monthly average electric power demand on the island (2015-2017)

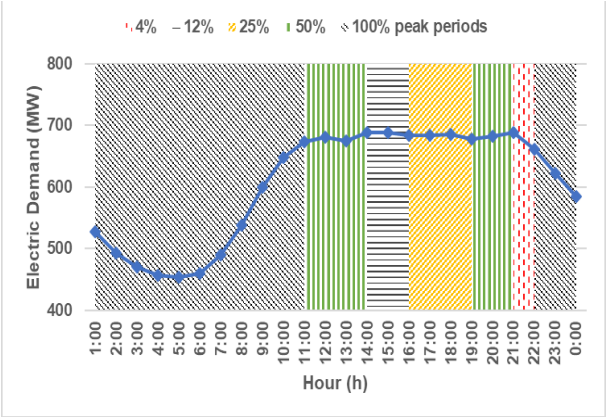


Figure 4: Classification of the % power peak periods in summer (Jul-Aug)

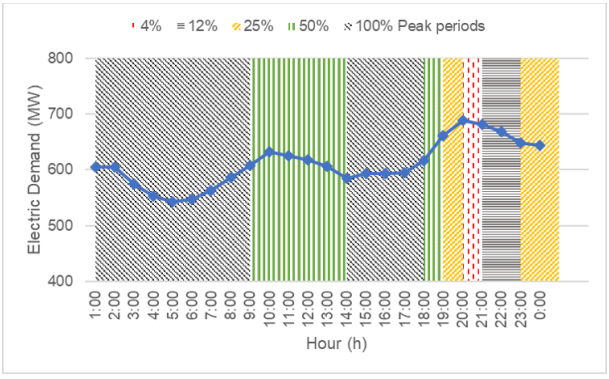


Figure 5: Classification of the % peak periods in winter (Dec-Jan)

2.5. Calculation of the capacity factor

The utilization rate of an electric power plant can be estimated as the parameter, capacity factor. The capacity factor of an electric power plant, *CF* is obtained following equation:

$$CF = \frac{GE(Wh)}{T(day)\times 24(hour/day)\times C(W)} \times 100(\%) \quad (1)$$

where T is the period that the plant has been operated, C is its electric power capacity and GE is the electric energy generated during the time interval T by it. The capacity factor of the N electric power plants, $CF(N)$ is obtained by the equation:

$$CF(N) = \sum_1^N \frac{GE_i(Wh)}{T_i(day)\times 24(hour/day)\times C_i(W)} \times 100(\%) \quad (2)$$

where T_i is the period that the i plant has been operated, C_i is the electric power capacity of the i plant and GE_i is the electric energy generated during the time interval T_i by the i plant. The capacity factor of multiple plants is calculated as the weighted average value and is not as simple an arithmetic average.

3. Results & Discussion

3.1. Capacity factors of wind power plants and solar power plants during 2015-2017 in Jeju Island.

Yearly-average capacity factors of WPPs and SPPs in the respective hour interval are shown in Figure 6 and Figure 7, respectively. The capacity factor of an electric power plant means the utilization rate of the plant. The operation of wind or solar power plants depends on weather.

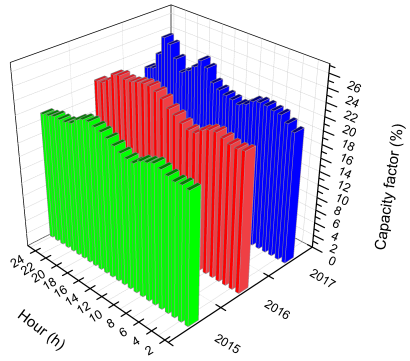


Figure 6: Average CF of WPP on the island (2015-2017)

Figure 6 shows some regular trend of generation time of WPPs, namely generation amount increasing in the morning and in the evening. This trend attributes to an island climate that regular wind flows from sea to land in the morning and from land to sea in the evening. Figure 6 shows the capacity factor range from 19% to 25%.

Figure 7 shows a daily trend of SPPs, the capacity factors of SPPs ranging from 0% to 50%. Generation occurs only daytime and reaches the peak at noon because 100% of SPPs in Jeju Island is the fixed type which does not track moving sun.

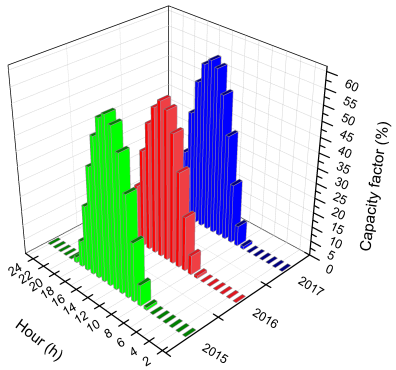


Figure 7: Average CF of SPP in the island (2015-2017).

Table 1 shows the yearly average capacity factors of regional WPPs and SPPs. The western region is shown to have better competitiveness for solar power plant business.

Table 1: Average CF of renewables on an island (2015-2017).

Region	North	East	South	West
Wind power plant	-	23.2%	22.2%	21.0%
Solar power plant	12.6%	12.8%	13.5%	17.1%

3.2. Capacity factor of the wind power plants during the electric power peak times

Figure 8 and Figure 9 show the capacity factors of the 3 regional wind power plants during summer and winter electric power peak times, respectively. The capacity factors are ranged from 9% to 15% during all summer % peak times in all regions in Figure 8, while those are ranged from 32% to 38% during all winter % peak times in all regions in Figure 9. The average annual capacity factor of WPPs over 3 years (2015-2017) in Jeju was calculated as 22.1%. Therefore, we can say WPPs in Jeju make a larger contribution to the electric power supply system in winter than in summer.

Generation of WPPs in summer is much below the annual average. Let's compare the difference of the 3 regional capacity factors of WPPs. In the summer peak time within 50% peak time, WPPs in the southern region makes a larger contribution to the electric power supply system than WPPs in the eastern and the western regions by about 1-3% in capacity factor. In the winter peak time within 50% peak time, WPPs in the western region make a larger contribution to the electric power supply system than WPPs in the eastern and the southern ones by about 1-3% in capacity factor. But, we can think that, in the winter peak time, there is no difference in the capacity factors of the 3 regions' WPPs because they have more than 30% capacity factor.

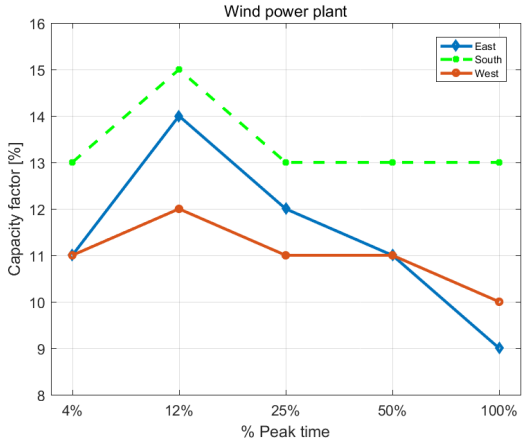


Figure 8: Capacity factor of the wind power plants during the summer electric power peak season (July-Aug)

Next, let's discuss whether WPPs generate more electric energy at peak time than at the nonpeak time. Figure 8 and Figure 9 show that the values of capacity factors at 4, 12 and 25% peak times are larger than those at 100% peak time. Generation by wind and solar cannot increase when we want or demand for electric energy increases. Also, it is necessary and important that WPPs and SPPs generate statistically more electric energy during peak time than during nonpeak time though their generation depends on the weather. WPPs in any region of Jeju Island, luckily, generated statistically more electric energy during electric power peak time than during electric power nonpeak time.

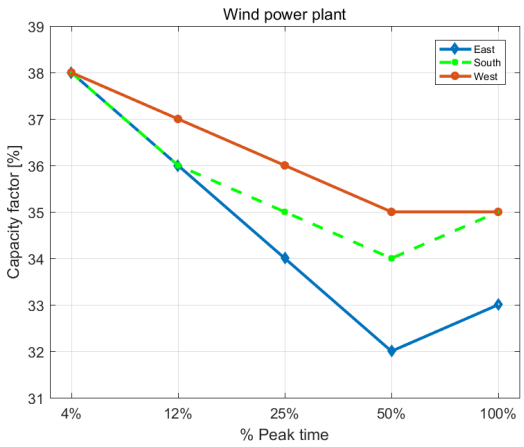


Figure 9: Capacity factor of wind power plants during winter electric power peak season (Dec-Jan)

3.3. Capacity factor of the solar power plants during the electric power peak times

Figure 10 and Figure 11 show the capacity factors of the 4 regional solar power plants during summer and winter electric power peak times, respectively. The capacity factors are ranged from 0% to 32% during all summer % peak times in all regions in Figure 10, while those are ranged from 0% to 12% during all winter % peak times in all regions in Figure 11.

The average annual capacity factor of SPPs over 3 years (2015-2017) in Jeju was calculated as 14.0%. Therefore, we can see that SPPs in Jeju Island contribute greatly to the electric power supply system during summer peak time except during 4% peak time with 1 hour duration, while they do not contribute to it and have the capacity factor below the annual average in winter peak time. Figure 11 shows the contribution of SPPs in all regions to an electric power supply system is zero within 25% peak time duration during the winter peak time.

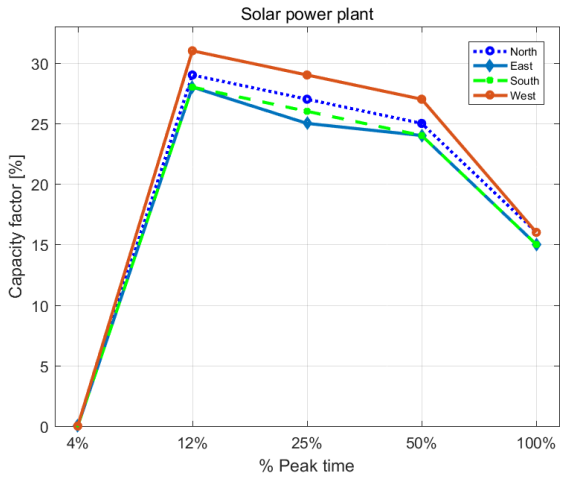


Figure 10: Capacity factor of the solar power plants during the summer electric power peak season (July-Aug)

In the case of SPPs, the regional difference of contribution to the electric power supply system is not shown during the summer and winter electric power peak times, but it is distinguished that the capacity factor of SPPs in the northern region is half of it in other regions during the winter peak time.

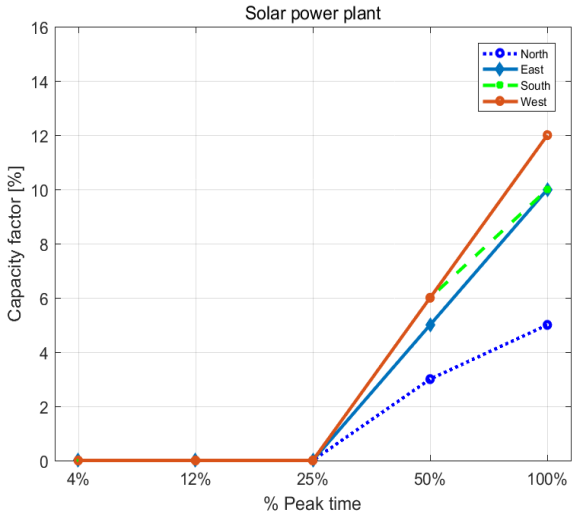


Figure 11: Capacity factor of solar power plants during winter electric power peak season (Dec-Jan)

3.4. Total discussion

In Jeju Island, there are two seasonal electric power peak times. They are the summer and winter peak times. Summer peak time attributes to cooling load and winter peak time to the heating load. We can see, from the electric demand patterns of Figure 4 and Figure 5, that, in the case of the summer peak time, considering the peak duration until 50% peak time needs and, in the case of the winter peak time, considering the peak duration until 25% peak time does.

WPPs showed a high contribution to the power grid only during the winter peak time and SPPs did high contribution to the power grid only during the summer peak time, while WPPs showed low contribution to the power grid during the summer peak time and SPPs did low contribution to power grid during the winter peak time. WPPs and SPPs have inter-compensation relation in the aspect to contribute to the power grid. We think this inter-compensation relation may usually take place in other regions of the world. It is lucky in aspect to penetrate WPPs and SPPs on large scale. Nevertheless, it will be a big problem, in the future that SPPs will have been penetrated on large scale, SPPs can't make any contribution to the power grid during the 4% peak time in summer in Jeju Island because the peak time is at evening.

It is one of the most important research themes in the renewable field to evaluate how much the specific renewable plants contribute to the power grid during the electric power peak time. This paper is to have examined the contribution of WPPs and SPPs to the power grid during the winter and summer power peak times in Jeju Island to be a carbon-free region by 2030. We were able to understand easily the contribution of renewable plants through introducing the concept of the % peak time in this paper.

4. Conclusion

The supply side must have sufficient electric power during even power peak times in summer and winter seasons. It is very important in designing a stable energy mix for a region to know the past statistical contribution of renewable power plants (SPPs and WPPs) during the peak times because the renewable power plants cannot be always operated when the demand increases. This study examined the contribution of SPPs and WPPs to the power grid by evaluating their capacity factors during the electric power peak times in Jeju Island.

The large scale solar and wind power plants in Jeju Island were grouped into four and three regions, respectively, depending on climates of their location. The % peak time was defined and is divided into the 4%, 12%, 25%, 50% to 100% peak times.

WPPs showed a high contribution to the power grid only during winter peak time and SPPs did high contribution to the power grid only during summer peak time, while WPPs showed low contribution to the power grid during summer peak time and SPPs did low contribution to power grid during winter peak time. WPPs and SPPs have inter-compensation relation in the aspect to contribute to the power grid. We were able to understand easily the contribution of renewable plants through introducing the concept of the % peak time in this paper.

References

This research was supported by the 2019 scientific promotion program funded by Jeju National University

References

[1] IEA, Key World Energy Statistics 2018.
[2] IRENA (2019), Renewable Energy Statistics 2019, The International Renewable Energy Agency, Abu Dhabi.
[3] Report citation Renewables 2019 Global Status Report REN21, 2019
[4] Renewables information: overview (2019 edition), International Energy Agency.
[5] A.Gusmao M.Grossbock, Sandi Aramco "Capacity value of photovoltaic and wind power plants in an isolated Mini-grid in the Kingdom of Saudi Arabia" conference paper.
[6] Korea wind and energy industry association (2018) http://www.kweia.or.kr/ (accessed 30 August 2018).
[7] Sang-Hyun Kang, "Analyzing the Capacity of an ESS for Mitigating the Fluctuations in the Dongbok-Bukchon Wind Farm" Jeju National University, the Degree of the master, 2017
[8] Seyed Hossein Madani, Ramteen Sioshansi, Paul Denholm" Comparison of Capacity Value Methods for Photovoltaics in the Western United States" Technical Report NREL/TP-6A20-54704, July 2012.
[9] Shuai Lu, Nader Samaan, Ruisheng Diao, Pevel Etingov "Capacity Value of PV and Wind Generation in the NV Energy System" Pacific Northwest national laboratory, September 2012.
[10] Navid Haghdahi, Anna Bruce, Iain MacGill, and Robert Passey "Impact of distributed Photovoltaic Systems on Zone Substation Peak Demand" IEEE Transactions on sustainable energy, Vol9, No 2, April.2018.
[11] Seyed Hossein Madaeni, Student Member, IEEE Ramteen Sioshansi, Member, IEEE, and Paul Denholm, Member, IEEE "Estimating the Capacity Value of Concentrating Solar Power Plants: A Case Study of the Southwestern United States" conference paper.
[12] Suchitra Subramaniyan, Mobi Mathew, Dr. Jami Hossain "Renewable Energy Capacity Value Estimation using Peak Period Method in Rajasthan" 2018 3rd International Conference for Convergence in Technology, India, Apr 06-08,2018.
[13] Mehdi Shahriari, Seth Blumsack "The capacity value of optimal wind and solar portfolios" Energy 148 (2018) 992-1005.
[14] Cai Jilin, Xu Qingshan, Wang Xudong, Jiang Ling "Estimation for the Capacity Value of PV and Wind Plants Considering Output Correlation" Published by Elsevier Ltd, 1876-6102, 2017.
[15] Chunhyun Paik, Yong Joo Chung, Young-Jin Kim "CF-based Approximation Approach for Calculating Capacity Credit of Solar Power in Korea" Journal of the Korean Solar Energy Society Vol37, No5, pp.13-26, October 2017

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 550-553 (2020)

www.astesj.com

Special Issue on Innovative Research in Applied Science, Engineering and Technology

ASTESJ

ISSN: 2415-6698

The Newtonian Model of the Smolensk Catastrophe

Józef Pawelec*

The European University in Warsaw, Warsaw, 03-199, Poland

ARTICLE INFO

Article history:
Received: 23 April, 2020
Accepted: 02 September, 2020
Online: 27 September, 2020

Keywords:
Suspicious fog
Misleading radar
3D model of crash

ABSTRACT

The pre-reason of the Smolensk catastrophe was a dense fog. The pilots took three trials to find the proper way to airfield. Each case the tower communicated: you are on the curse and path. Pilots, however, resigned. In third critical trial the co-pilot prolonged the response second ring to 8 seconds and the engines could not already take the plane up. It collided with a thick tree and made an upside down. Next it crashed on the ground and left a bloody trace of merely ~100 m long. This means that a mean acceleration at initial speed of 100 m/s and linear braking reached -50 m/s^2. The real values could be even higher as the peak slowdown is always higher the mean. The clue of the Smolensk crash was then a fog and high azimuth error of radar. If it was correct but the ceiling too low, the plane could lose the under-carriage but avoided the upside down and the bloody crash.

1. Introduction

The Smolensk catastrophe took 96 victims: the Polish chief political and military leaders including the current President Lech Kaczynski and the President on Exile - Richard Kaczorowski. The reasons of this tragic event are still under investigations. The Poland claims it was assassination [1]-[3], while the Russia rejects it [4]. So, our statements – although math-physically correct – should be recognized as some hypothesis. The final verdict waits for an international legal process. The plane crashes in civil transport occur one per a million of flights and they are caused by the pilot faults, the engine failures or the foggy weather. It is obvious that Polish Side has chosen the best crew for this high rank visit: the high educated and experienced pilots, familiar with Tu-154. They knew the airfield and speak Russian. Also, the plane has undergone to the major overhaul 2 years ago. So, the technical problems or the pilot faults seem to be unlikely.

2. The Weather

In UTC 10:40:52 of 10-th April 2010 at 2 km to the airport the crew was informed by the tower: *you are in the curse and path!* [3], [4]. The Tu-154 in that moment was 90 m over the local terrain but merely 30 m over the far airfield plane, Fig.1. The speed was 78 m/s and the descent rate –6.2 m/s. Hence, the proper level should be

$$h_{proper} = \left(\frac{2000}{78}\right) * 6.2 = 160\ m,\ h_{choosen} = 30m \quad (1)$$

Moreover, the curse of flight was ~8^0 left the runway. It is defined by the airfield tower and the collision point to come. The direction error was more dangerous than the low level of flight because the last one causes usually the loss of the undercarriage, while the false direction results in dangerous collisions with trees or buildings. And Tu-154 just collided with a birch of 40 cm thick and has lost ~6 m of the left wing! This caused the plane to upside down within several seconds and to crash in half-backside position 425 m before and 150 m aside the airfield in a hostile environment, Fig.1. We will show that the speed of hitting the ground could reach 100 m/s. The crater was also ~100 m long. Hence, at linear braking the negative acceleration reached -50 m/s^2. The real cases are always non-linear and bring higher values, so none could survive.

If there was no directional error, some accident could also take place, but it would surely draw less victims - if any – thanks to the braking action and normal position of plane.

So, we claim that main determining factors of the Tu–154 crash in Smolensk were the extremely foggy conditions and the high directional error of the radar reaching -8^0!

The visibility along the horizon line was ~300 m, so in case of an obstruction there was a few seconds for an escape. In such conditions the Head Command was not entitled to give the permission for landing at all!

It is probably the historical precedence that the plane with President on board was taken down in dense fog with the radar of an angular inaccuracy reaching ±8^0.

*Corresponding Author: Józef Pawelec; Email: k.kosmowski@wil.waw.pl

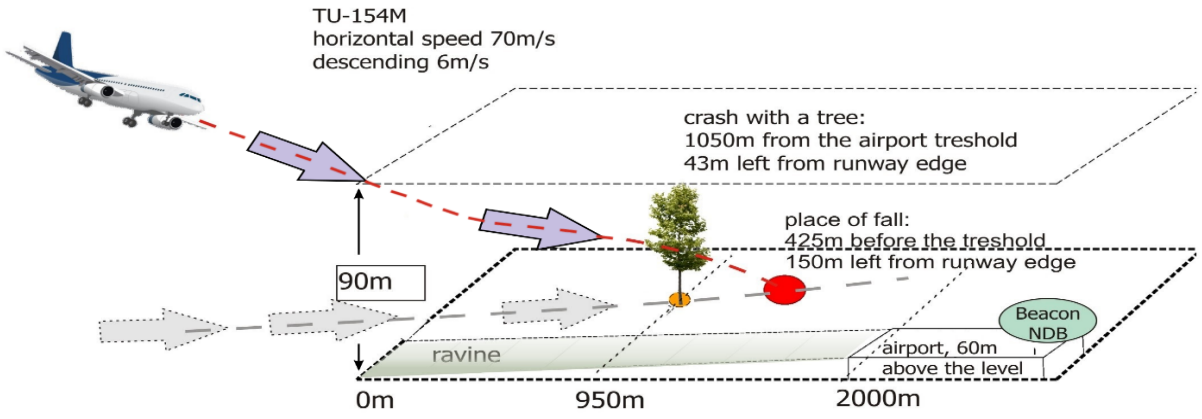


Figure 1: An illustration of the Tu-154 catastrophe: the plane flies much steep and much left the runway

The parts of crash have been dispersed over the area of ~300x300 m. This gave rise to some suspicious theories of assassination. The question is, however, who and when arranged a plot and put an explosive material on the board? So, we will trust further on the physical basis. The mass of Tu-154 reached 90 ton and the landing speed approached 100 m/s due to engines and gravity. If the collision is resilient, the reflection distance is given by eq. (2) [5], [6]

$$h = v^2 \sin(2\alpha)\,/2g \qquad (2)$$

where v – the speed, α – angle to the ground, g – acceleration.

For v = 100 m/s and $\alpha = 15^0$ this distance reaches h≈250 m.

If the collision is not resilient, the body is crashed but the maximal distance of parts dispersion reaches the comparable or even higher values (±).

For live beings the critical parameter is the negative acceleration. According to the witness reports the length of the landing crater reached ~100 m. The initial speed was also ~100 m/s due to the engines and gravitation. Assuming linear decreasing of speed, we obtain the mean of ~50 m/s and the time to run over the crater – 2s. However, at the beginning the speed decreases much higher than in the end, so we will use the total time to run only **1** s. Hence the maximal acceleration reached

$$a = \frac{\Delta V}{\Delta t} \approx -100\,\frac{m}{s} : 1s = -100\,\frac{m}{s^2} \approx -10G \qquad (3)$$

The real values could be even greater, while 3 G is already dangerous.

How these physical phenomena can be transferred onto the live body? The most sensitive is usually the heart. It can tolerate the acceleration up to 3 G. The level of 10 G is usually the lethal one.

4. Modelling of The Up-Side Down

The up-side down turnover played the critical role in all the catastrophe. We will model it using the Newtonian's theory.

The most important role in this theory plays the transversal distribution of the masses m_i along the wings line. If the mass m_i is at the distance r_i from the centre of turning, then – in case of breakdown – the energy and movement increases proportionally to $m_i r_i^2$. We could observe such scenes during the war if a plane tears a part of its wing in a battle. Then, it spins down and crashes in blow of fire.

The Tu-154 has lost one third of the wing in a crash with big tree (40 cm). This was the very dangerous disturbance. We should take into account that 90% of the plane masses are gathered in its very centre (±2 m), so in case of loss of a distant part, the plane is turned very quickly by ~180^0. We will calculate this turn-over using the real data.

The inertial moments will be defined for two mass distributions: the linear and exponential one. Next, we will take a mean. Below, the eq. (4) defines the square of the linear inertial moment k^2_l, while eq.(5) – the exponential moment k^2_e [5], [6]

$$k_l^2 = \frac{1}{M}\sum_{i=14}^{0} m_i r_i^2 = \frac{1}{45}\begin{pmatrix} 4*14^2 + 8*10^2 + \\ 12*6^2 + 16*2^2 + 0 \end{pmatrix} \approx \frac{2100}{45} \qquad (4)$$

hence $k_l \approx \sqrt{\frac{2100}{45}} \approx 6.8\ m$

The symbols used in eq. (4) mean:

M – the half mass of the plane body [45 000 kg =45 t],

m_i – its digitalized part at the distance of r_i

r_i – the arm from the middle of plane to m_i

k – the resultant arm of inertia [m].

The similar expressions for the exponential distribution of the wing masses bring the following results

$$k_e^2 = \frac{1}{M}\sum_{i=14}^{0} m_i e^{-r_i} = \frac{1}{45}\begin{pmatrix} 4e^{-14} + 8e^{-12} + \\ 16e^{-2} + \cdots \end{pmatrix} \approx 0.53 \qquad (5)$$

hence $k_e \approx 0.73\ m$.

Finally, we have chosen the square mean between (4)-(5) k_l=6.8 m, k_e=0.73 m; hence, square mean $k \approx 2\ m$.

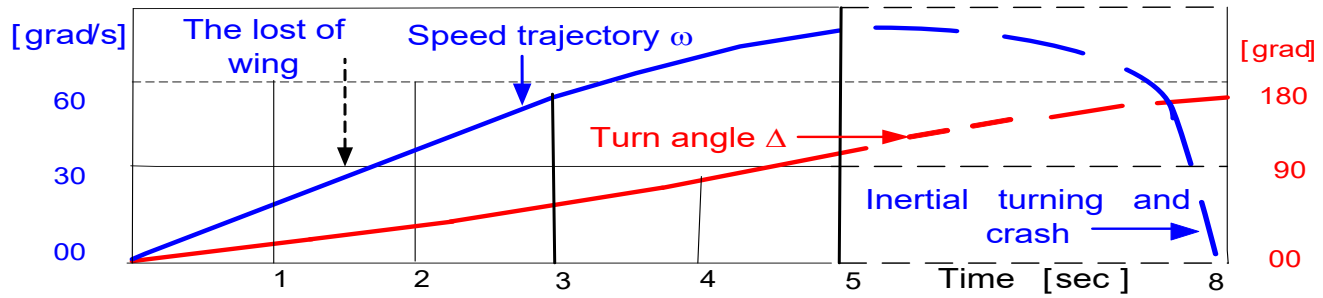


Figure 2: The simulated axial speed ω and the turn angle Δ of the Tu-154 plane after its collision with the tree vs. time

The angular frequency ω and the acceleration α of any rotation body is given by the following equations [5], [6]

$$\omega = Frt/Mk^2 \qquad (6)$$

ω – the angular speed of a body [degrees/s], t – the running time [s].

5. The Axial Turning

The force F, eq. (6), is the result of unbalance caused by the loss of a wing part and - next - the loss of total wing. We will take its mean as 10% of the gravitation force, $F=0.1*M$. Hence, for the shortened wing of r=13 m and the square of an inertial moment k^2 = 4 m^2, we obtain the approximated equation for the orbital speed ω in the first seconds, see (6)

$$\omega(t) = \left(\frac{F}{M}\right) * \left(\frac{r}{k^2}\right) t = 0.1 * \left(\frac{13}{4}\right) t = 0.33t \qquad (8)$$

This means that in the first second the plane turns around by 0.33 of radian, i.e. ~20^0!

One can observe in Fig.1 that a birch and the fall down points are distant away by 525 m. The plane used to cover it ~8 s, Fig.2. So, the mean speed was ~65 m/s and the max value could easily reach 100 m/s.

The red line shows the plane angle Δ relative to the horizon. We can see it is near 180^0.

6. The Discussion

The Russia accuses the Polish pilots for the crash [4]. This has no grounds. These pilots were the most experienced officers. They landed safely in extremely conditions, also on this same airfield and same plane and have made 3 trials preceding the crash! It is not excluded that a reason of the tragedy was an artificial fog and some falsification of the radar system without the knowledge given to the tower staff. No planes used this radar. The Putin's plane used modern ILS system.

The crashes in forests bring usually the fatal toll. Only in Poland there were two big ones: in Kabaty near Warsaw and in Miroslaviec. Nobody survived! The extensive literature on Tu-154 crash is given in [7-19].

In Smolensk crash, the large masses of Tu-154 were located in the very centre (±2 m), while the strange force appeared much outside (~13 m). These caused the plane to turn backside in ~8 sec. and in next few seconds it crashed on the freeze woody ground without any control by the crew. As we have shown the speed obtained ~100 m/s due to the engines and the gravity.

Such a crash could be compared – within the energy domain (10^9 kgm) – with the point collision of 4 trucks at a traffic circle, each of 22 t and the speed of 180 km/h. Who can risk to stand inside such a circle?

Why the crashes in the air are so dangerous in comparison with the land crashes? This is because in the last case the cars have the support in the ground, while the planes are hanging in free space. If one engine is damaged, the pilot can use the power of the second one, but if the wing is damaged, especially its ending part, the pilot is helpless.

The second wing could not replace the first one, because its role is just contrary to the action of the first.

7. Conclusions

The math-physical model of the Smolensk catastrophe has been given. It takes into account the distribution of masses, the values of forces, their directions and the resulted movement of a plane in pseudo 3D space. This approach is based on the Newtonian theory. It explains while the plane has made an upside down and why it crashed on hundred parts over the area of ±300m. This analysis suggests also the pre-reasons of the crash: the dense fog (artificial?), the radar having no valid certifications (?) and the tower staff not informed about (?) These charges should be investigated in the further legal process.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author is obliged to phd K. Kosmowski and Mr B. Grochowina for their valuable discussions and help in preparation the paper.

References

[1] J. Pawelec, "The Smolensk Catastrophe in the Light of Physics Lows", WORLDS4, London , July 30/31, 2019.
[2] J. Pawelec, "Misguidance under the Polish Tu-154 Crash", International Journal of Computers and Information Technology, September, 2018.
[3] The Ministry of Home Office, Poland, *The Final Report of Tu-154 Crash,*

Warsaw, July 31, 2011.
[4] The Interstate Aviation Committee, *The Final Report of Tu-154 Crash,* Moscow, February 29, 2012.
[5] D. Halliday at al., *Fundamentals of Physics,* John Wiley and Sons 2000.
[6] M. Jeżewski, *Physics,* Gov. Science. Ed., Warsaw 1991.
[7] Polish Pilots Saw Crash Coming, BBC News, 16.04.10
[8] Smolensk Air Disaster (Polish AF Tu-154 Crash), 2011
[9] Landing System Incompatibility Caused the Crash, 2012
[10] The Polish Pilots Asked for Russian Navigator, 2013
[11] N. Ojewska, Poland Polarized on Plane Crash, 2013
[12] P.Witakowski, Conferences on Smoleńsk Crash, 2013
[13] RiaNovosti: Polish Pilot was advised not to land, 2014
[14] The Tower Didn't Speak Eng., Rec. not landing, 2015
[15] Black Box Recording Made Public, *TV-Novosti,* 2015
[16] Apelblat M., British Experts Call for Future Inv.,.2015
[17] Poland to Russia over Withholding Wreckage, 2015
[18] Polish President Duda on Smolensk Crash, 30.11.2015
[19] Poland Re-open Investigations on Smolensk, 2016

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 557-566 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Application of EARLYBREAK for Line Segment Hausdorff Distance for Face Recognition

Chau Dang-Nguyen[*,1,2], Tuan Do-Hong[1,2]

[1]*Faculty of Electrical and Electronics Engineering, Ho Chi Minh University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Ho Chi Minh City, 72555, Vietnam*

[2]*Vietnam National University Ho Chi Minh City, Ho Chi Minh City, 71308, Vietnam*

ARTICLE INFO

Article history:
Received: 16 June, 2020
Accepted: 27 July, 2020
Online: 19 August, 2020

Keywords:
Hausdorff distance
EARLYBREAK
Similarity measure
Runtime analysis

ABSTRACT

The Hausdorff distance (HD) is defined as MAX-MIN distance between two geometric objects for measuring the dissimilarity between two objects. Because MAX-MIN distance is sensitive with the outliers, in face recognition field, average Hausdorff distance is used for measuring the dissimilarity between two sets of features. The computational complexity of HD, and also average HD, is high. Various methods have been proposed in recent decades for reducing the computational complexity of HD computing. However, these methods could not be used for reducing the computational complexity of average HD. Line Hausdorff distance (LHD) is a face recognition method, which uses weighted average HD for measuring the distance between two line edge maps of face images. In this paper, the Least Trimmed Square Line Hausdorff Distance method, LTS-LHD, is proposed for face recognition. The LTS-LHD, which is a modification of the weighted average HD, is used for measuring the distance between two line edge maps. The state - of - art algorithm, the EARLYBREAK method, is used for reducing the computational complexity of the LTS-LHD. The experimental results show that the accuracy of proposed method and LHD method are equivalent while the runtime of proposed method is 68% lower than LHD method.

1 Introduction

The Hausdorff distance (HD) is an useful parameter for measuring the degree of resemblance between two point sets. The advantage of HD is that there is no requirement of point-to-point correspondence. In many science and engineering fields, HD has particular attention of researcher, such as face recognition [1–5], image matching [6–8], image segmentation for medical image [9–11], image retrieval [12, 13], shape matching [14–16].

The HD is defined as MAX-MIN distance between two point sets, measuring how far two point sets are from each other. The HD computing contains two loops, the outer loop for maximization and the inner loop for minimization. Due to the sensitivity of the MAX-MIN HD with the outlier, average Hausdorff distance or partial Hausdorff distance (PHD) is used instead MAX-MIN HD in image matching or face recognition applications. Huttenlocher et al. The PHD was first proposed [17] for comparing the similarity between two shapes. In [18], the robust Hausdorff distance for face recognition was proposed, which uses PHD for measuring the distance between two sets of feature of face gray images. However, PHD is just an effective distance when the pollution of noise points is low. In face recognition research, average Hausdorff distance is commonly used.

The modified Hausdorff distance (MHD) was first presented [19] for image matching, where the directed Hausdorff distance is the mean of all distance from points to other sets. The 'doubly' modified Hausdorff distance (M2HD), which is the improved of the MHD, was proposed for face recognition [20]. Various algorithms using average weighted HD, an extension of MHD, were proposed for face recognition with the difference in weighted function., i.e., spatially weighted Hausdorff distance (SWHD) [21], spatially weighted modified Hausdorff distance (SW2HD) [22], spatially eigen-weighted Hausdorff distances (SEWHD) [23], edge eigenface weighted Hausdorff distance (EEWHD) [24, 25].

The extension of M2HD, similarity measure based on Hausdorff distance (SMBHD), was proposed for face recognition [3]. Another version of MHD, called modified Hausdorff distance with normalized gradient (MHDNG), was proposed for face recognition [8]. A new modified Hausdorff distance (MMHD) was presented for face recognition [26], which used average weighted Hausdorff dis-

[*]Corresponding Author: Chau Dang-Nguyen, Faculty of Electrical and Electronics Engineering, Ho Chi Minh University of Technology (HCMUT), Ho Chi Minh City, Vietnam, (84-8)38.647.256 Ext: 5700 & Email: chaudn@hcmut.edu.vn

tance for measuring the dissimilarity between two sets of dominant points of edge maps of face images. Based on edge map of face image, a novel face feature representation was proposed [27], line segment edge map (LEM). The line segment Hausdorff distance (LHD), which is the average weighted of distances between two line segments, was proposed for face recognition. Based on LHD, the extension of LHD, called spatial weighted line segment Hausdorff distance (SWLHD), was presented for face recognition [28] . In our previous work [1], we proposed a modification of LHD for face recognition, called Robust Line Hausdorff distance RLHD.

Supporting P and Q are the number of points, or elements, in two sets. For computing the directed distance of average HD, the distances from each point in the first set to all points in the second set have to be calculated to find the minimum value, which is the distance from the point in first set to its nearest neighbor in second set. The directed distance of average HD is the mean of all distances from points in first set to their nearest neighbors in second set. The computational complexities of methods using average HD are always $O(PQ)$, the same as MAX-MIN HD. In recent decades, many methods have been proposed for reducing the computational complexity of the HD computing, which is known very high. The key of reducing the computational complexity is reducing the average number of the inner loop. All of these methods uses a temporary HD *cmax* for quickly finding the non-contributed points in the inner loops and outer loops for final HD computing. However, the these methods, which is proposed with the purpose of reducing the running time of MAX-MIN HD computing, can not be used for reducing computing complexity the average HD computing because the temporary HD *cmax* does not exist in average HD computing. Due to the high computing complexity of average HD, the face recognition methods, which use average HD for measuring the distance between two sets of feature, are restricted from face recognition applications.

Here we propose an extension of the method in [1], called Least trimmed square Line Hausdorff distance (LTS-LHD) for face recognition. The LTS-LHD, an extension of weighted average HD, is used for measuring the dissimilarity between two line edge maps (LEMs). The LTS-LHD is the average of largest distances instead of all distances as average HD. The experimental results shows that the face recognition accuracy of the proposed method, LTS-LHD, is equivalent to the accuracy of LHD method, which used average HD for measuring distance between two LEMs, with a suitable parameter. Moreover, with LTS-LHD, the temporary HD *cmax* exist, and the methods, which are proposed with the purpose reducing the computational complexity of MAX-MIN HD computing, could be used for reducing the runtime of the proposed method. The EARLYBREAK [29], which is known as the state-of-art algorithm for reducing the computational complexity of HD computing, is used for reducing the computational complexity of proposed method. The runtime of proposed method is 68% lower than the LHD method.

The rest of the paper is structured as follows. The brief review of methods, which were proposed for reducing the computational complexity of the HD, is presented in Section 2. Section 3 presents the proposed method for face recognition, which uses LTS-LHD for measuring the distance between two LEMs. Section 3 also presents how to apply EARLYBREAK method for reducing the computational complexity of proposed method. Section 4 evaluates the performance of the proposed method and the performance of proposed method is also compared with the performance of LHD and RLHD method. Finally, conclusion is presented in Section 5.

2 Related works

Given two nonempty point sets $M = \{m_1, m_2, ..., m_p\}$ and $T = \{t_1, t_2, ..., t_q\}$, the directed Hausdorff distance $h(M,T)$ between M and T is the maximum distance of a point $m \in M$ to its nearest neighbor $t \in T$. The directed Hausdorff distance from M to T as a mathematical formula is

$$h(M,T) = \max_{m \in M} \min_{t \in T} \|m - t\| = \max_{m \in M} mindist(m,T) \quad (1)$$

where $\|.,.\|$ is any norm distance metric, e.g. the Euclidean distance. Note that, in general case, $h(M,T) \neq h(T,M)$ and thus directed Hausdorff distance is not symmetric. The Hausdorff distance between M and T is defined as the maximum of both directed Hausdorff distance and thus it is symmetric. The Hausdorff distance $H(M,T)$ is defined as

$$H(M,T) = \max(h(M,T), h(T,M)) \quad (2)$$

With two point sets M and T, if the Hausdorff distance between M and T is small, two point sets are partially matched, and if the Hausdorff distance is zero, two point sets are exactly matched.

Computing HD is challenging because its characteristics contain both maximization and minimization. Many efficient algorithms, in recent decades, have been proposed for reducing the computational complexity of the HD. We refer reader to the survey [17, 30] for general overview of this field. The efficient computing HD algorithms can be generally divided into two categories, approximate HD and exact HD. In the first category, which is approximation of HD, the algorithms try to efficiently find an approximation of the Hausdorff distance. These algorithms have been widely used in runtime-critical applications. On the other hand, the algorithms of the second category aim to efficiently compute the exact HD for point sets or special types of point sets like polygonal models or special curves and surfaces. Depending on data type of two sets, the HD algorithms can also be classified as polygonal models, curves and mesh surfaces, point sets.

With data type is polygonal models, a linear time algorithm for computing HD between two non-intersecting convex polygons was presented in [31]. The algorithm has a computational complexity of $O(m+n)$, where m and n are the vertex counts. In [32], an algorithm for computing the precise HD between two polygonal meshes with the complexity of $O\left(n^4 \log n\right)$ was presented. Due to the high computational complexity of exact HD calculation, approximate HD methods have been proposed. In [14], a method, that has the complexity of $O((m+n)\log(m+n))$, used Voronoi diagram to approximate the HD between simple polygons was presented. Another method for approximating the HD between complicated polygonal models was presented in [33]. By using Voronoi subdivision combined with cross-culling, non-contributing polygon pairs for HD are discarded. The method is very fast in practice and can reach interactive speed.

Many efficient algorithms were also proposed for calculating HD between mesh surfaces or curves. An efficient algorithm for calculating HD between mesh surfaces was presented in [34]. This algorithm is built on the specific characteristics of mesh surfaces, where the surface consists of triangles. To avoid sampling all points in the compared surfaces, the algorithm samples in the regions where the maximum distance is expected. The method for calculating the HD between points and freeform curves was presented in [35]. In [36], the improved method of [35] was presented for computing HD between two B-spline curves. For approximating the HD between curves, in [37], an algorithm was proposed, that converts the problem of computing HD between curves to the problem of computing distance between grids. In [38], an algorithm to compute the approximate HD between curves was presented. By approximating curves with polylines, the problem of computing HD between curves is converted to the problem of computing HD distance between line segments.

However, above methods are not general because the algorithms are based on specific characteristic of data types. Some general methods were proposed for point sets. An algorithm for finding the aggregate nearest neighbor (ANN) in database was proposed [39], that uses the R-Tree for optimizing the searching for ANN. The extension of [39], incremental Hausdorff distance (IHD) was proposed in [40] for efficiently calculating HD between two point sets. The algorithm uses two R-Trees for the same time, each for one point set, to avoid the interaction of all points in both sets. The aggregate nearest neighbor is simultaneously determined in both directions. However, complex structure of above algorithms makes the computation cost increase and the R-Tree is not suitable for general point sets.

In [29], a fast and efficient algorithm for computing exact Hausdorff distance between two point sets, which is known as a state-of-art algorithm, was proposed. The algorithm has two loops, with the outer loop for maximization and the inner loop for minimization. The inner loop can break as soon as the distance is found that is below the temporary HD (called *cmax*) because the rest iterations of inner loop do not make the value of *cmax* change, and the outer loop continues with the next point. Moreover, for improving performance, random sampling is also presented in this algorithm to avoid similar distances in successive iterations. Based on EARLYBREAK [29], an efficient algorithm, namely local start search (LSS) or Z-order Hausdorff distance (ZHD), for computing exact HD between two arbitrary point sets was presented [41]. The LSS method uses Morton curve for ordering points. The main idea of the LSS algorithm is that if the break occurs in current loop at the point x, it is quite possible that the break will occur at the position near x in the next loop. In the LSS algorithm, the variable *preindex* is used for preserving the location of break occurrence. In the next outer loop, the inner loop starts from *preindex* and scan its neighbor for finding the distance below the value *cmax*. In [42], an efficient framework, which contains two sub-algorithms Non-overlap Hausdorff Distance (NOHD) and Overlap Hausdorff Distance (OHD), for computing the HD between two general 3D point sets was proposed. For 3D point sets, [43] presented diffusion search of efficient and accurate HD computation between 3D models. This proposed method contains two algorithms for two types of 3D model, the ZHD for sparse point sets and the OHD for dense point sets.

In this study, we proposed the Least Trimmed Square Line Hausdorff Distance (LTS-LHD) for measuring the dissimilarity between two line edge maps (LEMs), which are the sets of line segments. The Hausdorff distance between two point set is based on the spatial locality of two point sets. Thus, the structure R-Tree of ANN and IHD or Z-order of LSS are just suitable for ordering the point sets [41]. However, the Hausdorff distance between two sets of line segments is based on both the spatial locality and the direction of line segment. Therefore, the structure R-Tree of ANN and IHD or Z-order of LSS are not suitable for the set of line segments as LEM, and thereby, the EARLYBREAK is used for reducing the complexity of computing the LTS-LHD.

3 Proposed method

3.1 LTS-LHD for face recognition

The original HD, the MAX-MIN distance, uses the distance of most mismatched points for measuring the distance between two sets. When the set is attacked with noise points, the original HD can not be used. The PHD was proposed for solving this problem by sorting the distance and taking the K ranked maximum value. However, PHD is an effective distance when the pollution of noise points is low. The MHD was also proposed for solving the problem of sensitivity of original HD with noise by taking the mean distance. As in [44], the LTS-HD was proposed for combining the advantage of the PHD and the MHD. The definition of directed LTS-HD from M to T is as follow

$$h_{LTS-HD}(M,T) = \frac{1}{K}\sum_{i=1}^{K}\left(\min_{t\in T}\|m-t\|\right)_{(i)} \tag{3}$$

where the $(\min\|m-t\|)_{(i)}$ represents the ith distance in the sorted sequence $(\min\|m-t\|)_{m\in M}$. The LTS-HD takes the mean of K minimum distances for measuring the distance between two sets.

In this paper, a new HD, called Least Trimmed Square Line Hausdorff Distance (LTS-LHD), for face recognition is proposed. Supporting $M^l = \left\{m_1^l, m_2^l, .., m_P^l\right\}$ and $T^l = \left\{t_1^l, t_2^l, .., t_Q^l\right\}$ are the LEMs of model and test images respectively; m^l and t^l are the line segments in the LEMs; P and Q are the number of line segments in model and test LEMs, respectively. The directed distance of the LTS-LHD from LEM M^l to LEM T^l, $h_{pLTS-LHD}\left(M^l, T^l\right)$, is defined as

$$h_{pLTS-LHD}\left(M^l,T^l\right) = \frac{1}{\sum_{i=K^{th}}^{P} l_{m_i^l}}\sum_{i=K^{th}}^{P} l_{m_i^l}\left(mindist\left(m_i^l,T^l\right)\right)_{(i)} \tag{4}$$

where $\left(mindist\left(m_i^l,T^l\right)\right)_{(i)}$ presents the ith value in the sorted sequence in ascending order $\left(mindist\left(m^l,T^l\right)\right)_{(m^l\in M^l)}$; and $K^{th} = f*P$, where f is a given fraction. The $mindist\left(m_i^l,T^l\right)$ denotes the distance from the line segment m^l to its nearest neighbor in T^l, and it is defined as below

$$mindist\left(m^l,T^l\right) = \min_{t^l\in T^l} d\left(m^l,t^l\right) \tag{5}$$

where $d\left(m^l, t^t\right)$ is the distance between two line segments m^l and t^l. Here, the directed distance of the LTS-LHD is used for measuring how far from LEM M^l to LEM T^l. So, different from the LTS-HD as in (3), where the directed distance is the average of smallest distances, the distance as in (4) is the weighted average of largest distances, which are greater or equal to the Kth ranked value, from line segment m^l to its nearest neighbor $mindist\left(m^l, T^l\right)$. However, the directed distance of the LTS-LHD as in (4) still has weakness. Supporting m_1^l and m_2^l are two line segments in M^l, d_1 and d_2 are two distances from these segments to their neighbors. Assuming d_1 is greater than the Kth ranked value and d_2 is less than the Kth ranked value of the sorted sequence $\left(mindist\left(m_i^l, T^l\right)\right)_{\left(m^l \in M^l\right)}$. As in (4), d_1 is used for the computing of $h_{pLTS-LHD}\left(M^l, T^l\right)$. However, it is possible that $l_{m_1^l}.d_1 \ll l_{m_2^l}.d_2$, because line segment m_2^l is much longer than line segment m_1^l. The miss match of the long line segment is more serious than the short line segment. So, the line segment m_2^l is much more important than the line segment m_1^l for computing the directed distance of LTS-LHD. Here, we modify the (4) for proposing the directed distance of the LTS-LHD as follow

$$h_{LTS-LHD}\left(M^l, T^l\right) = \frac{1}{\sum_{i=K^{th}}^{P} l_{m_i^l}} \sum_{i=K^{th}}^{P} \left(l_{m_i^l}.mindist\left(m_i^l, T^l\right)\right)_{(i)} \quad (6)$$

where $\left(l_{m_i^l}.mindist\left(m_i^l, T^l\right)\right)_{(i)}$ presents the ith value in the sorted sequence in ascending order $\left(l_{m^l}.mindist\left(m^l, T^l\right)\right)_{\left(m^l \in M^l\right)}$. The directed distance of the LTS-LHD is the weighted average of largest values of the product between length of line segment and the distance from that line segment to its nearest neighbor $l_{m^l}.mindist\left(m^l, T^l\right)$.

In general, $h_{LTS-LHD}\left(M^l, T^l\right) \neq h_{LTS-LHD}\left(T^l, M^l\right)$. Thus, the primary LTS-LHD is defined as

$$H_{pLTS-LHD}\left(M^l, T^l\right) = \max\left(h_{LTS-LHD}\left(M^l, T^l\right), h_{LTS-LHD}\left(T^l, M^l\right)\right) \quad (7)$$

In our previous work [1], a new data structure of LEM was proposed. Due to the angle between line segments and the horizontal axis, the line segments in LEM are grouped into N groups and $180/N$ degrees for each group. For example, in this paper, we use $N = 18$. An example of new data structure of LEM is shown in Fig 1.

The distance between two line segments $d\left(m^l, t^l\right)$ is defined as

$$d\left(m^l, t^l\right) = \begin{cases} \left\|m^l - t^l\right\| & \text{where } g_{t^l} - k \leq g_{m^l} \leq g_{t^l} + k \\ V & \text{otherwise} \end{cases} \quad (8)$$

where V is the value that is larger than the largest possible value of distance between two line segments; $\left\|m^l - t^l\right\|$ is the distance between two line segments and defined as

$$\left\|m^l - t^l\right\| = \sqrt{d_{pa}^2\left(m_i^l, t_j^l\right) + d_{pe}^2\left(m_i^l, t_j^l\right) + d_\theta^2\left(m_i^l, t_j^l\right)} \quad (9)$$

where d_{pa} is the parallel distance, which is the vertical distance between two lines; d_{pe} is the perpendicular distance, which is minimum displacement to align either the left end points or the right end points of two lines; and $d_\theta\left(m_i^l, t_j^l\right) = \theta^2\left(m_i^l, t_j^l\right)/W$ is the orientation distance. $\theta^2\left(m_i^l, t_j^l\right)$ is the smallest intersection angle between two lines and W is a weight, which could be determined by a training process.

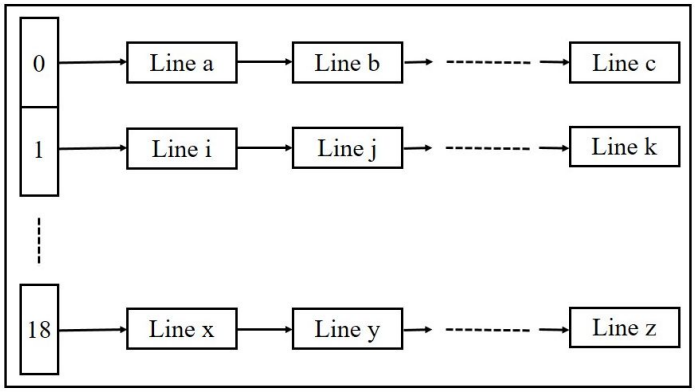


Figure 1: A novel data structure for LEM

It is possible that the line segment m^l could take a line t^l, that intersection angle between m^l and t^l is large, as it nearest neighbor. However, the line segment reflects the structure of human face, two corresponding line segments can not have large angle variation. For alleviating the undesire mismatch, the line segment finds its nearest neighbor if the group index (g_{m^l} and g_{t^l}) of two line segments are slightly different as in (8). On the other hand, the distance between two line segments takes a large value V.

The number of corresponding line pairs between the model and the test LEM can be used as other measure similarity. Because test and model images are aligned and scaled at the same size by preprocessing before matching, if the line segment m^l finds that line segment t^l is the nearest neighbor in T^l and line segment t^l locates in the position neighborhood N_p of m^l, such line segment m^l could be named as *high confident line*. *A high confident ratio*, as in [27], of an image could be defined as the ratio between number of high confident line segments (N_{hc}) and number of total line segments in the LEM of face image (N_{total}) as follow

$$R = \frac{N_{hc}}{N_{total}} \quad (10)$$

Hence, the number disparity between two LEMs has mathematical formula as follow

$$D = 1 - \frac{R_M + R_T}{2} \quad (11)$$

The complete version of Hausdorff distance between two LEMs is defined as

$$H_{LTS-LHD} = \sqrt{H_{pLTS-LHD}^2 + W_n^2 D^2} \quad (12)$$

where W_n is a weight that be determined by a training process.

3.2 EARLYBREAK for LTS-LHD

The directed distance of the LTS-LHD as in (6) is the weighted average of $\left(P - K^{th}\right)$ largest values of the product between length of

line segment and distance to its nearest neighbor $l_{m^l}.mindist\left(m^l, T^l\right)$. For computing the directed distance of the LTS-LHD, the distances from line segment m^l to all of line segments in T^l must be calculated for finding the minimum value, which is the distance from line segment m^l to its nearest neighbor $mindist\left(m^l, T^l\right)$. This process could be named as the inner loop. The inner loop must be performed with all line segments $m^l \in M^l$ and we call this is the outer loop. Assuming that $\left(P - K^{th}\right)$ temporary largest values are found and the minimum of these values is assigned to $cmax$. If a line segment m^l in the outer loop find out a line segment t^l in inner loop that makes the product between length of line segment m^l and distance $d\left(m^l, t^l\right)$ between two line segments is below the value of $cmax$, such line segment m^l is non-contributed line segment for the computing of the directed distance LTS-LHD. So, the computing distance from line segment m^l to the remaining line segments t^l is not necessary. Therefore, the computing of directed distance LTS-LHD could break and continuing with the next line segment in the outer loop as soon as a non-contributed line segment be found. Thus, the number of iterations of the inner loop is reduced. The lower of average number of inner iteration is, the lower of computational complexity of the LTS-LHD computing is. Here, we proposed a method using EARLYBREAK for reducing the computational complexity of the LTS-LHD by reducing the average number of inner iterations.

The Algorithm 1 describes our proposed method. Line 5 and line 13 are the outer loop and the inner loop, respectively. The function $DIST\,(.,.)$ in the Algorithm 1 is used for calculating the distance between two line segments as in (9).

The main steps of the Algorithm 1 are summarized as follows:

- A matrix h is created for saving the length of line segment m^l and the value of the product between length of line segment m^l and distance to its nearest neighbor.
- Adding line segments t^l having group index g_t into $list$.
- If there is at least one line segment in $list$, the inner loop will be executed. For each line segment $m^l \in M^l$, initializing distance to nearest neighbor $cmin = \infty$.
  - For each line segment in $list$, the distance from m^l to t^l is calculated. If a distance makes the product between it and length of line segment below the value of $cmax$, the algorithm will **break** and continue with the next line segment in the outer loop. In the other hand, this distance is used for updating the value of $cmin$.
  - The product between $cmin$ and length of line segment will be used for updating the matrix h.
- On the other hand, if there is no line segment in $list$, the matrix h will be updated with length of line segment m^l and the large value V, which is the distance from line segment m^l to its nearest neighbor.
- The matrix h will be sorted in ascending order for each interaction of outer loop according to the values of the first row of matrix h.

In the Algorithm 1, during first K_M iterations of the outer loop, the value of $cmax$, which is the minimum value of matrix h, is 0. The condition as line 15 in the Algorithm 1 does not appear, thus, the early break does not occur in first K_M iterations of the outer loop. The first K_M iterations of the outer loop, the value of K_M elements in the matrix h are updated. In the next iterations of the outer loop, the early break occurs if the value of product between the length of line and its distance to the nearest neighbor is below $cmax$.

Algorithm 1 : EARLYBREAKING for LTS-LHD

1: **Inputs**: Edge map M^l and T^l, fraction f
2: **Outputs**: Directed Hausdorff distance $h_{LTS-LHD}\left(M^l, T^l\right)$
3: $K_M = (1 - f) * P$
4: $h = zeros\,(2, K_M)$
5: **for** each line segment m^l in edge map M^l **do**
6: $cmax = h\,(1, 1)$
7: Get the group index g_m of line segment m^l
8: $cmin = \infty$
9: **for** $g_t = g_m - k : g_m + k$ **do**
10: Insert line segments t^l has group index g_t into $list$
11: **end for**
12: **if** $list$ is not empty **then**
13: **for** each line segment t^l in $list$ **do**
14: $d = DIST\left(m^l, t^l\right)$
15: **if** $d * lM \leq cmax$ **then**
16: $cmin = 0$
17: **break**
18: **end if**
19: **if** $cmin > d$ **then**
20: $cmin = d$
21: **end if**
22: **end for**
23: **if** $(cmin * lM) > cmax$ **then**
24: $h\,(1, 1) = cmin * lM$
25: $h\,(2, 1) = lM$
26: **end if**
27: **else**
28: $h\,(1, 1) = V * lM$
29: $h\,(2, 1) = lM$
30: **end if**
31: sort h in ascending order of the value of first row
32: **end for**
33: $h_{LTS-LHD}\left(M^l, T^l\right) = sum\,(h\,(1, :))\,/\,sum\,(h\,(2, :))$

By using the face database of Bern university [45] for training process, we obtain that: $W = 45$, $Np = 6$ and $Wn = 13$.

3.3 Analysis of computational complexity

Supporting P and Q are the number of line segments in LEM M^l and in LEM T^l, respectively. In the LHD method [27], each line segment $m^l \in M^l$ calculate the distance to all line segments in T^l for finding its nearest neighbor. The directed distance of the LHD is the weighted average of the distances from line segments $m^l \in M^l$ to their nearest neighbors in T^l. So, the complexity of computing directed distance of the LHD is $O\,(PQ)$.

The directed distance LTS-LHD computing as in (6) has computational complexity $O\,(PQ)$ in the worst case. In the worst case,

all of line segments of two LEMs are located in the same group, and each line segment has to calculate distance to all line segments in other LEMs for finding nearest neighbor. On the other hand, with the best case, if all line segments of each LEM are located in one group and two groups of two LEMs are much different, the computational complexity of directed Hausdorff distance is $O(P)$. If the line segments of LEM T^l are equally divided into 18 groups, the computational complexity of directed Hausdorff distance would be $O(PQ(2k+1)/N)$, where k is the difference of group index as defined in (8). The computational complexity of the LTS-LHD is always better than the LHD because the lower bound of the LTS-LHD computational complexity is the LHD computational complexity.

The Algorithm 1, in which the EARLYBREAK is applied for the LTS-LHD, has the computational complexities $O(P)$ and $O(PQ)$ for the best case and the worst case, respectively. In general case, assuming that line segments of LEM are equally divided into groups, the Algorithm 1 has a computational complexity of $O((1-f)PQ(2k+1)/N + fPX)$, where X is used to denote the average number of iterations in the inner loop. The lower value of X is, the lower of computational complexity of the method is and vise versa. The question is, in general, how high of the value of X is? In the formal way, the value of X in general case could be found through the analysis of probability theory.

Considering picking a random line segment t^l in the inner loop of the Algorithm 1, the distance d measured between line segment t^l and line segment m^l in current outer loop is a random variable. The event that meeting the condition that d is over $cmax$ is denoted as e. The probability of that event is that $P(e) = q$. The event e means non-appearance of the break in the algorithm. Obviously, the event $\overline{e}$, that d is less than $cmax$, occur with probability $P(\overline{e}) = p = 1 - q$.

Assuming that the inner loop has been implemented for X times before the break occurs. This is equivalent to that $(X-1)$ distances from line segment m^l in the outer loop to the line segments $t^l_1, t^l_2, ..., t^l_{X-1}$, namely $d_1, d_2, ..., d_{X-1}$, are over $cmax$ and one distance $d_X \leq cmax$. The probability density function of X is given by

$$\begin{aligned} f(x) &= P(d_1 > cmax, ..., d_{x-1} > cmax, d_x \leq cmax) \\ &= q * q * ... * q * p \\ &= q^{x-1}p \end{aligned} \tag{13}$$

Fig. 2 shows the probability distribution $f(x)$. The expectation of average number iterations of the inner loop X is equivalent to the expected value of $f(x)$

$$E[X] = \sum_{x=1}^{\infty} x f(x) = \sum_{x=1}^{\infty} x q^{x-1} p \tag{14}$$

The Eq. (14) could be rewritten in the form of a polynomial as follows

$$E[X] = p + 2qp + 3q^2p + 4q^3p + ... \tag{15}$$

By multiplying both side of (15) with q and subtracting the resulting equation from (15) , a simpler formula of (15) is given as

$$E[X](1-q) = p\left(1 + q + q^2 + q^3 + ...\right) = p\frac{1}{1-q} \tag{16}$$

Then, by using $p = 1-q$, the expectation of number of iterations in the inner loop X could be found as

$$E[X] = \frac{1}{1-q} = \frac{1}{p} \tag{17}$$

Equation (17) means that the number of iterations in the inner loop until the break depends on the value of p, which is the probability that a distance d is less than $cmax$. The higher p means the lower number of tries before early break and vice versa. The value of p depends on the value of $cmax$. For a larger value of $cmax$, it is easier for picking a line segment making the distance d to the current line segment of the outer loop less than $cmax$. Fig. 3 illustrates the relation between $cmax$ and the probability p. Here, it is assumed the distance d is a random variable with normal distribution for illustration. The value of p does not depend on the size of set, but rather on the value of $cmax$ and the distribution of pairwise distance d.

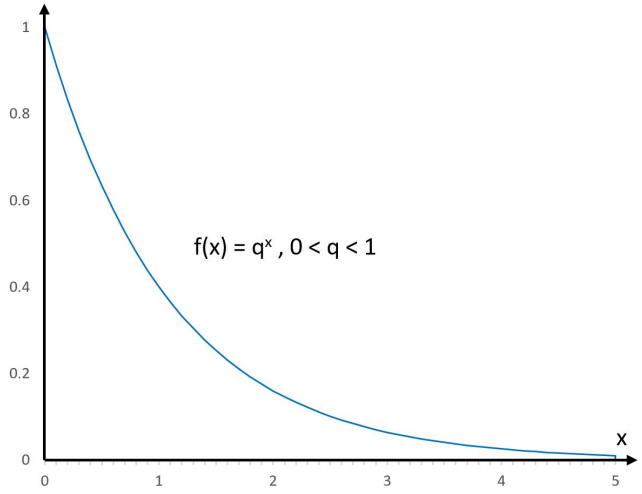


Figure 2: Probability density function $f(x)$

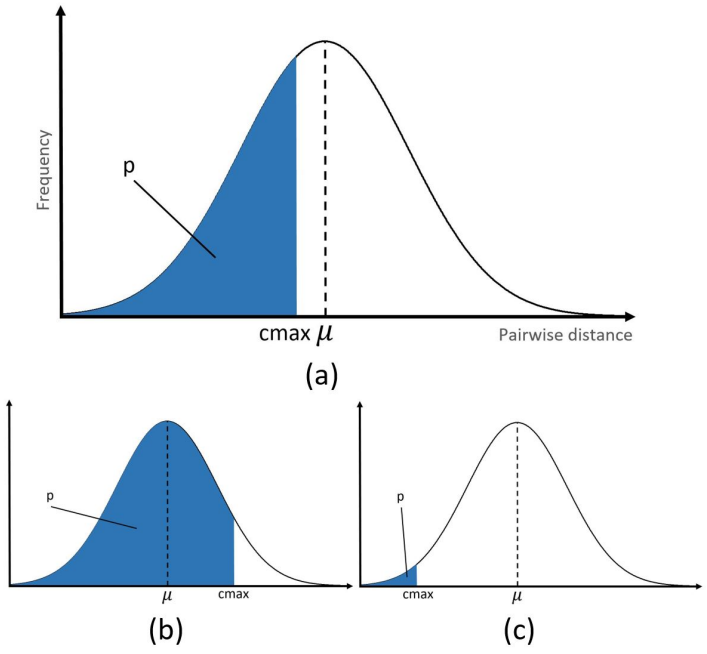


Figure 3: Normal distribution of pairwise distance is assumed. (a) Relation between $cmax$ and probability p. (b) Large value of $cmax$. (c) Small value of $cmax$.

4 Experiments

In this section, the performance of the proposed method, LTS-LHD, is evaluated for face recognition. It is done by measuring the distance from test image to all model images for finding the smallest distance. The recognition rate, which is the ratio of number of images correctly classified to the total number of images in the test set, is used for evaluating.

In this study, the face database from the University of Bern [45] and the AR face database from Purdue University [46] are used. Bern university face database contains frontal views of 30 people. Each person has 10 gray images with different head pose variations: two frontal pose images, two looking to the right images, two looking to the left images, two looking upward image and two looking downward images. The AR face database contains 2599 color face images of 100 people (50 men and 50 women), there are 26 images for each person and be divided into 2 sessions separated by two weeks interval. Each session has 13 images are the frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). However, one of frontal face image is corrupted (W-027-14.bmp) and only 99 pairs of face image are used for examining the performance of system for face recognition under normal conditions. In our experiment, a preprocessing before recognition process is used for locating the face. All image are normalized such that the two eyes were aligned roughly at the same position with a distance of 80 pixels. After that, all images are cropped with size 160×160 pixels. The experiments are conducted on the PC with 3GHz CPU and 4GHz RAM.

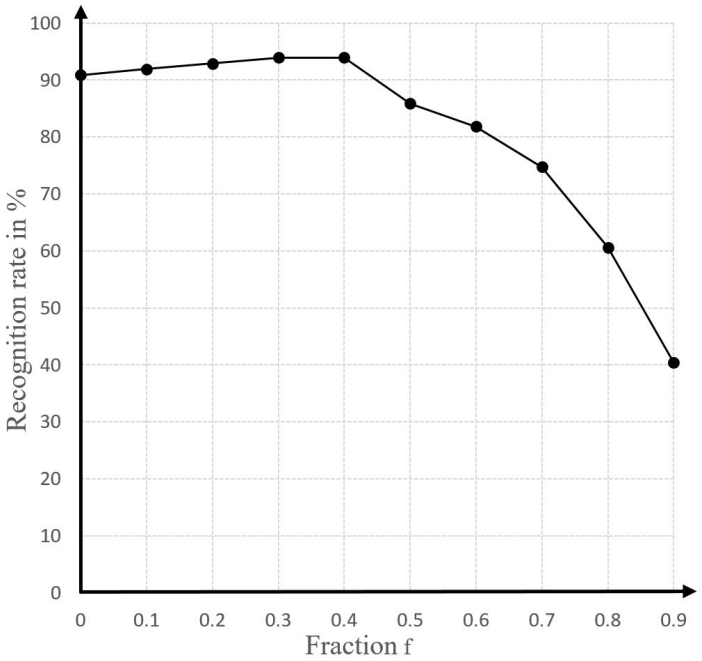


Figure 4: Influence of fraction f on recognition rate for AR database

4.1 Influence of fraction f on the performance

The recognition rate of the proposed method is expected to be low for both low and high values of fraction f. With high value of fraction f, a low number of line segments, which have largest values of the product between their length and distance to their neighbor, is used for computing the directed distance of the LTS-LHD. However, the outliers are commonly the line segments have largest values of that product. The high value of f means most of line segments used for computing the directed distance LTS-LHD are outliers, as in (6). On the other hand, when the value of fraction f is too low, a high number of line segments is used for computing the directed distance of the LTS-LHD. This means high number of line segments in similar regions of faces is used for the calculation of $h_{LTS-LHD}$. And thus, the contribution of line segments that discriminate the faces becomes low.

Fig. 5 shows the recognition rate of the proposed method for the AR database with various values of fraction f. It must be noticed that the value f of 1 is meaningless, because the value of directed distance of the LTS-LHD $h_{LTS-LHD}$ is zero for all face image pairs. The proposed method has highest recognition rate at the fraction f values 0.3 and 0.4.

The computational complexity of the proposed method is based on the average number of iterations in the inner loop, and thus is based on the value of *cmax*. The higher value of *cmax* is, the lower of runtime is and vice versa. The low value of fraction f means the large number of line segments is used for computing $h_{LTS-LHD}$, and the value of *cmax* becomes low. On the other hand, the high value of fraction f means the high value of *cmax*, and thus the computational complexity of the method is low. So, the value of fraction f is chosen at 0.4.

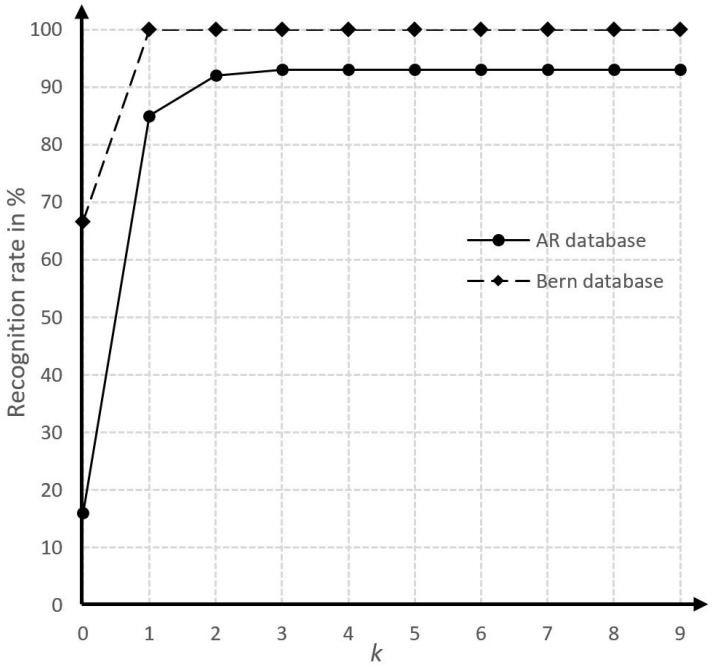


Figure 5: Influence of parameter k on recognition rate of proposed method

4.2 Influence of parameter k on the performance

The recognition rate of the proposed method is expected to be low for low value of parameter k and vice versa. As in algorithm 1, the value of k determines the number of line segments in the inner loop. The low value of k has strong effect on system performance. Supporting m^l is the line segment in current outer loop. The low value

of k means m^l has to find its nearest neighbor in a few number of line segments in *list*, as in algorithm 1. And thus it is possible that *list* does not contain the corresponding line segment of m^l, and m^l could take another non-corresponding line segment as its nearest neighbor. On the other hand, the higher value of k is not necessary because too much line segments, which are too far from m^l, are added into *list*. Such non-contributed line segments make the number of iterations of inner loop increase and thus the runtime of method increases. Fig. 5 shows the recognition rates of the proposed method with various values of k for both the Bern university database and the AR database. The recognition rate does not change for the value of k higher than 2 for the Bern university database and 3 for the AR database. So, the chosen value of k is 3.

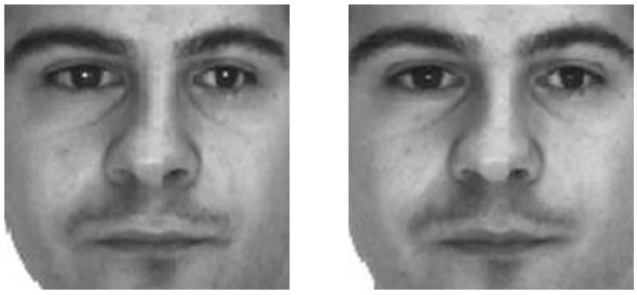
In the rest of this section, the recognition rate of the proposed method is compared with the LHD method in [27] and the RLHD method in [1], using average HD for measuring the dissimilarity between LEMs.

4.3 Face recognition under normal conditions

The frontal face images in normal conditions in the Bern university database and the AR database are used for evaluating the performance of the proposed method. Each person has two images, one for test set and one for the model set. The example of images in this experiment is shown in Fig. 6. The recognition rates of different methods are given in Table. 1. The recognition rates of all methods with Bern university database are higher than those with the AR database. The reason is the different between two images of each person in AR database is larger than Bern database. The illumination of model image and test image in AR database are also different. The recognition rate of the proposed method is equal to the RLHD method.



(a) Bern University Database



(b) AR Database

Figure 6: Example of two pairs frontal face image in normal lighting condition in the face database of Bern University and the AR face database

Table 1: Face recognition result

	Recognition rate	
Method	Bern database	AR database
LHD	100%	93%
RLHD	100%	94%
Proposed method	**100%**	**94%**

The matching time for the Bern university database of different methods are given in Table. 2. The proposed method has the runtime 68% lower than the LHD method and 17.5% lower than RLHD method. The improvement in runtime is achieved by using EARLYBREAK for reducing the average number of iterations in the inner loop.

Table 2: Matching time of different methods

Method	LHD	RLHD	Proposed method
Matching time (second)	337	131	108

4.4 Face recognition under varying lighting conditions and poses

The performance of the proposed method is also compared with the ones of the LHD and the RLHD methods for face recognition in the non-ideal conditions, e.g. face image with different poses or different lighting conditions. The AR database is used for evaluating the performance of different methods with varying lighting conditions of face image. Frontal face images of 100 people are used as model set. The face images with a light source on left side of face, with a light source on right side of face and with light sources on both sides of face are divided into three test sets with 100 images for each set. The recognition rates of different methods are given in Table. 3. The non-ideal lighting conditions make the recognition rate of all methods approximate 10% decrease. The face recognition accuracy of proposed method is 1%, on average, higher than the LHD and the RLHD methods. The interesting point of the experiment is that all three methods, in the condition of left light on, give the same recognition rates as the normal lighting condition in Table. 1 while the recognition rate in right light on condition is 6% - 9% lower than the ideal lighting condition. This could be due to the fact that the illumination of the right light is stronger than the left light. When both light on, the recognition rates of all methods are 12% lower than recognition rates in ideal lighting condition. The over-illumination has strong effect on recognition rates of all methods.

Table 3: Face recognition with varying lighting conditions

Lighting conditions	LHD	RLHD	Proposed method
Left light on	93%	94%	**94%**
Right light on	87%	85%	**88%**
Both light on	71%	72%	**72%**
Average	83.67%	83.67%	**84.67%**

The Bern university database is used for evaluating the performance of different methods with different poses of face image. The model set contains 30 frontal face images of 30 people. The test

set contains images of 30 people with different poses, e.g. looking to the left and right, looking up and down, for each person. The recognition rates of different methods are summarized in Table. 4. The pose variations have strong effect on recognition rates of all methods, where the recognition rates decrease 40% - 50% in comparing with the results in Table. 1. This could be explained that there are portions of face missing in comparing with frontal face. The recognition rate of the proposed method is lower than the RLHD method with the looking right image and higher than the RLHD methood in other conditions. On average, the proposed method has the recognition rate 2% higher than the RLHD method and 3% higher than the LHD method.

Table 4: Face recognition with varying poses

	LHD	RLHD	Proposed method
Looking left	46.67%	53.33%	**63.33%**
Looking right	53.33%	46.67%	**46.67%**
Looking up	66.67%	63.33%	**66.67%**
Looking right	60%	70%	**63.33%**
Average rate	56.67%	58.33%	**60%**

5 Conclusion

The Hausdorff distance, which is used for measuring the degree of resemblance between two geometric objects, has been widely used in various science and engineering fields. The computational of HD computing is high because the computing contain both maximization and minimization. Many methods have been proposed in recent decades for reducing the computational complexity of MAX-MIN HD computing. However, the proposed methods for reducing the computational complexity of MAX-MIN HD computing can not be used for reducing the computational complexity of average HD computing. In face recognition, average HD is widely used for measuring the distance between two sets of features, instead of MAX-MIN HD, which is known as a sensitive measure with noise. The computational complexity of average HD computing as high as the MAX-MIN HD computing. The high computational cost restricts face methods using he average HD from real time applications.

The LHD and he RLHD use average HDfor measuring the dissimilarity between two LEMs. In this paper, a modification of RLHD, called LTS-LHD was proposed for face recognition. The LTS-LHD uses only K_M line segments, not all of line segments as in RLHD, for calculating the directed distance. With a suitable parameter K_M, or suitable fraction f, the proposed method, LTS-LHD, has the performance slightly higher than the RLHD method, which is the average HD.

Moreover, in this paper, the EARLYBREAK is used for reducing the computational complexity of the proposed method. The early breaking can speed up the LTS-LHD by reducing the average number of iterations in the inner loop. The experimental results show that the runtime of proposed method is 68% lower than the LHD method and 17.5% lower than the RLHD method.

Acknowledgment We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.

References

[1] C. Dang-Nguyen, T. Do-Hong, "Robust line hausdorff distance for face recognition, in 2019 International Symposium on Electrical and Electronics Engineering (ISEE), HoChiMinh, Vietnam, 2019. https://doi.org/10.1109/ISEE2.2019.8921218

[2] J. Wang and Y. Tan, "Hausdorff distance with k-nearest neighbors, in Proceedings of the Third International Conference on Advances in Swarm Intelligence - Volume Part II, ser. ICSI12. Berlin, Germany, 2012. https://doi.org/10.1007/978-3-642-31020-1

[3] Y. Hu, Z. Wang, "A similarity measure based on hausdorff distance for human face recognition, in 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 2006. https://doi.org/10.1109/ICPR.2006.174

[4] S. Chen and B. C. Lovell, "Feature space hausdorff distance for face recognition, in Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR'10), Istanbul, Turkey, 2010. https://doi.org/10.1109/ICPR.2010.362

[5] J. Paumard, "Robust comparison of binary images," Pattern Recognition Letters, **18**(10), 1057 1063, 1997. https://doi.org/10.1016/S0167-8655(97)80002-5

[6] H. Zhu, T. Zhang, L. Yan, and L. Deng, "Robust and fast hausdorff distance for image matching," Optical Engineering, **51**(1), 2012. https://doi.org/10.1117/1.OE.51.1.017203

[7] X. Li, Y. Jia, F. Wang, and Y. Chen, "Image matching algorithm based on an improved hausdorff distance," in 2nd International Symposium on Computer, Communication, Control and Automation (3CA 2013) , Singapore, 2013. https://doi.org/10.2991/3ca-13.2013.61

[8] C.H. Yang, S.H. Lai, and L.W. Chang, "Hybrid image matching combining hausdorff distance with normalized gradient matching," Pattern Recognition, **40**(4), 11731181, 2007. https://doi.org/10.1016/j.patcog.2006.09.014

[9] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," IEEE Transactions on Medical Imaging, **39**(2), 499513, 2020. https://doi.org/10.1109/TMI.2019.2930068

[10] H. Yu, F. Hu, Y. Pan, and X. Chen, "An efficient similarity-based level set model for medical image segmentation," Journal of Advanced Mechanical Design, Systems, and Manufacturing, **10**(8), 01000108, 2016. https://doi.org/10.1299/jamdsm.2016jamdsm0100

[11] F. Morain-Nicolier, S. Lebonvallet, E. Baudrier, S. Ruan, "Hausdorff distance based 3d quantification of brain tumor evolution from mri images," 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, 2007. https://doi.org/10.1109/IEMBS.2007.4353615

[12] M. Grana, M. A. Veganzones, "An endmember-based distance for content based hyperspectral image retrieval," Pattern Recognition, **45**(9), 3472-3489, 2012. https://doi.org/10.1016/j.patcog.2012.03.015

[13] Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai, "3-D object retrieval with hausdorff distance learning," IEEE Transactions on Industrial Electronics, **61**(4), 20882098, 2014. https://doi.org/10.1109/TIE.2013.2262760

[14] H. Alt, B. Behrends, J. Blomer, "Approximate matching of polygonal shapes," Annals of Mathematics and Artificial Intelligence, **13**, 251265, 1995. https://doi.org/10.1007/BF01530830

[15] L. B. Kara and T. F. Stahovich, "An image-based, trainable symbol recognizer for hand-drawn sketches," Computer Graphics, **29**(4), 501517, 2005. https://doi.org/10.1016/j.cag.2005.05.004

[16] M. Field, S. Valentine, J. Linsey, and T. Hammond, "Sketch recognition algorithms for comparing complex and unpredictable shapes," in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Three (IJCAI '11), Barcelona, Spain, 2011.

[17] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge, "Comparing images using the hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, **15**(9), 850863, 1993. https://doi.org/10.1109/34.232073

[18] V. E. P and N. Sudha, "Robust hausdorff distance measure for face recognition," Pattern Recognition, **40**(2), 431442, 2007. https://doi.org/10.1016/j.patcog.2006.04.019

[19] M. P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in Proceeding 12th International Conference on Pattern recognition, Jerusalem, Israel, 1994. https://doi.org/10.1109/ICPR.1994.576361

[20] B. Takcs, "Comparing face images using the modified hausdorff distance," Pattern Recognition, **31**(12), 18731881, 1998. https://doi.org/10.1016/S0031-3203(98)00076-4

[21] B. Guo, K. Lam, W. Siu, S. Yang, "Human face recognition using a spatially weighted hausdorff distance, in The 2001 IEEE International Symposium on Circuits and Systems (ISCAS 2001), Sydney, Australia, 2001. https://doi.org/10.1109/ISCAS.2001.921027

[22] K. Lin, B. Guo, K. Lam, W. Siu, "Human face recognition using a spatially weighted modified hausdorff distance," in Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. (ISIMP 2001), Hong Kong, China, 2001. https://doi.org/10.1109/ISIMP.2001.925437

[23] K.H. Lin, K.M. Lam, and W.C. Siu, "Spatially eigen-weighted hausdorff distances for human face recognition," Pattern Recognition, **36**(8), 1827 - 1834, 2003. https://doi.org/10.1016/S0031-3203(03)00011-6

[24] H. Tan and Y. Zhang, "Computing eigenface from edge images for face recognition based on hausdorff distance," in Fourth International Conference on Image and Graphics (ICIG 2007), Sichuan, China, 2007. https://doi.org/10.1109/ICIG.2007.70

[25] H. Tan, Y.-J. Zhang, W. Wang, G. Feng, H. Xiong, J. Zhang, and Y. Li, "Edge eigenface weighted hausdorff distance for face recognition," International Journal of Computational Intelligence Systems, **4**(6), 1422-1429, 2011. https://doi.org/10.1080/18756891.2011.9727894

[26] Y. Gao, "Efficiently comparing face images using a modified hausdorff distance," IEE Procceeding Vision, Image and Signal Processing, **150**(6), 346350, 2003. https://doi.org/10.1049/ip-vis:20030805

[27] Y. Gao and M. K. H. Leung, "Face recognition using line edge map," IEEE Trans. on Pattern Analysic and Machine Intelligence, **24**(6), 764779, 2002. https://doi.org/10.1109/TPAMI.2002.1008383

[28] C. Du, G. Su, X. Lin, "Face recognition using a modified line segment hausdorff distance," in Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693), Xi'an, China, 2003. https://doi.org/10.1109/ICMLC.2003.1260095

[29] A. A. Taha and A. Hanbury, "An efficient algorithm for calculating the exact hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, **37**(11), 21532163, 2015. https://doi.org/10.1109/TPAMI.2015.2408351

[30] H. Alt and L. J. Guibas, "Discrete geometric shapes: Matching, interpolation, and approximation, Handbook of computational geometry," Elsevier, 2000.

[31] M. J. Atallah, "A linear time algorithm for the hausdorff distance between convex polygons," Information Processing Letters, **17**(4), 207-209, 1983. https://doi.org/10.1016/0020-0190(83)90042-X

[32] M. Barton, I. Hanniel, G. Elber, and M.-S. Kim, "Precise hausdorff distance computation between polygonal meshes," Computer Aided Geometric Design, **27**(8), 580-591, 2010. https://doi.org/10.1016/j.cagd.2010.04.004

[33] M. Tang, M. Lee, Y. J. Kim, "Interactive hausdorff distance computation for general polygonal models," ACM Transactions on Graphics, **28**(3), 2009. https://doi.org/10.1145/1531326.1531380

[34] M. Guthe, P. Borodin, R. Klein, "Fast and accurate hausdorff distance calculation between meshes." Journal of WSCG, **13**(2), 41-48, 2005.

[35] H. Alt and L. Scharf, "Computing the hausdorff distance between curved objects." International Journal of Computational Geometry and Applications, **18**(4), 307320, 2008. https://doi.org/10.1142/S0218195908002647

[36] X.D. Chen, W. Ma, G. Xu, and J.-C. Paul, "Computing the hausdorff distance between two b-spline curves," Computer-Aided Design, **42**(12), 1197 1206, 2010. https://doi.org/10.1016/j.cad.2010.06.009

[37] E. Belogay, C. Cabrelli, U. Molter, R. Shonkwiler, "Calculating the hausdorff distance between curves." Information Processing Letters, **64**(1), pp. 1722, 1997. https://doi.org/10.1016/S0020-0190(97)00140-3

[38] Y.B. Bai, J.H. Yong, C.Y. Liu, X.M. Liu, and Y. Meng, "Polyline approach for approximating hausdorff distance between planar free-form curves," Computer-Aided Design, **43**(6), 687698, 2011. https://doi.org/10.1016/j.cad.2011.02.008

[39] D. Papadias, Y. Tao, K. Mouratidis, and C. Hui, "Aggregate nearest neighbor queries in spatial databases," ACM Transactions on Database Systems, **30**(2), 529576, 2005. https://doi.org/10.1145/1071610.1071616

[40] S. Nutanong, E. H. Jacox, and H. Samet, "An incremental hausdorff distance calculation algorithm," Proceedings of the VLDB Endowment , **4**(8), 506517, 2011. https://doi.org/10.14778/2002974.2002978

[41] Y. Chen, F. He, Y. Wu, and N. Hou, "A local start search algorithm to compute exact hausdorff distance for arbitrary point sets," Pattern Recognition, **67**, 139 148, 2017. https://doi.org/10.1016/j.patcog.2017.02.013

[42] D. Zhang, F. He, S. Han, L. Zou, Y. Wu, and Y. Chen, "An efficient approach to directly compute the exact hausdorff distance for 3d point sets," Integrated Computer-Aided Engineering, **24**(3), 261277, 2017. https://doi.org/10.3233/ICA-170544

[43] D. Zhang, L. Zou, Y. Chen, and F. He, "Efficient and accurate hausdorff distance computation based on diffusion search," IEEE Access, **6**, 13501361, 2018. https://doi.org/10.1109/ACCESS.2017.2778745

[44] G. Chen, Q. Dai, X. Tang, and Z. Xu, "An improved least trimmed square hausdorff distance finger vein recognition," in 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, China, 2018. https://doi.org/10.1109/ICSAI.2018.8599439

[45] Bern University Database. http://www.fki.inf.unibe.ch/databases/iam-faces-database/FullFaces.tgz

[46] A. Martinez and R. Benavente, "The AR face database," CVC Technical Report, **24**, Jun 1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 568-575 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Fuzzy-PID Controller Combined with PSO Algorithm for the Resistance Furnace

Trinh Luong Mien[1,*], Vo Van An[2], Bui Thanh Tam[3]

[1]*University of transport and communications, 03 Cau Giay, Lang Thuong, Dong Da, Hanoi, Vietnam*

[2]*Binh Duong economics and technology university, 530 Binh Duong avenue, Thu Dau Mot, Binh Duong, Vietnam*

[3]*Thu Dau Mot university, 06 Tran Van On, Thu Dau Mot, Binh Duong, Vietnam*

ARTICLE INFO

Article history:
Received: 05 April, 2020
Accepted: 08 June, 2020
Online: 26 June, 2020

Keywords:
PID
PSO
Fuzzy-PID
Fuzzy logic
Resistance furnace
Temperature control

ABSTRACT

The paper presents a novel control strategy applying the particle swarm optimization (PSO) algorithm to optimize the scaling weights coefficients of the fuzzy-PID controller for the resistance furnace temperature control system, called PSO-based fuzzy-PID controller/ algorithm. The proposed PSO-based fuzzy-PID controller in this paper consist of the fuzzy-PID controller and the PSO algorithm. The proposed fuzzy-PID controller is combination of the advantage of PID control and fuzzy logic control. Firstly, the paper presents the mathematical model of the resistance furnace by identification method, based on the experimental data. Then, the design of the fuzzy-PID controller is given in this study. And then, the paper presents the design of the temperature control board using PIC16f with the installed PSO-based fuzzy-PID algorithm. Finally, the simulation and experimental results proved the stability of the proposed PSO-based fuzzy-PID controller with the disturbance, improved the furnace temperature control quality, through obtained major control criteria, such as overshoot, steady-state error, settling time, rising time.

1. Introduction

In industry and transportation, the resistance furnace is a very commonly device. Its temperature control has the characteristics of the one-direction temperature rise, large inertia, time delay and time-varying parameters. In the temperature control system, it is difficult to establish accurate models and determine parameters by mathematical methods, especially when it is disturbed.

The traditional control methods based on the exact object mathematical model, such as feedback control, PID control [1-3] can meet the system performance requirements, and has the advantage of eliminating the steady-state error, but its performance depends on the tuning of the parameters.

Fuzzy logic controller (FLC) has the advantage of inaccurate object model, rapidity and small overshoot, but the control process will have steady-state error. The research works [4-7] used only fuzzy logic for controlling temperature of the resistance furnace and achieved limited results. The quality of the temperature fuzzy logic control system depends on expert experiment and FLC's parameters These works only stopped at the computer simulation, did not solve the effects of disturbance.

The research works [8-16] were combined the advantages of the PID controller with fuzzy logic to design the intelligent controller for the resistance furnace. However, all most of them are still stopped in simulation on computer, but have not given the hardware control board with the intelligent control algorithm on microprocessor.

The research work [17] introduced the fuzzy PID based on genetic algorithm for the resistance furnace. This work [17] indicated that the temperature control system dynamic quality and stable precision is improved, but has not been conducted experimentally on physical equipment and has not assessed the impact of disturbance.

This paper presents a novel fuzzy-PID controller combined with PSO algorithm to control temperature of the resistance furnace with disturbance: from theory to practical experiments. The fuzzy-PID is designed based on the combination of PID control and fuzzy logic control, in which the basic parameters of PID are calculated by fuzzy logic control, and the parameters of fuzzy-PID are determined by multiplying the basic parameters of PID with the scaling weights. These multiplied scaling weights of the fuzzy-PID controller are calculated based on PSO algorithm, so that the dynamic response of the control system is better and the

*Trinh Luong Mien. Tel: +84904684595. Email: mientl@utc.edu.vn

steady-state error is eliminated. The part 2 of the paper presents the resistance furnace identification, determining the PID basic parameters. In part 3, a new PSO-based fuzzy-PID controller is given, here PSO algorithm is used to optimized the scaling weights of the fuzzy-PID controller. And then the simulation model of the resistance furnace temperature control system is built on Matlab in part 4 in order to verify the feasibility of the proposed PSO-based fuzzy-PID controller. Then, the design of the temperature control board using PIC16f microcontroller with PSO-based fuzzy-PID algorithms is introduced in part 5. Finally, part 6 is conclusions. The simulation and experimental results show that the proposed PSO-based fuzzy-PID controller improve the resistance furnace temperature control system quality.

2. The resistance furnace mathematical model by identification method

The resistance furnace, the studied object in this paper, has dimensions L375xW255xH345mm. The furnace uses resistance wire, powered by 220VAC/50Hz, maximum power of 1000W, maximum temperature of 100ºC. In this system, a 2-wire K-type thermocouple sensor with the setting temperature range 0-100ºC corresponding to the electrical signal 0-5VDC is adopted. Controlling the AC power supply capacity for the thermistor wire changes the furnace temperature here, the authors use the TRIAC BTA10-800B 10A 600V [18].

For resistance furnace identification, the control board here only plays the role of opening 100% the TRIAC to supply 220VAC power to the resistance wire. When the TRIAC was active with 100% capacity of the resistance furnace, at the time of measurement the authors realized that the furnace temperature was not changed immediately, it took a certain amount of time to convert electricity into heat energy, heat transfer in the furnace - delay time of the furnace. The temperature in the furnace will gradually increase to the maximum value, corresponding to the maximum power of the furnace.

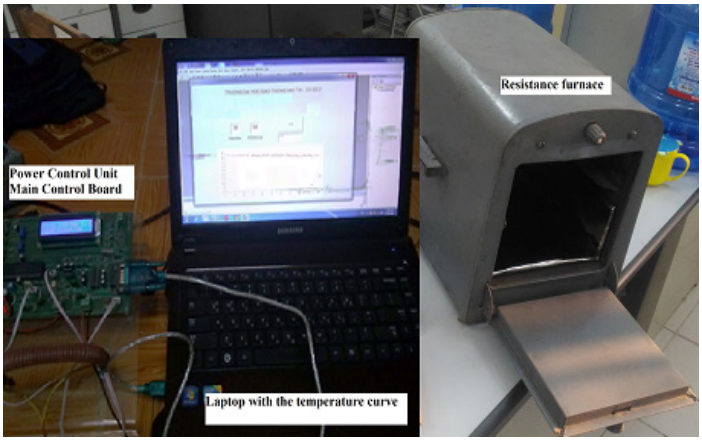


Figure 1. The resistance furnace identification process

Based on this experimental method, the resistance furnace mathematical model can be identified. The identification process was carried out at the ambient temperature about 27.8ºC. The measured steady temperature was about 91.4ºC. The time to reach the steady output temperature of the furnace was about 30 minutes.

Table 1. The experimental temperature values of the furnace

No.	Time [minute]	Temperature [ºC]	Note
1	1	27.8	ambient temperature
2	2	29.1	
3	3	30.5	
4	4	32.0	
5	5	33.5	
6	6	35.9	
7	7	38.5	
8	8	41.3	
9	9	44.7	
10	10	47.8	
11	11	51.2	
12	12	54.9	
13	13	58.2	
14	14	61.6	
15	15	65.0	
16	16	68.2	
17	17	71.1	
18	18	74.2	
19	19	76.8	
20	20	79.4	
21	21	81.7	
22	22	84.0	
23	23	86.0	
24	24	87.7	
25	25	88.9	
26	26	89.9	
27	27	90.5	
28	28	91.0	
29	29	91.4	
30	30	91.4	

The measured temperature curve of the resistance furnace as shown in Fig.2.

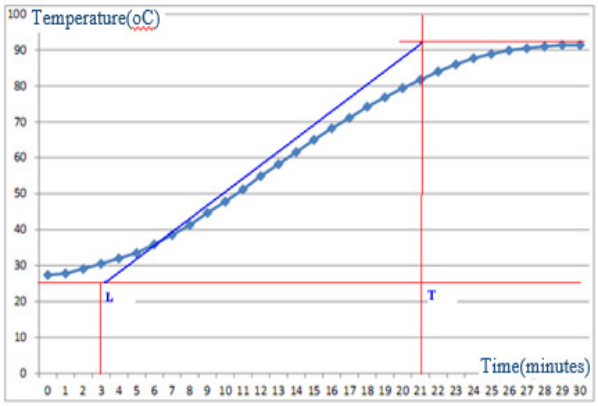


Figure 2. Measured furnace temperature curve

The measured furnace temperature curve has the S-shape, so the furnace mathematical model is described as below [1]:

$$P(s) = \frac{K}{1+sT} e^{-Ls} \quad (1)$$

where L is the time delay, period time that the output signal does not respond immediately; K is the transfer coefficient, is the limit output value as $t \to \infty$; T is the inertial time.

Based on the furnace temperature characteristic curve, the authors draw the tangent line with the inflection point of this curve, the parameters of the furnace as follows:

$$K = \frac{91.4 - 27.8}{100} = 0.64 \quad (2)$$

It is easy to identify $L=225$ second and $T=1230$ second.

Therefore, we obtain the transfer function of the furnace as:

$$G_{obj}(s) = \frac{0.64e^{-225s}}{1230s + 1} \quad (3)$$

It is easy to see that the mathematical model of the resistance furnace is nonlinear, large time delay.

According to Ziegler Nichols-1 (ZN1) [1], the PID controller parameters for the furnace, are determined as equations (5).

$$G_{PID}(s) = K_P(1 + \frac{1}{T_I s} + T_D s) = K_P + \frac{K_I}{s} + K_D s \quad (4)$$

$$K_{P0} = \frac{1.2T}{KL}, T_{I0} = 2L, T_{D0} = 0.5L \quad (5)$$

Therefore, the initial PID parameters are chosen as below:

$$K_{P0}=10.256;\ T_{I0}=450,\ T_{D0}=112.5$$

3. Design of PSO-based Fuzzy-PID controller

3.1. Proposed PSO-based Fuzzy-PID controller

To improve the performance efficiency of the fuzzy-PID controller for the resistance furnace, this work proposed a novel control approach applying the PSO algorithm to tune the scaling weights of the fuzzy-PID controller. The structure diagram describing the incorporate between the fuzzy-PID controller and PSO algorithms is presented in Figure 3.

The fuzzy-PID controller is synthesized based on the structure of the PID with K_P, T_I and T_D parameters, determined according to the fuzzy logic calculated blocks P, I, D corresponding to defuzzied output values K_{Pf} or T_{If} or T_{Df} and then multiply with the corresponding scaling weights, i.e. K_a, K_b, K_c, following formula as below:

$$K_P = K_{Pf}K_a;\quad T_I = T_{If}K_b;\quad T_D = T_{Df}K_c \quad (6)$$

The PSO algorithm is used to optimize three scaling weights, i.e. $K_a \to K_a^*$, $K_b \to K_b^*$, $K_c \to K_c^*$. So that after applying the PSO algorithms, the parameters' Fuzzy-PID is optimized as below:

$$K_P^* = K_{Pf}K_a^*;\quad T_I^* = T_{If}K_b^*;\quad K_D^* = T_{Df}K_c^* \quad (7)$$

For this proposed PSO-based fuzzy-PID controller, the parameters of the fuzzy-PID are continuously adjusted in a specified range. However, the scaling weights' fuzzy-PID is optimized by PSO algorithm. Therefore, the furnace temperature control quality will be greatly improved

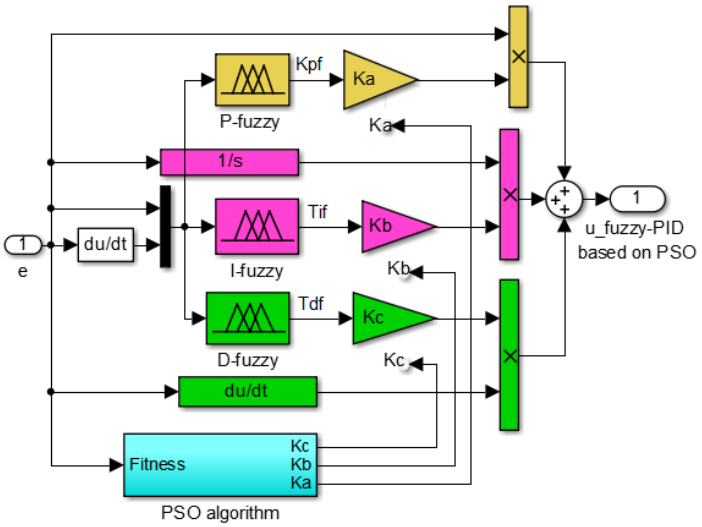


Figure 3. Structure diagram of a PSO-based Fuzzy-PID

3.2. Design of the fuzzy logic calculated blocks

Each fuzzy logic calculated block has two inputs and one output: first input as temperature error $(e=T_d^o-T^o)$, in which T_d^o-setpint temperature, T^o- measured temperature, and second input as temperature error derivation (de/dt), the output corresponding to P-fuzzy, I-fuzzy, D-fuzzy are the K_{Pf}, T_{If}, T_{Df}.

The P, I, D - fuzzy logic calculated blocks, corresponding to K_{Pf}, T_{If}, T_{Df}, are presented in Fig.4.

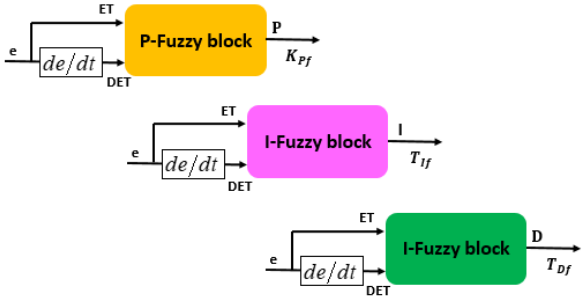


Figure 4. Structure diagram of the fuzzy calculated blocks

Two input fuzzy variables of each P-fuzzy, I-fuzzy, D-fuzzy are ET and DET, corresponding to e and de/dt. The output fuzzy variables of the P-fuzzy, I-fuzzy, D-fuzzy blocks are P, I, D, corresponding to parameters K_{Pf}, T_{If}, T_{Df}.

The physical value range of the input variables and the output variables are as follows: $ET \in [-20.0,20.0]$, $DET \in [-2.0,2.0]$, $P \in [0,20]$, $I \in [0,0.05]$, $D \in [0,2000]$. For D-fuzzy calculated block, $ET \in [-100.0,100.0]$, $DET \in [-10.0,10.0]$

In these fuzzy blocks, membership functions are selected triangles, and setting the input variables by 5 fuzzy sets: ET= {HQ, HD, HI, HV, HL}; DET= {TA, TZ, TI, TV, TL}, and output variables are equal to 5 fuzzy sets, corresponding to the language variables P, I, D we have P= {VD, VG, VI, VV, VL}; I= {VD, VG, VI, VV, VL}; D= {VD, VG, VI, VV, VL}.

With the number of fuzzy sets of five inputs and two fuzzy inputs for each fuzzy calculated block K_{Pf}, T_{If}, T_{Df}, so we have the total of 5 x 5 = 25 fuzzy rules for each fuzzy calculated block. Based on the resistance furnace specification and the principle of

PID parameters adjustment, the authors built the fuzzy rules table for the fuzzy block of calculated K_{Pf}, T_{If}, T_{Df}, as Table 2.

Table 2. Fuzzy rules of the P,I,D-fuzzy calculated blocks

P I D		ET				
		HQ	HD	HI	HV	HL
DET	TA	VD	VD	VD	VD	VD
	TZ	VD	VG	VD	VD	VD
	TI	VD	VD	VI	VI	VV
	TV	VD	VD	VI	VV	VL
	TL	VD	VD	VV	VL	VL

These P-fuzzy, I-fuzzy, D-fuzzy calculated blocks in this study use Max-Min inferential law and defuzzification according to the centroid point method, the output clear values of the P-fuzzy, I-fuzzy, D-fuzzy calculated blocks, corresponding to K_{Pf}, T_{If}, T_{Df}, are defined.

3.3. Building the PSO algorithm

The particle swarm optimization is one of the best effective optimization techniques. Firstly, all particles are assigned initial position and velocity values. Based on optimal value of fitness function, new position and velocity of the particles are updated. There are two fitness value are required in the update process of particles, that are p_{best} (personal best) and g_{best} (global best). The p_{best} value is trace in every iteration of particles and g_{best} value is computed among the best solution in p_{best} value [19]. In the PSO algorithm, the velocity value and position values of particles are continuously updated by following formulas.

$$v_{i,d}^{j+1} = \alpha v_{i,d}^{j} + C_1 r_1 (p_{best} - x_{i,d}) + C_2 r_2 (g_{best} - x_{i,d}) \quad (8)$$

$$x_{i,d}^{j+1} = x_{i,d}^{j} + v_{i,d}^{j+1} \quad (9)$$

where $v_{i,d}^{j+1}$ - new updated velocity of *i*-particle in *(j+1)* iteration. $v_{i,d}^{j}$ -the velocity of *i*-particle in *j*-iteration; $x_{i,d}^{j+1}$ - new updated position of *i*-particle in *(j+1)* iteration. $x_{i,d}^{j}$ -last position of *i*-particle in *j*-iteration; C_1 and C_2 are the acceleration factors for updating the particle velocity; r_1 and r_2 are initially random positive values which should be less than one; α is the assigned inertial weight coefficient to maintain the particles last velocity while the accelerations factors drive the flow of particles towards optimum solutions.

It is easy to see that the particles position is mainly depended upon the update velocity, and this update velocity is handled by the acceleration factors and assigned inertial weight coefficient. During the update process, the velocity and position of particles must satisfy the constraints, as shown below.

$$V_{d\_min} \le v_{i,d}^{j+1} \le V_{d\_max} \quad (10)$$

$$X_{d\_min} \le x_{i,d}^{j+1} \le X_{d\_max} \quad (11)$$

In this study, PSO algorithm is used to determine three the scaling weights of the Fuzzy-PID controller, i.e. K_a, K_b, K_c, And the fitness function can be selcted based on the combination of three goals, i.e. settling time – T_{STL}, quality index of integral of the absolute magnitude of the error - E_{IAE}, and peak overshoot - M_{POT}.

$$J(\theta) = \gamma_1 E_{IAE} + \gamma_2 M_{POT} + \gamma_3 T_{STL} \to min \quad (12)$$

$$E_{IAE} = \int_0^T |e(t)|\, dt \quad (13)$$

where $\theta = [K_a, K_b, K_c]$ are parameters to optimized. T denotes the simulation time. γ_1, γ_2, γ_3 are the assigned weights, here $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 0.5$.

The search boundary for the scaling weights of the Fuzzy-PID controller is assigned as follows:

$$0 \le K_a \le 30; 0 \le K_b \le 0.1; 0 \le K_c \le 10 \quad (14)$$

The PSO algorithm continuously varies the values of Fuzzy-PID controller parameters, following equation (6), until the objective function J is minimized to $J_{\min}$. The implementation of PSO algorithm involves the following steps:

Step 1. Initialize size of particles $N=10$, dimension search-space $D=3$ corresponding to K_a, K_b, K_c, maximum iteration $M=30$, inertial weight $\alpha=0.5$, acceleration factors $C_1=2$, $C_2=2$,

Step2. Initialize two velocity and position values vectors, and then compute the initial fitness p_{best} values vector for each particle.

Step 3. Minimize g_{best} value from the p_{best} values vector, following the formulas (12), (13), (14).

Step 4. Update new velocity and position values into the velocity and position vectors of particles in step 3, following the formulas (8), (9).

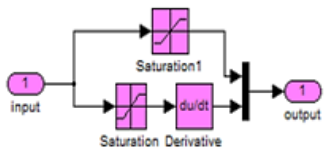
Step 5. Compute the values of fitness p_{best} and then p_{best} for new update position if these fitness values are improved the replaced last values with these new values (less than the previous fitness values).

Step 6. Repeat the step 3, 4, 5 until the last iteration or the desired value of the fitness function is satisfied.

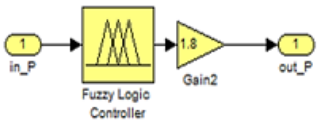
4. Simulating the furnace temperature control system using the proposed PSO-based Fuzzy-PID controller

4.1. Modeling fuzzy-PID controller on Matlab

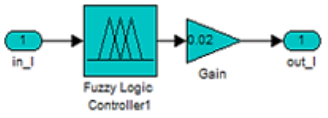
Based on Mamdani fuzzy method, the proposed fuzzy-PID in this study, was built on Matlab as show in Figure 5.



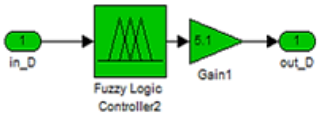
Fuzzy input block diagram



Fuzzy P block diagram



Fuzzy I block diagram



Fuzzy D block diagram

Figure 5. Modeling the new-type fuzzy-PID on Matlab

The triangle membership functions were selected. The fuzzy rules were built based on Table 2. The fuzzy inference mechanism was chosen Max-Min method and the defuzzification according to the centroid point method were adopted. The relational surface of the P-fuzzy, I-fuzzy, D-fuzzy blocks are described in Figure 6.

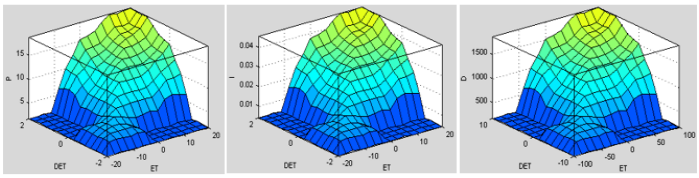


Figure 6. The relational surface of the P, I, D-fuzzy blocks

The simulation diagram of the heater temperature control system is built on Matlab, as shown Fig.7. We conducted the quality assessment of the control system with three designed controllers above: PID-ZN1, PID-CHR and new-type fuzzy-PID.

4.2. Determining the optimal parameters of the fuzzy-PID by the PSO algorithm

The initial scaling weights of the Fuzzy-PID controller, i.e. K_a, K_b, K_c, in simulation are selected randomly based on the initial parameters of PID, following the formular (5), so as to improve the quality of the furnace temperature control system. Here, the initial values K_a, K_b, K_c are as follows $K_a=10$, $K_b=0.02$, $K_c=5.5$

When applying the proposed PSO algorithm in part 3.3 to determine optimal parameters of fuzzy-PID controller, at this point, the optiaml values K_a, K_b, K_c are as follows $K_a=22.2$, $K_b=0.021$, $K_c=5.1$

4.3. Simulation results on Matlab

The simulation scenarios are carried out with 3 controllers: PID controller with the parameters determined by Ziegler-Nichols 1, denoted PID-ZN1; Fuzzy-PID controller with the unoptimized parameters, denoted FPID; and Fuzzy-PID controller with the optimal parameters, denoted PSO-FPID.

When there is not disturbance, the response of the system corresponding to three controllers as shown Figure 7.

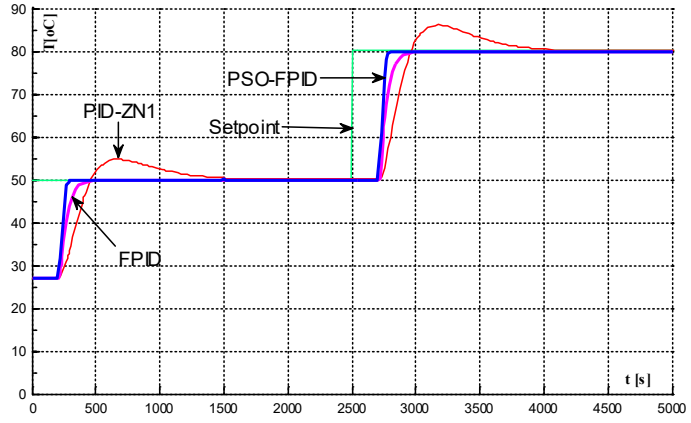


Figure 7. Response of the control system without disturbance

The overshoot of the control system with the proposed fuzzy-PID controller (FPID and PSO-FPID) is better than conventional PID controller (PID-ZN1). Specifically, when applying the PSO-based fuzzy-PID and the fuzzy-PID controller then the control system is not overshoot, meanwhile with PID-ZN1 controller then the control system is an overshoot of about 10%. The settling time of the system (included the time delay, 225 seconds) at the desired temperature value, $T_d=50^oC$, for the proposed PSO-based fuzzy-PID controller is so rapid, about 278 seconds, and for the proposed fuzzy-PID controller is about 401 seconds. The rising time of the system with the proposed fuzzy-PID controller is faster than PID controller. The steady-state error of the system with the proposed fuzzy-PID controller is almost eliminated.

Table 3. The system quality indexes without disturbance

Quality indexes	PID	Fuzzy-PID	PSO-based fuzzy-PID
Rising time (sec)	117.6	71.8	38.8
Settling time (sec)	1243.6	401.5	278.4
Overshoot (%)	<10	0	0
Steady-state error (%)	<1	~0	~0

Therefore, the two controllers proposed in this study (FPID, PSO-FPID) can bring improved quality of the furnace temperature control system.

When the sine disturbance impacts on the system: In fact, the operation environment is not ideal. The system operation is always affected by external disturbance, so that it makes to change the actual signals. Assume that the sine disturbance has the mixed frequency in range of 0.01-50Hz, and maximum amplitude of 50. At this time, the response of the furnace temperature control system corresponding to three controllers PID-ZN1, FPID, PSO-FPID, are shown in Figure 8.

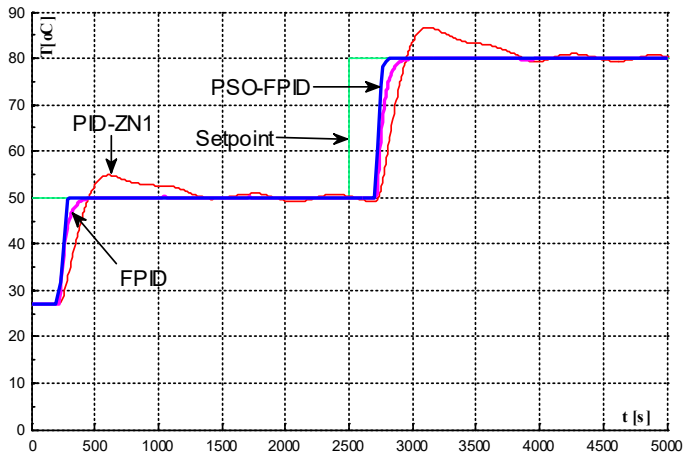


Figure 8. Response of the control system with sine disturbance

Figure 8 indicated the traditional PID controller (PID-ZN1) causes the system to oscillate around the set-point temperature value. This oscillated amplitude depends on the disturbance amplitude. Meanwhile, applying the proposed fuzzy-PID or PSO-based fuzzy-PID, the system quality is quite good, ensuring the system stability.

If the disturbance amplitude is so large, the furnace temperature control system may be unstable with the traditional

PID controller, but still ensure stability of the system with the proposed PSO-based fuzzy-PID.

Table 4. The system quality indexes with disturbance

Quality indexes	PID	Fuzzy-PID	PSO-based fuzzy-PID
Rising time (sec)	110.5	65.5	36.4
Settling time (sec)	1255	380.1	265.1
Overshoot (%)	<15	<1	~0
Steady-state error (%)	-2 ÷ 2	-0.5 ÷ 0.5	~0
Oscillated response	Yes	Small	No

5. Design of the temperature control board

5.1. Block diagram of temperature control board

In this section, the authors present the design of the PIC16F control board for controlling temperature of the resistance furnace. This temperature control board is designed by using PIC16F877A microcontroller [20], as shown in Figure 9.

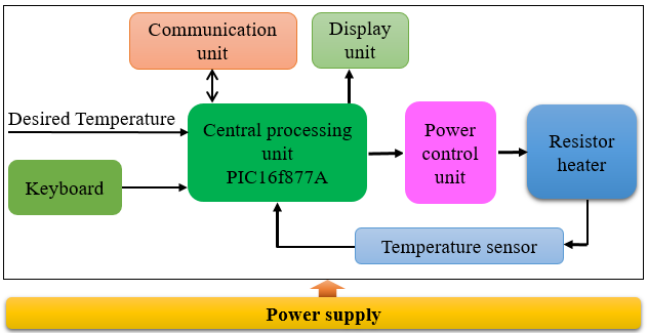


Figure 9. Block diagram of the temperature control board

The power supply block is responsible for supplying power to the entire system. The control object here is the resistance wire, wrapped around the copper pipe, causing the heat in the copper pipe to rise temperature in the furnace, it is controlled via the voltage signal of the power control unit. The temperature in the heater is measured by a temperature sensor. The display unit is responsible for displaying the actual temperature and set-point temperature. The keyboard is responsible for entering the set-point temperature. The central processing unit converts from analogue signal to digital values and performs the functions of fuzzification block, fuzzy-rules law table, defuzzification block and generating signals to command the power unit, thereby stabilizing the furnace temperature according to the set values. The power unit is responsible for controlling the AC power supplied to the resistance wire. The communication unit, using USB/RS232 port, is responsible for communicating between the board and the computer.

5.2. Layout circuit of the temperature control board

The principle schematic circuit diagram of the temperature control board includes the schematic circuit diagram of the PIC16f central processing unit, the LCD display block, the power control unit, the USB/RS232 communication block and the AC/DC converter block. The detail schematic circuit of the temperature control board is presented in [18].

Based on the principle schematic circuit, the layout circuit diagram is given in Figure 10 and the realistic layout circuit of the control board as shown in Figure 11.

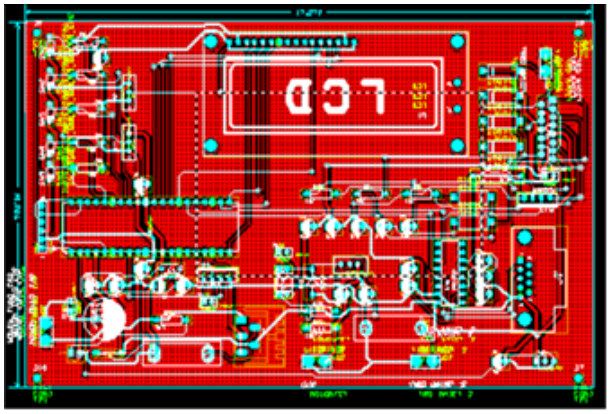


Figure 10. Layout circuit of the temperature control board

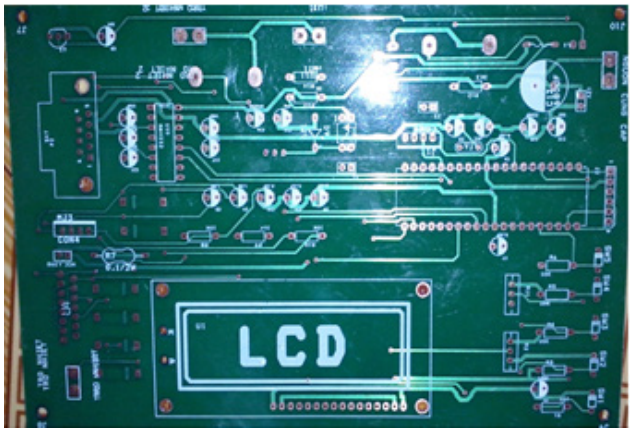


Figure 11. Realistic layout circuit of the control board

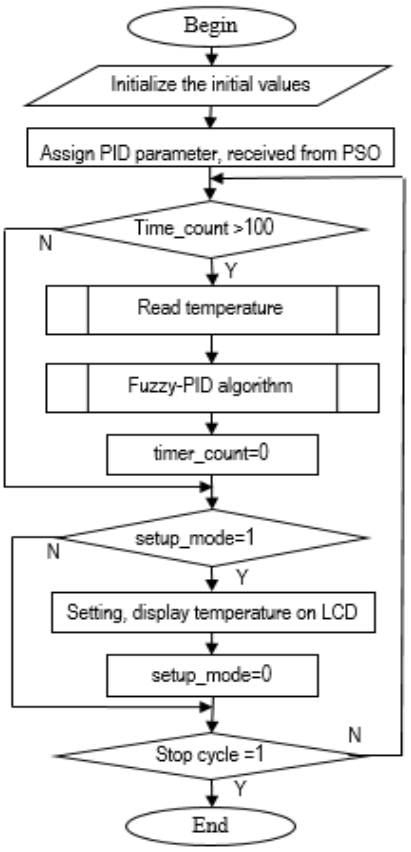


Figure 12. The main program flowchart for controlling the furnace temperature using PSO-based fuzzy-PID

5.3. Temperature control algorithm on PIC16f chip

The furnace temperature control algorithm is installed on the PIC16f microprocessor as described in Fig.12.

The detail of algorithm flowcharts corresponding to each subprogram, are presented in Fig.13.

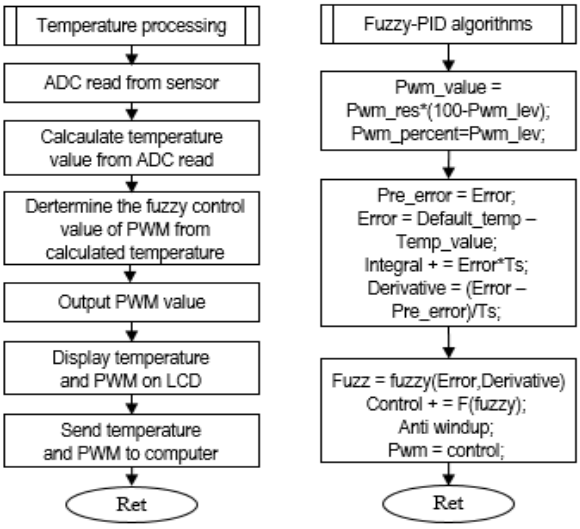


Figure 13. Algorithm flowchart of subprograms

5.4. Experimental control for the heater in the laboratory

Experiment of the resistance furnace temperature control system, using PIC16F control board with the proposed PSO-based fuzzy-PID algorithm, is presented in Figure 1.

The monitoring control interface based on Visual C# is developed on the computer to obtain the actual temperature value of the resistance furnace. From this interface, the set-point temperature value and the parameters of the PSO-based fuzzy-PID controller also can be assigned remotely and easily.

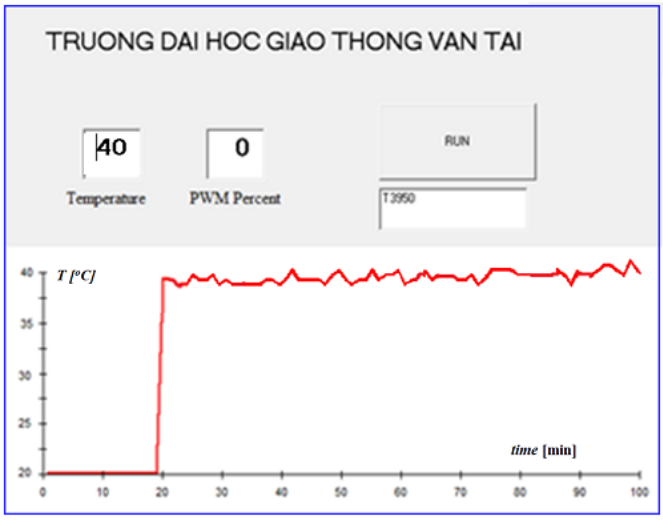


Figure 14. Real temperature response at setpoint 40°C

At the beginning of the experiment, the control board with installed PSO-based fuzzy-PID algorithm, will create the 100% PWM output signal to supplying the suitable AC power values to the resistance wire, after the delay time, the temperature in the furnace changes from environmental temperature $T_0=20^oC$ up to 39°C, and then reduce PWM signal to maintain the actual temperature value at setpoint value of $T_d=40^oC$. When the furnace temperature exceeds 40°C, the PIC16f microcontroller interrupts the PWM signal so that the temperature will reduce to the setpoint temperature value.

The experimental results in laboratory showed that the proposed PSO-based fuzzy-PID controller brings good control quality for the resistance furnace temperature control system. The proposed PSO-based fuzzy-PID controller with the optimal parameters usually gives better quality the proposed fuzzy-PID controller, because the parameters of the fuzzy-PID controller was not optimized. However, these two proposed controllers offer better quality than the traditional PID controller, especially when the disturbance impacts on the system.

6. Conclusion

This paper presented a novel control approach applying the PSO algorithm to tune the scaling weights coefficients of the fuzzy-PID controller which is combination of the advantage of the traditional PID control and fuzzy logic control. The fuzzy-PID controller is designed based on the typical fuzzy rules, and combined with the PID control, but the fuzzy-PID three scaling weights are tuned by the PSO algorithm. And then the simulation model of the resistance furnace temperature control system

without/with disturbance using the proposed controller was built on Maltab. The simulation results have confirmed the effectiveness of the two proposed controllers for the resistance furnace temperature control system, i.e. fuzzy-PID controller and PSO-based fuzzy-PID controller. Then the proposed PSO-based fuzzy-PID algorithm was implemented on the temperature control board which made of PIC16f microprocessor. Experimental results showed that proposed PSO-based fuzzy-PID algorithm provided better control quality than the traditional PID algorithm. In addition, the proposed PSO-based fuzzy-PID controller is consider to be an efficient control strategy applying for a class of nonlinear and uncertain complicated control objects.

From this research, the future work focusing on optimization structure of fuzzy logic calculated blocks, and apply adaptive control to tuning online/adaptive parameters of the fuzzy-PID controller. The future work also continues to develop the temperature control board using PIC16f in order to complete the equipment, and put the equipment into the actual application to control the temperature of materials in industry and transportation.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Aidan O'Dwyer, Handbook of PI and PID Controller Tuning Rules, Imperial College Press, 2003.

[2] Yu Feng Zhang, Ming Li and Jing Min Dai, PID Heating and Temperature Control Method Based On Dynamic Assignment, 5th International Conf. on Mechatronics, Materials, Chemistry and Computer Eng., 2017. https://doi.org/10.2991/icmmcce-17.2017.208

[3] V.Kabila, Glan Devadhas, Comparative Analysis of PID and Fuzzy PID Controller Performance for Continuous Stirred Tank Heater, Indian Journal of Science and Technology, Volume 8, Issue 23, Pages: 1-9, 2015, https://doi.org/10.17485/ijst/2015/v8i23/85351

[4] L.X. Wang, A course in Fuzzy Systems and Control, Prentice Hall International, Inc, 2002.

[5] Z.R. Radakovica et al., Application of temperature fuzzy controller in an indirect resistance furnace, Applied Energy, Volume 73, Issue 2, 2002. https://doi.org/10.1016/S0306-2619(02)00077-6

[6] Moh Syahrir Bin Ribuan, Development Of Fuzzy Logic: Temperature Controller Implementation - Microcontroller Based, University Teknikal Malaysia Melaka, 2009.

[7] P. Singhala, D. N. Shah, B. Patel, Temperature Control using Fuzzy Logic, International Journal of Instrumentation and Control Systems, Vol.4, No.1, 2014, https://doi.org/10.5121/ijics.2014.4101

[8] Nguyen Chi Ngon, Fuzzy PI controller: From design to application, Science magazine, Can Tho University, 2011.

[9] Shi Dequan, et al., Application of Expert Fuzzy PID Method for Temperature Control of Heating Furnace, 2012 International Workshop on Information and Electronics Engineering, Procedia Engineering, Volume 29, 2012, https://doi.org/10.1016/j.proeng.2011.12.703

[10] Sneha S Patole1, Shailendra K Mittal, Fuzzy PID Controller Design for Heating Control System, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 6, Issue 6, 2007, https://doi.org/10.15662/IJAREEIE.2017.0606140

[11] Qingjie Yang, Guohou Li, Xusheng Kang, Application of fuzzy PID control in the heating system, 2008 Chinese Control and Decision Conference, 2008, https://doi.org/10.1109/CCDC.2008.4597814

[12] Jiangjiang Wang, Dawei An, Chengzhi Lou, Application of Fuzzy-PID Controller in Heating Ventilating and Air-Conditioning System, 2006 International Conference on Mechatronics and Automation, June 2006, https://doi.org/10.1109/ICMA.2006.257656

[13] L. D. Vijay Anand, Lakshmi Asok,, Design of Fuzzy PID with Expert Control for a Temperature Process, International Journal of Engineering Research & Technology, Volume 03, Issue 02, 2014.

[14] Guofang G., Zhengzhong L., Design of Heating Furnace Temperature Control System Based on Fuzzy-PID Controller, Information Engineering and Applications. Lecture Notes in Electrical Engineering, Volume 154. Springer, https://doi.org/10.1007/978-1-4471-2386-6_196

[15] Trinh Luong Mien, Design of Fuzzy-PI Decoupling Controller for the Temperature and Humidity Process in HVAC System, International journal of engineering research & technology, Vol. 05, Issue 01, January-2016. https://doi.org/10.17577/IJERTV5IS010473

[16] Y.Zhang, et al., Application and Research of Fuzzy PID Control in Resistance Furnace Temperature Control System, IOP Conf. Series: Earth and Environmental Science, Vol.186, Issue 5, 2018. https://doi.org/10.1088/1755-1315/186/5/012051

[17] Liu Jing Yan, The Fuzzy PID Control of Resistance Furnace Temperature System Based on Genetic Algorithm, Applied Mechanics and Materials Vol. 273, 2013, https://doi.org/10.4028/www.scientific.net/AMM.273.678

[18] Trinh Luong Mien, Vo Van An, Lecture notes: Design of the temperature controller based on fuzzy logic using PIC16f microprocessor, University of transport and communications, 2018.

[19] A.H. Al-Mter and S. Lu, A Particle Swarm Optimization Algorithm Based On Uniform Design, International Journal of Data Mining & Knowledge Management Process, Vol. 6, No. 2, 2016. https://doi.org/10.5121/ijdkp.2016.6203.

[20] Microchip, PIC16F/LF1825/1829 Data Sheet DS41440A, Microchip Technology Inc., 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 576-582 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Warehouse Relocation of a Company in the Automotive Industry Using P-median

Zarate-Zapata, Aldo Cesar *, Garzón-Garnica, Eduardo Arturo, Cante-Mota, Román, Olmos-Álvarez, Fernando, Martinez-Flores, José Luis, Sánchez-Partida, Diana

Logistics and Supply Chain Management Department, Universidad Popular Autónoma del Estado de Puebla, 72410, México

ARTICLE INFO

Article history:
Received: 02 April, 2020
Accepted: 03 June, 2020
Online: 26 June, 2020

Keywords:
Logistics
P-median
Risk management

ABSTRACT

To have enough information on time can be helpful when companies try to reduce costs and operate more efficiently. An international company that supplies parts for the automotive industry is currently testing its new facilities in Mexico. The relocation of the raw materials and finished goods warehouses were tested using a P-Median model. The operating costs and risk factors were included in the model to provide a better solution and improve the operation of the warehouses and production lines. The research results compared different scenarios and indicated that the proposed better location isolates the forklift routes, mainly for finished products, and minimizes the cost of moving both raw materials and finished products to and from warehouses.

1. Introduction

The logistics activities have always been important for companies, because they represent various concepts, principles, and methods of different areas, like marketing, production, accounting, warehouse management, among others. With logistics, the chain value for the customers is increased, as well as for suppliers and stakeholders. An efficient logistics management could represent a decrease in costs, and continuous improvement in production [1].

In a warehouse, typically products arrive packaged and the warehouse personnel reorganizes, classifies, sorts and repackages such products. This function is of vital importance in any warehouse because it is needed to break down large chunks of products and redistribute them in small quantities. So, the supply chain is the sequence of the different processes in the organization within which the products move from their origin toward the customers [2].

In this paper, we propose a solution to a warehouse problem of a company in the automotive industry, using a p-median model to minimize the travel distance from the warehouses to all of the production stations.

2. Research objectives and research methodology

2.1. Research Objectives

This document aims to implement the p-median problem modeling with Lingo® software, which is a tool designed to solve linear, nonlinear, integer, stochastic, and optimization programming models as well as concentrate and model exact algorithms. It is done to explain the company's problems and decrease risk in the industrial process to obtain an exact solution that does not require model validation.

2.2. Research Methodology

The research methodology adopted for the research work has been represented as a theoretical concept. The application of the mathematical model is based on the p-median problem. Such a problem has been implemented in several documents, and by different authors, as cited in the literature review. Such literature review is based on studies carried out with the application of exact algorithms using the p-median problem as the basis.

The following criteria were used to select the literature:

- Literature published on risk industries, supply chain, and cases of study using the p-median problem.
- Literature published from 1975 to 2017.
- Articles published in refereed scholarly journals, working papers, and thesis.
- Journals explain p-median problem algorithms in their editorial scope
- Keywords used in the article: risk, logistics, p-median, risk management

3. Literature Review

In [3] it is mentioned that the p-median problem is useful to model many real-world situations such as the location of public or

* Zarate-Zapata, Aldo Cesar, aldocesar.zarate@upaep.edu.mx

industrial facilities, warehouses, and others. The p-median problem differs from the Uncapacitated Facility Location Problem in two respects:

1. There are no costs for opening facilities
2. There is an upper bound on the number of facilities that should be opened.

It models the problem of finding a minimum cost clustering and belongs to the class of NP-hard problems.

In [4], a set L of m facilities (or location points), a set U of n users (or customers or demand points), and a n x m matrix D with the distances traveled (or costs incurred) d_{ij} for satisfying the demand of the user located at i from the facility located at j, for all $j \in L$ and $i \in U$ are considered. The objective is to minimize the sum of these distances or transportation costs.

$$\sum_{(i \in U)} \mathrm{m}\, j \in \mathrm{J}\, dij$$

Where $J \in L$ and $|J| = p$. The p-median can be defined as a purely mathematical problem: given an n x m matrix D, select p, columns, or D so that the sum of minimum coefficients in each line within these columns is the smallest possible.

The p-median can also be interpreted in terms of cluster analysis; locations of users are then replaced by points in an m-dimensional space [5].

Besides this combinatorial formulation, the p-median is also an integer programming problem. Let us define two sets of decision variables: (*i*) $y_j = 1$ if a facility is opened in $j \in L$, and 0, otherwise; (ii) $x_{ij} = 1$, if customer i is served from a facility located in $j \in L$, and 0, otherwise. Then the integer programming formulation is as follows:

$$minimize = \sum^{n}_{i} * \sum^{m}_{j} d_{ij}\, x_{ij} \quad (1)$$

Subject to

$$\sum^{n}_{j} x_{ij} = 1,\ \forall I \quad (a)$$

$$x_{ij} \leq y_j,\ \forall i,j \quad (b)$$

$$\sum^{m}_{j} y_j = p, \quad (c)$$

$$x_{ij}, y_j, \in \{0,1\} \quad (d)$$

The objective function of the p-median model seeks to minimize the sum of the distances d_{ij} for each client. Constraint (a) is the one that assures that each client is assigned to a facility. Constraint (b) makes sure that each client is attended by precisely one server. Constraint (c) indicates that p facilities should be assigned. Equation (d) indicates that the total number of open facilities is set to p by constraint.

The research in [6] shows that the academic research on safety in logistics has mainly focused on transportation, and in particular, on safety concerning motor carriers [7]. Among others, studies research characteristics of professional drivers (e.g., personality, health, attitude), stress factors they face (time pressure, fatigue, stress) and how these relate to safety behavior and/or accidents [8], [9], [10], [11], [12]. Recently, warehouse safety has started to gain attention. For instance, In [13] it is analyzed which factors impact warehouse safety.

In addition, in [14] it is mentioned that the risk in an environment is very significant [15], and the companies that measure it are characterized by several risk factors. The term *supply chain vulnerability* [16] has been used to describe the dependence and risks that exist among organizations as they rise to the challenge of better, faster, cheaper. Two aspects connected to the risk assessment in supply chains can be found:

1) Risk exists inside the company and at the network level.
2) Risk evaluation is subjective because each people perceives differently a risk and the nature of the upstream and downstream relationships of such risk.

According to [17], the success of risk management will be based on the understanding of various categories of risks, that affect projects in an organization. Achieve an analysis of their mutual dependence, categorize them according to their importance, and develop strategies for risk management based on prioritization.

The physical risk can be further measured in terms of the potential damages in physical spaces such as warehouses and production lines. These measures, called *physical risk indexes*, are derived from inspection procedures by external experts. The risk measures are monitored from both the logistics and supply chain perspective and the top-management control perspective [18]. Furthermore, in the supply chain the objectives are considered in terms of the organization's goals concerning the final customer. Thus, the risk evaluation, in terms of weights and importance of indicators, should be guided by an awareness of the nature and importance of the market objectives [14].

4. Problem Description

An International Company of the Automotive Industry has a facility in Mexico; its primary purpose is the manufacturing of automotive cooling components, like radiators and heat exchangers. Currently, this company has only one production line in its planning stage. Their forecasted demand is very significant, and it is expected to grow further. Therefore, they will require a total of four production lines. The design of such lines is already available, as well as the preliminary data of the required and projected production levels.

The raw materials warehouse, as well as the finished goods warehouse, are currently operating on a small scale, working only at floor level, without racks. As the plant is not yet operating in full production mode, it is still possible to relocate the materials and finished goods warehouses, to a certain extent.

The cost of relocating the warehouse; at this stage, is minimal because production lines are not yet established. Still, not all of the building's spaces can be used for warehouses, as power, gas, water, and overhead lines for production lines are already established, close to the places where the machines would be installed.

The use of the current layout, or relocation to a new position within the plant, should consider the operating cost of the warehouses, but also the risk that comes with the location of those warehouses. If materials must be carried through the plant using a forklift, the cost is higher than if a manual cart can be used. Nevertheless, when using a cart, if the warehouse operators must cross the forklift's pathway, there is a higher risk of an accident

happening than if the route is free of such crossings. The risk of damaging finished goods with the forklift, or by crossing its roads, is also noted. The relationship between cost and risk should then be considered for the solution.

The company leases its forklifts, so the cost of ownership does not change whether they are used or not, except for the cost of fuel. If the distance traveled by the forklift can be kept small enough to meet the need for full production with the fewest number of forklifts, some savings can be made by that means. A shorter distance traveled by the forklift could mean that all production lines are serviced by just one forklift, saving on the fixed cost of the rent.

5. Methods and Procedures

The problem for this company was modeled as a p-median problem, using a distance matrix calculated from the warehouses to each production station.

The distances were computed as follows:

A map of the whole plant was used as a base, and divided, on the scale, to a square grid of 1 meter by side. The location of each working station, and those of the existing warehouses were marked in the map

The possible alternate locations of the warehouses were selected simply by finding the spaces within the plant that were currently not in use, and where the warehouses would fit. Such a method resulted in three possible locations for the warehouses, with little or no movement of the production lines from the current plant layout.

The distance from the warehouse entrance, in each of the proposed locations, to each one of the working stations, whether current or projected, was computed by counting the meters in the map, as if the cart or forklift would have traversed the route from the warehouse to the working station. The same process was done for the reception of raw materials and the finished goods.

If one of the warehouses proposed locations required the relocation of one or more production lines, each one of the distances was computed again, considering the new location of the production lines and warehouses.

As stated, three different scenarios were found possible, and those were evaluated to obtain a comparison against the current situation. The first scenario is the current situation. The second one involves moving the raw materials warehouse to a top-right position, in a space currently unused. The third scenario considers moving the raw materials warehouse to a down-center position, very close to the place where the production lines start, but this last scenario involves the location of the two middle production lines, a few meters offset to the top. Still, those production lines were not operating yet, so a small relocation was still possible.

The considerations to the problem at hand are as follows:

1. The cost of relocating the warehouses is minimal, because there is no hardware, like racks, in the warehouse at this point.
2. Although production lines 2-4 are not yet set up, some of the needed infrastructures are already prepared. It prevents such lines from being set anywhere. They must remain close to the originally intended position.
3. Forklift streets, walkways, isles, etc., can be relocated easily. However, there is no unlimited space.

Figure 1 shows the current layout of the plant, including only one production line, which is the current state of the company. However, they foresee the need to upscale their production levels to a full mode, needing four lines, in at most two years from now.



Figure 2 shows the possible layout of the plant when all four lines are in place, considering the current location of both the raw materials and finished goods warehouses. Two possible alternatives to the warehouse location have been found in the current plant layout.

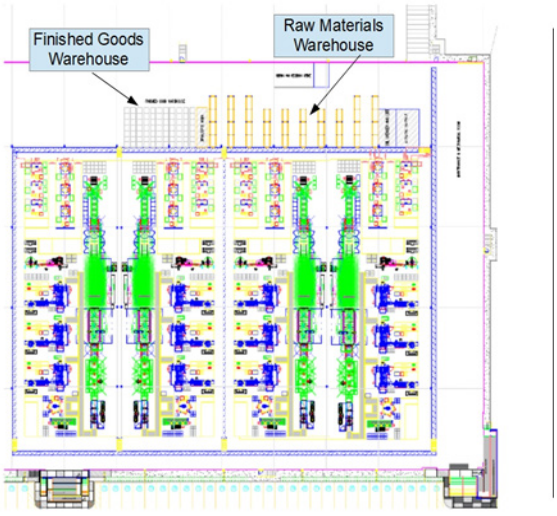


Figure 2. Current layout with no change in the warehouse locations.

Figure 3 shows the possible layout with all four lines in place, but with the warehouses relocated to a place currently unused in the plant, without the need to relocate any of the production lines.

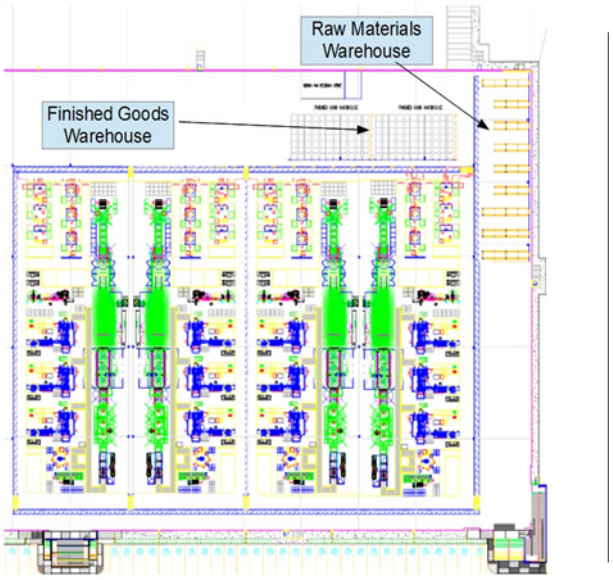


Figure 3. Proposed layout with warehouses in the top right corner.

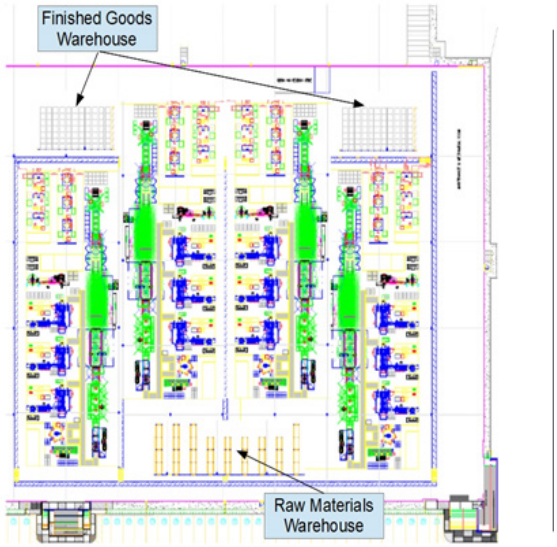


Figure 4. Proposed layout with raw materials warehouse in the bottom center.

6. Data Acquisition and Preparation

The primary way to compute the cost was to obtain the distance from the warehouses to each working station. More distance traveled means the operator would take more time to carry the items, whether by forklift or by manual cart. It was considered that for one production run in each production line, one shipment of raw materials would be required for each working station.

Table 1: Extract of the data table.

	Raw Material 1			Finished Goods 1			Raw Material 2			Finished Goods 2			Raw Material 3			Finished Goods 3			
Equipment	Distance	Risk	Forklift	Distance	Risk	Forklift	Distance	Risk	Forklift	Distance	Risk	Forklift	Distance	Risk	Forklift	Distance	Risk	Forklift	Arrival Factor
0-Raw Material Receive	17	0	1	0	0	0	47	0	1	0	0	0	121	0	1	0	0	0	1
1-Core Builder 1	62	1	1	0	0	0	34	1	1	0	0	0	70	1	1	0	0	0	0.033
1-Core Builder 2	54	1	1	0	0	0	26	1	1	0	0	0	78	1	1	0	0	0	0.033
1-Core Builder 3	46	1	1	0	0	0	18	1	1	0	0	0	86	1	1	0	0	0	0.033
1-Tube Mill	38	1	1	0	0	0	10	1	1	0	0	0	94	1	1	0	0	0	0.125
1-Laser Welding	78	1	0	0	0	0	51	1	1	0	0	0	65	1	0	0	0	0	0.125
1-Paint Flux	83	1	0	0	0	0	56	1	0	0	0	0	58	1	0	0	0	0	0.125
1-Furnace	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.125
1-Flame Brazing 1	14	0	0	0	0	0	21	1	0	0	0	0	129	1	0	0	0	0	0.125
1-Flame Brazing 2	18	0	0	0	0	0	16	1	0	0	0	0	135	1	0	0	0	0	0.125
1-Flame Brazing 3	23	0	0	0	0	0	12	1	0	0	0	0	141	1	0	0	0	0	0.125
1-EOL Pack 1	14	1	0	60	1	0	14	1	0	14	1	0	118	1	0	14	1	0	0.125
1-EOL Pack 2	18	1	0	64	1	0	9	1	0	18	1	0	112	1	0	18	1	0	0.125
1-EOL Pack 3	23	1	0	68	1	0	5	1	0	23	1	0	106	1	0	22	1	0	0.125

As stated, a grid was drawn over the map with the proposed layouts, with each square measuring approximately one meter. The same grid size was used for all the proposed locations. The distance within the warehouses was not computed, because currently there is not a definitive warehouse layout in place. The point of origin was set as the point within the warehouses closest to the forklift or cart roads, in the case of the current warehouse location, such point was the entrance of the warehouse. All the distances were then computed and stored in a single table. Each row in that table represents one working station of one production line. Each column represents one of the possible locations of both the raw materials and finished goods warehouses.

A column was added to each location to consider the cost if the type of material that has to be brought to the working station requires the use of a forklift, a value of 1 was added if the materials can be carried with a manual cart, or by hand, the value was 0. In the case of the risk, if the materials could be carried by hand or cart, but the route needed to cross the forklift pathways, a value of 1 was set.

Another value that had to be computed was the reception of the raw materials. This case had to also consider the cost of bringing the raw materials from the receiving docks to the warehouse. A modifier was added to transform this cost for each production round. The modifier consisted of the number of production lines that could be served with a single raw materials shipment. If an item arrives in a pallet containing 40 boxes, and three boxes would be needed for each production run, the modifier for that distance would be 3/40 boxes, or 0.075. Table 1 below shows an extract of the resulting data table, showing the risk, forklift, and arrival cost modifiers, as well as the distance computed for each scenario.

The model needed to include the risk and cost factors, apart from the distance from/to the warehouse. The model was modified to include those values as a weight. If a forklift works through all the shifts, and all of the weekdays, its cost is close to 8 times that of an employee. So, a cost factor of 8 to 1 is considered if the forklift is needed to bring raw material or finished goods to/from a working station. The risk of an accident would be considered to be 2 to 1, because, if an accident happens, the company would have to send the employee to the hospital, while paying his wages, and hiring an extra employee to cover for the recovery time.

Finally, the cost of receiving raw materials was also considered as the distance from the reception points to the raw materials warehouse. Shipments always arrive on pallets, so they must be moved with a forklift. However, as one pallet contains more than one box that could be sent to the production line, a factor was added, considering how many production orders could be fulfilled with that pallet for each working station. Two main factors were computed: For the Core Builder steps, one raw material shipment could cover 30 days' worth of orders; for the other steps, each shipment would cover eight days. So, factors of 1/30th (0.033) and 1/8th (0.125) were added to the data table.

7. Formulation

It was decided to implement the mathematical model of the p-median problem because the objective of the model is to find the midpoint for a group of clients and their location, so it was extrapolated to a group of workers in a space delimited by the production lines.

A slight modification had to be made to include all those factors in the model. So that each distance would be modified by the risk, cost, and reception factors. The change was made locally, only by computing the cost factors (risk, forklift use, raw materials) and adding it to the main cost. In general, the model remains as a weighted p-median problem, just as the one found in the literature.

$$minimize \sum^{n}_{(i=1)} * \sum^{m}_{(j=1)} x_i * ((d_{ij} (1+ (Cf \times Fl_{ij}) + (Rf \times LC_{ij})) + Rd_i Sf_j Cf + Fd_{ij} (1+Rf \times FLC_{ij} + Cf \times FFl_{ij}))) \quad (2)$$

Subject to

$$\sum^{n}_{(i=1)} x_i = 1 \quad (a)$$

$$\forall i \in \{1..n\}\ x_i \in \{1,0\} \quad (b)$$

$$\forall i \in \{1..n\}, \forall j \in \{1..m\} Fl_{ij},\ LC_{ij},\ FLC_{ij},\ FFL_{ij} \in \{1,0\} \quad (c)$$

Where

n is the number of different options to choose from.

m is the number of working stations or manufacturing steps.

Cf is the cost factor of the forklift.

Fl indicates if the materials for the working station *must* be carried with a forklift.

Rf is the risk cost of crossing a forklift's lane.

LC indicates if the path to the working station must cross or traverse a forklift's lane.

$d$is the travel distance from the warehouse to the working station.

Rd is the distance from the reception point to the raw materials warehouse.

Sf is the number of production runs that can be fed by a single pallet.

Fd is the distance from the working station to the finished goods warehouse.

FLC indicates if a forklift's lane must be crossed with finished goods.

FFL indicates if a forklift must be used for finished goods.

x indicates if the option will be used.

Restriction (a) indicates that one option must be selected.

Restriction (b) states that x_i can only have a value of 1 or 0.

Restriction (c) assures that FL_{ij} (Must use a Forklift), LC_{ij} (Lane Crossing), FLC_{ij} (Lane Crossing for Finished Goods), and FFL_{ij} (Use of Forklift for Finished Goods) have only a value of *1* or *0*.

The model was coded using Lingo® and solved. The code is the following:

1. Algorithm 1: LINGO ® Coded Model

```
2. 1. Model:
3. 2. Sets:
4.    Alternativa: Alt, Rd, x;
5.    Estaciones: WS, Sf;
6.    Arco(Alternativa,Estaciones):Dist, Fl, LC, FDist, FFl, FLC;
7.    EndSets

8. 3. Data:
9.    Alt = @OLE("C:\Tmp\WHR.xls");
10.  Rd =  @OLE("C:\Tmp\WHR.xls");
11.  WS =  @OLE("C:\Tmp\WHR.xls");
12.  Sf =  @OLE("C:\Tmp\WHdR.xls");
13.  Dist= @OLE("C:\Tmp\WHR.xls");
14.  Fl =  @OLE("C:\Tmp\WHR.xls");
15.  LC =  @OLE("C:\Tmp\WHR.xls");
16.  FDist=@OLE("C:\Tmp\WHR.xls");
17.  FFl = @OLE("C:\Tmp\WHR.xls");
18.  FLC = @OLE("C:\Tmp\WHR.xls");
19.  EndData

20. min=@Sum(Alternativa(i):@Sum(Estaciones(j):x(i)*
       (Dist(i,j)*(1+(Cf  *Fl(i,j))+(Rf*LC(i,j))))+(Rd(i) * Sf(i) *
       Cf ) + (FDist(i,j)*(1+ (Cf*FFl(i,j))+(Rf*FLC(i,j)))));
21.  Cf = 7;
22.  Rf = 2;
23.  @Sum(Alternativa(i): x(i) ) = 1;
24.  @For(Alternativa(i): @BIN(x(i)));
25.  @For(Alternativa(i): @For(Estaciones(j):
26.  @BIN( Fl(i,j) );
27.  @Bin(LC(i,j));
28.  @Bin( FFl(i,j));
29.  @Bin( FLC(i,j))
30.  ) );
31.  End
```

8. Results and Discussion

The model considered all the risk, cost, and incoming material cost modifiers. The solution obtained showed that the location that minimized the distance from the warehouse to each of the working stations was location number 2. An extract of the raw results obtained by solving the model is shown in table 2.

Table 2. Raw results of the model solver

Variable	Value	Reduced Cost
CF	7	0
RF	2	0
ALT(1)	1	0
ALT(2)	2	0
ALT(3)	3	0
RD(1)	17	0
RD(2)	47	0
RD(3)	121	0
X(1)	0	22328
X(2)	1	18161
X(3)	0	26156
WS(1)	1	0
WS(2)	2	0
WS(3)	3	0
WS(4)	4	0
WS(5)	5	0
WS(6)	6	0
WS(7)	7	0
WS(8)	8	0
WS(9)	9	0
WS(10)	10	0

9. Conclusions and future research

Logistics within the factory add no value to the finished goods, but wrong logistic management can become costly if accidents happen, whether to people, to the facilities, or the finished goods. The proposed location isolates the forklift pathways, mostly for finished goods, and minimizes the cost of moving both raw materials and finished goods from and to the warehouses. It is expected that, even in full production mode, the need for forklifts is kept below the line that requires more than one device. It is also expected that the manual cart movement is reduced and that the crossing of forklift lanes is also kept at a minimum. Both the company and the authors expect that both the raw materials and the finished goods warehouses can be operated by only one operator, as well as the raw materials receiving process. It would, in turn, keep the operating costs at a level similar to the present, where test runs are being performed, but when in full production mode.

Future work in this area could include more general computing of the costs associated with each physical movement. Nevertheless, the company still needs improving in many areas, like minimizing idle time, reducing cooling and heating of the furnaces to make more efficient use of energy, thus reducing costs, optimizing the use of significant items, reducing even further the need to return big items to the raw materials warehouse, and reorganizing the items within each warehouse.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] R. Ballou, *Logística administración de la cadena de suministro*. México: Pearson Educación de México, S. A. de C. V., 2004.

[2] J. Bartholdi and S. Hackman, "Allocating space in a forward pick area of a distribution center for small parts", *IIE Transactions*, vol. 40, no. 11, pp. 1046-1053, 2008. https://doi.org/10.1080/07408170802167662.

[3] P. Mirchandani and R. Francis, "Discrete location theory", *Discrete Applied Mathematics*, vol. 36, no. 1, pp. 93-94, 1992. https://doi.org/10.1016/0166-218x(92)90210-2.

[4] N. Mladenović, A. Alkandari, J. Pei, R. Todosijević and P. Pardalos, "Less is more approach: basic variable neighborhood search for the obnoxiousp-median problem", *International Transactions in Operational Research*, vol. 27, no. 1, pp. 480-493, 2019. https://doi.org/10.1111/itor.12646.

[5] P. Hansen and B. Jaumard, "Cluster analysis and mathematical programming", *Mathematical Programming*, vol. 79, no. 1-3, pp. 191-215, 1997. https://doi.org/10.1007/bf02614317.
[6] N. Hofstra, B. Petkova, W. Dullaert, G. Reniers and S. de Leeuw, "Assessing and facilitating warehouse safety", *Safety Science*, vol. 105, pp. 134-148, 2018. https://doi.org/10.1016/j.ssci.2018.02.010.
[7] D. Cantor, "Workplace safety in the supply chain: a review of the literature and call for research", *The International Journal of Logistics Management*, vol. 19, no. 1, pp. 65-83, 2008. https://doi.org/10.1108/09574090810872604.
[8] M. Douglas and S. Swartz, "A multi-dimensional construct of commercial motor vehicle operators' attitudes toward safety regulations", *The International Journal of Logistics Management*, vol. 20, no. 2, pp. 278-293, 2009. https://doi.org/10.1108/09574090910981341.
[9] M. Douglas and S. Swartz, "Career stage and truck drivers' regulatory attitudes", *The International Journal of Logistics Management*, vol. 27, no. 3, pp. 686-706, 2016. https://doi.org/10.1108/ijlm-11-2014-0180.
[10] R. Grytnes, H. Shibuya, J. Dyreborg, S. Grøn and B. Cleal, "Too individualistic for safety culture? Non-traffic related work safety among heavy goods vehicle drivers", *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 40, pp. 145-155, 2016. https://doi.org/10.1016/j.trf.2016.04.012.
[11] E. Kemp, S. Kopp and E. Kemp, "Six days on the road", *The International Journal of Logistics Management*, vol. 24, no. 2, pp. 210-229, 2013. https://doi.org/10.1108/ijlm-08-2012-0080.
[12] J. de Vries, R. de Koster, S. Rijsdijk and D. Roy, "Determinants of safe and productive truck driving: Empirical evidence from long-haul cargo transport", *Transportation Research Part E: Logistics and Transportation Review*, vol. 97, pp. 113-131, 2017. https://doi.org/10.1016/j.tre.2016.11.003.
[13] R. de Koster, D. Stam and B. Balk, "Accidents happen: The influence of safety-specific transformational leadership, safety consciousness, and hazard reducing systems on warehouse accidents", *Journal of Operations Management*, vol. 29, no. 7-8, pp. 753-765, 2011. https://doi.org/10.1016/j.jom.2011.06.005.
[14] B. Gaudenzi and A. Borghesi, "Managing risks in the supply chain using the AHP method", *The International Journal of Logistics Management*, vol. 17, no. 1, pp. 114-136, 2006. https://doi.org/10.1108/09574090610663464.
[15] G. Zsidisin, "Managerial Perceptions of Supply Risk", *The Journal of Supply Chain Management*, vol. 39, no. 1, pp. 14-26, 2003. https://doi.org/10.1111/j.1745-493x.2003.tb00146.x.
[16] G. Svensson, "A conceptual framework of vulnerability in firms' inbound and outbound logistics flows", *International Journal of Physical Distribution & Logistics Management*, vol. 32, no. 2, pp. 110-134, 2002. https://doi.org/10.1108/09600030210421723.
[17] R. Dandage, S. Mantha and S. Rane, "Strategy development using TOWS matrix for international project risk management based on prioritization of risk categories", *International Journal of Managing Projects in Business*, vol. 12, no. 4, pp. 1003-1029, 2019. https://doi.org/10.1108/ijmpb-07-2018-0128.
[18] Kirk, D.L., "Enforcement time nears for German Law," Business Insurance, Vol. 33 No. 36, pp. 19-20, 1999.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 3, 583-588 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Solutions for Building a System to Support Motion Control for Autonomous Vehicle

Quach Hai Tho[1,*], Huynh Cong Phap [2], Pham Anh Phuong [3]

[1]*University of Arts, Hue University, 530.000, Vietnam*

[2]*Vietnam - Korea University of Information and Communication Technologies, Da Nang University, 550.000, Vietnam*

[3]*Faculty of Information Technology, University of Education, Da Nang University, 550.000, Vietnam*

ARTICLE INFO

Article history:
Received: 12 May, 2020
Accepted: 14 June, 2020
Online: 26 June, 2020

Keywords:
Autonomous vehicle
Model predictive control
Path planning
Motion planning
Intelligent transportation systems

ABSTRACT

With a model predictive control approach including boundary analysis and uncertain prediction of activities of different road participants, this paper proposes solutions that support motion control by steering control and appropriate acceleration to create safe motion trajectories for an autonomous vehicle. The motion control support element is determined by the principle of minimal intervention and can handle complex situations, while building control model to predict real-time operation with speed factors, ability to control driving and limit the long period. The performance of this solution is assessed through simulation, then there are applied research orientations on practical autonomous vehicle accounting.

1. Introduction

Currently, the manufacturers have equipped with standard safety systems in researching the production of autonomous vehicles. However, in the context of traffic in complex environments, with the development of technology, the safety standards for autonomous vehicles have been raised, in which the driver assistance technology is a problem that needs to be cared about creating safe movements for vehicle. This technology can be mentioned by tracking driving conditions such as road conditions, vehicle status information, sensor technology, etc. to contribute to the development of advanced driver-assistance technology. For example, technology that supports the safety system in the adaptive cruise control system helps maintain distance from the vehicle ahead, collision avoidance assistance system helps predict collisions with obstacles with timely braking, as well as the lane keeping system helps the vehicle recognize the separator to adjust the vehicle direction so that the vehicle is always moving in the middle of the lane.

However, the scenario that these systems handle may be relatively simple compared to the diverse and complex situations we often encounter in traffic environment conditions. To solve this problem, we offer a motion control assistance solution to create a safe motion trajectory for the vehicle. The design of this control support system has two main goals: the first is minimal intervention - that is, applying autonomous control only when necessary, the second is to ensure safety - meaning the vehicle's collision-free state is clearly enforced through optimal constraints.

The control support solution we propose in this article is implemented by predictive control based on the non-linear model predictive control and optimize the steering support system by the steering system along with the acceleration of the vehicle. In the solution assuming the current position of the vehicle, road boundaries, vehicles ahead and uncertain predictions about the future state of the vehicle in the form of a parametrized posterior distribution by their mean and covariance are known. Specifically, for this solution, we will combine the uncertainty changes over time to the mobile obstacle predictions into the optimization problem, and also introduce constraints for boundary limits and moving obstacles while maintaining a vehicle's movement plan for a limited time.

The next section of the paper will introduce the basic principles of predictive control and thereby propose a control solution based on Non-linear Model Predictive Control. Next is the experimental section and conclusion with some suggestions for further research directions for the problem of autonomous vehicles.

2. Building solution to support motion control

Theoretically, in order for an autonomous vehicle to avoid collisions, we need to calculate to find the set of states in which the

Corresponding Author: Quach Hai Tho, Email: qhaitho@hueuni.edu.vn

vehicle may encounter a collision situation when participating in traffic and then control the movement so that the vehicle does not move into that state set. There are many studies [1]-[5] doing this task that have determined the inevitable collision state set or target motion set. However, these studies do not limit the applicability or make assumptions for simple traffic scenarios, so it is difficult to analyze calculations. In this paper, the solution that we propose is the idea of defining collision state set and thereby determining the set of probability constraints to avoid collisions. At the same time, in this solution, we combine the steering speed control and increase/decrease speed so that the collision avoidance effect can be realized better. The ideas have originated from the studies [6,7,8,9] in the process of determining the difference of the steering angle or the deviation of the front wheel to achieve a safe trajectory.

In many studies, the model predictive control (MPC) has been applied to control autonomous vehicle [7]-[11] as the MPC approach to plan motion based on the vehicle's basic movements and track the road to avoid obstacles, or the MPC approach without constraint conditions and determine the stability of the vehicle with the surrounding environment to provide a safe steering angle at constant speed in a discrete environment. In this article, the MPC approach that we use can handle complex situations with steering control, increase/decrease vehicle speed, and avoid mobile obstacles at a certain extent in an uncertain environment.

The cost function of most MPC methods [7], [11]-[13] often depends on the time and the constraints on the path, so it needs to be pre-optimal or specific time steps or generated out and track the fixed motion of the vehicle, this leads to differences in result from the initial conditions of optimization that could result in invalid linear constraints and unpredictable motion planning. Therefore, we apply the point processing method in [14] and directly solve the nonlinear model predictive control problem by focusing on providing all costs and constraints for the decoding set that does not need to be linear manually.

2.1. Statement of problems

This problem is built on two basic principles: the first is minimal intervention, the second is to ensure safety, meaning the probability of a collision involving the surrounding environment and other traffic objects must be below certain thresholds. And this problem is done in discrete time intervals $k \triangleq t_k$, with $t_k = t_0 + \sum_{i=1}^{k} \Delta t_i$ (t_0 is the current time, Δt_i is the i time step of the plan).

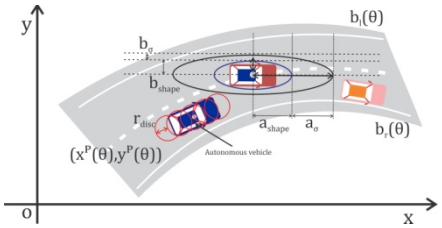


Figure 1. Modeling of autonomous vehicle and other vehicles

At interval k, with values of position $s_k = (x_k, y_k)$, linear velocity v_k, direction of vehicle $\emptyset_k$ and steering angle δ_k, the mathematical model of the vehicle is determined by the set $\boldsymbol{z}_k = [s_k, \emptyset_k, \delta_k, v_k] \in \mathbb{Z}$.

Assuming $\mathcal{B}(\boldsymbol{z}_k) \subset \mathbb{R}^2$ is the occupied area of the vehicle at $\boldsymbol{z}_k$ state (illustrated in Figure 1, with autonomous vehicle modeled by linked circles), the input values on the control using $\dot{\delta}_k$ steering speed and a_k acceleration is assigned by $\boldsymbol{u}_k = [u_k^{\delta}, u_k^{a}] \in \mathcal{U}$.

Thus, the future state of the vehicle is represented by a discrete dynamic system as follows:

$$\boldsymbol{z}_{k+1} = f(\boldsymbol{z}_k, \boldsymbol{u}_k) \tag{1}$$

With other objects in traffic such as different kinds of car, bicycles, pedestrians, etc. called other objects will be assigned by the index $i = \{1, \dots, n\}$, the input control and their configuration parameters are determined by the values of $\boldsymbol{z}_k^i \in \mathbb{Z}_i$ and $\boldsymbol{u}_k^i \in \mathcal{U}_i$.

In order to combine uncertainty, it is necessary to assume that the future state distributions of the objects at m future time intervals is known. They are parameterized to the average state value $\boldsymbol{z}_{1:m}^i$ and covariance $\boldsymbol{\sigma}_{1:m}^i$. Therefore, the degree of uncertainty in the forecast can be reflected through the covariance value $\boldsymbol{\sigma}_{1:m}^i$.

At each given state, each object will occupy one space $\mathcal{B}^i(\boldsymbol{z}_k^i, \boldsymbol{\sigma}_k^i, p_\epsilon) \subset \mathbb{R}^2$ having probability greater than p_ϵ (p_ϵ is the acceptable probability of collision that may happen), the model of the object and this occupied space are shown in Figure 1.

The free space defined in this problem is the workspace $\mathcal{W} = \mathbb{R}^2$ and the location map of obstacles $\mathcal{O} \subset \mathcal{W}$ containing static obstacles such as limit of roads and systems of seperator, etc., at the same time, the surrounding environment $\mathcal{E}(k)$ is determined to be the state of other objects (including means of traffic, obstacles) at the time k.

In this study, with the set of states $\boldsymbol{z}_{0:m} = [\boldsymbol{z}_0, \dots, \boldsymbol{z}_m] \in \mathbb{Z}^{m+1}$ and input set $\boldsymbol{u}_{0:m-1} = [\boldsymbol{u}_0, \dots, \boldsymbol{u}_{m-1}] \in \mathcal{U}^m$, we will build a general discrete time constraint optimization at m time steps with a time limit $\tau = \sum_{k=1}^{m} \Delta t_k$.

Thus, the goal of the solution is to calculate the optimal input values $\boldsymbol{u}_{0:m-1}^*$ for autonomous vehicle with minimizing cost function $\hat{J}_h(\boldsymbol{u}_{0:m-1}, \boldsymbol{u}_0^h) + \hat{J}_t(\boldsymbol{z}_{0:m}, \boldsymbol{u}_{0:m-1})$.

In which: $\hat{J}_h(\boldsymbol{u}_{0:m-1}, \boldsymbol{u}_0^h)$ is the minimum cost to minimize deviations from the input value $\boldsymbol{u}_0^h$, and $\hat{J}_t(\boldsymbol{z}_{0:m}, \boldsymbol{u}_{0:m-1})$ is the cost depending on the properties of the trajectory according to the motion plan.

This optimum problem follows a set of constraints: the first is to use the vehicle transition model, the second is the non-collision constraint with static obstacles and the third is the probability that no collision will occur p_ϵ with other traffic participants.

And the optimal trajectory of the vehicle is given as follows:

$$\boldsymbol{u}_{0:m-1}^* = \arg\min_{\boldsymbol{u}_{0:m-1}} \hat{J}_h(\boldsymbol{u}_{0:m-1}, \boldsymbol{u}_0^h) + \hat{J}_t(\boldsymbol{z}_{0:m}, \boldsymbol{u}_{0:m-1}) \tag{2}$$

Where:

$$\boldsymbol{z}_{k+1} = f(\boldsymbol{z}_k, \boldsymbol{u}_k)$$

$$\mathcal{B}(\boldsymbol{z}_k) \cap \mathcal{O} = \emptyset$$

$$\mathcal{B}(\boldsymbol{z}_k) \cap \cup_{i \in \{1,\dots,n\}} \mathcal{B}^i(\boldsymbol{z}_k^i, \boldsymbol{\sigma}_k^i, p_\epsilon) = \emptyset$$

$\forall k \in \{0, \dots, m\}$

$\boldsymbol{z}^i_{0:m}$, $\boldsymbol{\delta}^i_{0:m}$ with $i = 1, \dots, m$: are parameters for all other traffic objects, $\boldsymbol{z}_0$ is the initial state of vehicle.

2.2. Building a solution

The solution was built to calculate the creation of a safe motion trajectory with a predefined horizon. Optimal problems are constrained by cost values, transition models, maintaining vehicle movement within the boundaries of the lane median, and avoiding collisions with other road participants that must ensure the probability of collision below the value p_ϵ.

In solutions using model predictive control of previous studies [7,11,13], the researchers often use factors such as constant longitudinal speed and small angle assumptions in scenarios avoiding obstacles on the straight road. In this study, we will consider the impact of longitudinal speed to ensure overall safety in complex traffic environments.

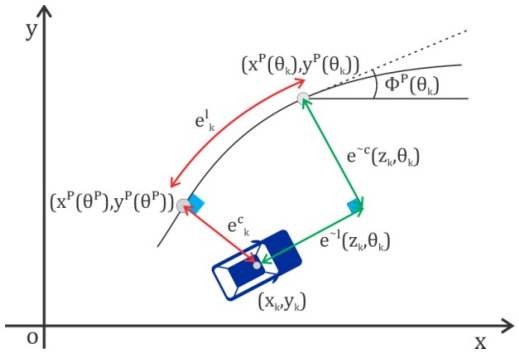


Figure 2. Vehicle dynamics model

The vehicle model in this solution has been introduced with a fixed rear wheel and the motion control is located on the front wheel with $\boldsymbol{z}$ and $\boldsymbol{u}$ states set here, and the control of rear wheel movement is via center bridge with distance L and a continuous dynamic model. This motion model is described by a discrete time model with integrals $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \int_k^{k+\Delta t_k} \dot{\boldsymbol{z}} \mathrm{d}t = f(\boldsymbol{z}_k, \boldsymbol{u}_k)$, as follows:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\emptyset} \\ \dot{\delta} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v\cos(\emptyset) \\ v\sin(\emptyset) \\ \frac{v}{L}\tan(\delta) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u^\delta \\ u^a \end{bmatrix} \tag{3}$$

Where: $\dot{z} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\emptyset} \\ \dot{\delta} \\ \dot{v} \end{bmatrix}$ and $u = \begin{bmatrix} u^\delta \\ u^a \end{bmatrix}$

In this model, the limits applied include steering angle $\|\delta\| \leq \delta_{max}$, steering speed $\|u^\delta\| \leq \dot{\delta}_{max}$ and longitudinal speed $v \leq v_{max}$. The introduction of these limits is consistent with vehicle performance and road traffic rules, as well as a number of restrictions will be set to ensure safety such as speed limits when moving into the corners by limiting the slip coefficient $\|\emptyset\| \leq \emptyset_{max}$ and limiting the maximum speed quickly by the acceleration limit $u^a \in [a_{min}, a_{max}]$.

The model predictive control method in this study is based on model predictive contouring control method [15,16,17] and is used to solve problems, in this study, it is not required that the motion trajectory of the vehicle must be exactly according to the reference trajectory, but the motion trajectory must be in the safe range.

On the reference trajectory, the process takes place as follows:

At time k, the vehicle is in position $s_k = (x_k, y_k)$, the process of following the reference path is limited and continuously differentiable in geometric space $\left(x^P(\theta), y^P(\theta)\right)$ of path parameter θ, with the tangential vector $\boldsymbol{t}$ and the normal vector $\boldsymbol{n}$ as follows:

$$\boldsymbol{t} = \begin{bmatrix} \frac{\partial x^P(\theta)}{\partial \theta} \\ \frac{\partial y^P(\theta)}{\partial \theta} \end{bmatrix}, \boldsymbol{n} = \begin{bmatrix} -\frac{\partial y^P(\theta)}{\partial \theta} \\ \frac{\partial x^P(\theta)}{\partial \theta} \end{bmatrix} \tag{4}$$

and the path instructions are described as follows:

$$\emptyset^P(\theta_k) = \arctan\left(\frac{\partial y^P(\theta)}{\partial x^P(\theta)}\right) \tag{5}$$

the path parameterized according to the arc length $\left(\frac{\partial \theta}{\partial s} = 1\right)$ will allow to estimate the vehicle's progress with v_k on the reference path and the actual path $s = \int v \, \mathrm{d}t$ if during the parameterization of the curves the arc length is negligible and the distance between the points is small compared to the arc length. At the same time, since the motion trajectory of the vehicle will follow a certain path and have a slight deviation from the reference trajectory determined by the road boundary, we can assume that the trajectory remains the same offset, so: $\Delta\theta \approx \Delta s = v\Delta t$, with this additional assumption will generate an approximate process according to the path parameter, as follows:

$$\Delta\theta_{k+1} = \Delta\theta_k + v_k \Delta t_k \tag{6}$$

where $v_k \Delta t_k$ describes the approximation process in time step k.

In the general case, finding the path parameter $\theta^P(x_k, y_k)$ of the nearest point $S(x_k, y_k)$ on the reference path is not feasible and not suitable for fast optimization. Therefore, the value of $\theta^P(x_k, y_k)$ will be approximated according to equation (6).

During the motion planning process, if the actual path deviates from the reference path, the delay error from the first approximation point in time progression to the next point and the position error referenced to the horizontal roads θ_k along the tangential path $\boldsymbol{t}_k$ are defined as follows:

$$e^{\sim l}(\boldsymbol{z}_k, \theta_k) = \frac{\boldsymbol{t}_k^T}{\|\boldsymbol{t}_k\|} \begin{bmatrix} x_k - x^P(\theta_k) \\ y_k - y^P(\theta_k) \end{bmatrix} = -cos\emptyset^P(\theta_k)\left(x_k - x^P(\theta_k)\right) - sin\emptyset^P(\theta_k)\left(y_k - y^P(\theta_k)\right) \tag{7}$$

If the delay error $e^{\sim l}(\boldsymbol{z}_k, \theta_k)$ is small, the process of constructing an approximate path will be as close to the horizontal asymptote as $\Delta\theta \approx \Delta s = v\Delta t$ and $\theta_k \approx \theta^P(x_k, y_k)$. Also, in the process of optimizing the prediction control, the delay error should be actively handled so that the error of the predictive process θ_k is small enough according to the motion planning process.

When projected on the standard path between the actual position and the predicted position, we will determine the contouring error by the deviation of these two positions, as follows:

$$e^{\sim c}(\boldsymbol{z}_k, \theta_k) = \frac{\boldsymbol{n}_k^T}{\|\boldsymbol{n}_k\|}\begin{bmatrix} x_k - x^P(\theta_k) \\ y_k - y^P(\theta_k) \end{bmatrix} = sin\emptyset^P(\theta_k)\big(x_k - x^P(\theta_k)\big) - cos\emptyset^P(\theta_k)\big(y_k - y^P(\theta_k)\big) \quad (8)$$

and the contouring error is a standard for establishing good motion planning when the vehicle is in motion not deviated from a given reference path. Therefore, the cost function of predictive state control is built based on the balance between the contouring error factors $e^{\sim c}(\boldsymbol{z}_k, \theta_k)$, delay error $e^{\sim l}(\boldsymbol{z}_k, \theta_k)$ and the process of building approximate path v_k to achieve the best combination, as follows:

$$J_{av}(\boldsymbol{z}_k, \theta_k) = \boldsymbol{e}_k^T Q \boldsymbol{e}_k - v_k \quad (9)$$

with path error vector formed from delay error and contouring error as follows:

$$\boldsymbol{e}_k = \begin{bmatrix} e^{\sim l}(\boldsymbol{z}_k, \theta_k) \\ e^{\sim c}(\boldsymbol{z}_k, \theta_k) \end{bmatrix} \quad (10)$$

For the performance of the path, all reference paths are parameterized by C^1 - a continuous clothoid path in a system of paths through predetermined points. The clothoid path estimate will be replaced by the third-order spline function of equally spaced nodes and parameterization of the spline functions along the arc length is sufficiently accurate, as well as achieves good performance in the calculation. Because the spline functions provide an analytical parameter about the reference path, the limit of the path and the derivatives needed to solve the nonlinear optimization problem.

At position $S(x_k, y_k)$, from the projection along the normal of the reference path at the actual culvilinear abscissa θ^P, we get the lateral distance $d(\boldsymbol{z}_k, \theta)$ with the reference path and approx slag equal to θ_k such that $d(\boldsymbol{z}_k, \theta_k) = e^{\sim c}(\boldsymbol{z}_k, \theta_k)$. Just as the movable space of autonomous vehicles at the crossroads θ_k is limited by the left boundary $b_l(\theta_k)$, the right boundary $b_r(\theta_k)$ of the lane and other static obstacles will be parameterized by the third order spline function to allow analytical evaluation and derivation.

As such, the horizontal traverse for the path is limited as follows:

$$b_l(\theta_k) + w_{max} \le d(\boldsymbol{z}_k, \theta_k) \le b_r(\theta_k) - w_{max} \quad (11)$$

where: w_{max} is the upper limit of the vehicle position projection on the standard reference path and this value is greater than $\frac{1}{2}$ of the vehicle width. At the same time, maintaining the validity of w_{max} as a limit just taking the radius of the vehicle's operating range as an upper limit is enough to ensure safety.

Because the relationship between the path and the direction of the vehicle is relative, we need to establish a constraint between the direction of the path $\emptyset^P(\theta_k)$ and the movemen direction of the vehicle $\emptyset_k$, as follows:

$$\|\emptyset_k - \emptyset^P(\theta_k)\| \le \Delta\emptyset_{max} \quad (12)$$

In the solution of this study, other traffic participants will be described by an orientation ellipse $\emptyset$ with a_{shape} being a semi-major axis and b_{shape} being a semi-minor axis of the ellipse in the longitudinal and horizontal directions of the objects.

The development of future trajectories of these objects is assumed to have some uncertain positions of posterior distribution and is parameterized by the average trajectory $\boldsymbol{z}_{0:m-1}^i$ and degree uncertainty $\boldsymbol{\sigma}_{0:m-1}$, as follows:

$$\boldsymbol{\sigma}_{k+1} = \boldsymbol{\sigma}_k + \boldsymbol{\sigma}\Delta t_k \quad (13)$$

In the general case, we propose a model for generating indeterminate positions of vehicles with uncertainty $\boldsymbol{\sigma}_k = \left[\sigma_k^a, \sigma_k^b\right]^T$ at the time k and $\boldsymbol{\sigma} = [\sigma^a, \sigma^b]^T$ is the uncertainty incurred. Therefore, the value of the variance is determined to approximate to adjust the direction of the vehicle motion aligning the main axis of the surrounding ellipse. The generation of indeterminate positions in the horizontal direction of the vehicle is limited by a maximum value to consider the maximum rationality of the maximum flow of vehicles currently in the current lanes.

The level-set of Gaussian $\mathcal{N}\big(0, diag(\boldsymbol{\sigma}_k)\big)$ describe the indefinite position of other road users at p_ϵ degree and ellipses formed with coefficients are set as follows:

$$\begin{bmatrix} a_{\boldsymbol{\sigma}_k} \\ b_{\boldsymbol{\sigma}_k} \end{bmatrix} = \begin{bmatrix} \sigma_k^a \\ \sigma_k^b \end{bmatrix}\left(-2log\left(p_\epsilon 2\pi\sigma_k^a\sigma_k^b\right)\right)^{1/2} \quad (14)$$

Therefore, we can use the main axis direction and add coefficients to the cross axles of the vehicle to identify the area of the obstacle with the probability of occupancy above the threshold p_ϵ. At the same time, the rectangular area used to determine the position and occupied area of autonomous vehicles will be replaced by a set of circulars with radius r_{av}. The use of circles instead of ellipses represents obstacles because the presence of an autonomous vehicle does not need to be aligned in a straight axis, and Minkowski's sum calculation efficiency is not possible if shown by ellipses whose axes do not fit tightly. The Minkowski sum of the sets of circles surrounding the vehicle and the collision constraints shown by the previous displaced ellipses are calculated as follows:

$$c_k^{obs,i}(\boldsymbol{z}_k) = \begin{bmatrix} \Delta x_j \\ \Delta y_j \end{bmatrix}^T R(\emptyset)^T \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{bmatrix} R(\emptyset) \begin{bmatrix} \Delta x_j \\ \Delta y_j \end{bmatrix}\Bigg|_{k,i} > 1 \quad (15)$$

$$\text{with } \forall j \in \{1, \dots, 4\}$$

where Δx, Δy are the distances from the circle set covering the area around the vehicle to the center of obstacle i at the time k and $R(\emptyset)$ is the rotation matrix corresponding to the motion direction of the obstacle, and the semi-major axis of the constraint ellipse. The result is calculated as follows:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a_{shape} + a_{\boldsymbol{\sigma}_k} + r_{disc} \\ b_{shape} + b_{\boldsymbol{\sigma}_k} + r_{disc} \end{bmatrix} \quad (16)$$

Thus, we have a collision-free constraint with a higher probability p_ϵ than other means.

As our goal of the solution in this research is to minimize the intervention of the driving, that is, the control system only intervenes with the steering operation when really necessary with the minimum intervention time:

$$J_h(\boldsymbol{z}_k, \boldsymbol{u}_k, \boldsymbol{u}_0^h) = \begin{bmatrix} u_k^a - a_0^h \\ \delta - \delta_0^h \end{bmatrix}^T K \begin{bmatrix} u_k^a - a_0^h \\ \delta - \delta_0^h \end{bmatrix} \quad (17)$$

where $\boldsymbol{u}_0^h = [\delta_0^h, a_0^h]^T$ is the inconsistent value of system state, δ^h is the steering angle value and a^h is the acceleration value at the time t_k.

In the process of setting values for control inputs, we only determine the steering angle value δ^h and acceleration a^h without determining the steering speed value $\dot{\delta}^h$. However, in the general case if the front view angle can be controlled, the speed value is still used as the input value, and the control process $\boldsymbol{u}_k$ maintains the steering speed and acceleration of the vehicle.

The calculation of the trajectory cost will include the cost of the model predictive control calculated by the equation $\boldsymbol{e}_k = \begin{bmatrix} e^{\sim l}(\boldsymbol{z}_k, \theta_k) \\ e^{\sim c}(\boldsymbol{z}_k, \theta_k) \end{bmatrix}$, and at the same time add input control deviations and slip coefficients to create smooth and comfortable trajectories when the vehicle is in motion. The weights R and A allow for sorting based on different priorities, so the trajectory cost is calculated as follows:

$$J_t(\boldsymbol{z}_k, \boldsymbol{u}_k, \theta_k) = J_{av}(\boldsymbol{z}_k, \theta_k) + \boldsymbol{u}_k^T R \boldsymbol{u}_k + \dot{\emptyset}_k A \dot{\emptyset}_k \quad (18)$$

where $J_{av}(\boldsymbol{z}_k, \theta_k)$ has transformed the inconsistency from the reference path error into a better direction of motion.

Finally, the solution proposed is the optimization problem. It is done by the minimum factor combining linear between the cost of intervention into the system and trajectory cost, as follows:

$$J_{av}(\boldsymbol{z}_k, \boldsymbol{u}_k, \theta_k, \boldsymbol{u}_0^h) = \beta\omega(t_k) J_h(\boldsymbol{z}_k, \boldsymbol{u}_k, \boldsymbol{u}_0^h) + \big(1 - \omega(t_k)\big) J_t(\boldsymbol{z}_k, \boldsymbol{u}_k, \theta_k) \quad (19)$$

in which, the weight β and exponential decay function $\omega(t_k) = exp(-\alpha t_k)$ are used to enhance the input value for the system.

We have chosen the solution to make the weight β reach a high value, so that when moving forward in the predictive model, the system will be able to respond well to inputs but still depend on J_t. By doing so, the solution presented in this study can plan a full implementation without having to predict the planned motion trajectory. Therefore, the nonlinear optimization problem with constraints on state, dynamics, paths and obstacles is constructed as follows:

$$\boldsymbol{u}_{0:m-1}^* = \underset{\boldsymbol{u}_{0:m-1}}{\arg\min} \sum_{k=1}^{m} J_{av}(\boldsymbol{z}_k, \boldsymbol{u}_k, \theta_k, \boldsymbol{u}_0^h) \Delta t_k \quad (20)$$

where:

$$\boldsymbol{z}_{k+1} = f(\boldsymbol{z}_k, \boldsymbol{u}_k) ;$$

$$\theta_{k+1} = \theta_k + v_k \Delta t_k ;$$

$$\boldsymbol{z}_k \in [\boldsymbol{z}_{min}, \boldsymbol{z}_{max}] ;$$

$$\boldsymbol{u}_k \in [\boldsymbol{u}_{min}, \boldsymbol{u}_{max}];$$

$$\|\dot{\emptyset}_k\| < \dot{\emptyset}_{max} ;$$

$$\|\emptyset_k - \emptyset^P(\theta_k)\| < \Delta\emptyset_{max} ;$$

$$d(\boldsymbol{z}_k, \theta_k) \in [b_l(\theta_k) + \omega_{max}, b_r(\theta_k) - \omega_{max}] ;$$

$$c_k^{obs,i}(\boldsymbol{z}_k) > 1, \; i = \{1, \dots, n\}, \forall k \in \{0, \dots, m\}$$

with initial initialization values for the path $\big(x^P(\theta), y^P(\theta)\big)$, the left boundary $b_l(\theta)$ and the right boundary $b_r(\theta)$ set by the arc and the static obstacles that the vehicle is moving. At the same time, in each loop when the system executes the initial states $\boldsymbol{z}_0$ and θ_0, the input variables $\boldsymbol{u}_0^h$ and predicts other objects of traffic $\boldsymbol{z}_{0:m}^i$ $\boldsymbol{\sigma}_{0:m}^i$ will be given a predictive model for the control system. After that, the optimal solution given by equation (20) and the optimal control $\boldsymbol{u}_0^*$ will be executed by the system.

3. Experimental results

To test and evaluate the proposed solution, we have conducted empirical simulation of processes in the matlap environment. At the same time, in order to ensure the objectivity and reliability when evaluating, we have conducted simulations with different scenarios and autonomous vehicles controlled moving with steering angle δ_0^h and desired acceleration a_0^h. Input variables will be handled using model predictive control to ensure generating safe movement, the reference path and the left boundary b_l, right boundary b_r are designed and determined accordingly with the road system. During the experiment, the calculations uses SI measurement system with sampling interval of 0.1s, the trajectory is transferred to I/O controller with simulation time of 0.005s.

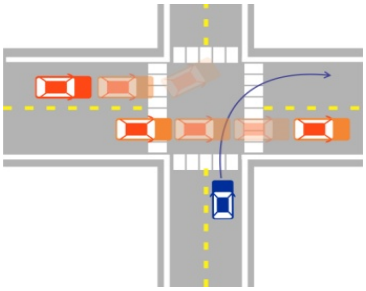


Figure 3. Simulating the scenario 1

Scenario 1: In this scenario, the autonomous vehicle will move into a corner, with the input values for the control system will be able to make the vehicle's direction of travel out of the limit of traffic lanes. At this point, the driver assistance system will perform braking to reduce the vehicle speed to the safe speed limit with the coefficient constraint before the car enters a corner, then increase the vehicle speed after going out of the corner to ensure the progress in the motion planning. With such a driver assistance process, it shows the advantages of controlling the longitudinal force, horizontal force and acceleration the vehicle, as well as implementing the vehicle's movement plan that is complete when entering corners with the brake and increase/decrease speed operations.

Scenario 2: In this scenario, the autonomous vehicle will implement the movement plan from the slip road and turn left to enter the main traffic lane, the initial positions and the speed of

other traffic participants are initialized randomly. To increase the time span of the planning without adjusting the calculations, we will take the approach of changing the distance at each step as follows: The first 40 steps have a distance of $\Delta t_k = 0.1s$ and the next 50 steps have $\Delta t_k = 0.4s$, which leads to a planning time of 24 seconds for all calculations to be performed in real time.

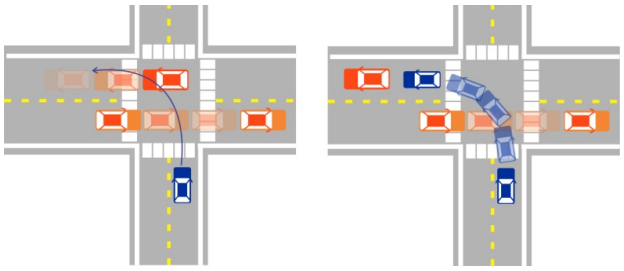


Figure 4. Simulating the scenario 2

The process of following this scenario will be repeated many times with randomly generated input values and there are times when the input values are insecure input to the system, which will lead to accident when the vehicle is moving as planned. However, with the control support system that our solution offers, it is possible to prevent incidents in the most likely cases of accidents.

Specifically: The planning process of movement when the vehicle is moving in a slip road before entering the main road, a collision may occur with the right boundary of the lane. After that, to move into the main lane, the vehicle turns left, at this time on the main lane, other vehicles are moving complicatedly, and unsafe situations may occur. To handle this movement of the vehicle, the control support system will activate the brakes to bring the vehicle's motion status to a stop state so that other vehicles in traffic can move through it, until the safe distance between the autonomous vehicle and other vehicles is large enough, the system activates the vehicle velocity so that the vehicle can move over the cross and merge into the main lane.

During the experiment, we can see that the uncertainty in the predictive of other vehicles is very important because the future states of the predictive model can deviate from the desired predictive point. In the case of omitting uncertainty, the motion planning process needs to provide more precise and specific constraints to ensure the vehicle safety.

4. Conclusions

This paper has proposed a motion control assistance solution to ensure the safety of an autonomous vehicle. The optimal feature of this solution is to shorten the motion planning cycle to minimize deviations from inputs of the prediction while ensuring movement safety. This controller support solution is only implemented in situations where a complex vehicle movement scenario is likely to have a collision with realistic warning elements.

The main idea of this technical solution will support designing an autonomous vehicle with a safe stop whatever the current vehicle control. The experimental simulation with the given scenarios shows that the safety factor can be achieved by calculating to consider all possible possibilities of other vehicles.

In the future, in order that this solution will be more reliable, we have experimented the settings by transferring the simulation to the real environment with experimental vehicle equipped with sensors. When experimenting on reality, it will add a number of factors to analyze the stability of the system so that the behavior of traffic participants is more accurately forecasted. The extensive implementation of this driver-assistance solution for semi-autonomous or autonomous vehicles in vehicle control systems will be able to minimize the amount of damage and create a safe movement plan for the future.

References

[1] A. Bautin, L. Martinez-Gomez, T. Fraichard, "Inevitable Collision States: A probabilistic perspective," *IEEE International Conference on Robotics and Automation 2010*,pp.4022–4027,2010,DOI: 10.1109/ROBOT.2010.5509233.

[2] D. Althoff, M. Althoff, D. Wollherr et al, "Probabilistic collision state checker for crowded environments," *IEEE International Conference on Robotics and Automation - 2010*, pp. 1492–1498, 2010, DOI: 10.1109/ROBOT.2010.5509369.

[3] D. Hoehener, G. Huang, D. D. Vecchio, "Design of a lane departure driver-assist system under safety specifications," *IEEE 55th Conference on Decision and Control (CDC)-2016*, pp. 2468–2474, 2016, DOI: 10.1109/CDC.2016.7798632.

[4] M. Forghani, J. M. McNew, D. Hoehener et al, "Design of driver-assist systems under probabilistic safety specifications near stop signs," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 43–53, 2016, DOI: 10.1109/TASE.2015.2499221.

[5] T. Fraichard, H. Asama, "Inevitable collision states. A step towards safer robots?" *Advanced Robotics*, vol. 18, no. 10, pp. 1001–1024, 2003, DOI: 10.1109/IROS.2003.1250659.

[6] J. Alonso-Mora, P. Gohl, S. Watson et al, "Shared control of autonomous vehicles based on velocity space optimization,"*IEEE International Conference on Robotics and Automation (ICRA)-2014,* pp. 1639–1645, 2014, DOI: 10.1109/ICRA.2014.6907071.

[7] S. Erlien, S. Fujita, J. C. Gerdes, "Shared steering control using safe envelopes for obstacle avoidance and vehicle stability," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 441–451, 2015, DOI: 10.1109/TITS.2015.2453404.

[8] V. A. Shia, Y. Gao, R. Vasudevan et al, "Semiautonomous vehicular control using driver modeling," *IEEE Transactions on Intelligent Transportation Systems*,vol.15,no.6,pp.2696–2709,2014, DOI: 10.1109/TITS.2014.2325776.

[9] Y. Gao, A. Gray, A. Carvalho et al, "Robust nonlinear predictive control for semiautonomous ground vehicles," in *2014 American Control Conference*, pp. 4913–4918, 2014, DOI: 10.1109/ACC.2014.6859253.

[10] A. Gray, Y. Gao, T. Lin et al, "Predictive control for agile semi-autonomous ground vehicles using motion primitives," *American Control Conference (ACC)-2012*, pp. 4239–4244, 2012, DOI: 10.1109/ACC.2012.6315303.

[11] S. J. Anderson, S. B. Karumanchi, K. Iagnemma, "Constraint based planning and control for safe, semi-autonomous operation of vehicles," *IEEE Intelligent Vehicles Symposium-2012*, pp. 383–388, 2012, DOI: 10.1109/IVS.2012.6232153.

[12] B. Paden, S. Z.Yong, D. Yershov et al, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, 2016, DOI: 10.1109/TIV.2016.2578706.

[13] S. J. Anderson, S. C. Peters, T. E. Pilutti et al, "An optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios," *International Journal of Vehicle Autonomous Systems*, vol. 8, no. 2-4, pp. 190–216, 2010, DOI: 10.1504/IJVAS.2010.035796.

[14] A. Domahidi, J. Jerez ,"FORCES Pro," embotech GmbH (http://embotech.com/FORCES-Pro), 2014.

[15] A. Liniger, A. Domahidi, M. Morari, "Optimization-based autonomous racing of 1:43 scale RC cars ," *Optimal Control Applications and Methods*, vol. 36, no. 5, pp. 628–647, 2017, DOI: 10.1002/oca.2123.

[16] D. Lam, C. Manzie, M. C. Good, "Model predictive contouring control for biaxial systems," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 2, pp. 552–559, 2013, DOI: 10.1109/TCST.2012.2186299.

[17] T. Faulwasser, B. Kern, R. Findeisen, "Model predictive pathfollowing for constrained nonlinear systems," in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 8642–8647, 2009, DOI: 10.1109/CDC.2009.5399744.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 592-600 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Review of Pedagogical Principles of Cyber Security Exercises

Mika Karjalainen*, Tero Kokkonen*

Institute of Information Technology, JAMK University of Applied Sciences, 40100, Jyväskylä, Finland

ARTICLE INFO

Article history:
Received: 10 August, 2020
Accepted: 15 September, 2020
Online: 05 October, 2020

Keywords:
Cyber Security Exercise
Expert Performance
Simulation Pedagogy

ABSTRACT

Modern digitalized cyber domains are extremely complex ensemble. Cyber attacks or incidents against system may affect capricious effects for another system or even for physical devices. For understanding and training to encounter those effects requires an effective and complex simulation capability. Cyber Security Exercises are an effective expedient for training and learning measures and operations with their outcomes in that complex cyber domain. Learning in cyber security exercises is relevant for different level actors in organisation hierarchy. Technical experts are able to train the technical capabilities whereas decision makers are able to train the decision-making capabilities under hectic cyber incident. In this paper, the pedagogical aspects of cyber security exercises are discussed in accordance with the law of the lifecycle of the cyber security exercise: planning phase, implementation phase, and feedback phase.

1 Introduction

This research is an extension of work originally presented in 2019 workshop on Cyber Range Technologies and Applications (CACOE 2019) organized in conjunction with 2019 IEEE European Symposium on Security and Privacy (EuroS&P 2019) [1]. This research is expanded from the original as follows: discussion about pedagogical theories and cyber-arena concept for complex environment simulation with the more detailed extended analysis of pedagogical aspects of the cyber security exercises and assessment of the exercise target audience.

Global digitalisation and networked systems have raised new threats. Modern digitalised cyber domains are extremely complex and forms incalculable reliance. That change in digital environment has reflected to the requirements of training and education. Traditionally, exercises are used in a military context to gain better performance for certain tasks. In the cyber domain and especially in the context of cyber resilience, the most valuable assets are personal skills. Those skills are trained efficiently with cyber security exercises.

Cyber security strategy of Finland [2] states that in the critical cyber competence areas, the high level of required training is confirmed by both national and international exercises. The significance of cyber security exercises is also observed in the cyber strategy of the United States of America [3] and the cyber security strategy of the European Union [4]. In addition, cyber security exercises are recognized as an important part of personnel training in commercial organisations, especially in the critical infrastructure organisations, for example, electricity companies [5]. There are several different cyber security exercises conducted globally, for example, the report [6] consists of a dataset of more than 200 cyber security exercises and the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) highlights several exercises they organise or contribute [7].

There exist frameworks categorizing the required skills of personnel in organisations . National Institute of Standards and Technology (NIST) has published document called National Initiative for Cyber security Education (NICE) Cyber security Workforce Framework (NICE Framework) as a reference structure that designates the complex essence of the knowledge, skills, and abilities (KSAs) required in the different roles and tasks of the work within cyber security [8, 9]. There are also frameworks for curricula of education: Computing Curricula 2020 (CC2020) forms a guideline for academic degree programs in computing) [10]. When discussing cyber security, the CC2020 refer to the Cyber security Curricula 2017 (CSEC2017) that form curriculum guidelines for degree programs in cyber security [11].

Simulations have been widely used for study experts [12]. Education in the engineering sciences relies heavily on hands-on training, i.e. applying learned phenomena in practice. In that sense, different learning environments, simulators and test-beds have a remarkable role in the engineering education. When discussing cyber security exercises, the extremely important component is the exercise platform that simulates the cyber domain. Traditionally that kind of platform is called cyber range. Cyber range is executed as a technical platform for exercises that mimic the required networks and systems. As exercise platform, a cyber range, is required

*Corresponding Authors: Mika Karjalainen, mika.karjalainen@jamk.fi and Tero Kokkonen, tero.kokkonen@jamk.fi

to be closed and totally controlled to allow risk free usage of real attacks and intrusions [13]–[14]. The term cyber range originates from the similarity of the kinetic ranges with potential to improve competence or capability with weapons, operations, tactics and techniques [15]. As stated in [16], there exist several diverse cyber ranges globally that vary from enormous virtual-Internets to simple laboratory based test-beds. Because the spectrum of cyber ranges is so multifarious, authors of [16] introduced the concept of Cyber Arena for the simulation of realistic complex cyber-physical domain with unexpected dependencies between networks and systems.

JYVSECTEC (Jyväskylä Security Technology) is JAMK University of Applied Sciences Institute of Information Technology based cyber security focused research, development and training center that offers information and cyber security services [17]. JYVSECTEC has extensive experience for organising cyber security exercises for both national security authorities and private companies of critical infrastructure. Since 2013, Finland's national cyber security exercise has been organised annually by JYVSECTEC [18]. JYVSECTEC has also been Finland's representative in the Cyber Defence Pooling & Sharing Project of European Defence Agency (EDA) [19]–[20].

JAMK University of Applied Sciences has organised several different cyber security exercises. During those exercises, more than 1,500 experts have been involved in those learning experiments. In addition, there is an annual course of cyber security exercise for the cyber security students of bachelor's and master's programs of JAMK University of Applied Sciences. The data for this research of multiple-case design originates from observations, notes and questionnaires collected from the numerous cyber security exercises organised by JAMK University of Applied Sciences. The focus of this research is to characterize pedagogical principles of cyber security exercises as the educational framework for understanding the complex and interdependent cyber domain in the individual or organisational level.

Albeit, the importance of the cyber security exercises is widely recognized, there is a deficiency in the research of pedagogical aspects, especially in the viewpoint of competence development. In high level, cyber security exercise is a three-phase process consisting of different components of exercise life-cycle: planning phase, implementation phase and feedback phase. Those three phases can be divided into smaller steps of process. This study presents a competence development oriented view on that lifecycle of cyber security exercises. As part of that competence development oriented view, those three different components of exercise life-cycle are explored with the perspective of learning outcomes.

2 Cyber Security Exercises and the Pedagogical Principles

In recent years, cyber security exercises have established their position as a tool for developing the skills of cyber security professionals and as an operating environment for teaching. As business environments and the using of ICT in business have evolved, they have also become more complex at the same time. Consequently, the requirements for teaching environments have also changed.

In order to be able to teach skills that meet the needs of working life, it is necessary to understand the needs and be able to teach them in such a way that teaching builds skills that are needed in working life [21]. According to Ericsson's deliberated practices (DP) theory, the development of specialist skills must take into account the need to set well-defined learning objectives for students and the need to take into account the level of students' existing skills [22]. According to the deliberated practice theory, students do not benefit from the training if the tasks are at a level that they can perform routinely or if the goal setting of competence development has not been done with sufficient accuracy to mirror the student's level of competence.

Modern ICT teaching must therefore be able to mirror the changes in the operating environment to the change in competence requirements. When a modern cyber range is used in a cyber security exercise, the aim must be to make the operating environment as realistic as possible. The comprehensiveness and complexity of the teaching environment places demands on the student's level of competence. Thus, if cyber security training is used as a pedagogical tool for competence development, it should be noted that according to the Miller pyramid, the student's level of competence should be at the top of the pyramid [23]. This argument is supported by the andragogy, known as a theory of adult learning. According to andragogy theory an adult as a learner is often motivated, capable of self-direction and reflection on one's own existing competence [24]. Thus, for the adult learner learning experiment should be able to cause cognitive dissonance that allows the learner to update existing knowledge with new knowledge created in the learning event [25]. It must be possible to build a path of competence development, where in accordance with the constructive methodology, the student's developing competence enables the student to achieve new levels of competence through developing cognitive abilities [26]. Constructive methodology identifies problem-based learning as one of the key learning methods, in which the student develops his or her own skills by solving problems that lead to the learner's new knowledge [27].

The cyber security exercise can be seen as a complex problem field where the student solves the problems ahead and thus generates new knowledge for himself. This theory is supported also by experiential learning theory [28]. A key element in cyber security practice is working as a member of a team. This models real-life work where a person acts as part of, for example, a security operation team (SOC). In the exercise, all individuals are placed as a member of Teams (Blue Team) whose job is to defend and sustain the business in the operating environment assigned to them. The student can also act as part of a red team functionality that simulates threat activities. This role is to train, for example, the skills required for penetration testing. In the exercise, learners act in the role assigned to them, and communicate as part of a team of events that they perceive in their own operating environment. Thus, students share knowledge, solve problems and build new knowledge collectively.

To sum up, the cyber security exercise combines several pedagogical theories. The pedagogical framework of the cyber security exercises is shown in Figure 1. The exercise is also demanding from the point of view of pedagogical implementation and often requires a significant investment in pre-exercise planning, where the operating environment is constructed so that the required technical elements can be modeled and operational functionalities are designed so that pedagogical objectives can be achieved.

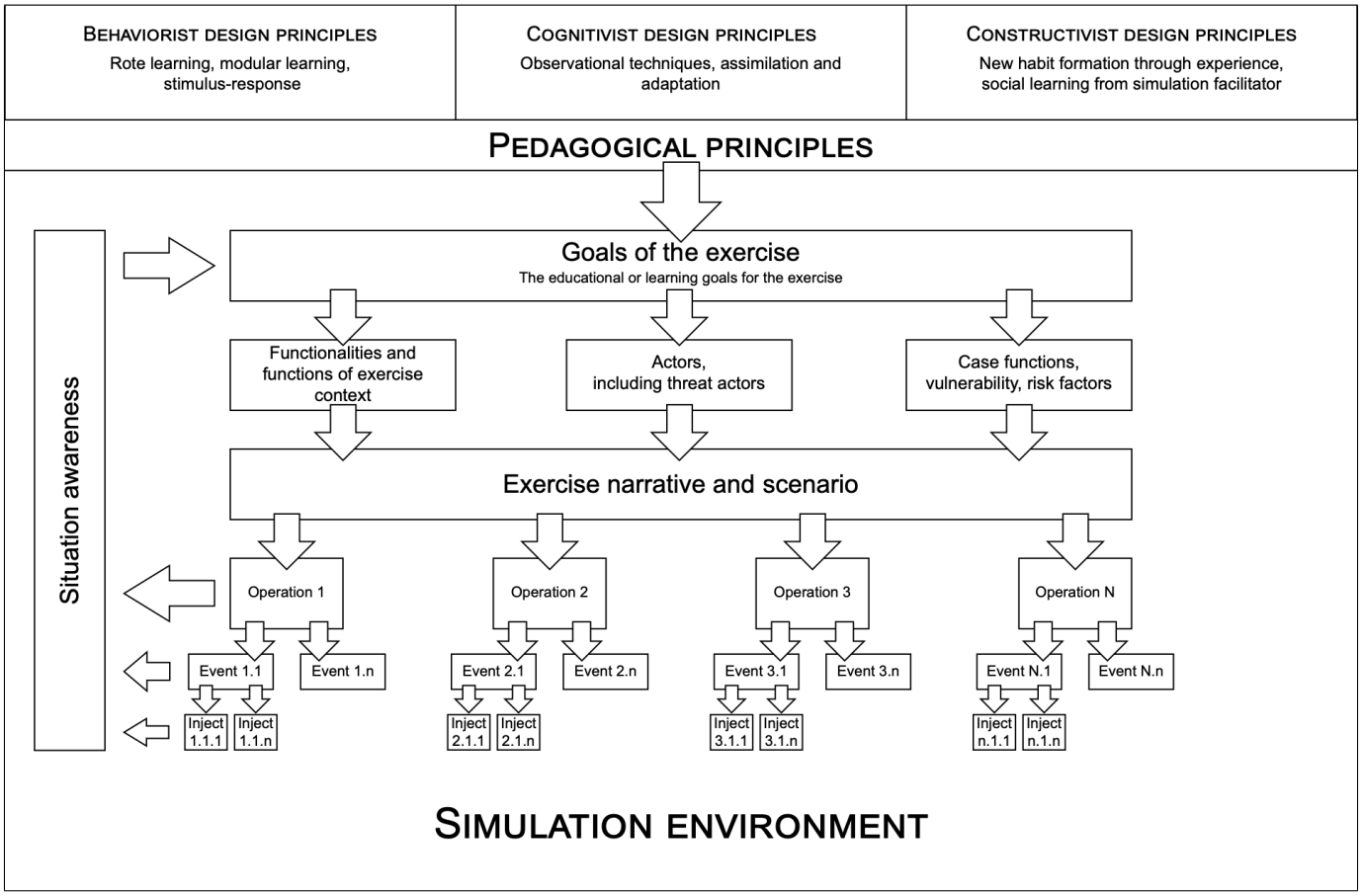


Figure 1: Pedagogical framework of the cyber security exercises [1]

2.1 Cyber Arena

According to complexity thinking, it should be possible to form an understanding of the functionality or entity under investigation as a whole, which is more than the sum of its parts [29]. The operating environment of cyber security can be seen as a complex entity consisting of different parts that interact with each other. Interaction takes place at different levels, such as in the technical operating environment, at the level of activities and processes, and as human interaction. The interconnection between different parts of the operating environment is partly defined and partly undefined. The key elements of complexity thinking are the recognition of the unpredictability of the environment, the difficulty of predicting cause consequences, and the self-organisation of the operating environment [30]. It is thus a matter of utilizing complexity thinking in accordance with the neo-reductionist school to model and simulate the subordination laws of the research object [31].

When the above is applied to the cyber environment, it is noteworthy to recognize the difficulties of applying traditional legislation, technological incompatibilities and the very rapid technical renewal of the environment. Thus, the student should be able to develop an understanding of unpredictability of the environment, unpredictable cause-and-effect relationships, and the risks of misuse of the technological element. In order for this entity to be embodied as part of cyber security education, a sufficiently realistic learning environment should be in place, such as Cyber Arena the overall high-level presentation of which is shown in Figure 2. It can be seen from the figure that the environment extensively models the cyber security domain. In order to be able to implement sufficient realism and the understanding of complexity, the teaching environment should model key functions and entities, as well as the interdependencies between the functions and or entities. In accordance with the authentic learning environment theory [32], the environment implements an operating environment in which the skills and competencies learned in cyber security practice will be applied.

3 Exercise Life-cycle

There are different definitions for the phases of cyber security exercise life-cycle. Wilhelmson and Svensson introduce three phases; planning, implementing and processing feedback, which are divided into into ten steps (exercise preparations, the master plan, the mission statement, exercise planning, practical preparations, implementation, evaluation, feedback, reporting, and the after action review) [33]. Consistently, MITRE describes three stages (Exercise Planning, Exercise Execution and Post Exercise) [34]. Vykopal et al. defines five stages for exercise; preparation, dry run, execution, evaluation, and repetition [35]. The Homeland Security Exercise and Evaluation Program (HSEEP) that provides a set of foundational concepts for exercise programs defines the management cycle with four stages: exercise design and development, exercise conduct, ex-

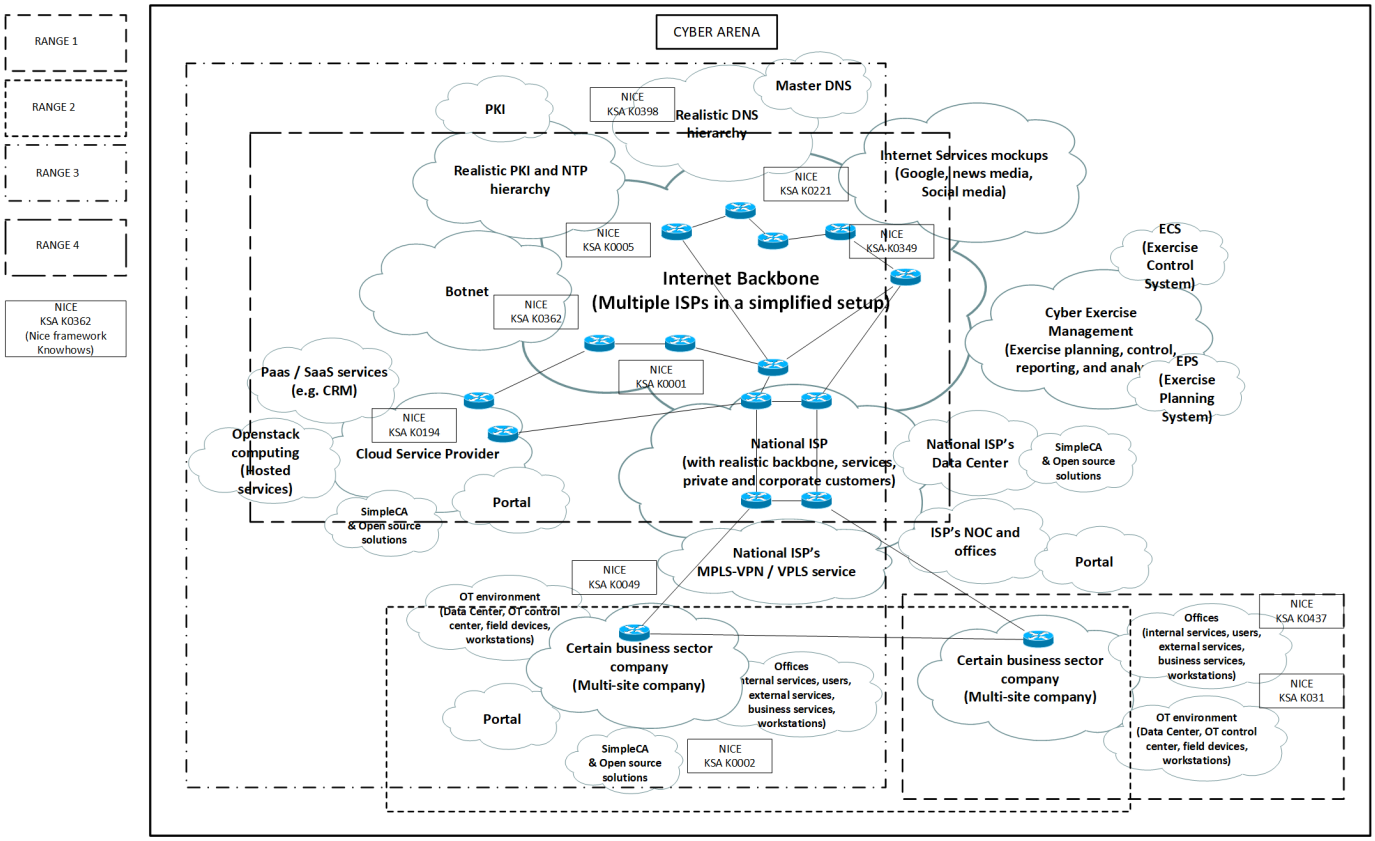


Figure 2: Comprehensive Cyber Arena [16]

ercise evaluation, and improvement planning [36]. As an integration of different definitions and a viewpoint for pedagogical aspects of this research the following three phases are selected *(i) planning phase, (ii) implementation phase, and (iii) feedback phase*. The extended view for pedagogical aspects of the cyber security exercises is shown in Figure 3. It illustrates whole process from planning phase to the implementation of the exercise.

3.1 Planning Phase

The first step of exercise life-cycle is the planning-phase. It is an extremely critical phase, because it determines the effectiveness of the whole exercise. From the viewpoint of competence development, the content of the exercise e.g. scenario, actors, and events shall be fitted to the requirements of the exercise target audience.

Based on the required learning outcomes and the target organisation, exercise parameters are derived including simulated operational environment of scenario for example technical functionalities, threat actors, risks and vulnerabilities. That scenario encompasses discrete events and injects of exercise describing the totality of simulated activities. If the scenario created during the planning phase includes obscurities, the exercise including technical environment may not increase the performance of the exercise target audience or organisation. All the elements mentioned supports the achievement of the set learning objectives. The learning objectives can be seen at several levels, the goals can be set from the perspective of organisational competence development, on the other hand, learning objectives can be set from the perspective of individual competence development. When learning goals are set from an organisational perspective, an individuals learning goals should be set so that they put the organisations goals into practice. In Figure 3, the goals of the exercise phase are opened, allowing us to look at an example of what kind of practical sections or tasks in the planning phase should be planned in order to be able to achieve the set goals in the exercise.

3.2 Implementation Phase

There are several differences between the phases of the exercise life-cycle. The most hectic phase is the implementation phase where the exercise target audience is acting in a simulated complex cyber domain under cyber deviation actions (attacks and intrusions). When acting under hectic and stressful cyber deviation circumstances, there is a requirement to maintain the understanding (situation awareness, SA) of the valuable assets' status in cyber domain. According to Endsley [37] *"Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future"*. The expertise of the individual has a remarkable outcome for the SA [38]. In this context, the term situation awareness refers to both, the understanding of the progress of the operational situation in the exercise, and the understanding of the monitoring and evaluation of the pedagogical objectives of the exercise.

During the exercise, also the decision making has a remarkable role in what incident handling actions shall be done and how to

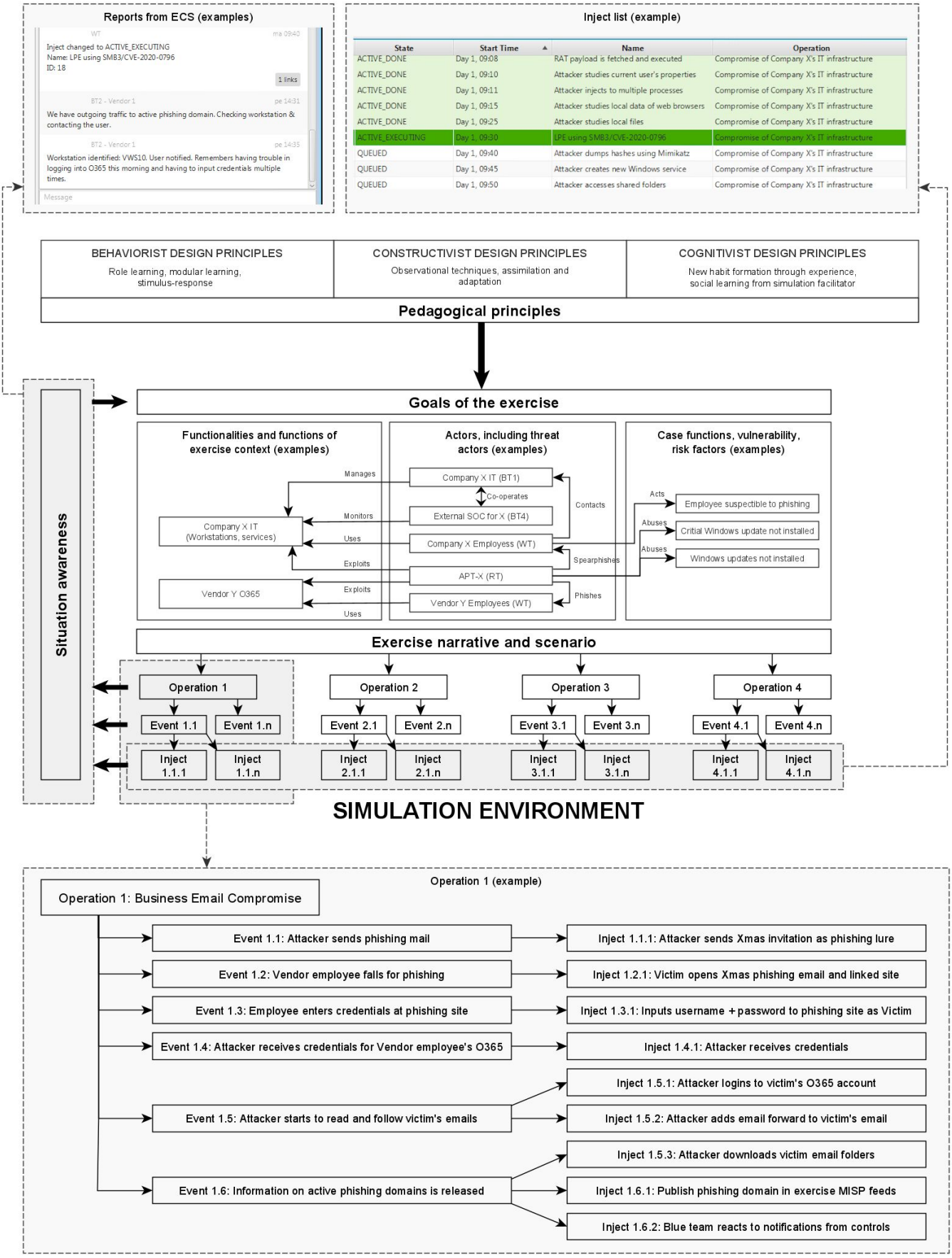


Figure 3: Detailed pedagogical framework of the cyber security exercises

categorize order of the required actions. Mostly, those decisions are based on SA as comprehension on two models of decision making cycle; Gartner's Adaptive Security Architecture (Predict-Prevent-Detect-Respond) [39] and OODA-loop (Observation-Orientation-Decision-Action) [40, 41].

Figure 3, illustrates that the operation line according to the exercise scenario, is opened to illustrate the practical actions of one operation performed during the exercise. The operation includes a series of events that are divided into injects with which the exercise is practically carried forward according to a planned scenario. An inject list has also been opened in the figure to show how in the practice the exercise proceeds with injects. The exercise management team (white team, WT) can use the information obtained through information systems to assess the situation, but it is often also necessary to monitor visually and interview students. This will ensure that the exercise proceeds as planned and that the set learning objectives can be achieved. If the WT notices that students are taking actions that are not realistic or the focus of the exercise begins to shift from the set goal, the WT should guide the course of the exercise. This can be done through information system injections or verbally by instructing students.

3.3 Feedback Phase

From the perspective of the development of an individual's competence, the feedback phase is the most important part of the exercise. Scheduling the feedback phase should be planned carefully with too long an interval of exercise and feedback may cause a decrease in learning intensity for the individual. The emotions and experiences raised by the exercise are alleviated and learning outcomes may suffer as a result. In the feedback phase, the pedagogical goals of the exercise are recalled and the events of the exercise are reflected against them by reviewing all operations performed and related events and injects in the exercise. This happens so that all operations performed in the exercise, related events and injects are reviewed. By doing so, the student can reflect the experience they have had during the exercise and thus deepen their own learning. It is also important to tie individual events through operations into an exercise scenario. This allows the student to increase the understanding of the bigger picture, for example, the threat actors' motive and the tools used for the attack. This is important and enables in the future the exercise event to be reflected in real situations of working life.

It is a good to set aside time for the feedback session so that the interaction between students and teachers can be enabled as widely as possible. Normally, each defensive team (blue team, BT) has the opportunity to open up their own observations and experiences in this section. This enables collegial and collaboratively learning. The offensive team (RT) also goes through its own operations, thus allowing BT to reflect its own observations in relation to the operations that took place.

4 Assessing Performance and Results

The Kirkpatrick four-level assessment framework can be used for the assessment of the exercise. Kirkpatrick divides assessment into four levels: (i) reaction, (ii) learning, (iii) behaviour, and (iv) results [42]. The Kirkpatrick framework is useful in assessing a larger entity such as an organisation or team, but it can also be used for the individual experience of learning in exercise. In the Kirkpatrick framework, the goals for the development of individual competence are set at level one and two. At level one the reactions caused by the exercise are assessed. At level two, the learning that has achieved by the individual is assessed. Kirkpatrick et al. recommends the use of control groups and tests for assessing the learning. The goals set for organisational competence development utilize the Kirkpatrick assessment model's, level three, which assesses the change in individual behaviour through the achieved learning outcomes, and level four which assesses the effects and implementation of advanced competence on organisational performance. Kirkpatrick model is a widely used evaluation model, the study [43] presents a framework for competence development and assessment in hybrid cyber security exercises, and the authors of [44] introduce one adoption of the Kirkpatrick Model.

Other methods for evaluating exercise can also be used, the authors of [45] have monitored communication during the cyber security exercise for understanding the behaviour of the exercise target audience. Their conclusion is that communication monitoring can be used as a resource in measuring the performance during the cyber security exercise.

When focusing on assessing the learning of an individual who has participated in the exercise, Brown and Pickford [46], have created a model that looks at the assessment of learning event as a whole. Brown and Pickford divide the assessment into the following subsections the significance and implementation of these sections should be planned in advance: why, what, how, who, when.

Why- why the assessment is made? What is the purpose of assessment in this particular learning event? In the context of a cyber security exercise, the aim of assessment is in some respects to control the individual's performance, facilitate the student's adaptation to the exercise, be able to assess the student's motivation, measure the competence, skills and know-how and to provide the student information about mistakes and inappropriate practices.

What- what are we assessing? In a cyber security exercise the processes of work, individual performance and success of team work can be assessed. In the exercise, the assessment should be performed at all stages of its life cycle.

How- how are we assessing? As discussed, Figure 3 illustrates the role of the situation awareness and all the inputs where the information for assessing will be collected. So, part of the information for the assessment can be collected via information systems and the reports from the BT that they are delivering to exercise control system. In addition to this, the teacher must monitor classroom activities. In this way, information can be obtained, for example, about an individual's performance in a specific role as part of a team. Visual observation can also provide information about the team's internal activities and role support and its possible functioning.

Who- who is suitable for making the assessment? In the exercise, students often work as a part of a team throughout the exercise. This provides an opportunity to implement the evaluation as a peer review, whereby the internal functionalities and inclusions of the team become more clearly assessed. In education leading to a degree, students are also often asked to have a learning diary in which the student can make a self-assessment of the exercise throughout its

life cycle. The role of teachers in the assessment may therefore be more aggregate.

When- when should assessment take place? In the cyber security exercise the assessment needs to be done in all phases of exercise. This is because the assessment plays a very important role as a function of guiding learning. The importance of formative or guiding assessment in cyber security exercises is emphasized. The theory of formative evaluation has been built specifically by Scriven [47]. According to Scriven, the concept of formative assessment became conceptualized. Formative assessment emphasizes that assessment should take place at all stages of the teaching and learning, and not just at the end. Several studies verify that learning outcomes are significantly improved when formative assessment that guides learning is included in the assessment as well as summative assessment. Thomas et. al [48] and Leahy et. al [49] have stated that learning outcomes improve when assessment includes formative assessment that guides learning in addition to the assessment learned skills.

Formative assessment emphasizes the importance of feedback. According to Hattie [50], the purpose of the feedback can be divided into three sections: *Feed up, feed back* and *feed forward* sections. The feed up gives the learner an answer to the question of where he or she is going. The purpose of feed up feedback is to continuously clarify and specify the learning objectives. Feed up feedback also aims to engage and motivate the learner to pursue to the set goals. A feed back, tells to the learner where he or she is at the moment. The feed back feedback is used to provide the student the information on how he or she has progressed in relation to the set learning objectives. In order to give exact feedback on the learner's position, the learning objectives must be precisely defined, also so that the prerequisites for progressive learning are perceptible and acceptable. Feed forward feedback, tells the learner what he or she should do next. In practice, guidance can be sought, for example, through questions that broaden the student's understanding, or with advice and tips on, for example, new ways for approaching to the set goals.

According to Hattie, each feedback question works on four levels: level of the task, level of process, level of self-regulation and level of person. The level of the task, i.e. how well the learner understands the set tasks and how he or she performs them. In practice, for example, feedback indicates whether an individual task has been solved correctly or incorrectly. Feedback should also be directed to correcting any malfunctions or performing the task correctly. Level of process indicates the process required to understand and perform a given task in the context of a cyber exercise, for example, what kind of operation is needed to bring an into the exercise so that the student learns the methods and technology used for a phishing campaigns. Level of self-regulation guides the learner to self-assessment and self-direction of action. At this level, feedback can also be used to guide the learner's motivation and adaptation to the teaching environment used. The level of the person includes assessments of the learner. This section often contains elements for assessing and providing feedback on a learners personality traits. In a cyber exercise, guidance can be given, for example, on a persons participation and activity as part of a team, what is a key part of a persons performance in the exercise.

The purpose of the feedback is to reduce the gap between existing competence and the target competence. Due to the complexity and scope of the cyber security exercise, special attention should be paid to the continuous feedback throughout the exercise life cycle.

5 Conclusion

There are elements in the cyber security operating environment in line with complexity thinking, however, it is a philosophical question of whether the cyber security environment is ultimately a complex entity. There is complexity in the operating environment. According to the definition of complexity, the phenomena are intertwined and the whole cannot be understood by disassembling the whole into parts and looking at the parts one by one. Unlike complex entities, the cyber operation environment can be controlled, although there may also be self-directed elements in the operation environment. When implementing cyber security education, the complexity of the operating environment should be taken into account. Therefore, the teaching environment should be a Cyber Arena style operating environment mimicking realistic operative cyber domain. In the Cyber Arena, several functional entities are combined forming an ensemble with complex cause-and-effect relationships manifested. Pedagogically, however, the constructive construction of competence development must be taken into account, in which case teaching starts from the parts or details of the operating environment and culminates the teaching for understanding the whole environment, including the interdependences of different entities.

As presented, the pedagogical objectives of the exercise should be taken into account at all stages of the life cycle. In the planning phase, goals are set for competence development. In education leading to a degree, the objectives are defined in the curriculum. In the exercise for the other target audience, the goals of competence development should be defined together with the representatives of the organisation. In this way, the objectives of the exercise are adapted to the current maturity and operations of the organisation. In the implementation phase, the realization of the goals must be monitored and, if necessary, the focus of the exercise must be directed towards the set goals. In the feedback phase, participants in the exercise are given the opportunity to interact through the exercise, in which the operations performed in the exercise are opened in detail. To support post-practice learning, it is also important to provide material to be distributed that allows students to return to the details of the exercise afterwards.

Generally accepted content frameworks, such as the NICE framework,can be used to design the content's learning objectives. This makes it possible to set structured teaching goals, through which the exercise scenario can be constructed in such a way that the technical functions do not remain separate events without causal relationships. The student who has done this is are able to form an entity from the exercise, at the latest at the feedback stage, through which he or she can learn about the effectiveness of the sub-entities of the operating environment as a whole.

Exercise evaluation is a challenging whole consisting of evaluating an individual, as well as evaluating the performance of an organisation or part of it. In order for assessment to serve the set competence development goals as well as possible, formative assessment that guides learning should be used where possible. In this way, the evaluation of the activities serves as a guiding element of the exercise, helping to ensure that the set learning objectives

are achieved. With regard to formative assessment, the importance of feedback is emphasized. It should be possible to deliver it at all stages of the exercise life cycle. Feedback should take into account interactivity and, where possible, make use of peer feedback from learners.

Future research should build on the understanding of the pedagogical requirements of cyber security exercise in relation to the teaching environment and individual learning gained in this and other studies and move towards assessing the development of organisational competence in cyber security exercise.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is funded by the *Cyber Security Network of Competence Centres for Europe (CyberSec4Europe)* project of the Horizon 2020 SU-ICT-03-2018 program.

References

[1] M. Karjalainen, T. Kokkonen, S. Puuska, "Pedagogical Aspects of Cyber Security Exercises," in "2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&p)," 103–108, 2019, http://dx.doi.org/10.1109/EuroSPW.2019.00018.

[2] Secretariat of the Security Committee, "Finland's Cyber security Strategy, Government Resolution 3.10.2019," https://turvallisuuskomitea.fi/wp-content/uploads/2019/10/Kyberturvallisuusstrategia_A4_ENG_WEB_031019.pdf, 2019.

[3] The White House, signed by President Donald J. Trump, "National Cyber Strategy of the United States of America," https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf, 2018.

[4] European Comission, "Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace," https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013JC0001&from=EN, 2013.

[5] FINGRID magazine, "Cyber security is ensured with genuine exercises," https://www.fingridlehti.fi/en/cyber-security-ensured-genuine-exercises/, 2017, Accessed: 12 May 2020.

[6] B. Uckan Färnman, M. Koraeus, S. Backman, "The 2015 Report on National and International Cyber Security Exercises : Survey, Analysis and Recommendations," Technical report, Swedish Defence University, CRISMART (National Center for Crisis Management Research and Training), 2015, http://dx.doi.org/10.2824/627469.

[7] The NATO Cooperative Cyber Defence Centre of Excellence, CCDCOE, "Exercises," https://ccdcoe.org/exercises/, Accessed: 12 May 2020.

[8] K. Saharinen, M. Karjalainen, T. Kokkonen, "A Design Model for a Degree Programme in Cyber Security," in "Proceedings of the 2019 11th International Conference on Education Technology and Computers," ICETC 2019, 3–7, Association for Computing Machinery, New York, NY, USA, 2019, http://dx.doi.org/10.1145/3369255.3369266.

[9] W. Newhouse, S. Keith, B. Scribner, G. Witte, National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework, 2017, http://dx.doi.org/10.6028/nist.sp.800-181.

[10] Association for Computing Machinery (ACM) and IEEE Computer Society (IEEE-CS), "Computing Curricula 2020, CC2020, Paradigms for Future Computing Curricula (Draft, Version 36)," https://cc2020.nsparc.msstate.edu/, 2020.

[11] Association for Computing Machinery (ACM) and IEEE Computer Society (IEEE-CS) and Association for Information Systems Special Interest Group on Information Security and Privacy (AIS SIGSEC) and International Federation for Information Processing Technical Committee on Information Security Education (IFIP WG 11.8), "Cybersecurity Curricula 2017, (CSEC2017), Curriculum Guidelines for Post-Secondary Degree Programs in Cybersecurity (Version 1.0)," https://www.acm.org/binaries/content/assets/education/curricula-recommendations/csec2017.pdf, 2017, Accessed: 12 May 2020.

[12] P. Ward, A. M. Williams, P. A. Hancock, Simulation for Performance and Training, 243–262, Cambridge University Press, New York, NY, US, 2006, http://dx.doi.org/10.1017/CBO9780511816796.014, iD: 2006-10094-014.

[13] P. Nevavuori, T. Kokkonen, "Requirements for Training and Evaluation Dataset of Network and Host Intrusion Detection System," in Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo, eds., "New Knowledge in Information Systems and Technologies," 534–546, Springer International Publishing, Cham, 2019.

[14] B. Ferguson, A. Tall, D. Olsen, "National Cyber Range Overview," in "2014 IEEE Military Communications Conference," 123–128, 2014, http://dx.doi.org/10.1109/MILCOM.2014.27.

[15] Z. Tian, Y. Cui, L. An, S. Su, X. Yin, L. Yin, X. Cui, "A Real-Time Correlation of Host-Level Events in Cyber Range Service for Smart Campus," IEEE Access, **6**, 35355–35364, 2018, http://dx.doi.org/10.1109/ACCESS.2018.2846590.

[16] M. Karjalainen, T. Kokkonen, "Comprehensive Cyber Arena; The Next Generation Cyber Range," in "2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&P)," , 2020.

[17] JAMK University of Applied Sciences, Institute of Information Technology, JYVSECTEC, "Jyväskylä Security Technology," https://www.jyvsectec.fi, Accessed: 12 May 2020.

[18] Ministry of Defence Finland, "Kansallinen kyberturvallisuusharjoitus KYHA18 järjestetään Jyväskylässä, Official Bulletin 11th of May 2018," https://www.defmin.fi/ajankohtaista/tiedotteet/2018?9610_m=9314, 2018, Accessed: 12 May 2020.

[19] European Defence Agency, EDA, "Cyber Ranges: EDA's First Ever Cyber Defence Pooling & Sharing Project Launched By 11 Member States," https://www.eda.europa.eu/info-hub/press-centre/latest-news/2017/05/12/cyber-ranges-eda-s-first-ever-cyber-defence-pooling-sharing-project-launched-by-11-member-states, 2017, Accessed: 12 May 2020.

[20] European Defence Agency, EDA, "EDA Cyber Ranges Federation project showcased at demo exercise in Finland," https://www.eda.europa.eu/info-hub/press-centre/latest-news/2019/11/07/eda-cyber-ranges-federation-project-showcased-at-demo-exercise-in-finland, 2019, Accessed: 12 May 2020.

[21] H. Collins, R. Evans, A Sociological/Philosophical Perspective on Expertise: The Acquisition of Expertise through Socialization, 21–32, Cambridge Handbooks in Psychology, Cambridge University Press, 2 edition, 2018, http://dx.doi.org/10.1017/9781316480748.002.

[22] K. Anders Ericsson, "Deliberate Practice and Acquisition of Expert Performance: A General Overview," Academic Emergency Medicine, **15**(11), 988–994, 2008, http://dx.doi.org/10.1111/j.1553-2712.2008.00227.x.

[23] G. E. Miller, "The assessment of clinical skills/competence/performance," Academic medicine, **65**(9), S63–7, 1990.

[24] S. B. Merriam, L. L. Bierema, Adult learning: Linking theory and practice, John Wiley & Sons, 2013.

[25] M. S. Knowles, Designs for adult learning: Practical resources, exercises and course outlines from the father of adult learning., Alexandria, Va: American Society for Training & Development, 1995.

[26] S. Lindblom-Ylänne, A. Nevgi, "The effect of pedagogical training and teaching experience on approach to teaching," in "11th EARLI conference, Padua," , 2003.

[27] J. R. Savery, T. M. Duffy, "Problem based learning: An instructional model and its constructivist framework," Educational technology, **35**(5), 31–38, 1995.

[28] D. A. Kolb, R. E. Boyatzis, C. Mainemelis, et al., "Experiential learning theory: Previous research and new directions," Perspectives on thinking, learning, and cognitive styles, **1**(8), 227–247, 2001.

[29] T. Hanén, "Faced with the Unexpected - Leadership in Unexpected and Dynamic Situations: an Interpretation Based on Complexity Theory (Orig: Yllätysten edessä: kompleksisuusteoreettinen tulkinta yllättävien ja dynaamisten tilanteiden johtamisesta)," Ph.D. thesis, National Defence University, 2017, http://urn.fi/URN:ISBN:978-951-25-2870-7.

[30] R. Geyer, S. Rihani, Complexity and public policy: a new approach to twenty-first century politics, policy and society, Routledge, 2010.

[31] K. A. Richardson, "MANAGING COMPLEX ORGANIZATIONS: COMPLEXITY THINKING AND THE SCIENCE AND ART OF MANAGEMENT," Corporate Finance Review, **13**(1), 23–29, 2008, https://search-proquest-com.ezproxy.jyu.fi/docview/198834948?accountid=11774.

[32] J. Herrington, R. Oliver, "An instructional design framework for authentic learning environments," Educational Technology Research and Development, **48**(3), 23–48, 2000, http://dx.doi.org/10.1007/BF02319856.

[33] N. Wilhelmson, T. Svensson, Handbook for planning, running and evaluating information technology and cyber security exercises, The Swedish National Defence College, Center for Asymmetric Threats Studies (CATS), 2014.

[34] J. Kick, "Cyber Exercise Playbook," The MITRE Corporation https://www.mitre.org/sites/default/files/publications/pr_14-3929-cyber-exercise-playbook.pdf, 2014, Accessed: 12 May 2020.

[35] J. Vykopal, M. Vizvary, R. Oslejsek, P. Celeda, D. Tovarnak, "Lessons learned from complex hands-on defence exercises in a cyber range," in "2017 IEEE Frontiers in Education Conference (FIE)," 1–8, 2017.

[36] The U.S Department of Homeland Security, "Homeland Security Exercise and Evaluation Program (HSEEP)," https://www.fema.gov/media-library-data/1582669862650-94efb02c8373e28cadf57413ef293ac6/Homeland-Security-Exercise-and-Evaluation-Program-Doctrine-2020-Revision-2-2-25.pdf, 2020, Accessed: 12 May 2020.

[37] M. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," Human Factors, **37**(1), 32–64, 1995, http://dx.doi.org/10.1518/001872095779049543.

[38] M. R. Endsley, Expertise and Situation Awareness, 633–652, Cambridge Handbooks in Psychology, Cambridge University Press, 2006, http://dx.doi.org/10.1017/CBO9780511816796.036.

[39] R. van der Meulen, "Build Adaptive Security Architecture Into Your Organization," https://www.gartner.com/smarterwithgartner/build-adaptive-security-architecture-into-your-organization/, 2017, accessed: 3 April 2020.

[40] G. L. Rogova, R. Ilin, "Reasoning and Decision Making under Uncertainty and Risk for Situation Management," in "2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)," 34–42, 2019, http://dx.doi.org/10.1109/COGSIMA.2019.8724330.

[41] B. Brehmer, "The Dynamic OODA Loop: Amalgamating Boyd's OODA Loop and the Cybernetic Approach to Command and Control," in "10th International Command and Control Research and Technology Symposium, The Future of C2," , 2005.

[42] D. L. Kirkpatrick, J. D. Kirkpatrick, Evaluating Training Programs, Berrett-Koehler Publishers, Inc., San Francisco, 2006.

[43] A. Brilingaitė, L. Bukauskas, A. Juozapavičius, "A framework for competence development and assessment in hybrid cybersecurity exercises," Computers & Security, **88**, 101607, 2020, http://dx.doi.org/10.1016/j.cose.2019.101607.

[44] A. Ahmad, C. Johnson, "A Cyber Exercise Post Assessment: Adoption of the Kirkpatrick Model," Advances in Information Sciences and Service Sciences (AISS), **7**(2), 2015.

[45] T. Kokkonen, S. Puuska, "Blue Team Communication and Reporting for Enhancing Situational Awareness from White Team Perspective in Cyber Security Exercises," in O. Galinina, S. Andreev, S. Balandin, Y. Koucheryavy, eds., "Internet of Things, Smart Spaces, and Next Generation Networks and Systems," 277–288, Springer International Publishing, Cham, 2018.

[46] S. Brown, R. Pickford, Assessing skills and practice, Routledge, 2006.

[47] M. Scriven, "SOCIAL SCIENCE EDUCATION CONSORTIUM. PUBLICATION 110, THE METHODOLOGY OF EVALUATION." , 1966.

[48] L. Thomas, C. Deaudelin, J. Desjardins, O. Dezutter, "Elementary teachers' formative evaluation practices in an era of curricular reform in Quebec, Canada," Assessment in Education: Principles, Policy & Practice, **18**(4), 381–398, 2011.

[49] S. Leahy, D. Wiliam, "From teachers to schools: scaling up professional development for formative assessment," Assessment and learning, **2**, 49–71, 2012.

[50] J. Hattie, "Teachers Make a Difference, What is the research evidence?" , 2003.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 594-599 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques

Moulay Smail Bouzakraoui[*,1], Abdelalim Sadiq[1], Abdessamad Youssfi Alaoui[2]

[1]*Department of Informatics, Faculty of sciences, Ibn Tofail University, Kenitra, 14000, Morocco*

[2]*Laboratory of Advanced Digital Enterprise Modeling and Information Retrieval, National School of Computer Science and Systems Analysis, Mohammed V University, Rabat, 10000, Morocco*

ARTICLE INFO

Article history:
Received: .14 June, 2020
Accepted: 21 July, 2020
Online: 19 August, 2020

Keywords:
Facial expression
customer emotion
Emotion recognition

ABSTRACT

Nowadays, Customer satisfaction is important for businesses and organizations. Manual methods exist, namely surveys and distributing questionnaires to customers. However, marketers and businesses are looking for quick ways to get effective and efficient feedback results for their potential customers. In this paper, we propose a new method for facial emotion detection to recognize customer's satisfaction using machine-learning techniques. We used a facial landmark point; we extract geometric features from customer's emotional faces using distances from landmarks points. Indeed, we used distances between the neutral side and the negative or positive feedback. After that, we classified these distances by using different classifier, namely Support Vector Machine (SVM), KNN, Random Forest, Adaboost, and Decision Tree. To assess our method, we verified our algorithm by using JAFFE datasets. The proposed method reveals 98,66% as accuracy for the most performance SVM classifier.

1. Introduction

This paper is an extension of work originally presented in world conference on complex systems (WCCS'19) [1]. We decide to develop this work because, we observe in the last decade, business primarily has an important impact based on the way, how customers observe their products and services [2]. Customer satisfaction can be measured using manuals methods such as the satisfaction survey, the interview, and the focus group. These methods are not efficient and effective to the cost, time and reliability of the data. Facial expressions are used to communicate non-verbally. They are a special way to express our emotions and appreciations. In the context of customer satisfaction, a negative feedback emotion is often related to a lower perceived quality of service. The facial expression contributes 55% to communicate with a speech. More than that, 70-95% of negative feedback can be understood verbally [3]. Companies have long been interested in understanding the purchasing decision-making process of consumers [4]. In this work, we aim at detecting the positive or negative emotions customers from the analysis of facial expressions. In this way, this type of information is useful. For example, we can calculate the statistics about the products as well as the replacement of their exposures. Appreciated products must be brought to the fore, while negatively appreciated products must be replaced or otherwise requested. For this reason, we are so existed to propose a new method for facial emotion detection to recognize customers' satisfaction using facial expression and machine learning techniques.

The remainder of this paper is structured as follows: Firstly, we offer an outline of the related work. Secondly, we introduce our approach that contains a system for customer satisfaction using facial expression and machine learning techniques. We reserved a separate section to the experimental results. In the last section, we sum up our work.

2. Related work

Currently in the field of marketing and advertising, facial recognition is used to study consumers' emotions in two forms: positive and negative emotions. Charles Darwin was the first to provide robust basis for emotions, representing their significance, their usefulness and communication. Already, many of facial expressions of emotions have an adaptive significance that serve to communicate something, most emotions are expressed alike on the human face nerveless of culture or race [5], his work "Expression of emotions in humans and animals", remains a

[*]Corresponding Author: Moulay Smail Bouzakraoui, Department of Informatics, faculty of sciences, university Ibn Tofail, Kenitra, Morocco, bouzakraoui.smail14@gmail.com

reference for many scientists. In [6], authors proposed method to classify muscle movement to code the facial expression. The facial movements are determined through action units (AU). Each AU is based on the affinity of the muscles. This method, used mostly in classifying facial movement named Facial Action Coding

System (FACS).We can define six universal basic emotions: happy, sadness, anger, surprise, fear and disgust. Various researchers have supported the universality of the expression of emotions. Generally, in reaction to similar stimuli, people prompt similar expressions with local variations [7]. Frequently, we used two methods: motion-based method and deformation of face. The change of face is taken into account for the motion-based method [8, 9]. On the contrary, we take into account a neutral image and another image for the deformation-based method [10, 11]. In [10], authors use FACS (Face Action Code System) as features to classify emotions. Their model consists to count images, which classified correctly and weighted by the system. They found that, 2% of images (6 image) are failed completely in tracking. They also reported 91% as average correct recognition ratio. Otherwise, in [11] a model which able to use also speech content is developed. They create a system to compute the Human Computer Interaction (HCI). In order to test their model, they utilize 38 subjects of affect recognition approach with 11 HCI-related affect states. Their model a gave 90% for the bimodal fusion as an average recognition accuracy.

In [12] feature extraction contains geometric and appearance-based models. The first model involves feature extraction about eye, mouth, nose, eyebrow…However; the second model covers the specific part of the face. They also evaluate their model by employing SVM, ANN, KNN, HMM, Bayesian network, and sparse representation-based classification. The experiment results report that, sparse representation-based classification is the more performant classifier for their model.

Deep learning algorithms are also used to extract pattern from facial expressions. More specifically, we develop a convolutional neural network to detect features from faces, and to classify these features to different emotions. This type of network contains two parts. The first part contains convolutional layers, which are based on applying mathematic convolution operation. The result of this part is features extracted from faces. The second part contains feed-forward network. This part of network is able to classify features extracted to different facial expression. In [13, 14] the based results researches of this algorithm is proposed. For example, authors, in [13], apply KNN to classify emotions. After that, they find 76.7442% and 80.303% as accuracy in the JAFFE and CK+, respectively, which illustrate viability and success of their model. In [14], authors employ also Decision Tree and Multi-Layer Perceptron to classify emotions. However, they find that, Convolutional Neural Network report the greatest recognition accuracy. An Affdex by Affectiva [15] is one of the most implements for visual emotional analysis. Which permits giving the emotional trend of a subject, through detection emotions [16], and the Microsoft Cognitive Services based on the Azure platform. These two studies are able to perceive age, sex and ethnicity and depend on Convolutional and/or Recurrent Neural Network [17]. In addition, they find that, the proposed model is yields performance increases for facial expression recognition using CK+ dataset. A multimodal affect recognition system was developed to determine if the customer exhibits negative affect such as being unhappy, disgusted, frustrated, angry or positive affect such as happy, satisfied and content with the product being offered [18]. Furthermore, authors apply Support Vector Machine an emotion template to evaluate their algorithm. They test the real-time performance, which evaluate the feedback assessment in order to compute accuracy and viability for multimodal recognition. To appreciate customer satisfaction special facial emotion namely" happy"," surprised" and "neutral" is proposed in [19]. They evaluate their model by using datasets as Radboud Faces and Cohn-Kanade (CK+). Output results report that their proposed algorithm a gave an important recognition accuracy utilizing Action Unit (AU) features, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). In the last years, eye-tracking system has been developed to analyze the research of customer [20]. Companies invest intensely within the advancement and publicizing of products, the return on investment (ROI) must be justified by the organization. Hence, the specific observing of the fascination of a product by the customer amid its promotion, and how to develop their promotion campaigns. Eye tracking is very promising in navigation advertising. In [21], authors developed an algorithm aims to compute sales assistant, which depends specially on emotion detection. Otherwise, [22] present a study, which contains products features impact the costumer's emotional. Furthermore, they used 21 emotional categories of products.

3. Our Approach

In this section, we present an algorithm to recognize the emotions of customers towards a given product. In order to analyze the facial expressions of a client to derive satisfaction from the product offered, our algorithm is based on three essential steps. The first is to acquire the image of the face and the second allows extracting the expressions in the form of geometric features. To obtain these features, our algorithm transforms the input image into geometric primitives such as points and curves. These primitives are located via distinctive patterns such as the eyes, mouth, nose and chin by measuring their relative position, their width and other parameters such as the distance between the nose, eyes, mouth. Our contribution is therefore to propose a method for selecting the optimal distances that can make the difference between facial expression linked to emotion and customer satisfaction. In the last step, these different distances are used to classify a client's emotions towards a product, into three types; positive, negative or neutral appreciation. These results will be analyzed and provided to the company to understand how customers perceive the product.

Otherwise, the pipeline of our algorithm begin by detecting the face into an input image. After that, we compute the key point of the face by applying a mask, in order to compute the distances between key points. We aim to measure the variation of face's shape in order to classify appreciation of customer's toward a product. For this raison, our algorithm is based on computing

face's key points to calculate distances between those landmarks. Finally, we employ a set of classifier to evaluate our algorithm

3.1. Features extraction (Geometric features)

In order to represent the emotions of a face, this step consists in locating the facial landmarks (for example the mouth, the eyes, the eyebrows, the nose and the chin). The model presented in figure (1) allows us to locate the face marks. This model has a drawback; the deformation of these marks modifies the facial expressions.

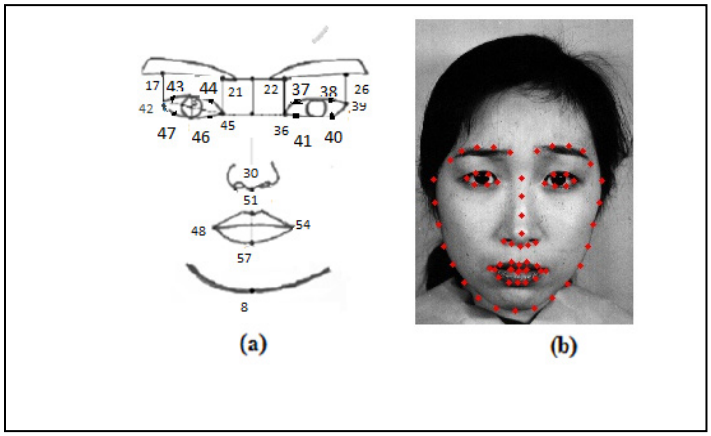


Figure1: (a) Facial model depicting the position of the chosen landmark point, (b) Landmark point on the test subject face

However, we used twenty-two facial landmarks points to calculate eighteen distances. Therefore, a vector of eighteen values will represent each face. The selected distance are:

For the eyes: six features
Left eye width, Right eye width, Right eye height, left eye height, Right eye height, left eye height

For the mouth: two features
Mouth Width, mouth height

For the chin: two features

Chin to nose distance, Chin to mouth distance

For the eyebrow: six features
Left eyebrow width
Right eyebrow width
Outer left eyebrow distance to eye
Inner left eyebrow distance to eye
Outer right eyebrow distance to eye
Inner right eyebrow distance to eye

For the nose: two features
Nose distance to left eye
Nose distance to right eye

Table1: Definitions of Geometric features

Distance	Description	Formula	Distance	Description	Formula
D1	Mouth Width	d(X48,X54)	D12	Right eye height	d(X38,X40)
D2	Mouth height	d(X51,X57)	D13	Outer left eyebrow distance to eye	d(X17,X42)
D3	Chin to mouth distance	d(X8,X57)	D14	Outer right eyebrow distance to eye	d(X26,X39)
D4	Chin to nose distance	d(X8,X30)	D15	Inner right eyebrow distance to eye	d(X22,X36)
D5	Left eyebrow width	d(X17,X21)	D16	Inner left eyebrow distance to eye	d(x21,x45)
D6	Right eyebrow width	d(X22,X26)	D17	Nose distance to left eye	d(X45,X30)
D7	Left eye width	d(X42,X45)	D18	Nose distance to right eye	d(X36,X30)
D8	Right eye width	d(X36,X39)			
D9	Left eye height	d(X43,X47)			
D10	Left eye height	d(X44,X46)			
D11	Right eye height	d(X37,X41)			

3.2. Training and classification

Using the most used algorithms in machine learning, namely: SVM, RANDOM FOREST, Decision Tree KNN, and Adaboost classifiers, the geometric features presented in the previous sections will be classified into three classes according to customer's satisfaction.

- **SUPPORT VECTOR MACHINE (SVM):** SVM is a discriminative algorithm, which used to develop a supervised model. It used to classify or to predict regression [23]. In this algorithm, we aim to find a separator, which can classify neural, positive, and negative emotion class. This separator is baptized hyperplane. To find a very performant hyperplane, we tested kernel methods. We found that RBF Kernel is the best kernel for our model. We utilized xi, as a vector support, and σ, as positive float, which is, specify a priori, to compute RBF Kernel. Alternatively, RBF$K(x_i, x) = exp\left(\frac{-\|x_i - x\|^2}{2\sigma^2}\right)$. We used also S as a set of vectors support, which contains xi,b as bias, and Lagrangian coefficient αi, is a vector support, to compute hyperplane equation:

$$H(x) = \sum_{i \in S} \alpha_i \, y_i K(x_i, x) + b. \quad (1)$$

- **RANDOM FOREST:** Random forest [24] is an algorithm constructed by combination between trees. This algorithm used decision trees concept to classify. In decision Tree, we calculate Gini index or Gain ratio to construct the tree. In Random Forest, we compute just Gini index to find the most popular attributes founded by Decision Trees [25]. To compute Gini index, we used the probability, which link between class k and selected case $p\left(q/p\right)$. Or, Gini index can written as follow:

$$Gini(p) = 1 - \sum_{q=1}^{C} p\left(q/p\right)^2. \quad (2)$$

- **K NEAREST NEIGHBORS (KNN):** KNN [26] used k closest neighbors to classify. It can be used for classification or clustering. The concept of this algorithm is finding the class, which have the majority of k closest neighbors. Precisely, we compute distances between examples and we take the examples, which have the smaller distances.
- **ADABOOST (ADAPTIVE BOOSTING):** This classier is baptized also Adaptive Boosting. It is one of the classifiers, which based on the use of boosting concept. It combines between multiple classifiers. It is an iterative algorithm. Adaboost consists to set up a solid classifier by combining a various inadequately performing classifiers in order to product high performant classifier. Any classifier accepts weight can be used as base classifier of Adaboost classifier. Adaboost should meet two conditions:

1. The training should be done by using examples weighted.
2. In every iteration, it attempts to give a very good fit to these examples by minimizing the value of error.

3.3. Experimental Results

In this work, the Jaffe dataset is used to test the performance of our method. The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. 60 Japanese subjects have rated each image on 6 emotion adjectives. The database was planned and assembled by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. To use this dataset for our algorithm; we have classified its images into three classes. The first one contains images for satisfied persons, the second is for non-satisfied persons and the third class is for neural persons. Indeed, each face is represented in the form of a vector of 68 elements. Each element is the coordinate of a key point. All key points are calculated using a mask. These operations are applied to all Jaffe images. Next, the distances, shown by Table 1, are calculated. Therefore, the input of our classifiers SVM, RANDOM FOREST, KNN, decision tree and Adaboost classifiers will be vectors, which contain eighteen distances. Alternatively, a distance vector will represent each face. On the other hand, we computed confusion matrix of each classifier. We had also calculated accuracy, sensitivity, and specificity of each classifier. The following equations present the formula of specificity, sensitivity, and accuracy:

- Specificity is the ratio between non-satisfied and neutral faces, which are not classified as a satisfied face and the number total of non-satisfied and neutral faces.

$$\text{Specificity} = \frac{\text{non - satisfied \& neutral faces correctly classified}}{\text{All non - satisfied \& neutral faces}}. \quad (3)$$

- Sensitivity is the ratio between the number of satisfied faces, which are classified correctly, and the number of all satisfied faces (i.e. the satisfied faces, which are classified correctly, satisfied faces, which are classified neutral faces, and satisfied faces, which are classified as non-satisfied faces).

$$Sensitivity = \frac{\text{satisfied faces correctly classified}}{\text{Total of satisfied faces}}. \quad (4)$$

- Accuracy is the ratio between the number of satisfied, neutral, and non-satisfied faces are correctly classified and the number of all faces.

$$Accuracy = \frac{\text{Faces are correcty classified}}{\text{Total of all faces}}. \quad (5)$$

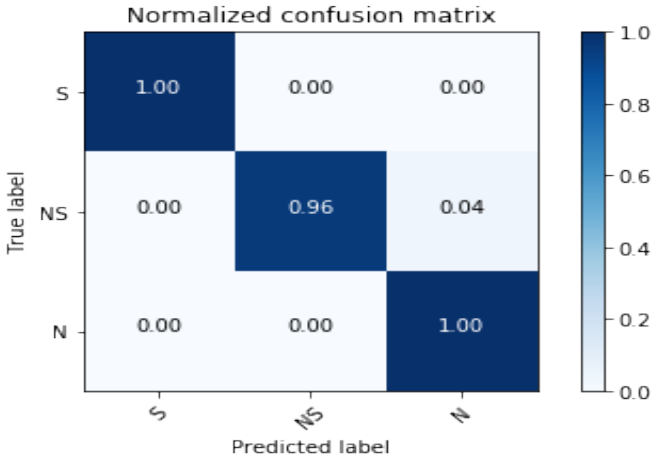


Figure. 2. Confusion matrices of our algorithm by applying SVM a JAFFE dataset

We also calculated the ROC curve (Receiver Operating Characteristics curve). ROC curve determinate the probability of distinguishing capability of our algorithm, i.e. how much our algorithm is able to separate between satisfied, neutral, and non-satisfied faces.

Table- 2: The sensitivity, specificity, and accuracy of applying our algorithm by using SVM, KNN, Decision Tree, Random Forest, and AdaBoost, and by utilizing JAFFE datasets

Dataset		Accuracy	Sensitivity	Specificity
JAFFE	SVM(RBF Kernel)	98.66%	100%	98%
	KNN	88.87%	90%	75.75%
	Decision Tree	98.66%	100%	98%
	Random Forest	77.76%	97%	62.77%
	AdaBoost	98.66%	100%	98%

We found that our algorithm detected 97% of satisfied faces correctly. In addition, our method extracted 91% non-satisfied faces correctly. Finally, it detected 100% of neutral faces correctly.

After computing confusion matrices of our algorithm, we computed the sensitivity, specificity, and sensitivity of our algorithm. We found the results showing into table 2. We calculate these statistic criteria for all classifier. We calculate the sensitivity of applying our algorithm by using SVM with RBF kernel, KNN, AdaBOOST, and Random Forest. We found that the most performant classifier is SVM with RBF Kernel. Table 2 shows the results of each classifier by using JAFFE datasets.

Our algorithm report that, geometrics features a gave a great model to measure satisfaction customers. Our algorithm consists to measure the variation of face's shape, by computing distances between face landmarks. Since, we based our classification on using the different variation between satisfied, neutral, and non-satisfied faces. For this reason, our algorithm a gave a great model to detect the satisfaction of customers by using emotions of customers faces.

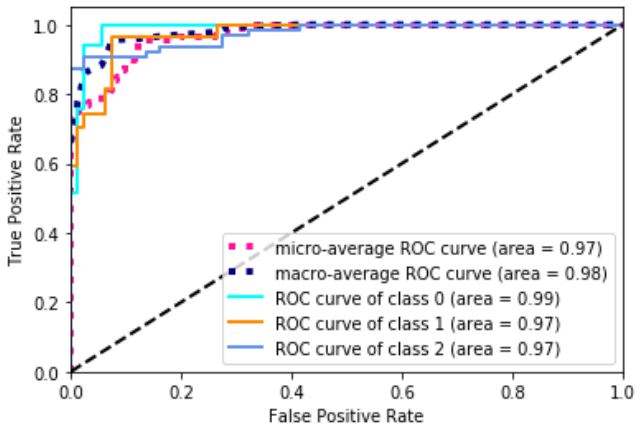


Figure. 3.ROC curve computed by applying our algorithm using SVM and JAFFE dataset.

4. Conclusion

One of the greatest challenges for marketing era is to ensure customer satisfaction. We have proposed a new method based on the classification of various facial expressions, to assess customer satisfaction using different classifiers: SVM, RANDOM FOREST, KNN and AdaBOOST. Distances between facial landmarks points represent the geometrics features. A vector of eighteen values represents customer's face expression. We tested our method on the JAFFE database and we obtained a performants result. In the future work, we will develop a multimodal algorithm, which contains speech and motion recognition.

References

[1] M.S. Bouzakraoui, S.Abdelalim, Y.A. Abdssamed, "Appreciation of Customer Satisfaction Through Analysis Facial Expressions and Emotions Recognition" in 2019 4th Edition of World conference on Complex Systems, Ouarzazate, Morocco, 2019. https://doi.org/10.1109/ICoCS.2019.8930761.

[2] F. Conejo, "The Hidden Wealth of Customers: Realizing the Untapped Value of your Most Important Asset", Journal of Consumer Marketing, **30**(5), 463-464, 2013. https://doi.org/10.1108/JCM-05-2013-0570.

[3] A. Mehrabian, Communication without words, Psychology today, 1968.

[4] S. M. C. Loureiro, " Consumer-Brand Relationship: Foundation and State-of-the-Art", Customer-Centric Marketing Strategies: Tools for Building Organizational Performance, 21, 2013, https://doi.org/10.4018/978-1-4666-2524-2.ch020

[5] C. Darwin, The expression of the emotions in man and animals, D. Appleton and Company, London, 1913

[6] J. Hamm, C.G. Kohler, R.C. Gur, R. Verma, "Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders", Journal of neuroscience methods, **200**(2), 237-256, 2011. https://doi.org/10.1016/j.jneumeth.2011.06.023

[7] A. A. Marsh, H. A. Elfenbein, N. Ambady - "Nonverbal "accents" cultural differences in facial expressions of emotion", Psychological Science, **14**(4), 373-376, 2003. https://doi.org/10.1111/1467-9280.24461

[8] M. Nishiyama, H. Kawashima, T. Hirayama, "Facial Expression representation based on timing structures in faces", in 2005 workshop on Analysis Modeling of faces and Gestures, Beijing, China, 2005. https://doi.org/10.1007/11564386_12

[9] K.K. Lee, Human expression and intention via motion analysis: Learning, recognition and system implementation, The Chinese University of Hong Kong Hong, 2004.

[10] M. Pantic, L.J.M. Rothkrantz, "An expert system for multiple emotional classification of facial expressions" in 1999 IEEE International Conference on Tools with Artificial Intelligence, Chicago, IL, USA, 1999. https://doi.org/10.1109/TAI.1999.809775

[11] Z. Zeng, J.T., M. Liu, T. Zhang, N. Rizzoto, Z. Zhang, "Bimodal HCI-related Affect Recognition" in 2004 International Conference Multimodal Interfaces, State College, PA, USA, 2004. https://doi.org/10.1145/1027933.1027958

[12] X. Zhao, S. Zhang, "A review on facial expression recognition: feature extraction and classification" IETE Technical Review, **33**(5), 505-517, 2016. https://doi.org/10.1080/02564602.2015.1117403

[13] K. Shan, J. Guo, W. You, D. Lu, R. Bie, "Automatic Facial Expression Recognition Based on a Deep Convolutional-Neural Network Structure" in 2017 IEEE 15th Int. Conf. Softw. Eng. Res. Manag. Appl. 123–128, London, United Kingdom, 2017. https://doi.org/10.1109/SERA.2017.7965717

[14] T.A. Rashid, "Convolutional neural networks based method for improving facial expression recognition" in 2016 Intelligent Systems Technologies and Applications, Jaipur, India, 2016. https://doi.org/10.1007/978-3-319-47952-1

[15] S. Ceccacci, A. Generosi, L. Giraldi, M. Mengoni, "Tool to Make Shopping Experience Responsive to Customer Emotions" International Journal of Automation Technology, **12**(3), 319-326, 2018. https://doi.org/10.20965/ijat.2018.p0319

[16] D. Matsumoto, M. Assar, "The effects of language on judgments of universal facial expressions of emotion" Journal of Nonverbal Behavior, **16**, 85-99, 1992. https://doi.org/10.1007/BF00990324

[17] R. Cui, M. Liu, M. Liu, "Facial expression recognition based on ensemble of multiple CNNs" in 2016 Chinese Conf. of Biometric Recognition, Chengdu, China, 2016. https://doi.org/10.1007/978-3-319-46654-5_56

[18] L. Huang, T. Chuan-Hoo, W. Ke, K. Wei, "Comprehension and Assessment of Product Reviews: A Review-Product Congruity Proposition" Journal of

Management Information Systems, **30**(3), 311-343, 2014. https://doi.org/10.2753/MIS0742-1222300311
[19] M. Slim, R. Kachouri, A. Atitallah, "Customer satisfaction measuring based on the most significant facial emotion", in 2018 15th IEEE International Multi-Conference on Systems, Signals & Devices, Hammamet, Tunisia, 2018. https://doi.org/10.1109/SSD.2018.8570636
[20] M. Wedel, R. Pieters - Review of marketing research, 2008.
[21] G. S. Shergill, A. Sarrafzadeh, O. Diegel, A. Shekar, "Computerized sales assistants: the application of computer technology to measure consumer interest – a conceptual framework", Journal of Electronic Commerce Research, 9(2), 176 – 191, 2008. DOI:10.1.1.491.5130
[22] P. M. A. Desmet, P. Hekkert, "The basis of product emotions", Pleasure with products: Beyond usability. ed. WS Green; PW Jordan, 61-67, 2002. DOI:10.1201/9780203302279.ch4
[23] C. Cortes, V. Vapnik, "Support-Vector Networks", Machine Learning, 20, 273-297, 1995. https://doi.org/10.1023/A:1022627411411
[24] B.Leo, "Random Forests", Machine Learning, 45, 5-32, 2001. https://doi.org/10.1023/A:1010933404324
[25] Raileanu, L. Elena, K. Stoffel, "Theoretical Comparison between the Gini Index and Information Gain Criteria," Annals of Mathematics and Artificial Intelligence 41, 77-93, 2004. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6
[26] G. Laencina, J. Pedro, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation" Neurocomputing, 72(7-9), 1483-1493, 2009. https://doi.org/10.1016/j.neucom.2008.11.026

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No.5, 613-619 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Typological Study of Portuguese Mortality from Non-communicable Diseases

Ana Paula Nascimento[1,*], Cristina Prudêncio[2], Mónica Vieira [3], Rui Pimenta [4], Helena Bacelar-Nicolau[5]

[1]*Biomatemática, Bioestatística e Bioinformática, Escola Superior de Saúde-Instituto Politécnico do Porto (ESS-IPP), Centro de Investigação em Saúde e Ambiente da Escola Superior de Saúde do Instituto Politécnico do Porto (CISA), Porto, 4000, Portugal*

[2]*Ciências Químicas e das Biomoléculas, Escola Superior de Saúde-Instituto Politécnico do Porto (ESS-IPP), Instituto de Investigação e Inovação em Saúde da Universidade do Porto (i3S), Centro de Investigação em Saúde e Ambiente da Escola Superior de Saúde do Instituto Politécnico do Porto (CISA), Porto, 4000, Portugal*

[3]*Ciências Químicas e das Biomoléculas, Escola Superior de Saúde-Instituto Politécnico do Porto (ESS-IPP), Centro de Investigação em Saúde e Ambiente da Escola Superior de Saúde do Instituto Politécnico do Porto (CISA), Instituto de Investigação e Inovação em Saúde da Universidade do Porto (i3S), Porto, 4000, Portugal*

[4]*Biomatemática, Bioestatística e Bioinformática, Escola Superior de Saúde-Instituto Politécnico do Porto (ESS-IPP), Centro de Estudos e Investigação em Saúde da Universidade de Coimbra, Porto, 4000, Portugal*

[5]*Faculdade de Psicologia da Universidade de Lisboa (FPUL), Instituto de Saúde Ambiental da Faculdade de Medicina da Universidade de Lisboa (ISAMB/FMUL), Lisboa, 1000, Portugal*

ARTICLE INFO

Article history:
Received: 10 July, 2020
Accepted: 19 September, 2020
Online: 05 October, 2020

Keywords:
Public Health
Non-communicable diseases
Date visualization
Cluster analysis

ABSTRACT

The most common non-communicable diseases, such as cardiovascular diseases and cancer, are a problem in global and national growth. The World Health Organization considers it a priority to study the specific causes of these diseases for trend monitoring. The aim of this paper is to identify a hierarchy of clusters of Portuguese mortality by non-communicable diseases using the agglomerative hierarchical cluster analysis. The Euclidean distance with complete linkage and average linkage criteria are used. These methods identify six clusters with both criteria, indicating some order of disease severity in the way clusters joint together. Special attention should be given to diseases in the last two clusters, where the last one is formed by ischemic heart disease, cerebrovascular diseases and larynx / trachea / bronchi and lung malignant tumor, all for males. In fact, these clustering results show that male gender seems to be a risk factor for at least two groups of the non-communicable diseases. Other suggested risk factors and / or pathophysiological mechanisms that in a direct or indirect way may enhance the common development of the pathologies found in the clusters arising from this study should also be an object of priority study.

1. Introduction

An important challenge of Public Health in Portugal is to decrease the number of deaths from Non-communicable diseases. In order to generate additional information in this domain, this paper is an extension of work originally presented in 14th Iberian Conference on Information Systems and Technologies [1].

According to the World Health Organization (WHO), the number of people who will suffer from the most common non-communicable diseases, including cardiovascular disease, cancer, or chronic respiratory diseases, will increase [2]. Non-communicable diseases are a worldwide and national problem and Public Health concerns [2], [3]. The study of non-communicable diseases is a matter of major importance in the world and in particular in Portugal [4], [5]. According to the Directorate General for Health (DGS) cardiovascular diseases and major cancers are mostly responsible for the lost years of life in the Portuguese population aged 35 and over [6]. The mortality rates of lung tumor malignant and rectal tumor malignant are increasing in Portugal

*Corresponding Author: Ana Paula Nascimento, ananascimento@ess.ipp.pt

[6]. Decreasing mortality from cancer is the biggest challenge for the next generations because cancer is the leading cause of early mortality [7]. The decrease in early mortality is relevant due to its individual and social impact [2]. Thus, according to WHO, it is of great importance to generate additional knowledge on the specific causes of these diseases, in order to monitor their trends [4]. Therefore, different studies are fundamental to analyze diseases evolution that might help to formulate hypotheses about risk factors and / or common pathophysiological mechanisms that, directly or indirectly, may enhance the common development of pathologies found in clusters, resulting from the hierarchy of non-communicable disease partitions, as this may be a working tool in health [8][9]. Unsupervised data analysis, namely cluster analysis, then can be useful to raise hypotheses about clusters of non-communicable diseases to be studied.

The development of the clustering analysis methodology is truly interdisciplinary. Taxonomists, social scientists, health scientists, psychologists, biologists, statisticians, mathematicians, engineers, medical researchers, computer scientists and others all contribute to the development of this methodology [10], [11]. Clustering analysis is a set of exploratory multivariate data analysis methods for identifying natural clusters in the data, based on a coefficient of similarity or dissimilarity between individuals or between variables, or more generally between statistical units of data [12]-[14]. This methodology aims to find homogeneous groups of statistical units in the data, where similarities between statistical units belonging to the same group are high and similarities between statistical units belonging to distinct groups are low [10], [12]-[14]. Hierarchical cluster analysis algorithms, both agglomerative and divisive, provide a hierarchy of partitions. Agglomerative hierarchical classification analysis methods are the most used [14][15]. These methods usually start with all statistical units separated into single-element groups (unit set - singletons), forming a cluster partition where the number of clusters is equal to the number of statistical units, and successively grouping together the most similar groups (according to the measure of similarity or dissimilarity between statistical units and the criterion of aggregation between groups) in the same cluster, until it forms a partition of a single cluster. It is recognized that agglomerative hierarchical cluster analysis can give different results for the same data, depending on the choice of the measure of comparison between statistical units and the aggregation criterion between groups [13], [16], [17]. One of the most used measures between individuals is the Euclidean distance. Moreover, the most commonly used aggregation algorithms are the farthest neighbor (or complete linkage) and the nearest neighbor (or single linkage) ones [10][12]. The complete linkage algorithm produces compact clusters, but is sensitive to outliers [16]-[18]. The single linkage algorithm is censored because it may disregard the structure of clusters, and it has the tendency to construct hierarchical partitions with chain effect [13], [16]-[18]. The average linkage method produces more balanced hierarchies, with no chain effect. This algorithm is an intermediate method between the two previously mentioned, and furthermore uses more data information [13], [16]-[18]. Therefore, the three aggregation criteria are often used over the same data sets, with the aim of knowing about clustering results robustness. Each agglomerative hierarchical cluster analysis is usually based on information concerning some "relative best partitions" of the hierarchy, namely immediately preceding relative maximum levels of increase / decrease in cluster dissimilarities [13], [19]. Moreover, by choosing a cut-off where there is the largest increase / decrease in cluster dissimilarities [13], it allows us to find an "absolute best partition". This suggests that, at the stage where the greatest increase / decrease occurs, the joining clusters are comparatively far apart. Therefore, the best number of groups in the data should be the number of groups present in the immediately preceding step [13], [16], [17].

Goal: The aim of this study is to obtain a hierarchy of cluster partitions of Portuguese mortality from non-communicable diseases, using agglomerative hierarchical cluster analysis, in order to identify clusters of diseases associated to different degrees of severity as well as clusters of diseases explicitly associated with gender. Clustering results might help either to identifying or to searching for common causes or risk factors to improve preventive medicine.

2. Materials and Methods

2.1. Data

The classification of diseases in this paper is according to the Tenth Revision International Classification of Diseases (ICD-10). This classification makes it possible for countries compile comparable national mortality [20].

Table 1: Non-communicable diseases (ICD-10)

Non-communicable diseases	Code
Lip/mouth and pharynx malignant tumor for males	1
Esophageal malignant tumor for males	2
Stomach malignant tumor for males	3
Colon malignant tumor for males	4
Rectal and anus malignant tumor for males	5
Liver / biliary / intrahepatic malignant tumor for males	6
Pancreatic malignant tumor for males	7
Larynx / trachea / bronchi and lung malignant tumor for males	8
Skin malignant tumor for males	9
Breast malignant tumor for males	10
Kidney malignant tumor for males	11
Bladder malignant tumor for males	12
Lymphatic tissue malignant tumor for males	13
Ischemic heart disease for males	14
Other heart diseases (I30-I33) for males	15
Other heart diseases (I39-I52) for males	16
Cerebrovascular diseases for males	17
Pneumonia for males	18
Chronic lower respiratory diseases for males	19
Asthma for males	20
Lip/mouth and pharynx malignant tumor for females	21
Esophageal malignant tumor for females	22
Stomach malignant tumor for females	23
Colon malignant tumor for females	24
Rectal and anus malignant tumor for females	25
Liver / biliary / intrahepatic malignant tumor for females	26
Pancreatic malignant tumor for females	27
Larynx / trachea / bronchi and lung malignant tumor for females	28
Skin malignant tumor for females	29
Breast malignant tumor for females	30
Kidney malignant tumor for females	31
Bladder malignant tumor for females	32
Lymphatic tissue malignant tumor for females	33

Ischemic heart disease for females	34
Other heart diseases (I30-I33) for females	35
Other heart diseases (I39-I52) for females	36
Cerebrovascular diseases for females	37
Pneumonia for females	38
Chronic lower respiratory diseases for females	39
Asthma for females	40

The non-communicable diseases used here are listed in Table 1. In this paper, it is presented an evolution of a previous work presented at [1]. In fact, a results analysis is performed considering diseases separated by gender. Table 2 displays the shorts names of the diseases, where M represents male gender and F represents female gender.

Table 2: Short names of non-communicable diseases

Short names of non-communicable diseases	Code
Lip_M	1
Esophageal_M	2
Stomac_M	3
Colón	4
Rectal_M	5
Liver_M	6
Pancreatic_M	7
Lung_M	8
Skin_M	9
Breast_M	10
Kidney_M	11
Bladder_M	12
Lymphatic_M	13
Ischemic_M	14
I30-I33_M	15
I39-I52_M	16
Cerebrovascular_M	17
Pneumonia_M	18
Chr.lowerrespiratory_M	19
Asthma_M	20
Lip_F	21
Esophageal_F	22
Stomac_F	23
Colon_F	24
Rectal_F	25
Liver_F	26
Pancreatic_F	27
Lung_F	28
Skin_F	29
Breast_F	30
Kidney_F	31
Bladder_F	32
Lymphatic_F	33
Ischemic_F	34
I30-I33_F	35
I39-I52_F	36
Cerebrovascular_F	37
Pneumonia_F	38
Chr.lowerrespiratory_F	39
Asthma_F	40

Data were requested from Statistics Portugal. Non-communicable diseases separate by gender are considered as individuals and annual age-standardized mortality rates in Portugal from 1994 to 2012 for non-communicable diseases are considered as variables. Annual age-standardized mortality rates are given by the ratio of the total number of deaths expected in the specific region for a given gender for a given disease over a year and the number of standard populations per 100 000 inhabitants. In theory any standard population could be used but frequently are used the Segi world, the European or the WHO world standard populations. Here it is used the European standard population. Age-standardized mortality rates enable comparisons to be made between populations that have different age structures. More details on methods of standardizing mortality rates can be found in [21], [22]. The mortality rates considered are based on individuals up to 65 years of age.

2.2. Methodology

In this paper agglomerative hierarchical cluster analysis of non-communicable diseases is applied. The age-standardized mortality rates for each of the years from 1994 to 2012 are used as variables to perform the analysis and the non-communicable diseases separated by gender are used as individuals. Since annual standardized mortality rates are quantitative variables, the Euclidean distance to measure disease dissimilarity is used. Euclidean distance depends less from the magnitude of values than the quadratic Euclidean distance, used in [1]. As aggregation criteria the three methods mentioned above were used. The clustering results were very similar, so in this presentation are only considered the average linkage and the complete linkage criteria.

3. Results

For all diseases listed in Table 1 in the case of male gender are presented in Table 3 the mean, standard deviation, minimum and maximum of annual age-standardized mortality rates between 1994 and 2012.

Table 3: Statistics summary of annual age- standardized mortality rates between 1994 a 2012 given by non-communicable diseases for males

Diseases	Mean	Standard Deviation	Min.	Max.
Lip_M	7.06	0.57	6.00	8.20
Esophageal-M	4.83	0.41	4.20	5.60
Stomac_M	10.48	1.49	8.50	12.90
Colon_M	6.33	0.34	5.50	6.80
Rectal_M	3.17	0.25	2.60	3.60
Liver_M	3.74	0.62	2.80	5.10
Pancreatic_M	3.84	0.37	3.30	4.60
Lung_M	25.21	0.91	23.60	27.60
Skin_M	0.86	0.21	0.50	1.20
Breast_M	0.15	0.05	0.10	0.20
Kidney_M	1.36	0.19	1.10	1.70
Bladder_M	1.81	0.26	1.40	2.30
Lymphatic_M	6.85	0.70	5.60	8.10
Ischemic_M	24.62	6.91	13.80	33.60
I30I33_M	0.44	0.11	0.30	0.60
I39I52_M	6.44	1.77	4.10	8.90
Cerebrovascular_M	20.54	6.55	12.10	30.30
Pneumonia_M	6.04	1.55	4.00	9.00
Chr.lowerrespiratory_M	4.22	1.42	2.40	6.80
Asthma_M	0.49	0.32	0.20	1.30

As shown in Table 3 the individuals of male gender died more by larynx / trachea / bronchi and lung malignant tumor, ischemic

heart disease and cerebrovascular diseases in Portugal between 1994 and 2012. Note that the standard deviation of annual age-specific standardized mortality rates in the case of larynx / trachea / bronchi and lung malignant tumor is very low so the mean is well representative of the mortality rates over the years. The standard deviation of annual age-specific standardized mortality rates in the cases of ischemic heart disease and cerebrovascular diseases is bigger than in all other cases. In fact, the mortality rates in these cases presented a decrease over the years while in the case of larynx / trachea / bronchi and lung malignant tumor the mortality rates are very similar. It can also be seen that in every other disease the standard deviation is low, so that does not exist a great difference between the mortality rates over the years. Table 4 presents the statistics summary in case of female gender.

Table 4: Summary statistics of annual age-standardized mortality rates between 1994 a 2012 given by non-communicable diseases for females

Diseases	Mean	Standard Deviation	Min.	Max.
Lip_F	0.64	0.14	0.40	0.90
Esophageal_F	0.34	0.12	0.10	0.50
Stomac_F	4.72	0.76	3.60	6.20
Colon_F	4.11	0.26	3.80	4.70
Rectal_F	1.74	0.15	1.40	2.00
Liver_F	0.95	0.15	0.70	1.20
Pancreatic_F	1.70	0.21	1.40	2.30
Lung_F	4.13	0.70	3.00	5.60
Skin_F	0.68	0.12	0.60	1.00
Breast_F	13.78	1.74	11.60	17.30
Kidney_F	0.48	0.08	0.40	0.70
Bladder_F	0.32	0.10	0.10	0.50
Lymphatic_F	4.56	0.58	3.60	5.50
Ischemic_F	6.13	2.05	3.00	9.40
I30I33_F	0.17	0.07	0.10	0.30
I39I52_F	2.82	1.08	1.40	4.40
Cerebrovascular_F	10.39	3.83	5.40	16.00
Pneumonia_F	2.18	0.59	1.30	3.40
Chr.lowerrespiratory_F	1.38	0.57	0.70	2.40
Asthma_F	0.32	0.20	0.00	0.80

As shown in Table 4, female gender individuals died more by breast malignant tumor, cerebrovascular diseases and ischemic heart disease in Portugal between 1994 and 2012. Note that the annual age-specific standardized mortality rates in the case of breast malignant tumor presents a decrease until 2004 while their values are very similar after that year. In the case of cerebrovascular diseases the mortality rates present a decrease until 2006 and are very similar after. For ischemic heart disease the mortality rates decrease over the years, nevertheless cerebrovascular diseases and ischemic heart disease seem to have a much higher mortaly rate in male than in female. In other cases the standard deviation of annual age-specific standardized mortality rates is very low, showing stability of mortality rates.

From the application of the agglomerative hierarchical cluster algorithms, the following results were obtained.

3.1. Agglomerative hierarchical cluster analysis with Euclidean distance and average linkage criterion

Table 5 contains the aggregation order matrix from the application of the agglomerative hierarchical cluster analysis with the Euclidean distance and average linkage aggregation criterion between groups. In Table 5 the columns Cluster 1 and Cluster 2 indicate the disease groups aggregating at each stage. The aggregation coefficient value in each step is noted by “Aggr. Coef.” and the increasing aggregation coefficient between successive steps is noted by “Dif.”.

Table 5: Aggregation order matrix of the cluster analysis with the Euclidean distance and average linkage aggregation criterion

Stage	Cluster1	Cluster2	Aggr. Coef.	Dif.
1	10	35	,332	-------
2	15	31	,592	0,260
3	22	32	,721	0,129
4	21	29	,775	0,54
5	15	22	,797	0,22
6	15	40	,901	0,104
7	10	15	1,124	0,223
8	25	27	1,140	0,016
9	12	25	1,289	0,149
10	9	21	1,334	0,045
11	10	20	1,490	0,156
12	9	26	1,499	0,009
13	23	33	1,982	0,483
14	11	12	2,011	0,029
15	9	10	2,220	0,209
16	7	24	2,486	0,266
17	6	28	2,516	0,03
18	16	34	3,013	0,497
19	11	39	3,081	0,068
20	6	7	3,204	0,123
21	11	38	3,838	0,634
22	1	4	3,878	0,04
23	2	23	4,184	0,306
24	5	6	4,196	0,012
25	1	13	4,553	0,357
26	2	5	5,439	0,886
27	9	11	5,542	0,103
28	16	18	5,854	0,312
29	19	36	6,519	0,665
30	2	19	7,273	0,754
31	1	16	8,281	1,008
32	3	37	10,544	2,263
33	1	2	12,334	1,79
34	3	30	16,129	3,795
35	1	9	18,145	2,016
36	14	17	19,928	1,783
37	8	14	32,412	13,132
38	1	3	39,753	7,341
39	1	8	89,941	50,188

Figure 1 presents the dendrogram associated to Euclidean distance and average linkage aggregation criteria.

From Table 5 and Figure 1 it may be observed the dendrogram has 39 levels. Reading the tree from top to bottom we find six well separated groups of diseases. The first group includes breast malignant tumor for males, other heart diseases (I30-I33) for females and males, kidney malignant tumor, esophageal malignant

tumor, bladder malignant tumor and asthma for females, asthma for males, lip/mouth and pharynx malignant tumor, as well as skin malignant tumor for females, skin malignant tumor for males and liver / biliary / intrahepatic malignant tumor for females. Thus the first group includes a mixture of diseases, where some paired diseases for females and males appear.

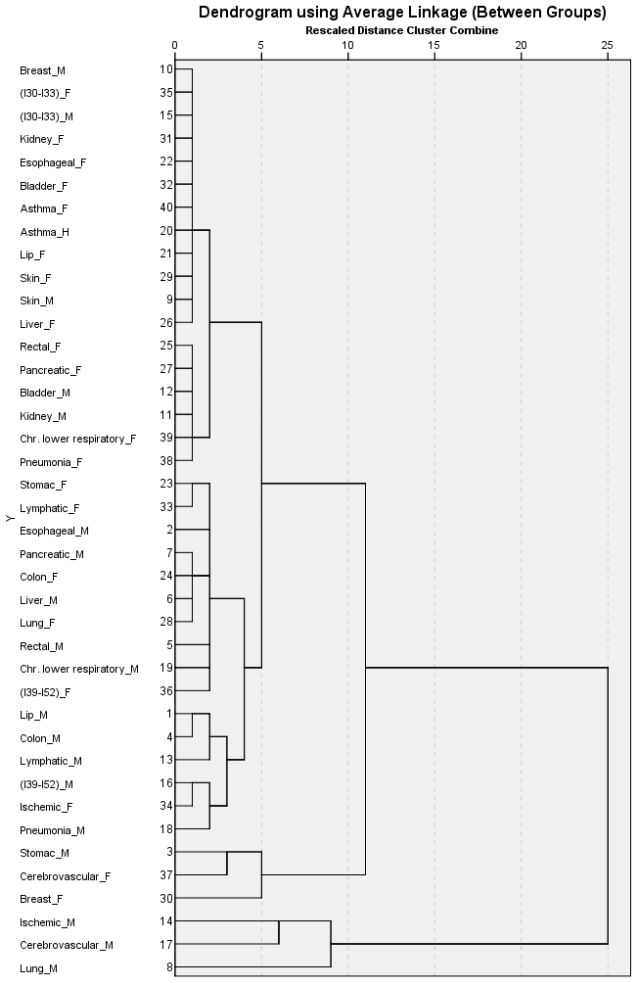


Figure 1: Dendrogram with Euclidean distance and average linkage aggregation criterion

The second cluster comprises: rectal and anus malignant tumor as well as pancreatic malignant tumor for females, bladder malignant tumor and kidney malignant tumor for males and chronic lower respiratory diseases and pneumonia for females. Note that the large cluster resulting from the union of these two groups mainly consists of diseases for females. The third group contains: stomach malignant tumor and lymphatic tissue malignant tumor for females, esophageal malignant tumor and pancreatic malignant tumor for males, colon malignant tumor for females, liver / biliary / intrahepatic malignant tumor for males, larynx / trachea / bronchi and lung malignant tumor for females, chronic lower respiratory diseases for males, other heart diseases (I39-I52) for females and rectal and anus malignant tumor for males. The fourth group contains: lip/mouth and pharynx malignant tumor, colon malignant tumor, lymphatic tissue malignant tumor, other heart diseases (I39-I52) and pneumonia, all for males and ischemic heart disease for females. Thus, this group is mainly constituted by diseases in male individuals. The fifth group contains: stomach malignant tumor for males, cerebrovascular diseases and breast malignant tumor for females. Finally, the sixth group includes: ischemic heart disease, cerebrovascular diseases and larynx / trachea / bronchi and lung malignant tumor, all for males.

3.2. Agglomerative hierarchical cluster analysis with Euclidean distance and complete linkage criterion

Figure 2 presents the dendrogram associated to Euclidean distance and complete linkage aggregation criterion.

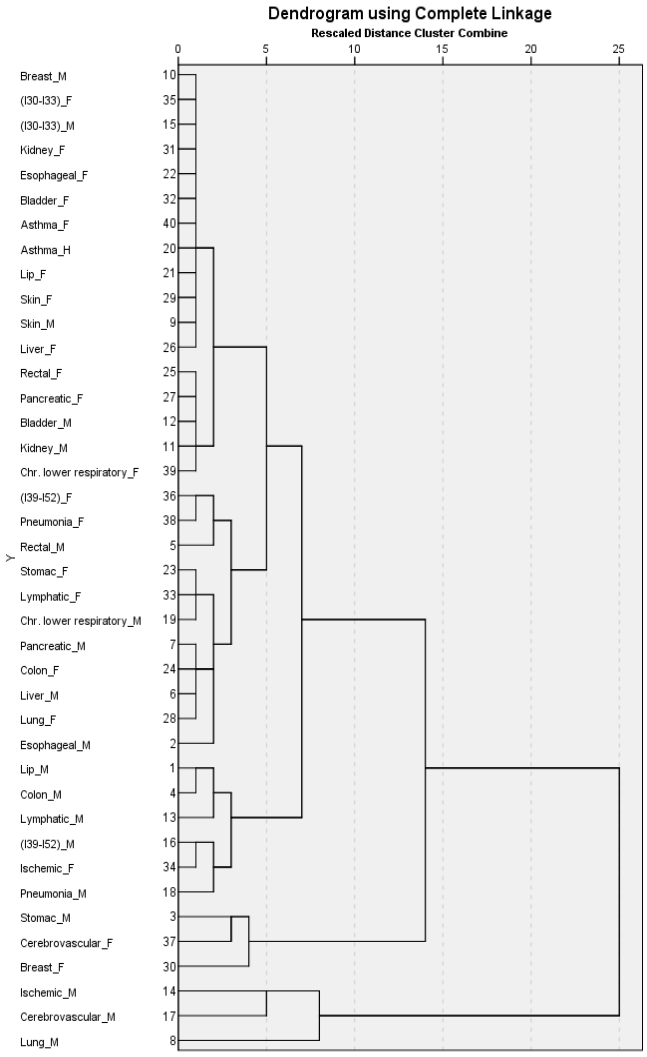


Figure 2: Dendrogram with the Euclidean distance and complete linkage aggregation criterion

As we point out above, the aggregation complete linkage and average linkage criteria gave very similar clustering results over our data. Comparing Figure 1 and Figure 2, it can be seen they both display dendrograms with 39 levels, showing six well separated clusters of diseases, grouped following a similar order in the hierarchy. Moreover, reading the tree from top to bottom in Figure 2, and comparing each cluster to the corresponding cluster in Figure 1, we find that: the first group is the same obtained with the average linkage aggregation criterion; the second group differs from the corresponding in average linkage criterion only because it doesn't include the pneumonia for females; the third group contains all the diseases included in the third cluster of average

linkage criterion plus the pneumonia for females; the last three groups are the same obtained with the average linkage criterion.

4. Discussion

Let's point out again that for analyzing and understanding clustering results obtained from hierarchical agglomerative methods based on Euclidean distance applied to diseases, it is expected that: 1- diseases in the same cluster will have similar mortality rates and similar behavior of the mortality rates over the years; 2- clusters of diseases indicate that the order of magnitude of mortality rates and the behavior over the years of mortality rates show larger differences between diseases in different clusters than diseases inside each cluster. Both aggregation criteria produce similar hierarchies with six well separated clusters (differing only by the classification of pneumonia for females), joining in same way. If one refers to the average linkage aggregation criterion the first group is formed at level 15, the second one at level 21, the third one at level 30, the fourth group at level 31 the fifth at level 34, and the six group at level 37, corresponding to some kind of chain effect in the dendrogram, indicating some order of disease severity in the way clusters joint together. The cut-off occurs at level 38, where the sixth cluster is separated from the cluster of every other diseases. Reading the tree from top to bottom, the first three groups include diseases with lower mortality rates and the last three put together the diseases with highest mortality rates. Then the diseases that seem to need more attention are the diseases in clusters four to six. Looking at Figures 1 and 2 it is clear that groups formed by diseases with lower mortality mainly occur for females while groups formed by diseases with higher mortality tend to be associated to male gender, so male gender appears to be a risk factor in many of listed diseases. The fourth cluster reveals that lip/mouth and pharynx malignant tumor, colon malignant tumor, lymphatic tissue malignant tumor, other heart diseases (I39-I52) in male gender, ischemic heart disease for females, and pneumonia for males need the same attention by health organizations in Portugal since they are in the same cluster. This suggests that may be there are common causes for these diseases that affect more the male gender than female gender. Looking again at Figure 1 and Figure 2, it can be seen that the diseases that cause more deaths are found in fifth and sixth clusters. In the case of breast malignant tumor and stomach malignant tumor, mortality rates have been decreasing [6], possibly due to preventive measures; however, it can be seen in the fifth cluster that the stomach malignant tumor for males provoke so many deaths like cerebrovascular diseases and breast malignant tumor for females and they come together with high values of mortality rate. Note that in the case of breast malignant tumor being woman is a risk factor. Diet and obesity possibly being risk factors for the diseases in the fifth cluster more preventive measures are necessary in that direction. The sixth cluster, very well separated from all the other clusters, includes ischemic heart disease, cerebrovascular diseases and larynx / trachea / bronchi and lung malignant tumor for male gender, diseases which present the largest mortality rates, showing that they deserve special attention. This cluster also appears very well separated from all the other clusters in previous work [1]. The results shown in the present work, point that the male gender may be a risk factor for some clusters of diseases. One explanation may be habits more associated to the male gender as smoking. Additionally, the role of hormonal and metabolic characteristics of the male versus female gender is yet to understand. The results confirm that mortality rates of the respiratory tract malignant tumors and cerebrovascular diseases are the highest and it is necessary to reinforce preventive measures in case of these diseases even that Portugal is significantly better than the mean of the high-middle Socio-Demographic Index group for ischemic heart disease and lung cancer [6]. The fact that ischemic heart diseases and cerebral vascular diseases are in the same cluster it is expected. In both pathologies there are common underlying causes such as the formation of atheromatous plaques and their relationship with obesity and hypercholesterolemia. The similarity of these pathologies to malignant tumors of the respiratory tract is may be related with the known increase of thromboembolic events secondary to tumors but this requires further studies to understand and elucidate this result. However, risk factors such as smoking are common to the three pathologies which may further relate those pathologies [23].

5. Conclusion

Public Health as a science for studying and preventing diseases, prolonging life and improving quality of life through organized efforts and informed choices needs to analyze the factors of health of a population [24]. Unsupervised data analysis can introduce new knowledge allowing organizations to take measures to provide the best health warranties of the general population.

The aim of this study was to search for an agglomerative hierarchical cluster analysis of Portuguese mortality from non-communicable diseases, in order to first of all identify clusters of diseases associated to different degrees of severity as well as clusters of diseases explicitly correlated with gender.

The obtained hierarchy provides six main clusters of diseases. Moving up the hierarchy it is found that these clusters are sequentially formed, some increasing order of disease severity appearing correspond to the increasing order of levels. The first groups include diseases with lower mortality rates while the last ones put together diseases with higher mortality rates.

Special attention should be given to the last three clusters since they contain diseases showing high degrees of mortality rates compared with other clusters. Moreover, the fourth cluster mainly includes diseases in male individuals and the sixth one, grouping the highest severe diseases, refers only to males. Thus, male gender seems to be a risk factor for these two clusters of diseases. Other suggested risk factors and / or pathophysiological mechanisms that in a direct or indirect way may enhance the common development of the pathologies found in the clusters arising from this study should also be an object of priority study.

Note that taking gender in account in the present study clearly improve results obtained in previous work [1]. Clustering results might help either to identifying or to searching for common causes or risk factors to improve preventive medicine.

One limitation of this work is the fact that the time period for study does not include data from recent years yet. It happens the work is included in a larger project and it is important to get these results. In future, recent data will be added, as well as countries comparison and other multivariate data analysis (e. g., factor analysis and other cluster analysis methods, like fuzzy hierarchy techniques). Data collection is in process.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors would like to thank Professor J. Pereira Miguel for introducing us to the analysis of these data, highlighting the importance of its study in preventive medicine and public health. The authors would also like to acknowledge Professor Brígida Mónica Faria for her comments and suggestions that greatly valued this work.

Authors would like to acknowledge Escola Superior de Saúde – P.Porto and Centro de Investigação em Saúde e Ambiente da Escola Superior de Saúde do Instituto Politécnico do Porto (CISA).

References

[1] A.P. Nascimento, C. Prudêncio, M. Vieira, R. Pimenta, H. Bacelar-Nicolau, "Cluster analysis of noncommunicable diseases in Portugal," in Iberian Conference on Information Systems and Technologies, CISTI, 2019, doi:10.23919/CISTI.2019.8760798.

[2] WHO, World Health Statistics 2012, 2012.

[3] R. Beaglehole, R. Bonita, R. Horton, C. Adams, G. Alleyne, P. Asaria, V. Baugh, H. Bekedam, N. Billo, S. Casswell, M. Cecchini, R. Colagiuri, S. Colagiuri, T. Collins, S. Ebrahim, M. Engelgau, G. Galea, T. Gaziano, R. Geneau, A. Haines, J. Hospedales, P. Jha, A. Keeling, S. Leeder, P. Lincoln, M. McKee, J. Mackay, R. Magnusson, R. Moodie, et al., "Priority actions for the non-communicable disease crisis," The Lancet, **377**(9775), 1438–1447, 2011, doi:https://doi.org/10.1016/S0140-6736(11)60393-0.

[4] A. Alwan, T. Armstrong, M. Cowan, et al, Noncommunicable Diseases Country Profiles 2011, 2011.

[5] WHO, Noncommunicable Diseases Country Profiles 2014, 2014.

[6] Direção-Geral de Saúde, Portugal : The Nation's Health 1990 – 2016. An overview of the Global Burden of Disease Study 2016 Results, Seatle, WA, 2018.

[7] Direção-Geral de Saúde, A Saúde dos Portugueses 2016, 2017, doi:ISSN: 2183-5888.

[8] T.N. Haregu, F.M. Wekesah, S.F. Mohamed, M.K. Mutua, G. Asiki, C. Kyobutungi, "Patterns of non-communicable disease and injury risk factors in Kenyan adult population: A cluster analysis," BMC Public Health, **18**(Suppl 3), 2018, doi:10.1186/s12889-018-6056-7.

[9] G. Pistolla, P. Prastacos, N. Tzanakis, A. Philalithis, "Clustering of mortality rates in Greeces prefectures," Scientific Research and Essays, **7**(45), 3860–3876, 2012, doi:10.5897/SRE11.1754.

[10] A.K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, **31**(8), 651–666, 2010, doi:10.1016/j.patrec.2009.09.011.

[11] H. Bacelar-Nicolau, F.C. Nicolau, A. Sousa, L. Bacelar-Nicolau, "3rd Stochastic Modeling Techniques and Data Analysis International Conference," in: Skiadas, C. H., ed., in One Cluster Analysis of Complex and Heterogeneous Data, International Society for the Advancement of Science and Technology (ISAST), Lisboa: 99–108, 2014.

[12] M.G.H. Omran, A.P. Engelbrecht, A. Salman, "An overview of clustering methods," Intelligent Data Analysis, **11**, 583–605, 2007, doi:10.3233/IDA-2007-11602.

[13] H. Bacelar-Nicolau, F. Nicolau, Á. Sousa, L. Bacelar-Nicolau, "Measuring similarity of complex and heterogeneous data in clustering of large data sets," Biocybernetics and Biomedical Engineering, **29**(2), 9–18, 2009.

[14] H. Bacelar-Nicolau, F.C. Nicolau, Á. Sousa, L. Bacelar-Nicolau, "Clustering of variables with a three-way approach for health sciences," TPM - Testing, Psychometrics, Methodology in Applied Psychology, **21**(4), 435–447, 2014, doi:10.4473/TPM21.4.5.

[15] F. Murtagh, P. Contreras, "Algorithms for hierarchical clustering: An overview," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **0**, 86–97, 2011, doi:10.1002/widm.53.

[16] C. Chatfield, A.J. Collins, Introduction to Multivariate Analysis, Chapman & Hall/CRC, 2018.

[17] B.S. Everitt, Cluster Analysis, 5th ed., Wiley, 2011, doi:10.1007/BF00154794.

[18] F. Murtagh, L. Pierre, "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," Journal of Classification, **31**, 274–295, 2014, doi:10.1007/s00357.

[19] O. Silva, A. Sousa, H. Bacelar-Nicolau, F.C. Nicolau, "Encontro Nacional da Sociedade Portuguesa de Matemática, Probabilidades e Estatística," in: Antunes, N., ed., in Comparação de pares de particões em análise classificatória, Sociedade Portuguesa de Matemática: 175–178, 2012.

[20] H. Quan, V. Sundararajan, P. Halfon, A. Fong, "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 Administrative Data," Medical Care, **43**(11), 2005.

[21] R. Beaglehole, R. Bonita, T. Kjellstrom, Basic Epidemiology, WHO, Geneva, 1993.

[22] S.K. Lwanga, C.Y. Tye, O. Ayeni, Teaching health statistics: lesson and seminar outlines, 1999.

[23] M. Palmer, J. Sutherland, S. Barnard, A. Wynne, E. Rezel, A. Doel, L. Grigsby-Duffy, S. Edwards, S. Russell, E. Hotopf, P. Perel, C. Free, "The effectiveness of smoking cessation, physical activity/diet and alcohol reduction interventions delivered by mobile phones for the prevention of non-communicable diseases: A systematic review of randomised controlled trials," PLoS ONE, **13**(1), 1–18, 2018, doi:10.1371/journal.pone.0189801.

[24] Elena Andresen, E.D. Bouldin, eds., Public Health Foundations: Concepts and Practices, John & Son, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No.5, 620-629 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Design Process in the Improvement of The Experience Between a Brand and its Target Audience Through a Digital Product: The Lexus Portugal's used Car Website Case Study

Nuno Martins[1,*], Juan-Ramon Martin-Sanroman[2], Fernando Suárez-Carballo[2]

[1]*Polytechnic Institute of Cavado and Ave / ID+, Barcelos, 4750-810, Portugal*

[2]*Facultad de Comunicación de la Universidad Pontificia de Salamanca, Salamanca, 37007, Spain*

ARTICLE INFO

Article history:
Received: 24 July, 2020
Accepted: 21 September, 2020
Online: 05 October, 2020

Keywords:
Design Process
Lexus Portugal
UI & UX Design
Digital Design
Communication Design

ABSTRACT

The study aims to demonstrate how the use of the design process can align a brand's strategy with the interests of its target audience through a digital product based on a case study. Currently, Lexus internal studies show that there is a possibility to meet the needs of new audiences, beyond the traditional ones (men, 50+). In order to achieve this goal, the following objectives have been defined: communicate the brand and its cars' main values and philosophy, namely ecology, economy, safety and comfort; promote its services and products; and design an interface that guarantees users easy, pleasant and attractive navigation. The work process consisted of identifying the brand's strategy and values, identifying users, analyzing the main competing brands in the market, and designing a prototype within the framework of User-Centered Design, for which it had to address crucial issues such as the application of character models, UX and UI design, the creation of wireframes and user flows, interface design and the development of usability tests. The results demonstrate that it is possible to align the strategic interests of brands with the needs, objectives and expectations of users in a context of increasing global concerns of citizens related to reuse and sustainability. In this sense, it is vitally important that brands adopt Design processes in order to converge their own brand interests with people's demands.

1. Introduction

The research arose from a proposal by the Salvador Caetano Group (official representative of the Lexus brand in Portugal) for developing an official website meant specifically for promoting pre-owned Lexus cars in the Portuguese market.

According to Lexus Portugal, its clientele is mainly males, aged 50 and above. However, studies done by the brand regarding the used cars sector in the Portuguese market show that Lexus products and services have a high potential for growth in new market segments such as women and youth, aging 30 and above. This potential is associated with certain intrinsic factors of the brand, namely:

- Hybrid Technology: in the first half of 2020, 99% of Lexus used cars available for sale had hybrid engine technology. Cars with this technology are all equipped with an automatic gearbox, which facilitates the driving experience and ensures better safety. This technology also results in low pollutant emissions, thereby having a reduced impact on human health. These attributes are valued by the identified market segments (women and youth) that are typically more conscious about environmental concerns.
- The "Lexus Safety System" available on a significant sample of used cars: this system avoids a considerable volume of accidents through alerts and direct intervention in the driving process. Safety is an attribute particularly valued by women and families that have children.
- The reliability of the cars: there is a proven low probability of breakdown, especially of the type that renders the car non-functional and non-operable. The reliability factor in the used car sector specifically appeals to younger clients who have a greater probability of acquiring pre-owned cars at more accessible price points.

Additionally, certain extrinsic factors also affect purchase decisions, such as:

*Corresponding Author: Nuno Martins, nunomartins.com@gmail.com

- Lexus' main competitors in the premium segment (BMW, Mercedes-Benz, Audi, and Volvo) do not have a specific focus on the female audience in their used-vehicle programs. Therefore, there lies an opportunity for gaining advantage through leveraging a differentiating element.
- Citizens tend to be more aware of the more environment friendly motorizations (hybrid and electric). The supply of competing engines in their used-vehicle programmes is predominantly diesel, and there is little to no supply of more eco-friendly engines.

For this set of factors, Lexus Portugal has considered it timely and pertinent to develop a communication strategy that caters specifically to women and youth.

The current official Lexus Portugal website, in this regard, is presently the only source for accessing pricing and availability rated information of pre-owned Lexus vehicles. However, it does not provide the user with a pleasant and properly informed brand experience. According to Lexus Portugal, this has been a restrictive factor in its approach to the new market segments, and has, therefore, affected used car sales targets.

The research has accordingly required multidisciplinary skills and expertise to develop a communication strategy that answers to the concerns and expectations as outlined by to Lexus Portugal. The study took into account not only internal studies made available by Lexus Portugal, but also references related to car sales and drivers' consumption habits [1–3].

2. General and specific objectives

The main objective is to verify if the design process can make the users' needs converge with the brands' strategic lines by developing a digital platform through a case study.

To achieve the general objective, a further number of specific tasks were listed:

- To make an integrated list of the objectives proposed by Lexus for the new online platform, and systematize the entire work process to be developed — Together with the teams of Lexus Portugal and Portugal-based advertising agency Caetsu (responsible for advertising and communication at Lexus Portugal), a workplan was created to analyse all information and define the research aims and the stages of the development process.
- To develop a comparative analysis of the main online reference platforms in order to make a note of good solutions and potential problems. The objective of this analysis was to contribute to a better understanding of the identified market and to arrive at a starting set of guidelines towards designing the new solution.
- To use a design methodology that, close to the User-Centred Design framework, allows to design an online platform capable of aligning the interests defined from the brand strategy and the user experience in general terms.

3. Methodologies

The methodology adopted aimed to apply key practices associated with process design, namely UX and UI Design, the application of persona models, the creation of *wireframes* and *user flows*, the design of interfaces, and the development of usability tests [4–6]. The process consisted of analysis, discussion and resolution making in a collaborative manner among the various stakeholders regarding the different problems associated with the design and development of the online platform [7, 8]. The research process involved, besides the researchers (from the Digital Design area), teams from Lexus Portugal (namely, the General Management, the Marketing team and the Sales team), and the advertising agency Caetsu, alongside a select number of test users [9, 10].

Towards initiating the design process, the research team met with executives from Lexus Portugal and Caetsu with the purpose of: comprehending the philosophy of the Lexus brand; discussing the objectives of the new online platform to be developed; discerning the market and the target audience; and defining the strategies and the work process [11–13].

In order to better understand the respective market, a comparative analysis was developed taking into consideration the online platforms belonging to Lexus's main competitors such as Mercedes-Benz, BMW, Audi, and Volvo, alongside websites of multi-brand vendors such as StandVirtual and CarNext. This made it possible to map the brand's territory, and gain an understanding of the communication that has been developed with audiences online.

Later on, a further set of meetings were held with Lexus Portugal and Caetsu for presenting the analysis results and defining the structure of the information architecture pertaining to the online platform [14].

Correspondingly, the third phase of the project, saw the design of the online platform being supported through the building of an extensive collection of visual elements, in order to expedite the development of the project. Subsequently, visual prototypes were developed, wherein, the interface design incorporated concepts and graphic language previously defined by Lexus [15].

The last phase of the research was dedicated to the development of a high-fidelity prototype, and to carrying out usability tests in order to observe user behavior, detect problems, and make the required improvements towards guaranteeing the best possible experience [16].

4. Comparative analysis of the online platforms of the main competing brands

In order to arrive at an effective overview of the main competing brands' strategic approach, an analysis of the respective online platforms was first conducted. The platforms of Mercedes-Benz, BMW, Audi, and Volvo were hence analysed, alongside with two premier multi-brand used car platforms in Portugal, StandVirtual and CarNext.

The goal was to study the strengths and weaknesses of each of the online platforms, with the objective of obtaining relevant knowledge to be later applied in the development of the new Lexus online platform [17,18].

The development process began with an analysis of the navigation structure and the identification of the various functionalities of each of the online platforms. As can be seen in

Table 1, BMW's online platform was found to be the most complete in terms of functionalities and it also had the most extensive communication strategy. The brand's website was not limited to only providing search tools for its wide range of cars, but also provided the possibility of accessing further detailed information about its services – which was illustrated and animated with photos and videos, that were emotionally engaging, and emphasized the brand's core values and philosophy.

Table 1: Identification of the functionalities of the online platforms of competing brands (January, 2020). Source: own elaboration

Features	Mercedes -Benz	BMW	Audi	Volvo	Stand Virtual	Car Next
Login/Logout	X				X	
Favorites		X		X	X	
Comparator		X		X		
Configurator	X	X	X	X	X	X
News		X			X	X
Product Highlights		X	X	X	X	X
Testimonies						X
Image Galleries	X	X	X	X	X	X
Videos/ Animations		X				
Languages						X
Prices	X	X	X	X	X	X
Product Promotion		X			X	X
Contacts for + info.	X	X	X		X	X
Content sharing (email / social media)		X		X		
Contact directly by email		X		X	X	X
Responsive website	X	X	X	X	X	X

Such 'emotional' communication strategy alignment with brand values was also adopted by Volvo, although not emphasized to the same degree as BMW. Volvo's used-vehicle website was mainly designed around car search and selection tools, and it was careful to present the information in an appealing way through presenting eye-catching images and titles, wherein some of the brand's concepts and values were conveyed. Another interesting feature of Volvo's website was the use of thumbnails of car models in the configurator tool on the main page. This detail helps the user to perform an easier and more informed search. The least positive point of this website was the detailed presentation of each selected car, which revealed certain inconsistencies with the information organization and hierarchy.

Hallmark of Mercedes-Benz's used car website, on the other hand, was the simplicity of the search process, however, weaknesses pertaining to the content were apparent, particularly in terms of the lack of information about the advantages of acquiring a used car at the brand's official dealers. In addition, there was a lack of information about the brand's mission and vision, as well as images and graphics that could connect users emotionally to the brand's values and principles.

Of all the examples analysed, Audi's website was clearly the least interesting. The search tools were not optimally functional, and represented an obstacle in accessing information. In both the mobile and desktop versions of the website, several poorly distributed and organized information text blocks made it difficult to read the content. The website interface was also noticingly unattractive, replete with monochrome tones and outdated graphics. The photography of the vehicles though was of high quality which somewhat accentuated their perceived value.

Furthermore, the websites of *Stand Virtual* and *CarNext* were analyzed. These represent the largest and most popular online portals in Portugal for buying and selling used cars from most major brands. They were primarily directed at private entities and individuals.

Although the communication strategy of Stand Virtual and CarNext was different from what is intended by the Lexus group, they were considered important references since they dominate the used car sales market. Unlike Lexus, the target audience of the two companies is eclectic, seeking a diverse variety of cars from different brands.

Overall, both websites had a well-structured homepage with good content hierarchy. They did not have spaces dedicated to lifestyle concepts, wellness, or promotional texts. Both websites focused on two objectives: product search through filters, strategically positioned at the beginning of the page and easy to navigate; and inviting the user to make their product available for sale on the respective online platform. There was also a 'testimonials' section in CarNext's platform, which augmented the credibility of the service. This feature was considered a good reference and was correspondingly adopted in the development of the new Lexus platform.

In summary, it was found that, with the exception of BMW, the websites of the competitors were mainly limited to working as online platforms for searching and promoting vehicles available for sale.

BMW's website was the one that proved to be the closest to the intended solution for Lexus Portugal, providing relevant information about the brand and its services; and exploring images and videos with a more emotional outlook. However, the website's graphic and image component was not as refined as the brand's main website (bmw.pt), for the Portuguese market.

In accordance, the proposed prototype for Lexus Portugal intends to create a greater degree of integration between the main website of Lexus Portugal and its pre-owned division, presenting coherent solutions with the same level of graphic and image quality.

Following this competitor review, the corresponding developed work focused on the design and development of the Lexus used car online platform that was presented.

5. Online platform development

The design process of the online platform was developed along the following phases, in conjunction with regular feedbacks from the Lexus Portugal and Caetsu teams:

5.1. Interviews and definition of the Personas

The concept of *personas* was created to define the target audience of the digital solution under study. *Personas* are fictional personalities that can characterize potential users, representing the aspirations and attributes of a target group that a brand entity seeks to communicate with. Such idealization process can help understand the needs, experiences, behaviors, and ambitions associable with the target users. *Personas* are meant to simplify the design process, from the point of view that all processes are developed from the perspective of the user in order to ensure the best user experience for the target audience [19–21].

In order to define the *personas* and to generate a better understanding of the overall context of Lexus's used car market segment, including the aims and aspirations of the brand and its customer segment, a set of interviews was conducted with Lexus Portugal executives (CEO, Sales Manager, Brand and Product Director, and Digital Marketing Manager) and the creative lead from Caetsu responsible for Lexus' communication and advertising. The interviews were of a structured type, previously scripted, and consisted of open questions.

Through these interviews it was found that the Lexus client is mainly male, over 50 years old, and is positioned in the upper to upper-middle class social bracket. However, the brand's goal is to extend its target audience to younger age groups, from the age of 30 onwards, and particularly to the female public. It was also found that Portuguese customers are still unclear about the main advantages of hybrid technology. Moreover, when the customer visits a dealer of the brand, the vast majority have no knowledge about the available after-sales services or customization options for used cars.

These results reinforce our conviction about the need to develop an online platform that is not simply a search tool for used cars available for sale. Decision making when buying a car is increasingly influenced by previous online searches, before even visiting the physical sales stands [22]. For this reason, it is essential to make the complete information set available online: not only about the products, but also about the brand, the offered services, the sales conditions, and technological specifications, among other related data.

The *personas* were developed with the support of the Lexus Portugal's and Caetsu's teams. In defining the *personas* (as shown in Table 2), the research thus took into account not only the type of usual customer, but also the youth and female target segments that Lexus also wants to reach out to.

Three types of *personas* were defined, two female and one male. The majority are female, contrary to the results on the type of visitors that Lexus Portugal has had on its website (www.lexus.pt). In 2020, the majority of visitors to Lexus Portugal's website (approximately 65%) were male, with an average age of close to 45 years.

Table 2: Analysis grid of different *personas*.

JOÃO 45 years old	**LEONOR** 28 years old	**SOFIA** 55 years old
Architect in a studio in Lisbon and a technology lover.	Team coordinator, in an information technology company.	Medical specialist at a Coimbra Hospital
Features: João feels the need to buy a car that promotes safe, pleasant driving and, if possible, with a low environmental impact. He is looking for an excellent driving experience, but above all he wants a good deal. During his research, João realized that there are several technological solutions that cover sustainability concerns, and economic and technological consumption, without giving up comfort and luxury. Despite finding various platforms for car sales, and looking for the best price, he prefers to invest in a safe purchase. Besides all issues related to the product, João has doubts regarding the choice of the car brand and, for that, he tries to research the benefits and conditions offered by the various brands in their proprietary websites.	**Features:** Leonor lives far from her job and her old car is beginning to show the need for replacement with a more modern option that is safe and comfortable. As her everyday commutation is long, she is looking for an economical and durable solution. These constraints encourage Leonor to search for brands that have low maintenance cost, extended warranty services, and which offer excellent after-sales customer service. Beyond these concerns, Leonor is also aware that this is an important purchase and should be very well considered, which leads her to research, systematically and carefully, in trying to rationalize the choices as much as possible.	**Features:** Sofia usually leaves work late, and at the end of the day, enjoys her ride home relaxing in a car that offers unmatched comfort and luxury. She is looking to upgrade from her present one and promptly searches on her phone for the best deals available. She immediately quits from websites that feel clunky or too complicated, and gives clear preference to websites that are easy to navigate on her phone, and which provide the most relevant information upfront. Sofia considers it fundamental that the contents made available on the website are presented in a language and style that is simple, direct and easy to understand, and which does not get too technically complex. She knows the Lexus brand well as it is a frequent choice of her family members.

This type of majority audience is represented in the *persona* identified with the name João. This persona is also close to the type of Lexus customer, which is characterized by its focus on economy, the environment, security, reliability and technology.

However, the brand aims to change this trend and ensure that the younger and female audiences who visit the website do so successfully. The *personas* Leonor and Sofia intend to represent these customers' segments.

The *persona* Leonor represents the younger age group, which generally has less purchasing power. For this reason, buying a semi-new or used car can be a good option. They are people who buy cars for a long period of time and therefore prefer to make a slightly higher investment in the purchase in order to guarantee that it is a safe and reliable investment. Therefore, they prefer cars with high reliability, low maintenance costs and safety.

The *persona* Sofia represents the oldest age group, which usually has little availability and patience to seek information, especially in digital media. For this reason, it is fundamental to

have an online platform with good usability and that presents in a direct and brief way the main attributes of the brand's products and services. Although the typical client of the brand is male and above 50 years, the female gender is also important. According to internal Lexus Portugal studies, the female gender usually has a high influence on the purchasing decision of a relative, namely the partner (boyfriend or husband).

5.2. Information architecture

According to Garrett [23], the success of the user experience depends upon the designer's consideration of the following set of assumptions in the construction of information architecture [24, 25]:

- Strategy: represents what the user expects from the digital product.
- Objectives: requirements and specifications of functionalities that the product should meet.
- Structure: definition of hierarchies and content organization.
- Survey of technical limitations and usability of the product.
- Skeleton: design of the organization of the elements that will constitute the interface.
- Surface: corresponds to the final phase of the digital product, what the user interacts with.

In accordance with the needs as outlined by the Caetsu and Lexus team, a set of features and sections were defined for the digital product. The navigation experience was first established for the online platform to be well structured. During this initial stage, difficulties and possible obstacles regarding the user´s experience was also anticipated [26].

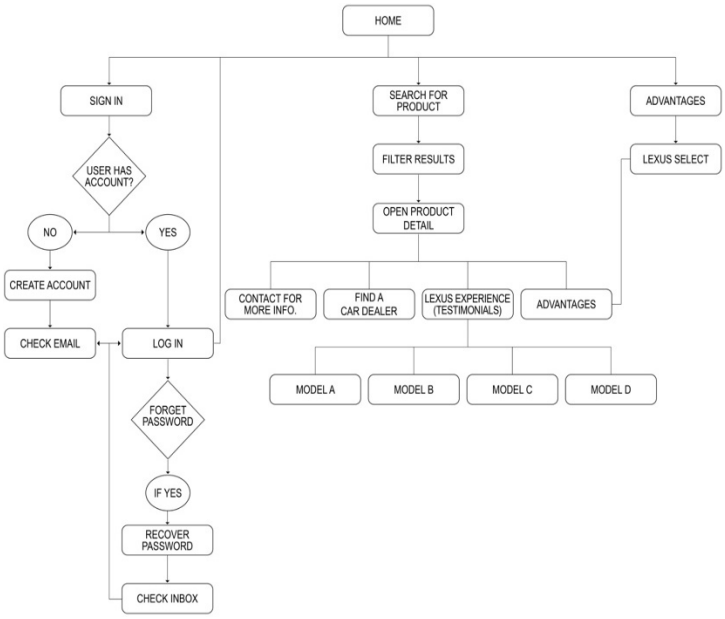


Figure 1: User flow diagram

5.3. Wireframes and user flows

After characterizing the *personas* and defining their main needs, the first low fidelity sketches, or *wireframes*, were created. These sketches were made by means of basic representation through shapes, with the objective of studying the disposition of contents and the functionality of each page. The sketches were at first drawn manually and then transferred digitally, allowing the skeleton of the online platform to be visualized.

Further on, in order to study the sequence of steps for the user to accomplish a certain task, *user flows* (as shown in Figure 1) were developed towards establishing performance parameters for each task. *User flows* are considered fundamental for the organization of *wireframes* in order to understand step by step the user's navigation intentions. For the development of the *user flows* the Sketch software was used, allowing to test the interaction between the different *wireframes* through simple graphic representations.

After this first phase, the first screens were designed, and later on, a set of detailed drawings, typography, icons, buttons, and other interactive elements were added, which completed the navigation paths and the linking between pages and sections.

5.4. Interface design

It was deemed fundamental for the interface to incorporate a visually pleasant layout, which could help users to read, concentrate, and perform tasks without hinderances. The interface was also finetuned to generate predictable user behavior and provide the user a sense of control when using the product.

When an interface has a nice layout, it helps users to read, concentrate, perform tasks better and increase real and perceived usability [27, 28]. The interface should generate predictable user behavior and give the user a sense of control when using the product.

5.4.1. Grid

Grids are practical tools in the structuring process and are responsible for the consistency and visual organization of the design and the elements that build the interface, and were thus considered essential in the design of the website [29].

Later on, a number of concepts linked with Responsive Web Design were used in order to create an interface that would be capable of adapting to various mobile devices, in accordance with the individual/situational preference of each user [30]. This was also done to create a cohesive, harmonious and balanced visual composition, towards ensuring optimal adaptability and to follow a structured logic throughout the procedure.

Due to the multiplicity of screen sizes, it was decided to define, in the desktop version, the current most used resolution in Portugal: 1366 pixels wide [31].

The grid used to build the desktop version of the platform was divided into twelve columns, each sixty pixels wide, with twenty pixels column gaps. At the end of each page, a margin corresponding to ten pixels was also incorporated.

In order to ensure a responsive layout that would adapt seamlessly to the interface format, similar parameters were considered: a screen width of 360 pixels (most prevalent locally), and a grid divided into twelve columns, each twenty-seven pixels wide and with a column gap of four pixels.

5.4.2. Color Palette

The design process in urderlined by the notion that the selection of the chromatic palette, during the UI Design process,

does not depend squarely on aesthetic sense, but on the intention behind the application of each element. This could be so in terms of capturing viewer attention to certain points, transmitting certain sensations; improving navigation; or creating a sense of harmony between all the elements that make up the interface [32].

Color, however, is a fundamental part of communicating Lexus' history. The selected color set (as shown in Figure 2) mainly emphasized on bringing softer tones in order to convey the brand's luxury positioning and subtle refinement. A predominance of neutral colors was also meant to highlight the contextual photographs of environment and lifestyle in conjunction with products of the brand. More luminous colors were additionally used in some details to highlight the aspect of technological contemporaneity in the brand, and capture the user's attention to certain informative or interactive elements.



Figure 2: Brand Lexus Colour Palette

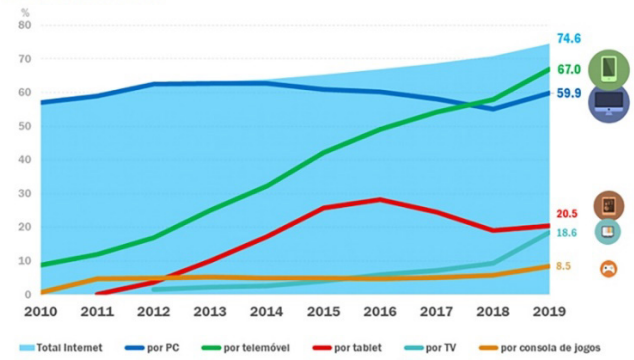
5.4.3. *Typography*

Bahat [33] views the use of the right typography as fundamental to guaranteeing readability, accessibility, hierarchy, and graphic balance. In digital products typography strongly influences the overall usability of the interface.

It is also one of the defining elements of a brand's visual identity. In the platform, the typographic family adopted was "Nobel" by Adobe Fonts, designed by Tobias Frere-Jones. It is a Sans Serif font, with several typographic styles and is suited to work correctly on digital media.

5.5. *Design of the online platform screens for the mobile and desktop versions*

According to Marketest [34], in the last 10 years, the use of the Internet through mobile devices has registered substantial growth in Portugal, especially through smartphones. As can be seen in Graph 1, in 2018 Internet use through smartphones surpassed desktop access. This trend has been determined to continue [34].



Graph 1: Internet use by platforms. Source: Marktest (2019)

For this reason, the design of the different screens of the platform focused on a mobile first approach and was then adapted to the desktop version. The strategic decisions that guided the design process behind the different screens were as follows:

- Homepage: Since it would be the first page that the user would have access to when visiting the website, the information and tools considered most relevant for the Lexus' target audience were highlighted, such as information on the advantages of the Lexus Select Service; information about the Lexus' experience (including access to testimonials from Lexus customers); and the possibility of searching immediately for available semi-new cars. As can be seen in figure 3, on the homepage there is a predominance of photographs and graphics that are positioned to accentuate the environment and the identity of the Lexus brand. In this way, the above mentioned objective of designing an online platform that reflected the Lexus philosophy and a user-friendly experience, was addressed.

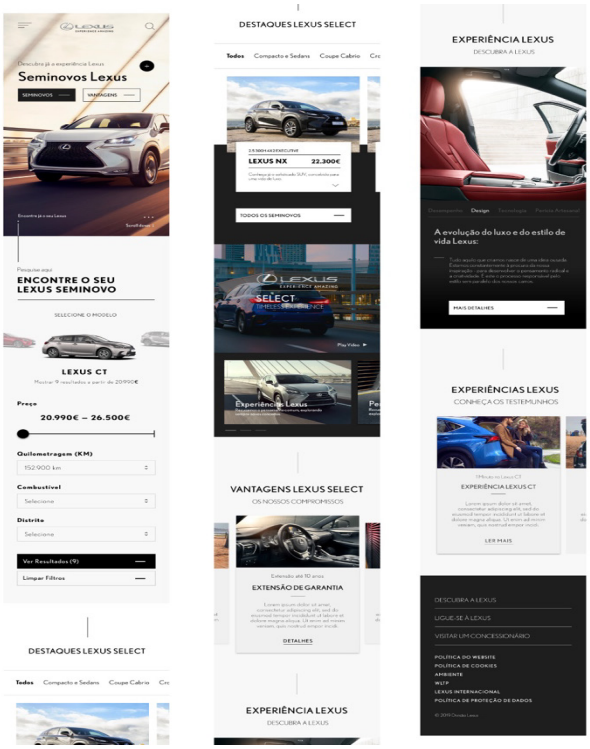


Figure 3: Lexus Homepage (mobile version)

- Configurator: Represented by the search and filter section of the homepage, the configurator was incorporated in a subtle manner, and not emphasized, with a view to provide the user greater freedom of choice in navigation. In the configurator the images of the car models were highlighted, allowing an easier, faster and more informed search.
- Testimonials: The "Testimonials" section was created with the aim of reaching out through personal narratives, towards building trust and empathy between the user and the brand. In this section testimonials from Lexus vehicle owners were presented, alongside promotional videos and informative texts about the cars.
- "Start" buttons: In the introduction to the "Lexus Advantages" page, ignition-style start buttons (similar to those found physically on Lexus cars and used to actually start the car's engine) were strategically applied. Through this type of elements, the intention was to involve the user with the visual ethos of the brand and its cars.
- Dark Mode: The introduction of a view mode with a predominance of dark tones was applied on pages such as "Lexus Select Advantages" and "Testimonials", towards ensuring eye comfort, and for garnering greater concentration in reading on higher text density sections.
- Quiz: on the online platform a Quiz component was created (as shown in Figure 4), entitled "Discover the ideal Lexus for you", with the aim of bringing communication between the user and the brand closer. In this section, the users have an option to answer a small questionnaire in order to identify and share their main preferences. Upon obtaining the users' inputs, the website employs a set of algorithms to suggest the most suitable car model. The analysis of the inputs also provides real-time availability of suggested options, and the respective store locations with available stock. The results that do not directly match the user's needs, are displayed towards the end of the search page. This strategy was intended to contribute to increasing empathy between the users and the brand.

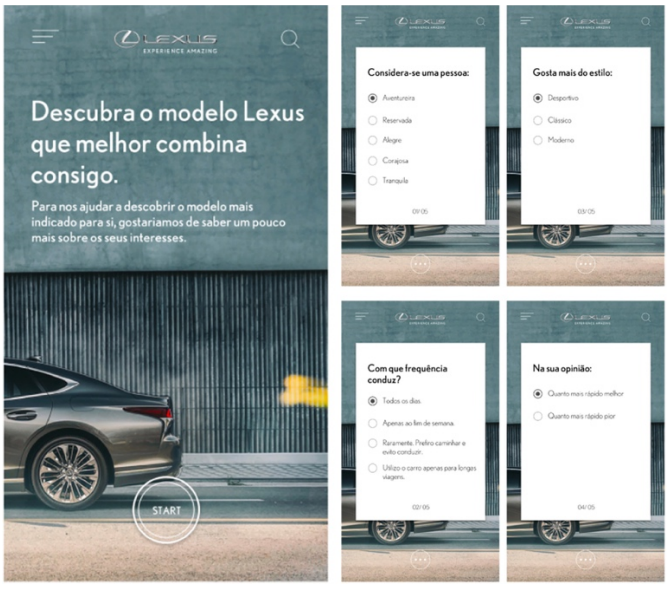


Figure 4: Quiz section (mobile version)

- Product listing: In this area the main characteristics of each car is presented, providing relevant information to viewers with illustrated icons that allow for an easier and faster reading experience. Also incorporated in the section are images of the respective cars, that are aligned to a predefined presentation standard, in order to ensure consistency and good image quality. Additionally, information pertaining to test-drive locations nearby is included.

As mentioned earlier, the mobile version of the platform was designed first, and then ported to the desktop version (as shown in Figure 5).

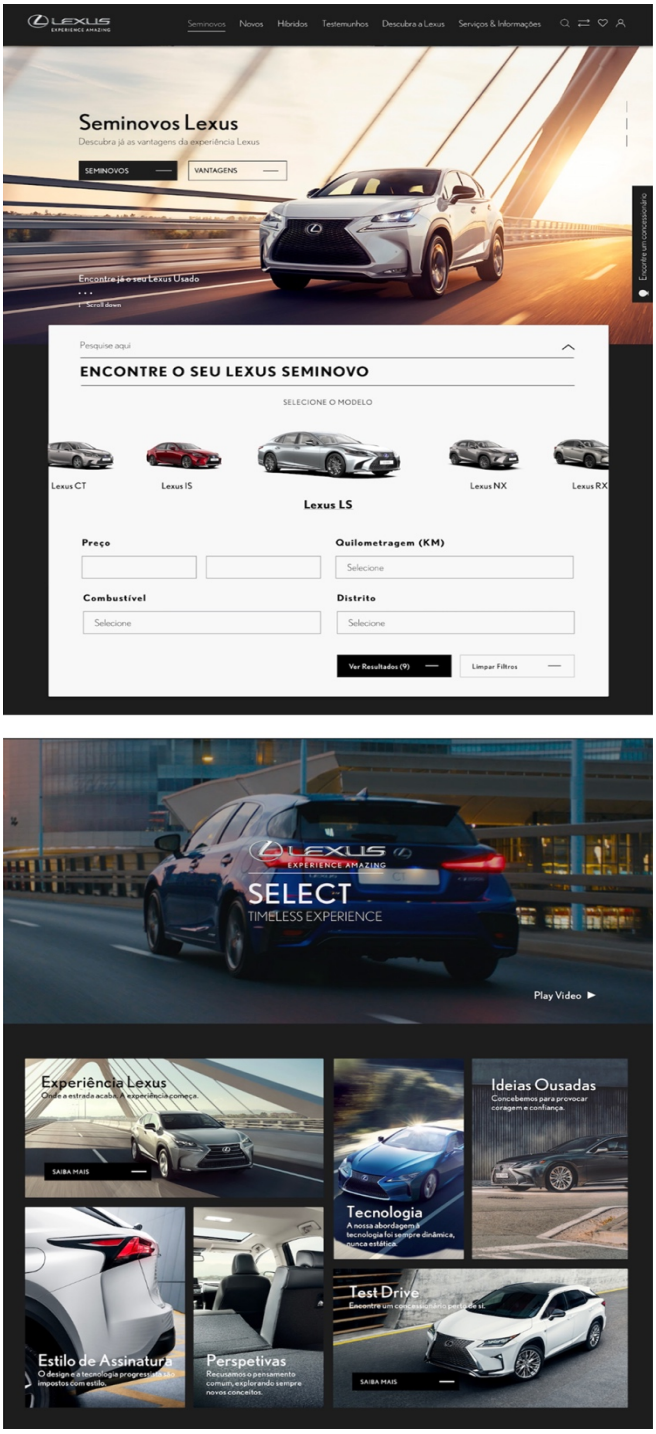


Figure 5: Homepage details (desktop version)

5.6. Prototype

Upon concluding the designing of each screen of the mobile and desktop versions in the 'Sketch' application, the software 'Invision' was used to prototype the online platform. Through this software, animations and transitions between screens were produced and applied. This allowed a better understanding of the flow and navigation of the website, before moving on to the usability testing phase of the prototype.

5.7. Usability tests

The usability tests were composed of a series of actions that were aimed to evaluate and verify the user's ease of use of the interface. The tests were conducted with users representing each persona for whom the product was being developed. The participants carried out a list of tasks and objectives, while being analyzed to provide observations on behavioral data [35].

There are several methods to study and evaluate usability. According to Nielsen [16], through 5 users, 85% of the usability problems of the product or service can be detected.

Throughout the development of the online platform several usability tests were performed in order to obtain a constant feedback from users and make the necessary improvements accordingly. This process helped reach the last phase of the work with a reduced number of errors and inconsistencies in the interface design.

For the last usability tests, towards attaining further rigor in the analysis method, six individuals (two for each of the three defined *personas*), men and women aged between 27 and 58, were selected for the sample.

The tests consisted of observing and evaluating user behavior in performing the following tasks:

- Enter the website and find the car configurator.
- Select a car and find the respective data sheet.
- Take the test to find out which car is right for you and ask for more information about it.
- Create a user account on the website.
- Find out more about Lexus Select and its benefits.
- Go to the testimonials page of former Lexus vehicle owners.

Users were also asked to indicate the difficulties, indecisions and doubts they encountered.

The tests lasted an average of 12 minutes per individual. According to the observations made, the following findings stood out:

- All users were able to accomplish the proposed tasks without demonstrating difficulty in finding the intended goal. However, in task 3, there was some indecision on the part of two of the users to start their task (by clicking on the button "The Lexus for you"). Taking this result into account, it was considered necessary to further highlight this functionality on the website, and subsequently improve it.
- The icons designed for the presentation of the car's technical data were highlighted by five users. These users considered this option an important aid in reading the contents.
- Older users took more time than newer users in performing the tasks. However, the difference was small.

After the usability tests, the same users were also asked to answer a short questionnaire towards assessing the level of satisfaction they experienced while using the online platform. The questionnaire was set to examine key areas such as the interaction, the perception of the information, the usefulness factor, and an overall impression of the website. In Table 3 the questions that have been asked to users are presented.

Table 3: The questions asked to users to evaluate the level of satisfaction experienced while using the online platform.

QUESTIONS
Perception and quality of the information
All the information on the website was readable?
Did you understand all the content of the website?
Is the information on the website well organized by the different parts of the website?
Usefulness
Has the website enabled you to become more enlightened about the brand's products?
Did the website clearly conveyed the advantages of buying a Lexus used car?
Has the website aroused your interest in getting to know the brand and its services better?
Interaction
Was the interaction with the website easy?
Did you find what you wanted easily?
Was the navigation through the website easy or did you ever felt lost?
Overall assessment
What is your overall assessment of the website?
Is there anything you would improve on the website?
If you answered "Yes", what would that be?

The results of these tests were reasonably positive. The tests demonstrated a high success rate in the overall fulfillment of tasks, identifying only minor problems that were correspondingly solved. The level of user satisfaction, based on the feedback obtained through the questionnaire, was high and thus determinative of the prototype's market-readiness.

6. Conclusions

As for the first of the specific objectives, the joint work with Lexus and the Caetsu agency allowed, in the first instance, to identify that the brand's strategy, beyond a better conversion ratio of the digital platform, was interested in promoting values related

to the circular economy, sustainability and responsible use of natural resources, which was a fully consistent action with offering a new life to used cars, even more so when these vehicles are almost entirely hybrid vehicles. Secondly, it allowed, in line with its strategy, to identify and model typical users based on the research data that the brand already had.

Secondly, the study concluded that, with the exception of BMW, the websites of the competing brands were mainly limited to serving as search tools for their respective used cars available for sale. The analysis of the competing brands, together with the collaborative research actions developed with the Lexus Portugal and Caetsu teams, provided a comprehensive understanding of the second-hand car segment in Portugal, and gave an idea of Lexus' target audience.

Regarding the third secondary objective, the methodological foundations of the design and development process of the online platform facilitated the realization of a viable solution by strictly adhering to the key principles of UX and UI design to achieve cohesion between the different elements of the process, such as the application of character models, information architecture, the creation of *frameworks* and *user flows* and the conceptualization of appropriate interfaces.

In addition, usability tests were conducted to evaluate user acceptance and identify areas for improvement. Once the necessary changes had been made, the prototype was submitted to the management of Lexus Portugal, which subsequently approved it and granted permission to move on to the execution phase, with a view to making it public in the coming months.

Thus, the results presented allow to claim that the online platform designed in this research can align the strategic interests of Lexus with those of its users, allowing Lexus Portugal to have a distinctive and even more complete solution compared to competing brands and entities, and to provide users with a solution more suited to their needs, goals and motivations.

Although user studies indicate a good result in terms of usability and user experience, in the future it's necessary to carry out further testing of the implemented product, in order to iteratively add improvements to it.

Based on the results, it is understood that the process employed in this research can be used in those cases where the strategic interests of the brands are to be aligned with those of the previously identified target audiences.

The process used in this research is intended to be the basis of a model that helps brands to align their strategic interests with the needs, expectations and objectives of their potential users. All of this is done from a design perspective and with the aim of turning their audience (potential customers) into satisfied users.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work is financed by national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/04057/2020.

References

[1] G. Pedros-Perez, P. Martínez-Jiménez, P. Aparicio-Martinez, "The potential of car advertising in pursuing transport policy goals: Code of good practices in the Spanish context" Transportation Research Part D: Transport and Environment, **72**, 312-332, 2019. https://doi.org/10.1016/j.trd.2019.05.010

[2] A. Boelhouwer, A.P. van den Beukel, M.C. van der Voort, C. Hottentot, R.Q. de Wit, M.H. Martens, "How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands" Transportation Research Interdisciplinary Perspectives, **4**, 2020. https://doi.org/10.1016/j.trip.2020.100103

[3] Paulo Dinis, Fernando Moreira da Silva, "The Passenger Car in Portugal: Features of Functional Modules" Procedia Manufacturing, **3**, 312-332, 2015. https://doi.org/10.1016/j.promfg.2015.07.313

[4] Z. Kahraman, "Using user-centered design approach in course design" Procedia - Social and Behavioral Sciences, **2**(2), 2071-2076, 2010. https://doi.org/10.1016/j.sbspro.2010.03.283

[5] J. Braga, "Is human-centered Design broken?". UX Collective, 2019. [Online]. Available: https://uxdesign.cc/is-human-centred-design-broken-cac130eecc48. [Accessed: 28-Aug-2019]

[6] A. Chammas, M. Quaresma, C. Mont'Alvão, "A Closer Look on the User Centred Design" Procedia Manufacturing, **3**, 5397-5404, 2015. https://doi.org/10.1016/j.promfg.2015.07.656

[7] J. Liedtka, "Why Design Thinking Works". Harvard Business Review, 2018. [Online]. Available: https://hbr.org/2018/09/why-design-thinking-works.

[8] C. F. Llema, C. M. Vilela-Malabanan, "Design and Development of MLERWS: A User-Centered Mobile Application for English Reading and Writing Skills" Procedia Computer Science, **161**, 1002-1010, 2019. https://doi.org/10.1016/j.procs.2019.11.210.

[9] C. Abras, D. Maloney-Krichmar, J. Preece, User-Centered Design. In Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, 2004.

[10] D. A. Noman, The Design of everyday things, New York: Basic Books, 2013.

[11] N. Martins, D. Brandão, D. Raposo (Eds.), Perspectives on Design and Digital Communication: Research, Innovations and Best Practices, Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-49647-0.

[12] D. Raposo, "From Brand Mark to Corporate Visual Identity Systems" Convergências Journal, **2**(3), 2009. Retrieved from journal URL: http://convergencias.esart.ipcb.pt/?p=article&id=40

[13] J. Park, "Developing a knowledge management system for storing and using the design knowledge acquired in the process of a user-centered design of the next generation information appliances" Design Studies, **32**(5), 482-513, 2011. https://doi.org/10.1016/j.destud.2011.05.001

[14] D. Brown, "Chapter 2 - Agile Methods + UX = Agile UX" In Diana DeMarco Brown (Eds.), Agile User Experience Design: A Practitioner's Guide to Making It Work. 39-69, 2013. https://doi.org/10.1016/B978-0-12-415953-2.00002-9

[15] M. Kouprie, F. Sleeswijk Visser, "A framework for empathy in design: stepping into and out of the user's life" Journal of Engineering Design, **20**(5), 437-448, 2009. https://doi.org/10.1080/09544820902875033

[16] J. Nielsen, "Enhancing the explanatory power of usability heuristics" in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94). Association for Computing Machinery, New York, NY, USA, 152–158, 1994. https://doi.org/10.1145/191666.191729

[17] N. Martins, M. Lemos, D. Brandão, D. Raposo, J. Neves, J. Silva, "Lexus used-vehicle online platform: comparative analysis of major competing brands' websites" in Advances in Design, Music and Arts: Proceedings of the 7th Meeting of Research in Music, Arts and Design, EIMAD 2020, May 14-15, 2020.Springer Series in Design and Innovation, **9**. Springer, Cham, 243-255, 2021. https://doi.org/10.1007/978-3-030-55700-3_17.

[18] D. Lachapelle, "A Designer's Guide to Competitor Research". Wayfair Experience Design, 2017. [Online]. Available: https://medium.com/wayfair-design/a-designers-guide-to-competitor-research-914b0deee25a. [Accessed: 23-Jul-2019]

[19] R. Dam & T. Siang, "Personas: A Simple Introduction". Interaction Design Foundation. [Online]. Available: https://www.interaction-design.org/literature/article/personas-why-and-how-you-should-use-them.

[20] A. Cooper, R. Reimann, D. Cronin, About Face 3: The essentials of interaction design. Information Visualization, **3**, Canada: Wiley Publishing, Inc, 2007.

[21] F. Anvari, D. Richards, M. Hitchens, M. Ali Babar, H. Tran, P. Busch, "An empirical investigation of the influence of persona with personality traits on conceptual design" Journal of Systems and Software, **134**, 324-339, 2017. https://doi.org/10.1016/j.jss.2017.09.020

[22] A. Teixeira, "Como está a evoluir o E-Commerce em Portugal em 2020". Digitalks, 2020. [Online]. Available: https://digitalks.pt/artigos/a-evolucao-do-e-commerce-em-portugal.

[23] J. J. Garret, The elements of the user experience. Berkeley: New Riders, 2003.

[24] R. Hartson, P. Pyla, "Chapter 16 - Designing the Ecology and Pervasive Information Architecture" In Rex Hartson & Pardha Pyla (Eds.), The UX Book (Second Edition). 341-358, 2019. https://doi.org/10.1016/B978-0-12-805342-3.00016-3

[25] A. Grilo, "O que é Arquitetura da Informação e como ela influencia a UX do seu produto". Comunidade UX Design Natal, 2016. [Online]. Available: https://medium.com/ux-design-natal/o-que-%C3%A9-arquitetura-da-informa%C3%A7%C3%A3o-e-como-ela-influencia-a-ux-do-seu-produto-b4f20881b2b4.

[26] F. Santana, "Arquitetura de Informação e o seu propósito". Coletivo UX, 2017. [Online]. Available: https://coletivoux.com/arquitetura-de-informação-e-o-seu-propósito-29cd278ebdfe.

[27] K. Kashimura, M. Kurosu, "Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability" in CHI '95: Conference Companion on Human Factors in Computing Systems. 292–293, 1995. https://doi.org/10.1145/223355.223680

[28] N. Tractinsky, A.S. Katz, D. Ikar, "What is beautiful is usable" Interacting with Computers, **13**(2), 127–145, 2000. https://doi.org/10.1016/S0953-5438(00)00031-X

[29] M. Soegaard, "The Grid System: Building a Solid Design Layout". Interaction Design Foundation, 2019. [Online]. Available: https://www.interaction-design.org/literature/article/the-grid-system-building-a-solid-design-layout.

[30] E. Marcotte, Responsive web design, Paris: Eyrolles, 2011

[31] StatCounter GlobalStats, "Screen Resolution Stats Portugal (Nov 2018 - Nov 2019)". StatCounter, 2019. [Online]. Available: https://gs.statcounter.com/screen-resolution-stats/all/portugal.

[32] T. Schlatter, D. Levinson, "Chapter 6 - Color" In Tania Schlatter, Deborah Levinson (Eds.), Visual Usability: Principles and Practices for Designing Digital Applications. 171-211, 2013. https://doi.org/10.1016/B978-0-12-398536-1.00006-1

[33] N. Bahat, "Improving UI Design Through Better Typography". Awwwards, 2015. [Online]. Available: https://www.awwwards.com/improving-ui-design-through-better-typography.html.

[34] Marktest, "Acesso à internet em Portugal é cada vez mais Mobile". Grupo Marktest, 2019. [Online]. Available: https://www.marktest.com/wap/a/n/id~2555.aspx.

[35] J. Nielsen, "Disruptive Workflow Design". Nielsen Norman Group, 2012. [Online]. Available: https://www.nngroup.com/articles/disruptive-workflow-design.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 645-651 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Finding Association Patterns of Disease Co-occurrence by using Closed Association Rule Generation

Panida Songram*, Phattanaphong Chompowiset, Chatklaw Jareanpon

POLAR Lab, Department of Computer Science, Faculty of Informatics, Mahasarakham University, 44150, Thailand

ARTICLE INFO

Article history:
Received: 03 August, 2020
Accepted: 25 September, 2020
Online: 05 October, 2020

Keywords:
Disease co-occurrence
Association disease
Association pattern of disease
Frequent disease
Closed itemset mining
Closed association rule mining

ABSTRACT

This paper proposes a closed association rule generation technique to investigate the association patterns of diseases that are frequent co-occurrence. Diseases records of 5,000 patients are studied to find the association patterns of disease co-occurrence. The CHARM algorithm is adapted to find frequent diseases that can cover all-important patterns with a small number. Then the association patterns of disease co-occurrence are created in a form of association rules from the frequent diseases. The rules represent diseases associated with other diseases. Accuracy and prediction ratio are defined to evaluate the generated association patterns. From the experimental results, the generated association patterns give 79.76% of accuracy and 84.03% of prediction ratio although the number of generated association patterns is small. Moreover, the top-10 association patterns of disease co-occurrence are investigated. Besides, the 5 most frequent diseases are found to deeply study the other related diseases of them. From the investigation, we found that diabetes mellitus, metabolic disorders, and renal failure are highly related to hypertensive diseases with 88.81% of confidence. In addition, we found that influenza and pneumonia, plastic and other anemias are highly related to metabolic disorders.

1. Introduction

The death of people is mostly caused by diseases. Diseases are social and economic problems in the world. A lot of money is paid for treating diseases. If a disease is early detected, people can prevent themselves from the disease. Detecting a disease can be found from other related diseases. Finding association patterns among these diseases is very challenging work in the domains of biology and medicine. The study of the association of diseases not only helps people to understand the relation of diseases but also leads to improvement in clinical manifestation, etiology, pharmacology, and epidemiology. Many techniques were proposed to study the association of diseases, such as network techniques, graph theory, network science, statistical methods, and mathematical modeling. They are studied based on microbes, disease-related genes, microRNAs (miRNA), disease-related metabolic reactions, and electronic medical records. [1].

Association rule mining is a technique that has been widely used in the clinical domain. It is applied to find association patterns of diseases, such as finding the relation of metabolic syndrome and other diseases [2], finding the relation of the disease and medicines [3], and finding the relation of factors and disease [4].

Unlike the previous works, we propose to investigate association patterns of disease co-occurrence based on closed association rule mining that will generate a small number of patterns with coverage of all-important patterns. The association patterns of disease co-occurrence are investigated from electronic medical records in Thailand. First, frequent diseases are generated based on closed itemset mining. Then they are used to generate closed association rules for representing association patterns of disease co-occurrence. Moreover, the sorting method is presented to select the top-k association patterns of disease co-occurrence for investigating diseases that are highly related to each other. Also, the most frequent diseases are found and explored other diseases that are related to them.

The rest of this paper is organized as follows: Section 2 mentions to related works. Section 3 gives the concept of closed association rule mining. Section 4 explains the overall proposed methodology. Section 5 gives the details of the experimental setup and experimental results. Finally, the conclusion is provided in Section 6.

2. Related Work

Finding the association of diseases has been widely studied in the domains of biology and medicine. It is investigated with

*Corresponding Author: Panida Songram, Email:panida.s@msu.ac.th

different datasets and different techniques, such as network techniques, graph theory, network science, statistical methods, and mathematical modeling. For example, in [5] proposed a microbe-based human disease network based on the text mining process. The network is constructed from the microbe-disease association dataset. It is investigated to find relationships between microbes and disease genes, symptoms, chemical fragments, and drugs. Cosine similarity is employed to identify the similarity between two diseases. In [6] proposed a disease-related gene mining method based on a weekly supervised learning model. The method consists of two parts, First, the differentially expressed gene set is screen based on the weakly supervised learning model. Second, a support vector machine is adopted to predict the disease-related genes in the differentially expressed gene set. The method verified the validity and accuracy of the method. In [7] proposed similarity computations to predict the associations between miRNAs and diseases. The similarity among miRNAs is computed based on the sequence and function information of miRNAs. The similarity among diseases is computed based on the semantic and function information of disease. Then the data sources are integrated by using the kernelized Bayesian matrix factorization method to infer potential miRNA-disease associations. The unknown miRNA-disease associations were effectively predicted from the method.

Association rule mining is a popular technique that was exploited in the medical domain. For example, in [2] adopted association rule mining to study metabolic syndrome that is related to other diseases and to understand the strength of association between diabetes mellitus, hypertension, and hyperlipidemia on patient's records in Taiwan. From the study, it was found that diabetes mellitus is related to oral diseases and blear eyes. Patients with metabolic syndrome have a higher connection with liver diseases than patients with diabetes mellitus. In [3] analyzed patient prescriptions to identify the relationship between the disease and medicines that are used to treat the patient's illness. The patient prescription datasets in 2015 and 2016 from two hospitals are collected to find the relationship. First, the top 10 diseases are clustered by the K-means algorithm. Next, the Apriori algorithm is applied to find the relationship between diseases and medicines. In [4] applied association rule mining to detect factors that contribute to heart disease for males and females on the UCI Cleveland dataset. Three algorithms, Apriori, Predictive Apriori, and Tertius, are investigated to identify the factors. From the investigation, females are at higher risk than males to be heart disease.

From previous works, association patterns of diseases are studied based on numerous factors such as genetics, metabolites, microbes, and miRNAs. In the real world, those factors are hard to understand and to access for people who are not in the domains of biology and medicine. Finding association patterns of diseases from a disease dataset is an easy way. The disease dataset can find out from the electronic medical records of patients. Besides, studying association patterns from medical records of patients in each area may get different knowledge. In this paper, association patterns of disease co-occurrence are investigated on a disease dataset that is retrieved from electronic medical records in Thailand. A closed association rule generation technique is proposed to investigate the association patterns of disease co-occurrence.

3. Closed Association Rule Mining

Association rule mining is a popular technique in data mining and has been widely used in many applications with several domains. It discovers the relationship between items in a large dataset. The basic definition of association rule mining can be explained as follows.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a finite set of items in a database and $D = \{d_1, d_2, \ldots, d_m\}$ be a set of transactions in the database, where each transaction represents a set of items. X and Y are itemsets, where $X, Y \subseteq I$. The support of X is the number of transactions containing X, denoted as $supp(X)$. The length of X is the number of items in X. An association rule r: $X \rightarrow Y$ is a relationship between itemset X and Y, where $X \cap Y = \varnothing$. X is called the antecedent of rule and Y is called the consequent of rule. The support of association rule r is defined as $supp(X \cup Y)$. The confidence of association rule r is defined as $conf(r) = supp(X \cup Y)/supp(X)$. The problem of association rule mining is to find all association rules passing minimum support threshold (*min_supp*) and minimum confidence threshold (*min_conf*).

The process for mining association rules consists of two main steps. The first step is to mine frequent itemsets having support no less than the minimum support threshold. The second step is the generation of association rules. Association rules are generated from frequent itemsets having length≥2. Frequent itemset X having length l will be possibly generated 2^l-2 rules that is the number of subsets of X. For example, X = (ACF). The length of X is 3 and the set of subsets of X is {A, C, F, AC, AF, CF}. Therefore, the number of association rules is $2^3 - 2 = 6$. They are A→CF, C→AF, F→AC, AC→F, AF→C, and CF→A. The confidence value of each rule will be calculated and a rule having confidence no less than the minimum confidence threshold will be selected as an interesting rule.

The important step of association rule mining is frequent itemset mining. Many algorithms were proposed for finding frequent itemsets, such as Apriori [8], FP-Growth [9], Ecat [10], DFIN [11], NegFIN [12]. However, a large number of frequent itemsets may be generated if a low minimum support threshold is given or a large dataset is mined. Then a large number of association rules are also generated. Closed itemset mining was proposed to reduce the number of frequent itemsets. It mines frequent itemset having no superset with the same support. Closed frequent itemsets are sufficient to mine association rules. All non-redundant association rules will be found from closed itemsets and cover the rules generated from frequent itemsets [13]. Thus, many redundant rules can be eliminated. The concept of closed itemset is based on the two following functions f and g as defined in Eq. (1) and Eq. (2).

$$f(T) = \{i \in I \mid \forall t \in T, i \in t\} \quad (1)$$

$$g(X) = \{t \in D \mid \forall i \in I, i \in t\} \quad (2)$$

Function f returns the set of itemsets included in all the transactions in T, where $T \subseteq D$

itemset if $c(X) = f(g(X)) = X$. The set of closed itemsets is defined as Eq. (3), where FI is a set of frequent itemsets.

$$CI = \{X \mid X \in FI \wedge \nexists Y \in FI, X \subset Y \wedge supp(X) = supp(Y)\} \quad (3)$$

For example, if a set of frequent itemset is FI = {(D):2, (FD):2, (B):2, (F):3, (AF):2, (CF):2, (ACF):2, (C):3, (AC):3, (A):3}. The set of closed itemset is CI = {(FD):2, (B):2, (F):3, (ACF):2, (AC):3} because (D):2 is subset of (FD):2 with the same support, so it is not a closed itemset. (AF):2 and (CF):2 are subsets of (ACF):2 with the same support, so they are not closed itemsets. (C):3 and (A):3 are subsets of (AC):3 with the same support, so they are not closed itemsets. In conclusion, closed itemsets are non-redundant patterns and cover all important patterns. Many algorithms were proposed to find closed itemsets, such as CHARM [14], DCI_CLOSED [15], and LCM [16].

4. Proposed Methodology

4.1. Data Collection and Preparation

The dataset is collected from a hospital database, Thailand. It is retrieved from disease records of patients who are over 30 years olds. The dataset consists of ICD-10 codes of 5,000 patients. Each transaction is ICD-10 codes of a patient. To reduce various ICD-10 codes, ICD-10 codes are grouped [17] and represented by numbers according to Table 1. For example, A00-A09 are grouped in the same category and represented as 1. The dataset is represented as number format because it is easily cleaned and computed for finding association patterns. The dataset is cleaned by removing duplicated numbers of each transaction. After the cleaning process, each transaction contains unique numbers that represent the disease occurrence of a patient. Finally, the characteristic of the dataset is shown in Table 2. An example dataset is shown in Figure 1.

Table 1: ICD-10 Category

ID	ICD-10	Meaning
1	A00-A09	Intestinal infectious diseases
2	A15-A19	Tuberculosis
3	A20-A28	Certain zoonotic bacterial diseases
4	A30-A49	Other bacterial diseases
5	A50-A64	Infections with a predominantly sexual mode of transmission
6	A65-A69	Other spirochetal diseases
…	…	…
222	Y90-Y98	Supplementary factors related to causes of morbidity classified elsewhere

Table 2: The Characteristic of the Dataset

Characteristic	Count
The total number of diseases	202
The total number of patients	5,000
The maximal number of diseases that occurs in a patient	56
The minimal number of diseases that occurs in a patient	6
The average number of diseases that occurs in a patient	11

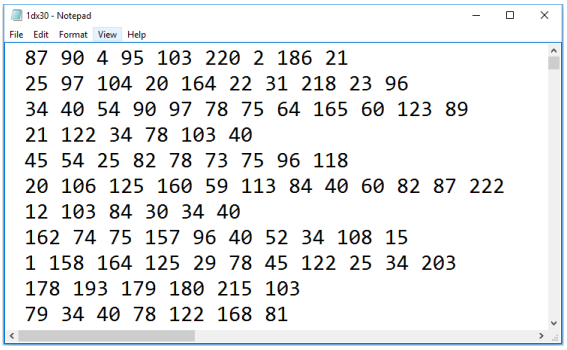


Figure 1: An Example Dataset

4.2. Finding Frequent Diseases

Closed itemset mining is adopted to find frequent diseases because it can generate non-redundant diseases with long disease co-occurrence and coverage of all-important patterns. The diseases with a certain frequency will be selected. The minimum support threshold is used as a filter to select interesting patterns of disease co-occurrence. A disease is considered as an item. All frequent diseases are found based on the CHARM algorithm [14] because CHARM can generate non-redundant frequent diseases with effective of computation time. It uses both itemsets and transaction ids to reduce the search space. Moreover, diffsets and a hash-based approaches are quickly exploited to remove redundant frequent diseases.

The CHARM algorithm firstly constructs an IT-tree that each node is represented by a pair of an itemset and a set of transaction ids. Then it performs a bottom-up depth-first search on the tree to find frequent itemsets. As soon as a frequent itemset X is generated, a set of transaction ids of X is compared with those of other itemsets having the same parent. If the set of transaction ids of X includes the set of transaction ids of the other itemsets, X and the other itemsets are merged to be closed itemsets because they are the same equivalence class. The idea for generating closed itemsets and eliminating non-closed itemsets is based on the following properties.

- If $g(X) = g(Y)$ *then* $c(X) = c(Y) = c(X \cup Y)$. This property implies that X can be replaced by $X \cup Y$ and Y is removed from further consideration.
- If $g(X) \subseteq g(Y)$ *then* $c(X) \neq c(Y)$ *but* $c(X) = c(X \cup Y)$. This property implies that very occurrence of X can be replaced by $X \cup Y$ but Y cannot be removed because it will generate a different closed itemset.
- If $g(X) \supset g(Y)$ *then* $c(X) \neq c(Y)$ *but* $c(Y) = c(X \cup Y)$. This property implies that very occurrence of Y can be replaced by $X \cup Y$ but X cannot be removed because it will generate a different closed itemset.
- If $g(X) \neq g(Y)$, X and Y are not the same equivalence class so that they will be considered to generate closed itemsets.

4.3. Finding Association Patterns of Disease co-occurrence

After finding all frequent diseases, frequent diseases having length no less than two will be used to find the association patters

of disease co-occurrence. An association pattern of disease co-occurrence is in form rule $X \rightarrow Y$, where X and Y are frequent diseases. The minimum confidence threshold is used to filter interesting association patterns. The association patterns of disease co-occurrence are discovered based on the Faster algorithm [8]. The idea of the Faster algorithm is trying to avoid the generation of rules that do not meet the minimum confidence threshold. If a rule $(I - X) \rightarrow X$ passes the minimum confidence threshold then all rules $(I - Y) \rightarrow Y$ will also pass the minimum confidence threshold, where $Y \subset X$. If a rule $(I - Y) \rightarrow Y$ does not pass the minimum confidence threshold, the rule $(I - X) \rightarrow X$ will not pass the minimum confidence threshold. This is because $supp(I - X) \geq supp(I - Y)$ and then the confidence of $(I - X) \rightarrow X$ is not more than the confidence of $(I - Y) \rightarrow Y$. For example, AC→F does not pass the minimum confidence threshold. A→CF and C→AF will not pass the minimum support threshold because of F⊂CF and F⊂AF. Therefore, A→CF and C→AF do not need to generate and compute their confidence.

4.4. Finding Top-k Association Patterns of Disease co-occurrence

After finding all association patterns that pass the minimum confidence threshold, the association patterns will be sorted and selected the top-k association patterns of disease co-occurrence. Given two association patterns, r_i and r_j, r_i has higher precedence than r_j if the following conditions hold:

- $conf(r_i) > conf(r_j)$; or
- if $conf(r_i) = conf(r_j)$, but $supp(r_i) > supp(r_j)$; or
- if $supp(r_i) = supp(r_j)$, but $size(r_i) > size(r_j)$, i.e., length of antecedent of r_i is larger than r_j.

The confidence value is considered as the first priority because it shows how much diseases are related to other diseases. High confidence shows that disease(s) Y is strongly related to disease(s) X. The support value is considered as the second priority because it shows how many patients occur an association pattern of disease co-occurrence. High support shows diseases that occur together in many patients. Next, the length of association pattern is considered as the third priority. The long pattern gives more information than the short one.

4.5. Evaluation Matrix

This paper aims to find association patterns of disease co-occurrence. To evaluate association patterns generated from the proposed method, the dataset is divided into a training set and a testing set by using the 10-fold cross-validation. For each fold, the training set is used to create a predictor that consists of association patterns, represented as rules. The consequent of the rule is considered as predicted diseases. The testing set is used to evaluate the predictor. Two matrixes, prediction ratio and accuracy, are defined to evaluate the effectiveness of the generated association patterns.

A prediction ratio is defined as $Predict = |P|/|A|$, where $|P|$ is the number of predicted diseases correctly and $|A|$ is the total number of the antecedent of rules appears in the testing set. Accuracy is defined as $Accuracy = |C|/|T|$, where $|C|$ the number of matching rules in the testing set and $|T|$ is the number of transactions in the testing set.

5. Performance Evaluation

5.1. Experiment Setup

To investigate the association patterns of disease co-occurrences on electronic medical records, four experiments are conducted. All experiments are implemented by JAVA and use the library in SPMF[18]. The details of the experiments are explained as follows.

The first experiment is conducted to compare the performance of CHARM with a well-known algorithm, call FP-Growth, when generating frequent diseases on the whole dataset with different minimum support thresholds. The minimum support thresholds are set from 1% to 10%. Both algorithms are evaluated by using the number of frequent diseases and computation time.

The second experiment is conducted to investigate the number of association patterns, accuracy, and prediction ratio when using different minimum support thresholds and different minimum confidence thresholds. To reliable results in the medical domain, the minimum support thresholds are set to 10%, 20%, 30%, and the minimum confidence thresholds are set to 60%, 70%, 80%, 90%. The dataset is divided into a training set and a testing set by using the 10-fold cross-validation. The number of association patterns, accuracy, and prediction ratio are reported on average.

The third experiment is conducted to discover the top-10 association patterns of disease co-occurrences from the whole dataset. The minimum support threshold is set to 10%. Then the top-10 association patterns of disease co-occurrences are selected by using the sorting method as explained in section 4.4. The fourth experiment is conducted to find the 5 most frequent diseases in the whole dataset. Then top-3 association patterns are selected to investigate other related diseases of the 5 most frequent diseases.

5.2. Experimental Results

Table 3 reports the number of frequent diseases that are generated on the dataset by using CHARM and FP-Growth. It shows that CHARM gives a smaller number of frequent diseases than FP-Growth when minimum support is set to 1%. Both algorithms generate the same frequent diseases when the minimum support threshold is more than 1%. However, the computation time of CHARM outperforms FP-Growth and almost steady although the minimum support threshold is small as shown in Figure 2. Therefore, the CHARM algorithm is selected for finding frequent diseases in our work.

Table 3: Comparing the Number of Frequent Diseases

min_supp(%)	The number of frequent diseases	
	CHARM	FP-Growth
1	7,015	7,025
2	1,885	1,885
3	861	861
4	502	502
5	327	327

6	216	216
7	158	158
8	128	128
9	105	105
10	81	81

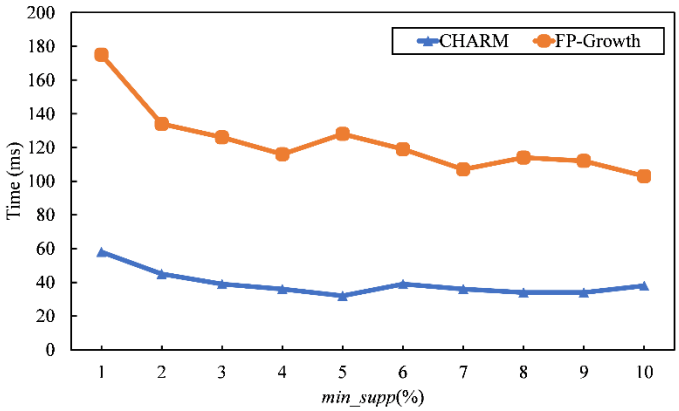


Figure 2: Computation Time of CHARM and FP-Growth

In Table 4, it reports the number of association patterns when different minimum support thresholds and different minimum confidence thresholds are given. The number of association patterns is reduced when the minimum support thresholds and the minimum confidence thresholds are increased. No association pattern is found when minimum confidence threshold is set to 90% so the accuracy and prediction ratio are not evaluated with 90% of minimum support threshold. Table 5 reports the accuracies. It shows that the highest accuracy is 79.76% when minimum support threshold and minimum confidence threshold are set to 10% and 60%, respectively. When minimum support threshold and minimum confidence threshold are increased, the accuracies will be reduced because the number of association patterns is also reduced and then the number of matching patterns in the testing set is small. In Table 6, the most prediction ratios are high although the number of association patterns is very small because the association patterns are created from the most frequent diseases that are strongly related together.

Table 4: The Number of Association Patterns of Disease Co-occurrence

min_supp(%) \ *min_conf(%)*	60	70	80	90
10	59	38	19	0
20	12	9	3	0
30	4	3	1	0

Table 5: Accuracy (%)

min_supp(%) \ *min_conf(%)*	60	70	80
10	79.76	74.32	56.82
20	71.42	71.20	41.26
30	59.23	58.08	34.03

Table 6: Prediction Ratio (%)

min_supp(%) \ *min_conf(%)*	60	70	80
10	81.04	80.03	82.56
20	77.53	82.77	79.30
30	69.70	84.03	80.58

Table 7: Top-10 Association Patterns of Disease Co-occurrence

Rank	Association pattern	No. of patients	Confidence (%)
1	Diabetes mellitus, Metabolic disorders, Renal failure ==> Hypertensive diseases	683	88.81
2	Diabetes mellitus, Renal failure ==> Hypertensive diseases	800	87.62
3	Hypertensive diseases, Influenza and pneumonia ==> Metabolic disorders	526	86.22
4	Diabetes mellitus, Hypertensive diseases, Renal failure ==> Metabolic disorders	683	85.37
5	Aplastic and other anemias, Diabetes mellitus ==> Metabolic disorders	584	85.25
6	Aplastic and other anemias, Hypertensive diseases, Renal failure ==> Metabolic disorders	504	85.13
7	Aplastic and other anemias, Renal failure ==> Metabolic disorders	640	84.76
8	Diabetes mellitus, Renal failure ==> Metabolic disorders	769	84.22
9	Diabetes mellitus, General symptoms and signs ==> Metabolic disorders	555	84.09
10	Aplastic and other anemias, Diabetes mellitus ==> Hypertensive diseases	574	83.79

Table 8: The 5 Most Frequent Diseases and Related Diseases

Disease	No. of patients	Association pattern
Metabolic disorders	3,288	1. Hypertensive diseases, Influenza and pneumonia ==> Metabolic disorders (support = 526, confidence = 86.22%) 2. Diabetes mellitus, Hypertensive diseases, Renal failure ==> Metabolic disorders (support = 683, confidence = 85.37%) 3. Aplastic and other anemias, Diabetes mellitus ==> Metabolic disorders (support = 584, confidence = 85.25%)
Hypertensive diseases	2,599	1. Diabetes mellitus, Metabolic disorders, Renal failure ==> Hypertensive diseases (support = 683, confidence = 88.81%) 2. Diabetes mellitus, Renal failure ==> Hypertensive diseases (support = 800, confidence = 87.62%) 3. Aplastic and other anemias, Diabetes mellitus ==> Hypertensive diseases (support = 574, confidence = 83.79%)
Diabetes mellitus	1,953	1. Metabolic disorders, Hypertensive diseases, Renal failure ==> Diabetes mellitus (support = 683, confidence = 74.72%) 2. Hypertensive diseases, Renal failure ==> Diabetes mellitus (support = 800, confidence = 71.49%) 3. Metabolic disorders, Renal failure ==> Diabetes mellitus (support = 769, confidence = 63.92%)
General symptoms and signs	1,734	1. Other diseases of the respiratory system ==> General symptoms and signs (support = 321, confidence =55.15%) 2. Metabolic disorders, Other diseases of the respiratory system ==> General symptoms and signs (support = 274, confidence =57.44%) 3. Influenza and pneumonia, Renal failure ==> General symptoms and signs (support = 252, confidence = 52.39%)
Renal failure	1,500	1. Aplastic and other anemias, Metabolic disorders, Hypertensive diseases ==> Renal failure (support = 504, confidence = 71.59%) 2. Aplastic and other anemias, Hypertensive diseases ==> Renal failure (support = 592, confidence = 68.91% 3. Metabolic disorders ==> Renal failure (support = 640, confidence = 60.09%)

Table 7 shows the top-10 association patterns of disease co-occurrence. From the top-10 association patterns of disease co-occurrence, most of top-10 association patterns are similar. For example, the first rank shows that if a patient has diabetes mellitus, metabolic disorders, and renal failure, then the patient has a chance to be hypertensive diseases with 88.81% of confidence. The second rank shows that if a patient has diabetes mellitus and renal failure, then the patient has a chance to be hypertensive diseases with 87.62% of confidence. We can conclude that diabetes mellitus, metabolic disorders, renal failure and hypertensive diseases are highly related together. In addition, we found that influenza and pneumonia, plastic and other anemias are highly related to metabolic disorders.

Table 8 shows the 5 most frequent diseases and other related diseases that are represented in association patterns. The 5 most frequent diseases are metabolic disorders, hypertensive diseases, diabetes mellitus, general symptoms and signs, and renal failure. The most frequent disease is metabolic disorders. 3,288 patients or 65.76% of the dataset have metabolic disorders. Hypertensive diseases, influenza and pneumonia, diabetes mellitus, renal failure, and aplastic and other anemias are highly related to metabolic disorders with more than 85% of confidence.

6. Conclusion

In this paper, we proposed a technique to find the association patterns of disease co-occurrence based on closed association rule mining. Closed itemset mining is applied to find frequent diseases. Then the frequent diseases are used to create association patterns of disease co-occurrence. The association patterns are sorted to select top-10 association patterns of co-occurrence. Moreover, the 5 most frequent diseases and other related diseases are discovered. From experiment results, they show that the association patterns of disease co-occurrence give high accuracy if the number of association patterns is large. The prediction ratio is high although the number of association patterns is very small because the association patterns are created from the most frequent diseases that are metabolic disorders, hypertensive diseases, diabetes mellitus, and renal failure. From the investigation of association patterns, we found that diabetes mellitus, metabolic disorders, and renal failure and hypertensive diseases are highly related together. Moreover, influenza and pneumonia, plastic and other anemias are highly related to metabolic disorders.

Acknowledgment

This research project was financially supported by Faculty of Informatics, Mahasarakham University 2020. We would like to thank Mahasarakham hospital for the dataset.

References

[1] Z. Batool, M. Usman, K. Saleem, M. Abdullah-Al-Wadud, F.-e. Amin, and A. Al-Eliwi, "Disease–disease association using network modeling: challenges and opportunities," Journal of Medical Imaging and Health Informatics, **8**, 627-638, 2018. https://doi.org/10.1166/jmihi.2018.2342

[2] C. Chan, C. Chen, and B. Liu, "Discovery of association rules in metabolic syndrome related diseases.", in IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 856-862, 1-8 June 2008. https://doi.org/10.1109/IJCNN.2008.4633898
[3] M. Harahap, A. M. Husein, S. Aisyah, F. R. Lubis, and B. A. Wijaya, "Mining association rule based on the diseases population for recommendation of medicine need," Journal of Physics: Conference Series, **1007**, 1-11, 2018. https://doi.org/10.1088/1742-6596/1007/1/012017
[4] I. T. T. K. S. Nahar Jesmin, and Y.-P. Phoebe Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," Expert Systems with Applications, **40**(4), 1086 - 1093, 2013. https://doi.org/10.1016/j.eswa.2012.08.028
[5] W. Ma, L. Zhang, P. Zeng, C. Huang, J. Li, B. Geng, J. Yang, W. Kong, X. Zhou, and Q. Cui, "An analysis of human microbe-disease associations," Briefings in bioinformatics, **18**, 85-97, 2017. https://doi.org/10.1093/bib/bbw005
[6] H. Zhang, X. Huo, X. Guo, X. Su, X. Quan, and C. Jin, "A disease-related gene mining method based on weakly supervised learning model,", BMC Bioinformatics , **20**, 169-174, 2019. https://doi.org/10.1186/s12859-019-3078-9
[7] W. Lan, J. Wang, M. Li, J. Liu, F. Wu, and Y. Pan, "Predicting microRNA-disease associations based on improved microRNA and disease similarities," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 15, No. 6, 1774-1782, 2018. https://doi.org/10.1109/TCBB.2016.2586190
[8] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules," in 20th International Conference Very Large Data Bases VLDB, **1215**, 487-489, 2000. https://doi.org/10.1007/BF0294884
[9] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach," Data Mining and Knowledge Discovery, **8**(1), 53-87, 2004. https://doi.org/10.1023/B:DAMI.0000005258.31418.83
[10] M. J. Zaki, "Scalable algorithms for association mining," IEEE Transactions on Knowledge and Data Engineering, **12**(3), 372-390, 2000. https://doi.org/10.1109/69.846291
[11] Z.-H. Deng, "DiffNodesets: An efficient structure for fast mining frequent itemsets," Applied Soft Computing, **41**, 214-223, 2016. https://doi.org/10.1016/j.asoc.2016.01.010
[12] N. Aryabarzan, B. Minaei-Bidgoli, and M. Teshnehlab, "negFIN: An efficient algorithm for fast mining frequent itemsets," Expert Systems with Applications, **105**, 129-143, 2018. https://doi.org/10.1016/j.eswa.2018.03.041
[13] M. Zaki, "Closed itemset mining and non-redundant association rule mining," Encyclopedia of Database Systems, 365-368, 2009. https://doi.org/10.1007/978-0-387-39940-9_66
[14] M. Zaki, and C.-J. Hsiao, "CHARM: An efficient algorithm for Closed Itemset Mining," in the 2002 SIAM International Conference on Data Mining, 457-473, 2002. https://doi.org/10.1137/1.9781611972726.27
[15] C. Lucchese, S. Orlando and R. Perego, "Fast and memory efficient mining of frequent closed itemsets," IEEE Transactions on Knowledge and Data Engineering, **18**(1), 21-36, 2006. https://doi.org/10.1109/ TKDE.2006.10
[16] T. Uno, M. Kiyomi, and H. Arimura, "LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets," in Proceeding of IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1, 1-11, 2004. https://doi.org/10.1.1.108.155
[17] W. H. Organization, "ICD-10 Version:2019," International Statistical Classification of Diseases and Related Health Problems, 2019.
[18] P. Fournier-Viger, J. C.-W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, and H. T. Lam, "The SPMF Open-Source Data Mining Library Version 2," Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol. 9853, 36-40, 2016. https://doi.org/10.1007/978-3-319-46131-1_8

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 652-663 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Automatic Stochastic Dithering Techniques on GPU: Image Quality and Processing Time Improved

Giorgia Franchini*, Roberto Cavicchioli, Jia Cheng Hu

University of Modena and Reggio Emilia, Department of Physics, Informatics and Mathematics, 41125, Modena, Italy

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 01 November, 2020
Online: 24 November, 2020

Keywords:
Dithering
Half-tone
Floyd-Steinberg
Stochastic Algorithm
Parallel Computing

ABSTRACT

Dithering or error diffusion is a technique used to obtain a binary image, suitable for printing, from a grayscale one. At each step, the algorithm computes an allowed value of a pixel from a grayscale one, applying a threshold and, therefore, causing a conversion error. To obtain the optical illusion of a continuous tone, the obtained error is distributed to adjacent pixels. In literature there are many algorithms of this type, to cite some Jarvis, Judice and Ninke (JJN), Stucki, Atkinson, Burkes, Sierra but the most known and used is the Floyd-Steinberg. We compared various types of dithering, which differ from each other for the weights and number of pixels involved in the error diffusion scheme. All these algorithms suffer from two problems: artifacts and slowness. First, we address the artifacts, which are undesired texture patterns generated by the dithering algorithm, leading to a less appealing visual results. To address this problem, we developed a stochastic version of Floyd-Steinberg's algorithm. The Weighted Signal to Noise Ratio (WSNR) is adopted to evaluate the outcome of the procedure, an error measure based on human visual perception that also takes into account artifacts. This measure behaves similarly to a low-pass filter and, in particular, exploits a contrast sensitivity function to compare the algorithm's result and the original image in terms of similarity. We will show that the new stochastic algorithm is better in terms of both WSNR measurement and visual analysis. Secondly, we address the method's inherent computational slowness: we implemented a parallel version of the Floyd-Steinberg algorithm that takes advantage of GPGPU (General Purtose Graphics Processing Unit) computing, drastically reducing the execution time. Specifically, we observed a quadratic time complexity with respect to the input size for the serial case, whereas the computational time required for our parallel implementation increased linearly. We then evaluated both image quality and the performance of the parallel algorithm on a exhaustive image database. Finally, to make the method fully automatic, an empirical technique is presented to choose the best degree of stochasticity.

1 Introduction

Most printing devices are constrained by a limited number of color intensities (typically 2/4/8/16 values), as consequence, before printing an image they must first convert the original color domain into a new set of values that can be represented with the color intensities at their disposal. A process we call quantization.

An intuitive and trivial quantization method is to map each input pixel to the nearest available color. However, most of the time this leads to a visually unpleasant and low quality output image due to a large number of artifacts. Here, dithering comes to help and it has been proved to be able of achieving a higher quality results. In particular, the best known and popular dithering algorithms is the Floyd-Steinberg (FS) [1] which is based on errors diffusion. Besides the FS, there are other algorithms that differ from this for the weight distribution matrix. The error scattering technique is straightforwardly described in Algorithm 1.

Algorithm 1 Dithering algorithm

```
for rows ∈ image do
    for pixels p ∈ row do
        find closest color τ to p
        err = p − τ
        diffuse err to neighbouring pixels according to the error diffusion scheme
    end for
end for
```

*Corresponding Author: Giorgia Franchini, University of Modena and Reggio Emilia, Email: giorgia.franchini@unimore.it

In case of binary dithering applied to a monotone image (e.g. grayscale to black and white), Algorithm 1 is just a simple threshold operation. Extending it to a larger set of colors (or tones) is a trivial change in *finding the closest color*. It was one of the most revolutionary methods and is widely used in digital printers, which prints small and isolated dots.

The other algorithms against which we will compare for the numerical experiments are: Atkinson, Burkes, Fan, Sieraa, Filter Lite, Jarvis, Judice and Ninke (JJN), Jarvis, Stucki and Shio Fan [2].

There are several error distribution scheme and different ways to scan the image, such as the traditional left-to-right direction and the top-to-bottom raster. Considering all the matrices of weights related to the various methods would be redundant, therefore we show only a few in Figure 1 to clarify the concept.

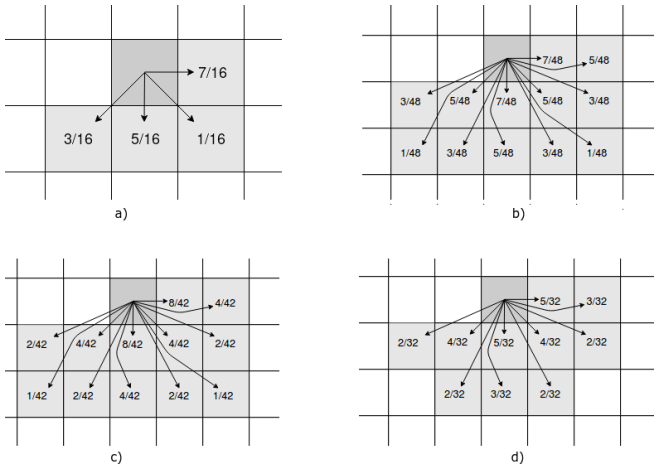


Figure 1: Error diffusion scheme for different dithering algorithms: Floyd-Steinberg (a); Jarvis, Judice and Ninke (b); Stucki (c); Sierra (d). The flux of the whole image remains the same.

We also took care of the borders, in particular, we ensured that the image flux or average is maintained through the entire image, Figure 2 shows the modified FS error distribution *weights* in all particular cases.

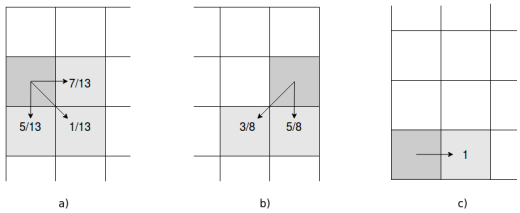


Figure 2: Error diffusion in the FS dithering in case of boundaries. a) first column, b) last column and c) last row.

The major shortcomings of the FS dithering algorithm are the generation of disturbing hysteresis artifacts or worm patterns at extreme gray-levels[3, 4] and the slowness [5].

2 Background and preliminaries

The purpose of image dithering, or halftoning, is the procedure that generates a pattern of quantized pixels able to create a continuous-tone image illusion. A comprehensive analysis about halftoning techniques can be found in [2]. Dithering is necessary for showing gray scale images on printed surfaces when the only available tones are black and white. Dithering algorithms will produce different results with respect to quality and image characteristics. To allow a comparison between different algorithms, we must therefore quantify the performance. Doing so by means of psychovisual tests is hard due to the need of controlled conditions and long duration of the tests. A possible solution to this problem is developing quality measures that can numerically express the perceived visual difference between the dithered image and the original one. The ideal target would be to include several objective aspects of image quality within one single, robust and reliable measure, so to evaluate and improve dithering algorithms.

We cans also describe digital halftoning as an artifact minimization problem, because we search for the dithered image that minimizes the measured visibility of artifacts. We therefore need a computational model that allows us to compute the visible error, so to automatically rank images in an optimization algorithm. Nevertheless, measuring in a quantitative way the visual quality of a dithered image is still a challenging task.

Objective quality measures can be divided in two classes: mathematical/statistical ones and human visual system (HVS) based.

The first ones could be used to evaluate dithered patterns of constant grey values. These metrics provide a great understanding of the possible relationships for a given point distribution. These kinds of measures are usually independent of viewing conditions or observer, easy to compute thanks to the low computational complexity. It is possible to find in literature different examples of this kind of measures, such as: Structural Similarity (SSIM) [6], Peak Signal-to-Noise Ratio (PSNR) [7] and Mean Squared Error (MSE) [8]; nevertheless, none of these measures consider the presence of artifacts as explained in [9, 10].

The second one tries to model the perceived visual quality and predict artifacts generated by the dithering procedure. The ideal halftone algorithm is able to minimize this visibility. Since the eye detects more distortions at certain spatial frequencies, the development of a halftone visibility metric is based on a model of the HVS.

HVS models are relatively simple and have proven to be quite successful when applied to algorithms that seek the best possible distribution scheme. All digital halftone techniques, whether they are based on screening algorithms, error scattering or iterative halftone, all adopt these models, either explicitly or implicitly. In fact, even those methods that are categorized as model-free because they do not explicitly include an HVS block in their block diagram, still replace it with a low pass filter. Moreover, not only is an HVS model fundamental to the design of most halftone techniques, but additionally, the shape of the HVS model can be modified to achieve increased texture quality in the dithering result.

Therefore, the performance of a certain halftone algorithm can be maximized by adequately designing improved HVS models.

Digital halftoning usually refers to methods and algorithms able

to convert continuous-tone images to binary ones, for displaying them in bilevel devices such as printers, both ink jet and laser ones. Within the computer industry, the market demand for fast, high resolution printing devices has recently escalated. Halftoning's contribution to this industry has been essential to its success. For this reason investigating image quality, and in particular finding an objective measure to evaluate the results is of fundamental importance.

2.1 A Quantitative Measure

As discussed, HVS models use human visual selectivity and sensitivity to model and improve the perceived image quality. HVS is based on the psycho-physical process that relates psychological phenomena (contrast, brightness, etc.) with physical phenomena (light intensity, spatial frequency, wavelength, etc.). It determines which physical configurations give rise to a particular psychological sensation (perceptual, in this case).

The two dimensional Discrete Fourier Transform (DFT) is applied on the image and then multiplied using Hadamard product with a Contrast Sensitivity Function (CSF), so that a component of the image at a particular angular frequency is weighted by the CSF value at that frequency. The result of this computation is the two dimensional DFT of an image that would lead to the same psychological response when viewed from a visual system with a flat CSF to which the original image leads when viewed by HVS [11]. The weighted signal-to-noise ratio (WSNR) of an $M \times N$ pixel binary image (y), given the original (x) of the same size is computed as:

$$WSNR(dB) = 101 \log_{10}\left(\frac{\sum_{u,v} |X(u,v)C(u,v)|^2}{\sum_{u,v} |(X(u,v) - Y(u,v))\, C(u,v)|^2}\right) \qquad (1)$$

where $X(u,v)$, $Y(u,v)$ and $C(u,v)$ are correspondingly the DFT of the input image, of the output image and the CSF.

2.2 Stochastic version

As already described in the previous sections, halftoning converts the input image into a black and white version of it to be reproduced on a binary output device, such as an ink-jet printer, which can only choose whether to print a dot or not in each position. The human eye, behaving as a low-pass filter, blurs the dots and spaces together and creates the illusion of many continuous shades of gray tones. Depending on the specific way the dots are distributed, a display device can produce different degrees of image fidelity with more or less granularity. According to the HVS, isolated and randomly arranged dots, if distributed in a proper manner, should produce images with the highest quality, maintaining fine details and sharp edges. Nevertheless, some displays and printing devices are unable to reproduce isolated dots consistently in their entirety and, as a result, introduce printing artifacts that greatly degrade the aforementioned details that the computed dot distribution is designed to preserve. For this reason, many printing devices create periodic patterns of grouped dots, which are easier to produce on the printed page in a consistent manner. For this reason, in our halftone study, the main objective is to determine the optimal dot distribution for that HVS, and then produce these models computationally efficiently.

Since the first article in which Floyd and Steinberg's algorithm was mentioned, many changes to the original error spread algorithm have been introduced to avoid the undesirable artifacts present in the original algorithm. But while these modifications eliminate disturbing artifacts at certain levels of gray, many do so therefore at the expense of other levels. Jarvis [12] and Stucki [13], for example, introduced a different mask model and weights, called 12-element error filter, but the artifacts remain during the algorithm because of the fixed behavior manifested by these algorithms. Hence we realize that the problem is not related to the number of elements involved but rather to the determinism of the method. Other approaches suggested to change the scanning path in which pixels are processed, including the serpentine as the most trivial up to the space-filling curves like Hilbert's curves. Unfortunately, these methods still create strong periodic patterns. To mitigate the problem of artifact our idea is to transform the FS algorithm, a deterministic algorithm, into a Stochastic (SFS) version inspired by [14]. A crucial point of our new approach is the preservation of the average image. As in all halftone methods, where everything and only the error is spread. Even in our method, while modifying the weight matrix at each pixel, the average of the values of the original image is not changed.

In particular we choose a real number p and, for each pixel, we generate 2 random number:

$$r_1 \in [-5/16, 5/16] \text{ and } r_2 \in [-1/16, 1/16]$$

and we compute

$$\frac{7}{16} + p \cdot r_1, \frac{5}{16} - p \cdot r_1, \frac{3}{16} + p \cdot r_2 \text{ and } \frac{1}{16} - p \cdot r_2 \qquad (2)$$

as error diffusion coefficients. This method, as a modification of Ulichney's one[15], is a good approximation of a blue-noise process. Blue-noise halftoning is characterized by a distribution of binary pixels where the minority pixels are spread as homogeneously as possible. This method of pixel distribution creates a pattern that is isotropic (or radially symmetric), aperiodic and does not contain any low-frequency spectral components. From the considerations made about HVS, it's not surprising that blue-noise creates the visually optimal arrangement of dots. Unfortunately, blue-noise is not trivial or fast to generate [16], but whit our method this can be approximated.

2.3 Optimal choice of p

By introducing our new method we have underlined the need to choose a p parameter.

In our previous work the choice of p was totally empirical and not automated, that is, we tried different values of p in a predetermined range and picked the best according to the image quality measured by WSNR. Here we introduce a method for the automatic setting of p. As pointed out above, these algorithms are often used for printers in different areas, but often as an integral part of industrial processes. For this reason and for the sake of completeness, finding a fully automatic method would make the method more effective and usable on a large scale. There are two critical issues to be addressed: the choice of an appropriate interval and the need for an intervention by the operator for the choice. First of all, let's discuss the choice of the interval. As we have defined p there are no

limitations, it simply has to be a real value. Since it is not a probability, it can certainly assume both values greater than 1 and negative values. Conceptually there is a symmetry, at least on average, between negative and positive values so we can only take into account positive values without losing in the generality. On the other hand, choosing values much greater than 1 means to substantially change the coefficients of the matrix used for the diffusion of the errors, for example choosing $p = 2$ the values, on average, are doubled. In addition to these preliminary considerations we conducted several experiments aimed at understanding where the maximum value of WSNR was in relation to the p choice. As it can be seen in Figure 3 , the maximum is usually included in the $[0, 2]$ interval.

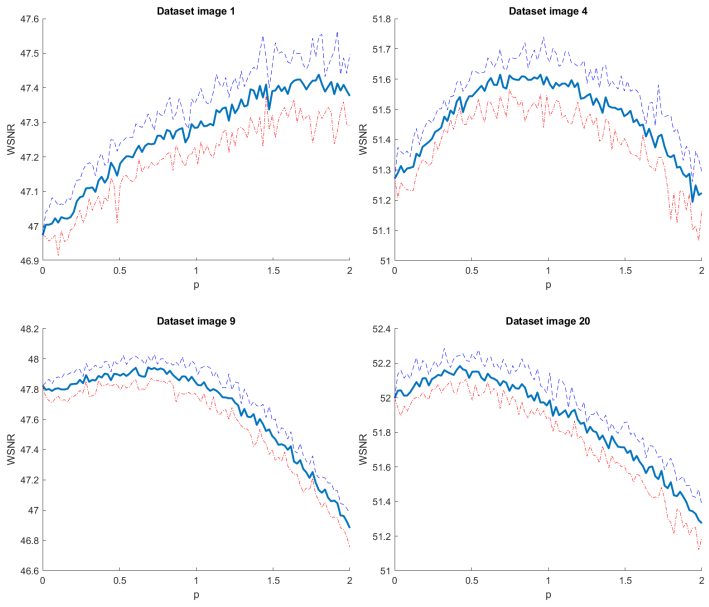


Figure 3: Sample of the algorithm behaviour for p in the interval $[0, 2]$ for some images of the dataset available in [2]. Due to the stochastic nature of the algorithm, we repeat the algorithm 10 times with different random seeds for each value of p, plotting minimum, average and maximum value for the resulting WSNR.

Once the range is fixed, we address the second issue by means of an optimization algorithm that provides the p value that maximizes the WSNR. The first idea, given the amount of references in the literature, was to use a first-order technique as a gradient method [17]. But we immediately noticed that, due to the lack of convexity of the function to be maximized, such methods are often not effective. There are certainly techniques to use gradient methods even in non-convex areas but they require at least the differentiability of the function. Since we have to maximize a stochastic process we see both the basic hypotheses of the gradient methods fall: convexity and differentiability. Fortunately, our implementation of the parallel algorithm on GPUs makes execution time particularly low, even for large images. This allows us to approach efficiently the problem in an empirical way without increasing the time excessively. Considering all these premises, our final idea is to calculate, first, the WSNR for different values of $p \in S_n$ where S_n is an equally spaced partition of the interval $[0, 2]$ with n elements. Due to the stochasticity of the method, for each $p \in S_n$ we consider the average over m runs.

Finally, choose the p with the highest average WSNR. We summarize our method in the equation 3 where m is the number of runs, I and $SFS(p)$ represent respectively the input image and the output image obtained from a single run of our stochastic dithering algorithm with parameter p.

$$p* = \arg\max_{p \in S_n} \frac{1}{m} \sum^{m} WSNR(I, SFS(p)) \tag{3}$$

We obtained the best results with the following configuration:

- Size of partition S_n, n = 100
- Number of repetitions for each p, m = 10

In this way, given any image in our experiments, the method lead to satisfying results in terms of similarity to the initial image.

3 GPU implementation

Dithering algorithm are infamously known for being "embarassingly" unparallel due to the inherently sequential nature of the dithering operation: the every pixel output depends on all the previously computed outputs. Achieving good speed ups in such problems is a challenging task and several work from the research community tried to overcome this issue, either proposing a "friendlier" dithering algorithm or a more efficient implementation. For example in [18]–[20] they introduce a block-based error diffusion algorithm which allows to simultaneously process blocks of pixels and produced good visual results, however is not clear which block dimension should be taken in order to get the best image quality; whereas in the work [21] the authors propose to parallelize over multiple pages in order to saturate the GPU utilization, in this work we focus on a single image case; in [22] they propose an efficient implementation that consists of a hybrid approach where CPU and GPU are both exploited but the CPU is used in the regions where GPU is expected to perform worse, in particular, towards image corners; finally [5] provided an optimal parallelization algorithm from a theoretical point of view, unfortunately, assigning each thread a diagonal portion of the image is currently not suited for general purpose devices as it causes more time spent for additional computational overhead than what parallel devices can manage to reduce.

In this work, we adopt the classical Floyd-Steinberg algorithm as it's still the dithering algorithm that yields the best result and propose an implementation that rely solely on GPUs. Graphics Processing Units (GPUs), in particular General Purpose GPUs, are massively parallel architectures that comprises thousands of cores designed to compute in an extremely efficient way repetitive and simple operations such as those involved in the rendering pipeline in Computer Graphics, hence "Graphics" in the name. Each core are less performing and simpler than state of art processors but the computational power in such devices lies in the number of cores rather than the capability of each single core. Nowadays GPUs are widely adopted for many tasks in data mining, physics simulation, medical imaging and machine learning, to name a few, hence the "general purpose" in the name. We propose an implementation [1] that

[1] https://git.hipert.unimore.it/hjiacheng/stochastic_floydsteinberg

is optimized for the NVIDIA GPUs, from which we will also get the results reported in section 4.1, and we program with the architecture's standard Application Programming Interface (API), called CUDA [23] developed by NVIDIA itself. Nonetheless, our implementation takes into account obstacles and design principles valid for almost all GPUs so it's not limited to NVIDIA architectures.

Our implementation consists of two parallelization levels we call stream level and pixel level. First of all we, describe the pixel level parallelism consists of a function, named Parallel-FS, that implements the Floyd-Steinberg on GPU and assumes the image being already stored in the GPU memory, this function will be then invoked by the stream level function.

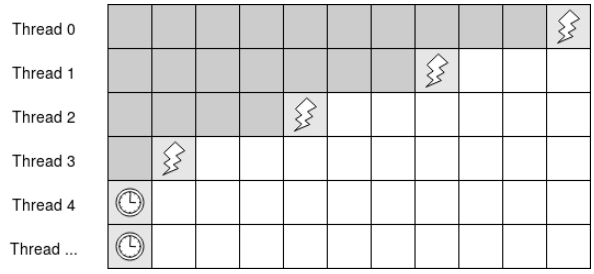


Figure 4: Each thread is assigned a row and waits for the previous one to be three pixels ahead instead of two. The thunder symbol on the diagonal represents the pixels processed simultaneously, whereas the clock indicates a waiting thread

In order to achieve good results one must take into account the of hardware on which the algorithm is executed. GPUs are SIMT (Single Instruction Multiple Thread) architectures partitioned in several Streaming Multiprocessors (SM), the latter are clusters of cores with multiple hardware schedulers for dispatching a fixed number of threads called warps. Each thread in the warp are committed to perform the same operation simultaneously, a feature called lock-step execution, and they are able to fetch multiple data at the same time if the addresses in the memory are aligned, a phenomena called "coalesced access". While these are good features in simple and regular memory accesses, it becomes actually a downside in case of Floyd-Steinberg because operations and accesses are neither simple nor regular. Dithering algorithm finds the closest color input and handle borders with many conditional, hence branching, instructions, which hurts the performance as the device must turn-off the portion of the threads not involved in the current branch because of the lock-step execution, moreover there's no coalesced access as the error diffusion scheme follows an irregular pattern leading to increased latency. This is the reason why the optimal theoretical algorithms presented in [5] cannot be implemented, nonetheless we try to approximately recreate the theoretical optimal situation in Parallel-FS. The function takes as input three data structures, an input buffer I_{buff} where the original image is stored, an output buffer O_{buff} where the result will be stored and error buffer E_{buff} for the temporary error diffusion values. Each row i is assigned a thread which, for each pixel (i, j), $j \in \{1, ..., I_{WIDTH}\}$ where I_{WIDTH} is the image width, reads from the input buffer and computes the output $O_{buff}(i, j)$ by finding the closest color according to the value $S = I_{buff}(i, j) + E_{buff}(i, j)$, and distribute the error $S - O_{buff}(i, j)$ to the neighbouring pixels according to the error diffusion scheme of Floyd-Steinberg, updating $E_{buff}(i, j+1)$, $E_{buff}(i+1, j)$, $E_{buff}(i+1, j-1)$ and $E_{buff}(i+1, j+1)$. The actual implementation carefully takes into consideration also borders compared to this simplified formulation. Each thread must synchronize with the previous thread in order to compute the assigned pixel result, in our implementation, in order to avoid update conflicts in the error buffer, which would require slow atomic operations, each thread waits for the previous one to be three pixels ahead with respect to his row instead of two, similarly to [21] (see Figure 4). The Figure 4 also illustrates how pixels are processed simultaneously along the diagonal (hence the name "pixel level parallelism").

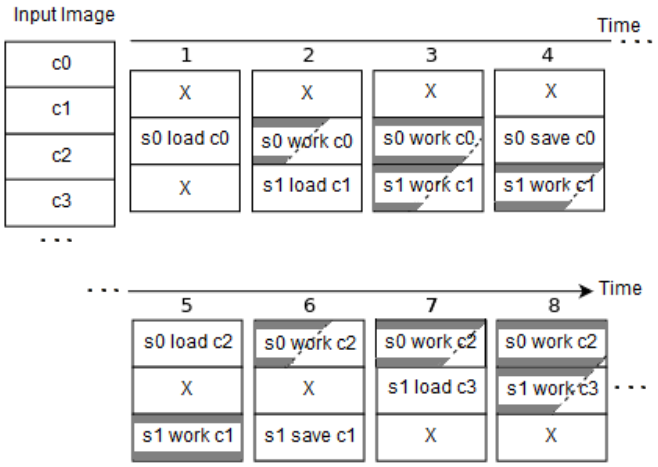


Figure 5: Visual illustration of the Double Buffering parallel execution scheme for the stream level parallelism, with two CUDA streams; sj and ci and denote CUDA stream respectively and image chunks. The "load" and "save" operations represent the data transfer from host to device input buffer I_{buff} and from device output buffer O_{buff} to host respectively and "work" consist of the invocation of Parallel-FS. It can be noted that, starting from time step 2, the memory transfer latency is completely hidden by GPU activity. "X" denotes the empty chunk, note that there is always an empty chunk because given $N_{STREAMS}$ there are $N_{STREAMS} + 1$ blocks in the data structure, so each iteration, every stream finds the results of the first row of the error buffer already calculated by the previous stream by simply sliding up.

The stream level parallelism addresses a problem more related to the physical devices limitation rather than the FS algorithm. In previous function we assumed the image to be already when the output is ready, the result is transferred back into host memory. Unfortunately, in most parallel implementations memory bandwidth represents one of the main bottleneck. The outer loop in our stored in GPU memory, however, for a task to be executed on GPU, every required data must be copied first from host memory (accessible by CPU only) to device memory (accessible by GPU only) and parallel implementation addresses this problem by hiding almost completely the memory transfer overhead between CPU and GPU wit the double buffering technique, which also removes the device memory size constraint for huge images.

Given an image of size $I_{HEIGHT} \times I_{WIDTH}$ and two input parameters N_{CHUNKS} and $N_{STREAMS}$, the first parameter define the number of chunks the image is divided into, whereas the latter represents the number of chunk to be processed simultaneously. First we define $Chunk_{HEIGHT} = \frac{I_{HEIGHT}}{N_{CHUNKS}}$ and we divide the image into N_{CHUNKS} of size $Chunk_{HEIGHT} \times N_{WIDTH}$, secondly we create three data struc-

ture of size $(Chunk_{HEIGHT} \cdot (N_{STREAMS} + 1)) \times I_{WIDTH}$ size for the input, output and error buffer respectively. Now, $N_{streams}$ CUDA stream are created, each of these concurrent execution stream are assigned an image chunk are responsible of computing the result by invoking the Parallel-FS function previously described on the assigned chunk. During the algorithm execution, every function invoked by each stream references to the same data structure but on a different memory address offset, the size of each buffer is $N_{STREAMS} + 1$ chunk blocks instead of $N_{STREAMS}$ because the the additional chunk allows to leave one chunk empty at every iterative step so Parallel-FS function avoids storing the error computed in the last row into an auxiliary buffer, minimizing the number of conditional operations. Because of the empty chunk, when a CUDA stream finishes the job, it can slide on the upper chunk in a toroidal fashion along the height, which means that the errors computed in the last row of the data structure are actually propagated in the first and the CUDA stream working in the first block moves to the last in the next outer loop iteration as it finds the error values ready to be used. The indexing and synchronization required to ensure the overall correctness of the FS algorithm details will not be discussed for the sake of simplicity. The final result is that when some CUDA streams are processing their chunk, others are transferring data from host to device memory and vice-versa, thus the memory bottleneck is hidden and the architecture is fully exploited. In Figure 5 we show the presented parallel execution scheme.

3.1 Theoretical improvement

Consider an image of size $P \times Q$, where P identifies the number of rows and Q the number of columns. We want to know the order of magnitude of the time that is theoretically needed to process the whole image. We use as a measure unit the time needed to process a single pixel.

In the serial version: $Time = P \times Q$ where $Time$ is the time for the whole process. In the parallel version, with the same image dimension, we have

$$Time = 3P + Q - 3 \text{ with } Q > P \tag{4}$$

and

$$Time = 3Q + P - 3 \text{ with } P > Q. \tag{5}$$

Considering, as said before, the time we use to process the single pixel we can assume to spend a M time to calculate the first row, but at this same time, always considering the parallel implementation, we have already calculated the triangle at the top left as shown in Figure 4. Going on to finish the second line, we only need to calculate another 3 pixels and we can proceed with a similar reasoning for all the rows in the image. So we have $Q+3\times(P-1) = Q+3P-3$. Even in an image where the number of rows and columns is swapped we have the same runtime. For these considerations we can conclude that we pass from a time that grows as $O(P \times Q)$ with the input size, in the serial case, to a time that grows linearly $O(P + Q)$ in the parallel case, all this from a theoretical point of view. This allows us to have a great saving in computational time, especially in industrial contexts where the images are of huge dimensions. This means an incredible acceleration, especially when the size of the image is large, as often happens in industrial cases.

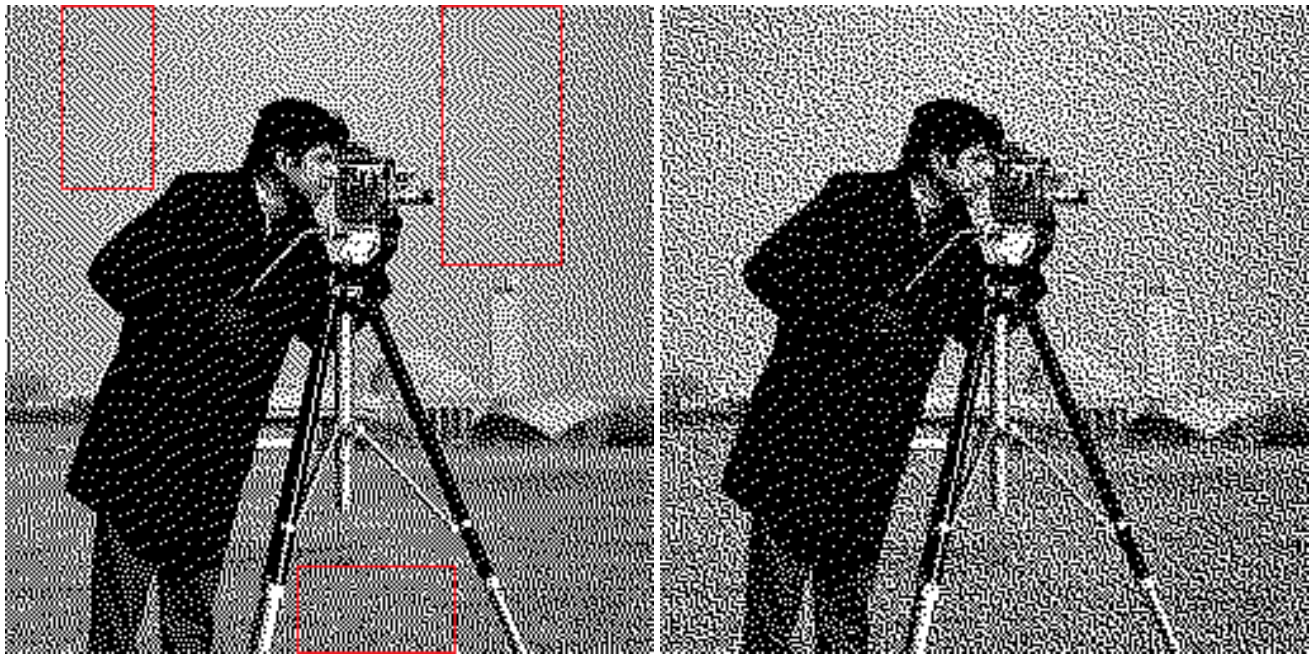


Figure 6: Cameraman image dithered with $p = 0$ on the left. The red rectangles highlight regions with visible artifacts. On the right, dithered image with optimal p value to maximize the WSNR value.



Figure 7: Lena image dithered with $p = 0$ on the left. The red rectangles highlight regions with visible artifacts. On the right, dithered image with optimal p value to maximize the WSNR value.

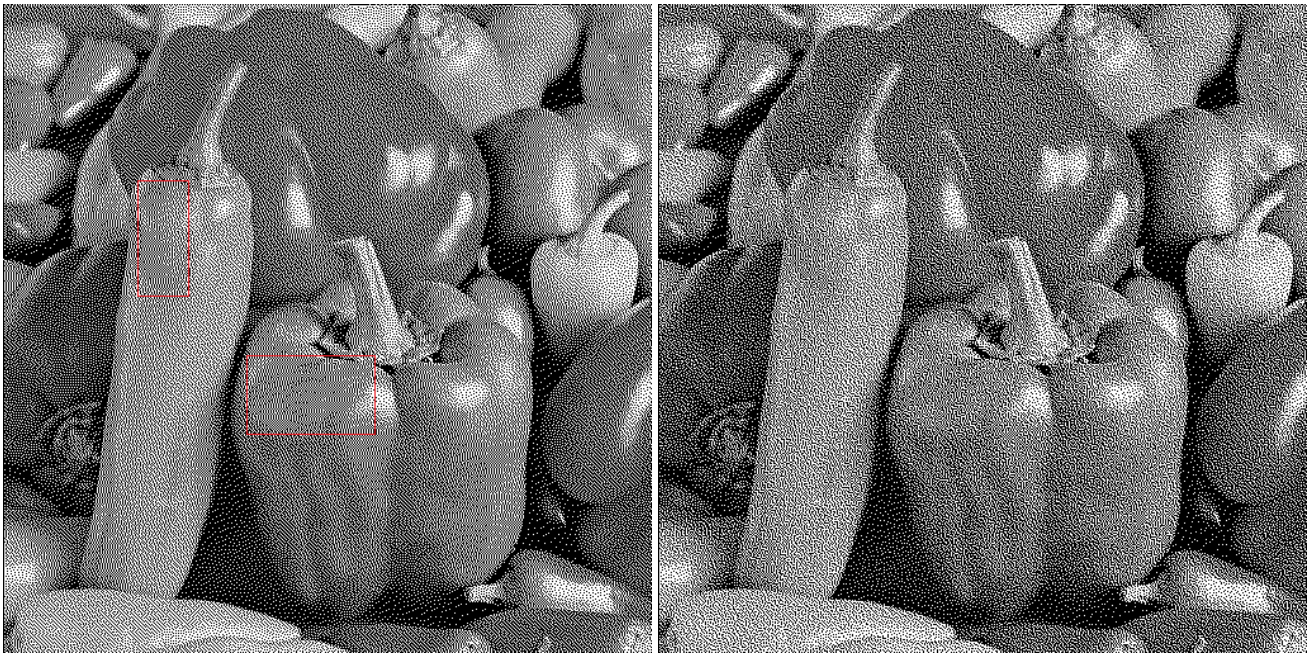


Figure 8: Peppers image dithered with $p = 0$ on the left. The red rectangles highlight regions with visible artifacts. On the right, dithered image with optimal p value to maximize the WSNR value.

4 Numerical experiments

The authors of [2] provided a sufficiently large dataset of images on which we applied our algorithm[2], the Digital Halftone Database (DHD). This dataset is composed of 196 reference images extracted from the CVG-UGR-Image database and the Genreal-100 dataset. In addition to these real images, we also assessed our approach on synthetic images. In particular, a particularly complicated image to process is the grayscale gradient, as in the Figure 10 or multi-tone chessboard. Moreover, the well known Lena, peppers and cameraman have been tested as well. Not all the images have same values of width and height since the algorithm can be applied to rectangular pictures as well. In the first experiments, we chose synthetic images. To verify the improvement with respect to the presence of artifacts,

[2] https://sites.google.com/view/sankarasrinivasan-s/research-publication/DHD-Database

we applied our algorithm to images composed of all shades of the (0 – 255) grayscale. This was done because some tones are more susceptible than others to artifact forming. In a second experiment we applied the new techniques to images often used in published works, in order to compare the results with the literature. Finally we compare against the results of [2], that are provided as results images in the same dataset. Obviously it can be observed that there is a considerable difference between the proposed stochastic approach and the simple FS. The experiments were performed on a desktop computer equipped with a CPU and GPU with the following specifications: Intel i7 7700K CPU, NVIDIA GeForce GTX 1080 GPU and 32GB RAM.

4.1 Results

With respect to image quality, we computed the measure described in Sec. 2.1 after applying FS, SFS and, for completeness, other dithering algorithms known in literature: Stucki[13], SIERRA[24] and Jarvis, Judice, and Ninke [12]. The results for the different images in Figure 10 (top row) and Figure 6, 7 and 8 are shown in Table 1 were we highlighted in bold the best WSNR values. The charts in Figure 13 show the behaviour of the WSNR value with respect to p for these images. Furthermore, we provided additional visual comparisons in Figures 11 and 12.

We compared our algorithm to the error diffusion ones presented in [2] on the same DHD dataset previously mentioned. We computed WSNR from the output images that are made available in the dataset and compared against our approach: over 196 different images we achieved the maximum WSNR value amongst all the algorithms 71% of the times, while the remaining 29% has been topped by the Filter Lite Dot Diffusion algorithm from [25].

In Table 2 we report the values of WSNR for some images of the dataset. We have to remark that, due to the different output size of the image, for some pictures we were unable to compute the WSNR value of comparison. The algorithms that are taken into account here are our FSF, Atkinson (ADD), Burkes (BDD), Fan (FDD), tradirional Floyd Steinberg (FSDD), Frankie Serra (FSIDD), Filter Lite (FLDD), Jarvis Judice and Ninke (JJNDD), Jarvis (JDD), Stucki (SDD) and Shio Fan (SFDD).

In addition to providing a theoretical discussion, we tested the different versions of the algorithm, serial and parallel, on images of different sizes. As we expected we obtained very good results: while serial time increases quadratically, parallel time increases only linearly. Let's better illustrate the trends in Figure 9.

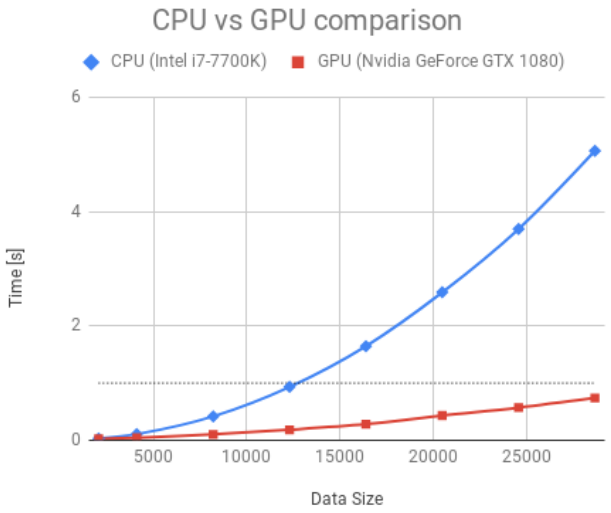


Figure 9: Comparison between the two versions: serial and parallel. Time in seconds on vertical axis, horizontally there is side dimension in pixels of a square image. In dotted line: all parallel execution time are less than one second.

Table 1: Dithering methods WSNR comparison

Image	FS	SFS			Stucki	Sierra	JJN
		$p = 0.25$	$p = 0.5$	$p = 0.75$			
Gradient	52.58	53.15	**53.34**	53.28	50.49	50.64	50.28
Lena	48.51	48.48	**48.57**	48.51	41.79	41.58	41.08
Cameramen	45.78	46.02	46.30	**46.43**	39.29	39.13	38.61
Pepper	46.13	46.39	46.93	**47.27**	39.45	39.13	38.55

Table 2: Error Diffusion WSNR comparison

Index	***FSF***	***ADD***	***BDD***	***FDD***	***FSDD***	***FSIDD***	***FLDD***	***JJNDD***	***JDD***	***SDD***	**SFDD**
1	47.50	20.00	43.61	44.33	48.23	40.90	**49.74**	26.35	40.44	41.44	44.48
3	**50.64**	24.09	44.40	44.22	46.92	41.28	47.99	28.72	40.79	41.51	44.63
5	45.49	18.80	41.72	41.08	44.91	38.80	**46.15**	23.56	38.27	39.11	41.28
7	**49.77**	20.99	43.65	45.49	47.57	41.22	49.57	27.41	40.73	41.60	45.97
17	**44.97**	15.81	39.52	40.40	42.48	36.30	43.72	20.87	35.80	36.50	40.79

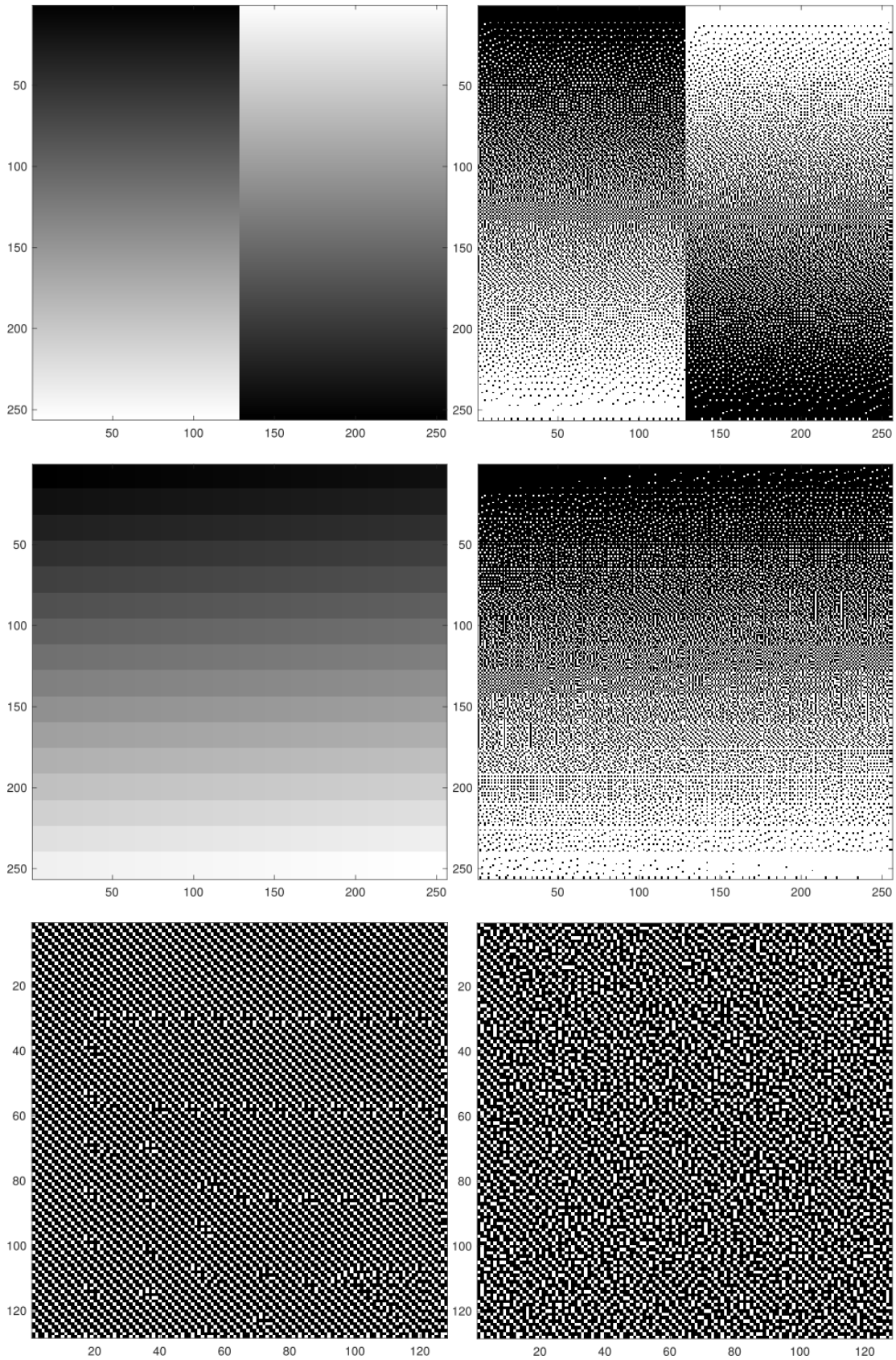


Figure 10: Inverted gradient pattern, original image (top-left) and halftoned using traditional FS (top-right): emerging artifacts are clearly discernible. In the middle row, 256 squares of all possible 1 Byte quantization values (left) ant their dithered version (right) Bottom line, particular of a single tone (value 128): on the left side artifacts arise distinctly with traditional FS, while on the right side the stochastic nature of the method allows no emerging pattern.

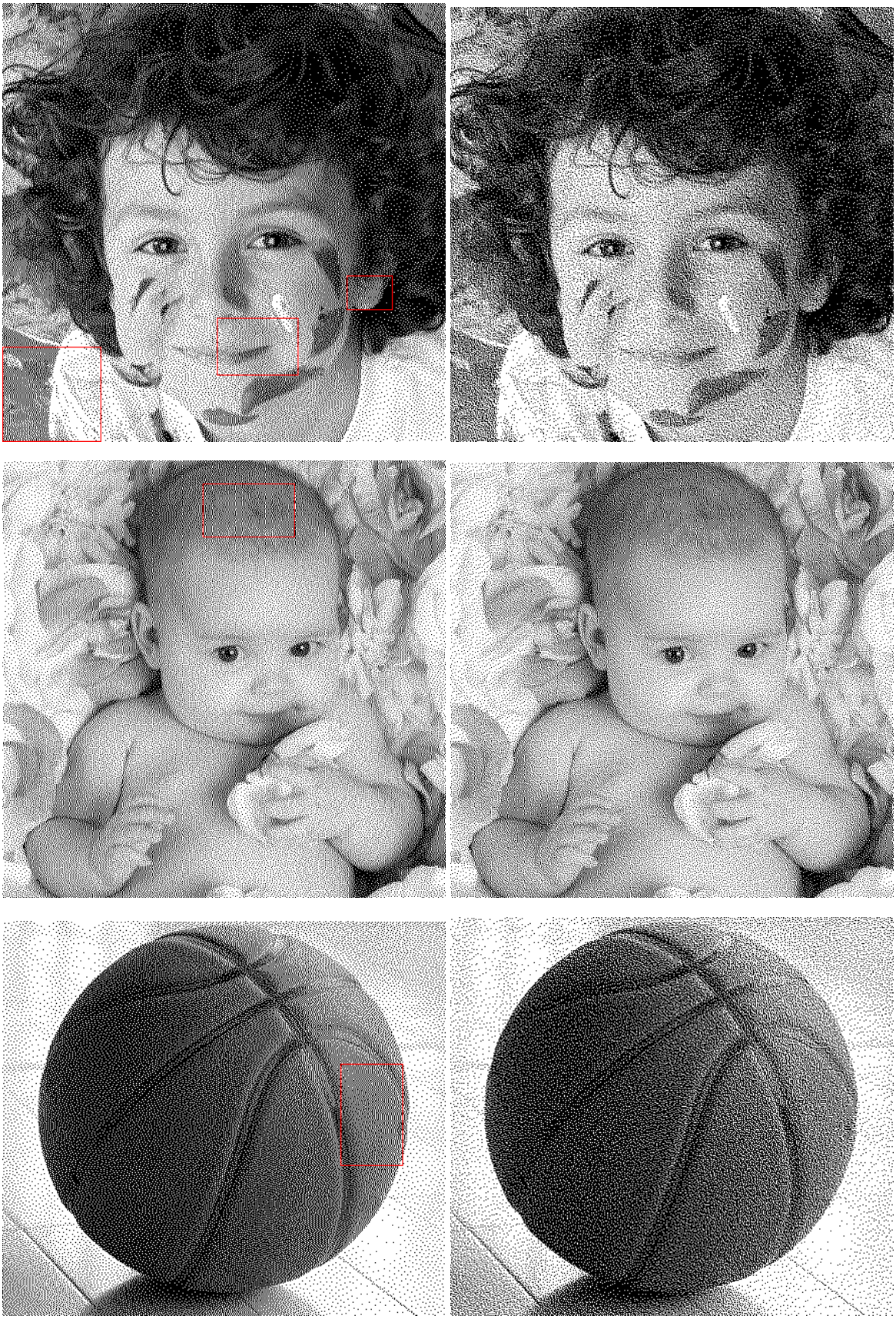


Figure 11: On the left: dithered images using FLDD for images 24, 4 and 6 of the Digital Halftone Database. The red rectangles highlight regions with visible artifacts. On the right, same images dithered using SFS with best p*.

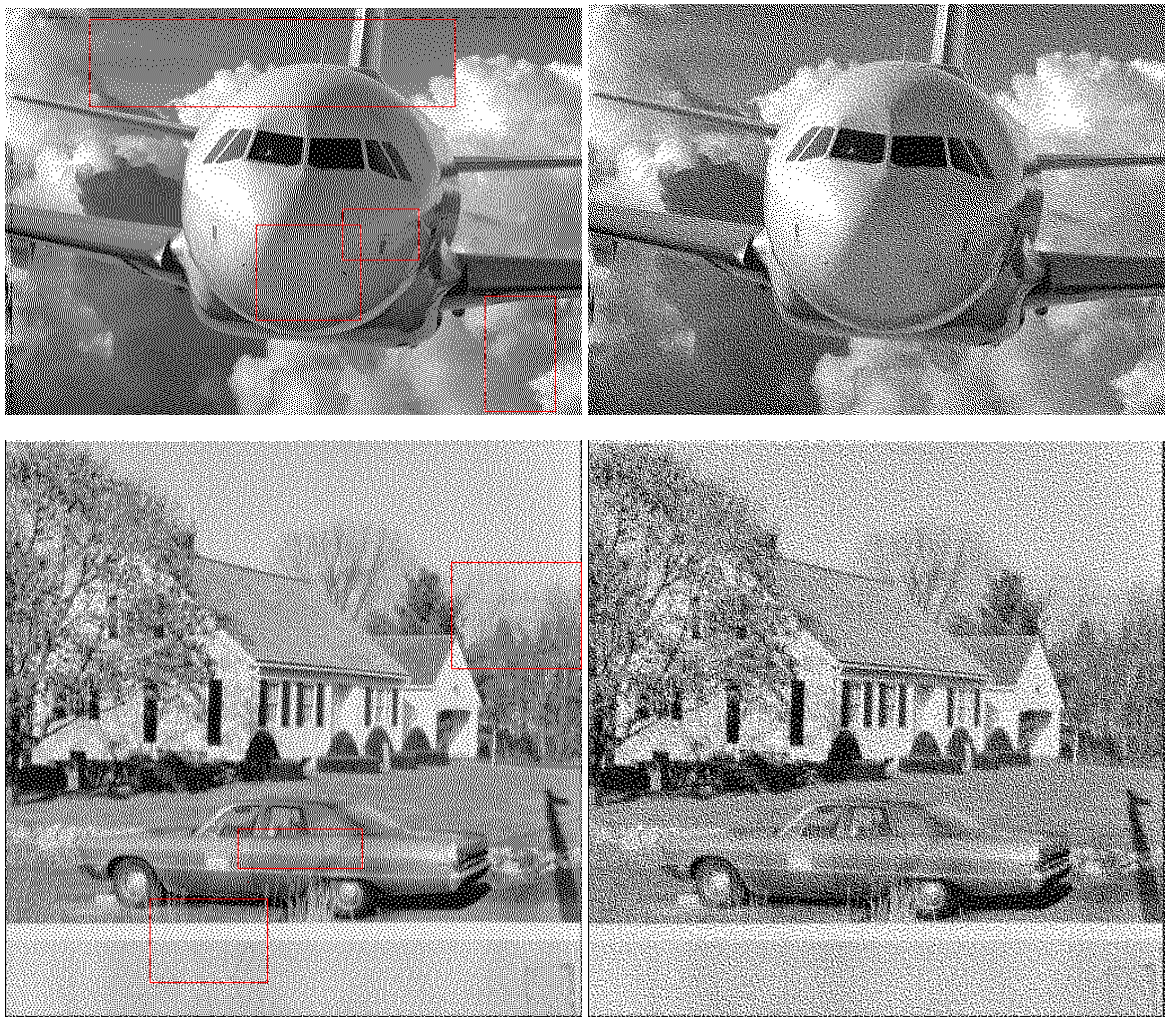


Figure 12: On the left: dithered images using FLDD for images 26 and 7 of the Digital Halftone Database. The red rectangles highlight regions with visible artifacts. On the right: same images dithered using SFS with best p*.

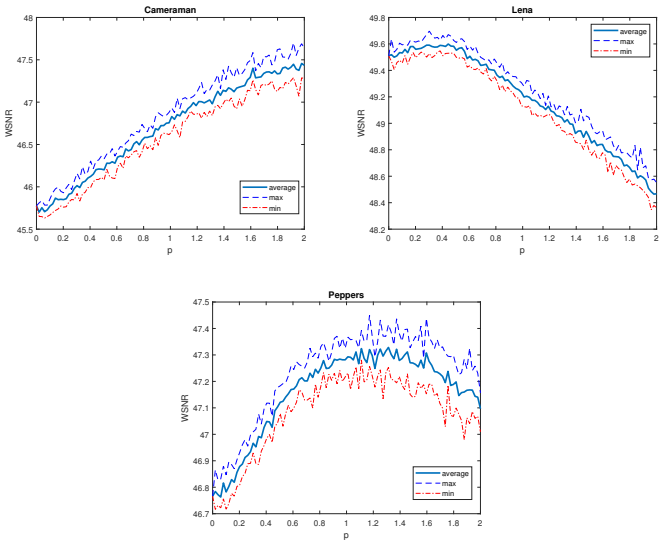


Figure 13: WSNR computed for our Stochastic Floyd-Steinberg algorithm with $p \in [0, 2]$ on the 3 well known images of Cameraman, Lena and Peppers. We proceed as for Figure 3.

5 Conclusions

In conclusion, the new algorithm we presented solves both the problems highlighted at the same time: the presence of artifacts and the slowness of execution. In particular, the transformation of the algorithm from deterministic to stochastic, with the introduction of a white noise, avoids the emergence of artifacts. The improvement can be evaluated through a specific measure that takes into account human perception called WSNR. Moreover, there is no increase in computational time; on the contrary, the parallel version we propose shows an excellent speed-up as the image size increases; which, especially in industrial applications, can be very large. Thanks to our parallel implementation, which guarantees speed to the method, we were able to fully automate the technique. So given an image we are capable of applying the best amount of stochasticity to the algorithm, thus obtaining an image as free of artifacts as possible. All the research part of the optimal parameter is done before the actual image printing process. This guarantees a method in which the operator does not have to make a decision and the printer is not occupied for a long time. In conclusion, the method is totally automatic, fast and guarantees a final image as close as possible to the initial one and suitable for printing. As future work some of the

possible extensions could be: more sophisticated methods to find optimal p, developing other error measures that take into account artifacts or geometric textures of different kinds. Another idea is to evaluate the 4, 8 and 16 gray tones versions of the dithering algorithm by proposing a technique that finds the best trade-off between image quality and number of tones.

References

[1] R. W. Floyd, L. S. Steinberg, "An adaptive algorithm for spatial gray scale," in Proceedings for the Society for Information Display, 75–77, 1976.

[2] J. M. Guo, S. Sankarasrinivasan, "Digital Halftone Database (DHD): A Comprehensive Analysis on Halftone Types," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1091–1099, IEEE, 2018.

[3] Zhigang Fan, R. Eschbach, "Limit cycle behavior of error diffusion," in Proceedings of 1st International Conference on Image Processing, 2, 1041–1045, 1994, doi:10.1109/ICIP.1994.413514.

[4] M. Pedersen, F. Albregtsen, J. Hardeberg, "Detection of worms in error diffusion halftoning," Proceedings of SPIE - The International Society for Optical Engineering, **7242**, 2009, doi:10.1117/12.805555.

[5] P. T. Metaxas, "Optimal Parallel Error-Diffusion Dithering," in Proceedings of SPIE - The International Society for Optical Engineering, volume 3648, 485–494, doi:10.1117/12.334593.

[6] P. Ndajah, H. Kikuchi, M. Yukawa, H. Watanabe, S. Muramatsu, "SSIM image quality metric for denoised images," in Proc. 3rd WSEAS Int. Conf. on Visualization, Imaging and Simulation, 53–58, 2010.

[7] A. Horé, D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," in 2010 20th International Conference on Pattern Recognition, 2366–2369, 2010, doi:10.1109/ICPR.2010.579.

[8] B. Girod, "Psychovisual Aspects Of Image Processing: What's Wrong With Mean Squared Error?" in Proceedings of the Seventh Workshop on Multidimensional Signal Processing, 207–220, 1991, doi:10.1109/MDSP.1991.639240.

[9] T. Mitsa, K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 5, 301–304, 1993, doi:10.1109/ICASSP.1993.319807.

[10] D. R. Mehra, "Estimation of the Image Quality under Different Distortions," International Journal Of Engineering And Computer Science, **8**, 17291–17296, 2016, doi:10.18535/ijecs/v5i7.20.

[11] R. Näsänen, "Visibility of halftone dot textures," IEEE Transactions on Systems, Man, and Cybernetics, **14**(6), 920–924, 1984, doi:10.1109/TSMC.1984.6313320.

[12] J. F. Jarvis, C. N. Judice, W. Ninke, "A survey of techniques for the display of continuous tone pictures on bilevel displays," Computer graphics and image processing, **5**(1), 13–40, 1976, doi:https://doi.org/10.1016/S0146-664X(76)80003-2.

[13] P. Stucki, "Image processing for document reproduction," in Advances in digital image processing, 177–218, Springer, 1979.

[14] R. Ulichney, Digital Halftoning, MIT Press, Cambridge, MA, USA, 1987.

[15] R. Ulichney, Digital halftoning, MIT press, 1987.

[16] I. Georgiev, M. Fajardo, "Blue-noise dithered sampling," in ACM SIGGRAPH 2016 Talks, 1–1, 2016.

[17] D. P. Bertsekas, J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," SIAM Journal on Optimization, **10**(3), 627–642, 2000, doi:10.1137/S1052623497331063.

[18] Y. Zhang, J. L. Recker, R. Ulichney, I. Tastl, J. D. Owens, "Plane-dependent error diffusion on a GPU," in Proceedings of SPIE - The International Society for Optical Engineering, 2012, doi:10.1117/12.906966.

[19] Y. Zhang, J. L. Recker, R. Ulichney, G. B. Beretta, I. Tastl, I.-J. Lin, J. D. Owens, "A parallel error diffusion implementation on a GPU," in Proceedings of SPIE - The International Society for Optical Engineering, 2011, doi:10.1117/12.872616.

[20] P. Li, J. P. Allebach, "Block interlaced pinwheel error diffusion," Journal of Electronic Imaging, **14**(2), 1–13, 2005, doi:10.1117/1.1900136.

[21] B. Trager, C. W. Wu, M. Stanich, K. Chandu, "GPU-enabled parallel processing for image halftoning applications," in IEEE International Symposium of Circuits and Systems (ISCAS), 1528–1531, 2011, doi:10.1109/ISCAS.2011.5937866.

[22] A. Deshpande, I. Misra, P. J. Narayanan, "Hybrid implementation of error diffusion dithering," in 18th International Conference on High Performance Computing, 1–10, 2011, doi:10.1109/HiPC.2011.6152714.

[23] Nvidia, "CUDA Programming Guide Version 10.0," .

[24] T. Helland, "Image Dithering: Eleven Algorithms and Source," http://www.tannerhelland.com/4660/dithering-eleven-algorithms-source-code/

[25] C. Labs, "Libcaca study: the science behind colour ASCII art," `http://caca.zoy.org/wiki/libcaca/study/introduction` [Accessed: Jul. 20, 2020].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 655-659 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Social-Interactive Learning Concept Used for Electronic Educational Resource "Post-Graduate Foreign Language" and the Obtained Learning Curve

Natalya Chernova[*,1], Victor Chernov[1], Margarita Emelianova[2], Raisa Akhunzianova[3], Danil Sukhopluev[1]

[1]*Kazan (Volga Region) Federal University, Naberezhnye Chelny Institute, Naberezhnye Chelny, 423812, Russian Federation*

[2]*I.N. Ulianov Chuvash State University, Faculty of Foreign Languages, Cheboksary, 428015, Russian Federation*

[3]*Naberezhnye Chelny State Pedagogical University, Chair of Foreign Languages, Naberezhnye Chelny, 423806, Russian Federation*

ARTICLE INFO

Article history:
Received:27 June, 2020
Accepted:04 August, 2020
Online: 25 August, 2020

Keywords:
Educational process
Second language acquisition
E-learning
Participation
Zone of proximal development

ABSTRACT

The article touches upon an urgent question of social-interactive learning. The resource under study deals with one of the possible ways to solve this problem, in the context of global socio-cultural transformations and a new paradigm for the development of society. It is the formation of a subject-subject dialogue between a teacher and a student. The second language research study showed how pre-and intermediate speakers' performance breaks down in the face of a difficult narrative task and self-regulation and control over the mediational means are lost. More advanced speakers are able to guide themselves through the task. The main idea of such education is that good learning leads to development. It seems to us important to refer the concept of the zone of proximal development to the development of the individual, which affects the boundaries of the zone of proximal development. Electronic educational resource with the help of mediating means or sign operations makes external social interactions become "internalized", namely, internally reconstructed psychological processes - ways of thinking and learning A student identifies active personality development prospect through the actual experience. The students' learning curve let us prove the idea that IT technologies intensify the process of studying, but should provide not only language accomplishments but active communication with a teacher as well.

1. Introduction

This paper is an extension of work originally presented in 2017 10th International Conference on Developments in E-Systems Engineering (DeSE). The desire to find the right balance between the usage of electronic educational resources, direct participation on the part of a student and a learning curve determined the problem of our research. In theoretical terms, this is the problem of substantiating social-interactive learning concept in the electronic-digital form. In practical terms, it is the problem of determining the relationship between the success of training and the formation of a student's personal qualities as a subject of educational activity [1]. We would like to develop the idea that constant transitions of external actions into internal ones and vice versa take place, not only constantly, but also contribute to penetration into the zones of proximal development, due to the expansion of the amplitude of oscillations, i.e. development and self-development of the subject itself (Popov L.M.) [2].

In Vygotsky's understanding, this means that the personality is realized and manifested in all spheres of mental development in ontogenesis [3]. The actualization of personal characteristics, in particular, social orientation and personal abilities and qualities contributes to the subjective position formation (Leontiev A.N.) [4]. The development of the semiotic function is concretized in the ability to highlight in the text (in the broad sense of the word) the signified and the signifier; in the signified - the object and meaning, in the signifier - the form and content (meaning), which, in its turn, is the development of new levels of means mastery. Thus, it is necessary not only for the formation of scientific knowledge, but also for any spheres [5]. The "Post-graduate Foreign Language" electronic educational resource was developed according to the idea implying that a student should act, first of all,

[*]Corresponding Author: Natalya Chernova, Naberezhnye Chelny 423812, Russian Federation, +79033190461, Email: natalchern@gmail.com

as a researcher of thinking, activity. Having mastered which, a student can successfully bring his/her random impressions about different phenomena of the surrounding world into a single system of verified judgments, grounded in understanding the essence of a particular sphere of reality (Luria A.R.) [6]. The main goal of a person is self-development, first of all, of personal qualities, without which the fulfillment of functional goals by a person is either impossible or of poor quality. This logic supposes the subjective activity of the student, who him/herself creates the teaching and makes him/herself, while the line between the processes of training and education is erased.

As it turns out, only instructional conversations relate human mental functioning to positioning themselves as individuals in carrying out a task in a foreign language (R. Donato). Computer assisted learning helps to move a focus from form to meaning. So, societal context participation (Zinchenko V.V.) [7] is implemented as social-interactive learning concept. The authors' aim is to provide evidence of the relationship between the success of training and the usage of "Post-graduate Foreign Language" electronic educational resource, which provides additional opportunity for a student to be a subject of educational activity when learning a foreign language. "Participation" metaphor is the main tool used in the process of the program creation according to Social-interactive Learning Concept.

The "Post-graduate Foreign Language" electronic educational resource is an automated learning system, providing a socio-cultural approach (Vygotsky L.S.) to language learning [3]. In the first place, it is the opportunity of extemporary communication (i.e. Figure 4). Further, a student does not only gain some new knowledge in the subject but receives skills of various kinds. As a result, a student has an opportunity of realizing in practice a fundamental way of learning [8].

The study shows that "activity theory" (Leontiev A.N.) finds its application in e-teaching as well as classroom language learning [4]. The paradigm of "zone of proximal development" is followed through: 1) communicative language studying as a part of cognition; 2) semiotic mediation; 3) private speech. In analyzing the recorded interaction between the students and the teacher, it shows that instructional conversations are developmentally sensitive to the students' growing ability to use the language required to carry out a specific task. The evidence of this is the shift from other-regulated to self-regulated error correction.

This computer training program can be presented as a systematic presentation of the above-mentioned approach to the process of studying, which compiles the fundamental provisions of the subject-activity approach and the socio-cultural theory with modern technologies in e-learning, such as Moodle software platform [1]. The choice of the studied subject is due to the necessity of fluent English for any scientist or a post-graduate.

The "Post-graduate Foreign Language" resource shows that Social-interactive Learning Concept can be reflected through metacognitive strategies such as predicting, planning, and monitoring. So, mediating learning with the help of IT technologies shows that the Learning Curve levels up in comparison with traditional learning strategies providing instructions alone. The succession and number of exercises help students and teachers as well help in the process of communicative and cognitive activities in an ESL class [8].

2. Background to the Problem

The program under study provides for independent work on the part of a student and work under the guidance of a teacher as well [2]. It helps in getting the necessary knowledge, gives an opportunity of providing effective dealing with the educational problem and allows active learning in natural environment. This computer-training program is a means of studying "Academic English" within the Post-graduate curriculum and meets all educational and methodological demands [7].

Referring to the private speech in his study, J. Smith (1996) operationalizes it as a 'verbal attempt of self-regulation during problem-solving tasks', which is, the result of stress that accompanies construction of the meaning in L1 or L2, the private speech is distinguishable from the interpersonal communication. Smith examines a grammar class of high intermediate ESL students and claims that private speech can arise in a discussion [9]. The thing is that classroom discourse is usually analyzed for its social, communicative value, neglecting the cognitive function of instructional talk.

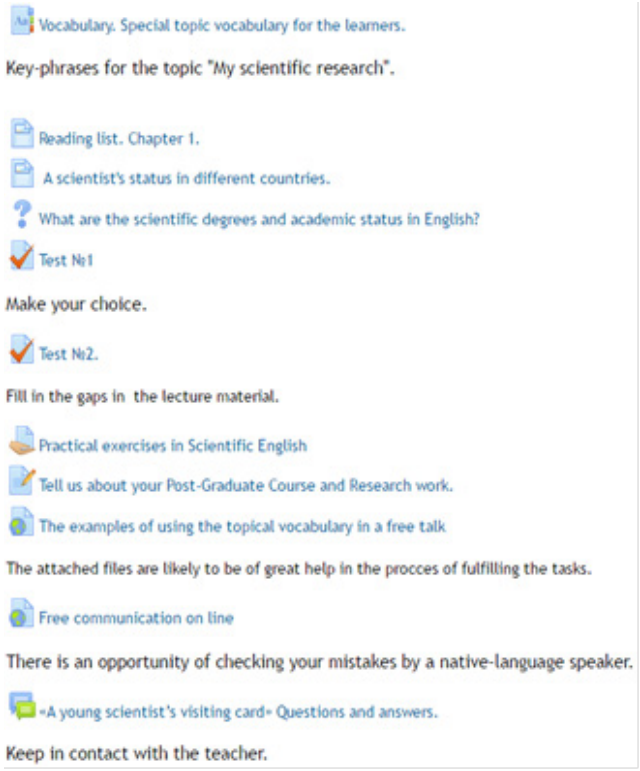


Figure 1: "My scientific research" item program

Classroom conversations depend on spontaneity, unpredictability and focus on new information. If the discussion is shaped toward a curricular goal, and teachers in their turns build or activate background knowledge in students, these conversations can also be instructional. They are relevant to language classrooms because they provide opportunities for experiencing language outside. The concept of instructional conversations was explored within the frames of a PhD study, compiled by N.A. Chernova. The motivation for the study was the following observation. It turned out that instructional conversations usually arose with fairly

advanced ESL learners. The question remained, was it possible to have an instructional conversation in a beginning foreign language class? Limited linguistic resources are not available for a topically coherent, extended, and meaningful conversation. The following data were received. If a teacher followed a predictable pattern a) setting up the vocabulary practice exercise; b) completing textbook exercises using the new vocabulary; and c) spontaneous questions and comments by students and teacher embedded within the practice exercise, conversational communication was out of the question (Chernova & Mustafina, 2016) [7]. The possibility of such conversation depended on management talk and extension activities, which show features of it most consistently and impressively [10].

Here is the screen shot to illustrate Chapter 1 with the topic "My scientific research". The tasks "Tell us about your Post-Graduate Course and Research work", "Free communication online", and "«A young scientist's visiting card» Questions and answers" provide for Instructional conversations. The last one let the students keep in contact with the teacher.

The tests are checked automatically, but two-way communication is possible when fulfilling training tests. So, you keep in contact with the teacher in English. The fact that every stage is under control is the most prominent achievement with this resource.

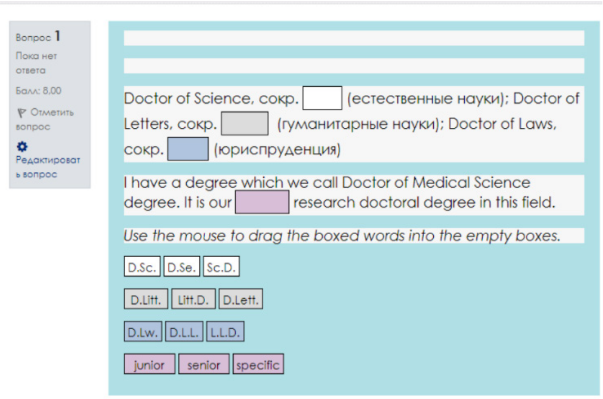


Figure 2: Chapter 1. Test №2. Fill in the gaps in the lecture material

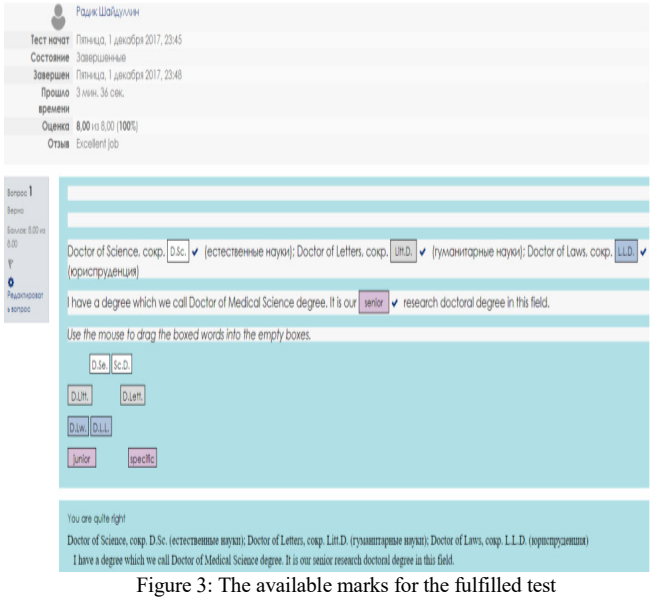


Figure 3: The available marks for the fulfilled test

There is some material in Russian in the test due to the need to check understanding. The items in the upper left-hand corner are given in Russian because of the application interface.

The "Post-graduate Foreign Language" resource makes it possible to estimate the test results at once. Thus, the process of training is intensified.

3. Methods of Research

The methodological basis for creating a program is the theory of D. Wood, J.S. Bruner, J. Ross and R. Donato about spontaneous communication in the classroom [11]. At the initial stage, students simply reproduce monosyllabic statements, repeating them after the teacher, and the invariable model of teaching in the classroom is "a teacher – students – a teacher", or otherwise "subject-object-subject". Over time, students have spontaneous responses to the teacher's statements, as well as mutual assistance and joint construction of statements, the teacher's help decreases, and students play an increasingly active role in the lesson, using the same techniques as the teacher [12].

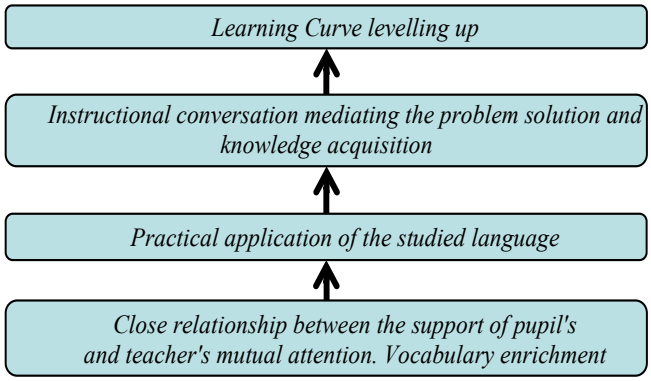


Figure 4: Educational technique used with "Post-graduate Foreign Language" electronic educational resource

The Instructional conversation is a dialogue in which the speakers are involved in solving the problem in the process of acquiring knowledge [4]. We would like to emphasize that it mediates both the solution of the problem and the acquisition of knowledge. Growth in the quality of knowledge is the main result achieved. This is the tracked indicator presented in the study.

In the framework of the experimental program, the study was carried out. The experimental group, within the framework of this strategy, discussed their actions in a dialogue and corrected each other's mistakes, explaining them. The results were compared with those of the control group the students of which studied the subject "Academic English" within the Post-graduate curriculum but didn't use the "Post-graduate Foreign Language" electronic educational resource. The topics and the volume of material studied was the same. The class periods in both groups are fully alike. The quantitative values of the quality of knowledge obtained after the experimental study are presented in Figures 5, 6.

The study took place in Kazan Federal University. The participants were the first-year post-graduate students. The level of language proficiency was rather high in the second term by the time of the study. By this time a student can already use the specific vocabulary on their specialty in spontaneous dialogues. The study was aimed at testing the influence of "Post-graduate Foreign Language" electronic educational resource on the correct use of special language forms in oral speech.

4. Results of the Study and Discussion

The obtained changes in the quantitative characteristics of the quality of knowledge are statistically significant. The system of evaluation begins with 1 (equals to "Bad" mark) and finishes with 5 (equals to "Excellent" mark).

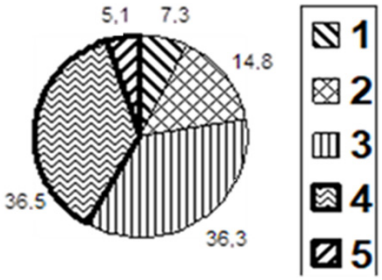


Figure 5: Quality of knowledge. Control Groups. Control Survey

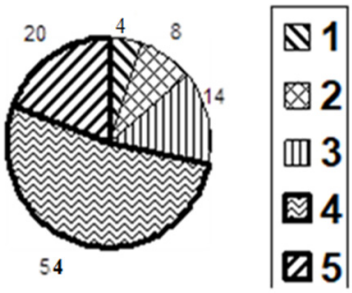


Figure 6: Quality of knowledge. Experimental Groups. Control Survey

Taking into account that levels 4 and 5 are sufficient, it can be concluded that 74% of students in the Experimental Groups achieved this indicator, and 41.6% in the Control groups. Thus, the quality of knowledge in the Experimental Groups is almost 1.8 times higher than in the Control Groups.

This distribution allows us to speak about an increase in knowledge among students of Experimental Groups, which is expressed in the ability to notice and correct errors with minimal or no hint, take full responsibility for correcting errors, and be consistent in the use of keywords in all contexts.

The study shifts the governing concept of learning away from the acquisition metaphor toward the participation one with an individual's activity. In other words, students are taking part in a collaborative dialogue. Communicative collaboration is based upon the active and purposeful agent on the part of a student. The authors offer the following model of ESL classes: 1) self-directed activity comprising self-dependence, activity, social direction, self-government, reflection → 2) reflective, problem-solving orientation with the help of the "Post-graduate Foreign Language" electronic educational resource → 3) internalization of social interactive process in the zone of proximal development with the English language as a cognitive tool for the individual under the guidance of a teacher → 4) the second language mediated process of an agent formation in learning and professional activities (direct solution of professional tasks).

This model allows to be sensitive to students' needs and abilities and support the overpowering and transformative agency embodied in the learner. When working with e-learning, it is important to arrange the teaching process based on the strengths of the students developing their confidence and ability to work and study independently and in a team. The sociocultural context of the teaching process contributes to the development of the students' identity within the society [10]. The diversity of sociocultural contexts used during the classes has to teach them to think critically and creatively, solve problems and recognize the advantages. These approaches applied during the teaching process provide students not only with the knowledge and skills, but also with emotional comfort and sufficient self-esteem.

By the end of the studying course, "free talk" technique is being practiced. The students in this study were tested individually. Firstly, they were asked a series of individual questions in the form of interviews, and secondly, they were asked three discussion questions, responding to which students had to express their opinion, tell the story and come up with a statement. The questions were formulated in such a way that, when Answering the questions provided for compiling complex sentences with specific vocabulary. The quality evaluation was carried out in three stages: ascertaining, control and post experimental ones. The last stage was held in four weeks. The ascertaining stage showed that the level of knowledge was alike at the end of the first term. The control stage determined significant improvements in the use of specific language forms and complex sentences at the end of the second term in the Experimental groups as a result of the "Post-graduate Foreign Language" electronic educational resource usage (Table 1). The students from Experimental groups showed a good progress in all four control tasks in comparison with the ascertained stage. The progress in Control groups was not so considerable. Moreover, the level of responses with the Experimental groups could be considered as a prolonged one after taking a post-experimental test within the period of four weeks.

Table 1: Levels of Specific Vocabulary and Complex Sentences Usage with "Free Talk" Technique

Groups		**Specific language vocabulary and Complex sentences Usage**			
		1	**2**	**3**	**4**
		Inter view	Opinion expression	Story telling	Discourse statement
Control Groups	Ascertaining stage	12.3%	17.7%	25.4%	18.2%
	Control stage	38.8%	38.3%	50.4%	38.9%
Measurement odds		**+26.5%**	**+20.6%**	**+25.0%**	**+20.7%**
Post-experimental stage		26.1%	27.4%	37.6%	27.5%
Experimental Groups	Ascertaining stage	13.4%	18.1%	24.6%	19.3%
	Control stage	64.2%	57.3%	73.8%	58.6%
Measurement odds		**+50.8%**	**+39.2%**	**+49.2%**	**+39.3%**
Post-experimental stage		60.1%	54.3%	70.6%	54.9%

Observations clearly show that students' learning curve proves the idea that IT technologies intensify the process of studying, but should provide not only language accomplishments but active communication with a teacher as well. Firstly, the help is provided only by the teacher with the help of Instructional conversations,

then gradually it is internalized by the students, and at last it is used as an instrument of mediation, which provides good progress in knowledge [8].

The concept of participation is used in this methodology [11]. The "Post-graduate Foreign Language" electronic educational resource was made by using these methods of teaching. The data received prove the idea, that fulfilling different tests and practical exercises only doesn't let students intensify the process of studying. They need active communication with a teacher as well.

5. Conclusion

To access the course, you need to follow the link provided http://edu.kpfu.ru/enrol/index.php?id=1494 on the site of Kazan Federal University. Teaching the discipline solves the following tasks: improvement and further development of speech and language skills and abilities in all types of a foreign language speech activity (reading, speaking, listening, writing). That includes the highly specialized field of foreign languages as well. Development of skills and abilities you need for independent work help post-graduate students (applicants) to improve the level of proficiency in a foreign language in order to use it for scientific and professional activities. They are able to work with the world information resources in a foreign language in the profile of the specialty in order to prepare written (abstracts, annotations, abstracts, articles, motivational presentation) and oral (reports) scientific texts. Technology-Enhanced Learning should include such IT technologies which promote triadic and dynamic interaction rather than dyadic interaction at the same time. Indication language process is made complicated in a dyadic format. Electronic educational resource has a good future ahead of him. And, it is necessary on the part of teachers to benefit from this perspective of educational process upgrading.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A.I. Bashmachnikov, I.A. Bashmachnikov, IT textbooks and educational systems development, Moscow: Philin, 2003.
[2] N. F. Talyzina, Upravlenie protsessom usvoeniya znaniy [Managing the acquisition of knowledge], Moscow State University Press, 1975.
[3] L. S. Vygotsky, Thought and Language, Cambridge MA: MIT Press, 1986.
[4] J.P. Lantolf, Sociocultural Theory and Second Language Learning, Oxford University Press, 2000.
[5] J. Lave, E. Wenger, Situated Learning: Legitimate Peripheral Participation, Cambridge University Press, 1991.
[6] A.R. Luria, Language and Cognition, New York: Wile, 1981.
[7] N. A. Chernova, J.N. Mustafina, "Instructional conversation as a socially-constructed cognitive tool serves second language learning by mediating its own construction, and the construction of knowledge about itself" TURK ONLINE J DES ART COMMUN, **6**, 3114-3117, 2016. https://doi.org/10.7456/1060NVSE/130
[8] W. Horton, K. Horton, Electronic education: instruments and technologies, Moscow: Kudits Publishing House, 2005.
[9] Smith, J., "A seven-minute slice of chaos or I am puzzling through now," Unpublished research report, University of Pittsburgh, 1996.
[10] J. V. Wertsch, Sociocultural Studies of Mind, Cambridge University Press, 1995.
[11] D. Wood, J.S. Bruner, G. Gross, "The role of tutoring in problem-solving" J CHILD PSYCHOL PSYC, **17*(2)***, 89–100, 1976. https://doi.org/10.1111/j.1469-7610. 1976.tb00381.x
[12] P. Zukow-Goldring, K. R. Ferko, "An ecological approach to the emergence of the lexicon: Socializing attention". In V. John-Steiner, C.3. Panofsky, & L. W. Smith (Eds.), Sociocultural approaches to language and literacy: An interactionist perspective, Cambridge University Press, 1994. https://doi.org/10.1017/CBO9780511897047.008

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 660-666 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Correlation-Based Incremental Learning Network for Gas Sensors Drift Compensation Classification

Panida Lorwongtrakool[*1], Phayung Meesad[2]

[1]*Department of Information Technology, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology, North Bangkok, 10800, Thailand*

[2]*Department of Information Technology Management, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology, North Bangkok, 10800, Thailand*

ARTICLE INFO

Article history:
Received: 19 June, 2020
Accepted: 04 August, 2020
Online: 25 August, 2020

Keywords:
Classification
Correlation Distance
Gas Sensor
Incremental Learning
Radial Basis Function

ABSTRACT

A gas sensor array is used for gas analysis to aid in an inspection. The signals from the sensor array are fed into machine learning models for learning and classification. These signals are characterized by time series fluctuating according to the environment or drift. When an unseen pattern is entered, the classification may be incorrect, resulting in decreased model performance. Creating a new model results in the problem of forgetting the old knowledge called Catastrophic Forgetting. Accordingly, this research proposes Correlation-Based Incremental Learning Network (CILN) using the Correlation Distance method to measure similarity and the Gaussian membership function to determine membership of each node. The gas sensor array data is used to verify the proposed algorithm by choosing 16 steady-state features (DR) from 13,910 records which are divided into 6 classes: 1) Ethanol, 2) Ethylene, 3) Ammonia, 4) Acetaldehyde, 5) Acetone, and 6) Toluene. The data are normalized and divided as the training sets into 10%, 20%, 30%, 40%, and 50%, respectively. The proposed algorithm was compared with well-known classifiers. CILN experiment results yield the highest accuracy of 98.96% using 50% of the training data set. It shows that CILN has the incremental learning ability and can be used with data that fluctuate according to the situation.

1. Introduction

The odor inspection, such as food quality check, environmental perfume check, chemical leak check or even weapons or drugs inspection require experts. However, there are limitations of smelling by human nose, and the effectiveness of smelling depends on individual health. Humans may feel tired, and importantly, the human nose is not suitable for smelling various toxins [1].

Many researchers have tried to devise a variety of inspection methods, such as chemical properties or flavor tests, and these methods require direct contact. Another interesting method is to test odors or gases using an electronic nose or array sensor since this method does not require immersion into the sample but inspect the gas response. Currently, the array sensor is extensively used for inspections, such as potential contagious or chemical contamination inspection during the production process, checking the freshness of pork [2], fungus inspection in strawberries [3], evaluating the quality of black tea [4], identify the type of wine and chinese liquor [5], distinguishing wine making techniques, [6] and medical diagnoses [7-8].

Machine learning is a useful tool for data analysis and data learning. Machine learning is divided into 2 approaches: Batch Learning and Incremental Learning. 1) Batch learning refers to a learning method which learn data at once. 2) Incremental learning (also called online learning) refers to learning methods which is applied to streaming collected over time. In this method, learning functions can be updated when new data is entered into the system [9].

In the real world, input data can be dynamic or streaming depending on the situation or environment, as data change or drift. Therefore, if the model is unable to learn new patterns, the performance of the model will be reduce [4]. Finally, all earlier data sets will no longer be available, and a new model must be created when there are new data. This process also leads to a phenomenon known as catastrophic forgetting [10].

[*]Corresponding Author: Panida Lorwongtrakool, panida.l@email.kmutnb.ac.th

Therefore, to improve the algorithm for learning latest information while keeping the old knowledge of earlier data sets, this research proposes Correlation-Based Incremental Learning (CILN). The proposed algorithm can learn new patterns and adapt itself automatically while keeping existing knowledge.

2. Literature Reviews

2.1. Gas Sensor array

The gas sensor array is important to the electronic nose and is used to detect gas molecules from electrical signals. Each sensor has a different sensitivity to gas. When the gas sensor contact with gas molecules, it responds to the gas, forming a spectrum of different gases called odor fingerprint. The response is recorded and sent to the signal processing system for analysis using proper analysis methods to determine the type of gas [11].

2.2. Incremental Learning

In general, creating a model for classification requires a training set for patterns recognition. When the model has an acceptable performance, it will be used for prediction. However, the general limitation of the model is that it will work well for a period only and, inevitably, a new model needs to be created because the algorithm is not created for incremental learning. Therefore, it is extremely sensitive to continuous data in the form of streaming. A new model needs to be created to solve this problem, and it is unable to use the existing knowledge.

In addition, data in the new situation are still a problem, and there is a need to incremental learning capability [12] by gradually learning knowledge without abandoning or forgetting the existing one without retraining. To solve this problem, incremental learning algorithms must be able to combine new knowledge with previously bought knowledge in a way like human learning methods which are based on earlier learning [13].

Therefore, algorithms that can learn from new data without having to access the earlier set of data and support prior knowledge would be a good method of classification to support both static data and data stream, especially when new data samples are added continuously.

1) Characteristics of Incremental learning

- Incremental learning algorithms handle with continuous data and non-stationary distributions.
- It adapts to new data without forgetting the existing knowledge; it does not need to retrain.
- It is compatible with data streams or big data to create machine learning models faster.
- It uses instance windows or instance weighting mechanism without making modifications to the algorithms. New models are calculated based on time periods of using windows or weights by considering new data received.
- All or some parts of the data are used to create an initial model, check for changes in data (using the detection function), and rebuild models as needed based on new data.
- It can automatically change the learning mechanism to increasingly learn new data. For example, the weights of an artificial neural network are adjusted every time there is a new pattern coming into the system.

2) Types of Incremental Learning

- Instance-incremental learning refers to the system that receives data at Step t - the input point $\mathbf{x} \in X^n$, where X represents the input domain in n dimensional space and predicts the output $y \in Y$. The output, Y, can be either continuous in regression or in classification in one dimensional space.
- Batch incremental learning receive batches of data $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K)$ and must specify the label $(y_1, y_2, \dots, y_K)$ for each input point, where K is the number of data points.

Traditional Artificial Neural Networks (ANNs) work well in jobs with static data, where there is no incremental learning needed. Adding new capabilities to ANNs often results in catastrophic forgetting [10]. Therefore, researchers have attempted to solve this problem. In [14] developed a hybrid system including supervised learning and unsupervised learning online by using Fuzzy and Neural Networks together with Euclidean distance.

In addition, learning algorithms have been developed based on Support Vector Machine (SVM) together with Mahalanobis distance which is an elliptical kernel method for multidimensional data. The ability to classify data was compared with the traditional method using Euclidean distance, and no difference was found [15]. However, scattering of data should be considered when adopting the similarity measurement method.

Using threshold is another way to help learning, adjust the model, and support the integrity of earlier knowledge, while only adding a small number of parameters [16]. The important for a creating an incremental learning model is that it must be done quickly by using a small amount of data and gradually adjusting the model according to the new data while keeping old knowledge without access to the initial training data set [17].

However, it was found that the attempt to improve the recognition system for greater accuracy often result in more complex problems. Hence, Incremental Similarity (IS) has been presented, as it yields high accuracy and low complexity. Incremental Similarity is used for incremental online learning. The system can learn from sample data received, and only some parameters need to be updated. It was found that the efficiency of Incremental Similarity was higher than that of the traditional model [18].

At present, deep learning has received great attention, and it has been used in recognition, such as face recognition. However, the face recognition model without incremental learning after training results in problems with new data during the operation. In [19] introduced the Incremental SVM method that allowed the system to update the classification model in real-time, resulting in the increased accuracy of the system and reduced training time.

However, deep learning processing consists of many connected parameters, and it takes quite a lot of time in the training process, as well as retraining if the structure is not enough. Adding incremental learning capabilities to the system with deep learning can be quite difficult. Therefore, researchers have attempted to invent and adjust the Broad Learning System (BLS) [20] which

was created in the form of a flat network where the input is transferred and mapped into feature node. Then, the structure will be expanded broadly to add nodes. The incremental learning algorithm has been developed for rapid change without retraining.

The shallow learning network is more suitable for incremental learning than the deep learning network. It was also found that Radial Basis Function (RBF) is used in incremental learning and is combined with other techniques, such as Neuro-Fuzzy [14]. In addition to RBF, SVM is combined with Mahalanobis distance [15], and Self Organizing Map (SOM) is combined with Euclidean distance [21]. The popular methods used to similarity measure are Euclidean distance and Mahalanobis distance. The Euclidean distance is simple but limited because it is sensitive to scales of variables and suitable for data scattered in circles. The Mahalanobis distance is suitable for elliptically scattered data by considering the covariance matrix to solve the problem of Euclidean distance. Therefore, the similarity measure should be selected based on data scattering. Another interesting similarity measurement method is Correlation distance, as it considers the relationship of variables and is suitable for continuous data. In this research, the researchers proposed incremental learning with Correlation distance, which is a way of measuring similarity to support data that change according to the environment.

2.3. Radial Basis Function

Radial Basis Function is a feed forward neural network with only 1 hidden layer. RBF does not have a complicated structure and is more flexible and faster than Multilayer Perceptron (MLP) architecture, each kernel function is connected to hidden nodes which means the connection to one cluster. Incremental learning by increasing hidden nodes and updating weight and relevant parameters [4]. The norm value between two data points can be calculated by using the general calculation formulas, such as Euclidean distance, Mahalanobis distance, Correlation distance or others as shown in Figure 1.

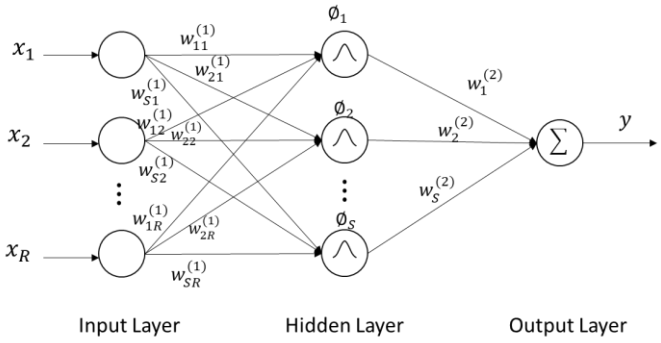


Figure 1: The Architecture of RBF

2.4. Correlation Distance

Correlation distance is a statistical measure used to measure the independence of two values or any two vectors. Correlation distance values are between 0 and 1, which can be measured by the variance or standard deviation. It can be calculated according to (1).

$$d_c(\mathbf{p},\mathbf{q}) = 1 - \frac{cov(\mathbf{p},\mathbf{q})}{(std_{\mathbf{p}})(std_{\mathbf{q}})} \quad (1)$$

$$cov(\mathbf{p},\mathbf{q}) = \sum_{j=1}^{n}(\mathbf{p}_j - \bar{\mathbf{p}})(\mathbf{q}_j - \bar{\mathbf{q}}) \quad (2)$$

$$std(\mathbf{p}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\mathbf{p}_j - \bar{\mathbf{p}})^2} \quad (3)$$

$$\bar{\mathbf{p}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{p}_j \quad (4)$$

where $d_c(\mathbf{p},\mathbf{q})$ is correlation distance from **p** to **q**; **p** and **q** are any data points and n is number of dimension.

3. Method

The objective of this research is to propose the Correlation-Based Incremental Learning Network by using the Correlation Distance and membership function. The operational structure of CILN, as shown in Figure 2.

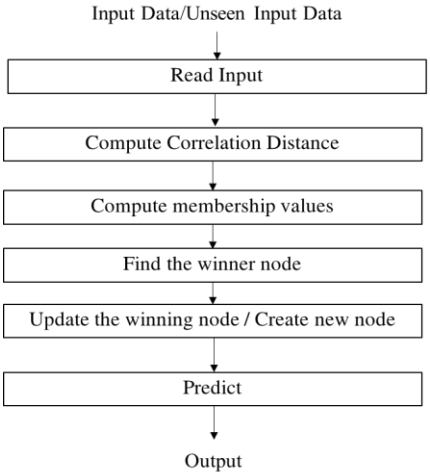


Figure 2: Operational Structure of CILN

Step 1) Set the membership threshold parameter (m_{th}) for determining new neuron, where m_{th} is a value between 0 and 1.

Step 2) Read input data

- Read in a pair of an input pattern **p** and target *t*.
- If new knowledge is found set a new neuron $\mathbf{W}_P$ using **p** and $\mathbf{W}_T$ according target *t* of the input pattern in $\mathbf{W}_P$.

Step 3) Read in the next input pattern and its corresponding target, if any.

Step 4) Measure the Correlation distance between the input **p** and the prototype $\mathbf{W}_P$, which are centroids using (5).

$$d_i(\mathbf{p},\mathbf{W}_{Pi}) = 1 - \frac{cov(\mathbf{p}-\mathbf{W}_{Pi})}{(std_{\mathbf{p}})(std_{\mathbf{W}_{Pi}})} \quad (5)$$

Step 5) Compute membership values of each node using the Gaussian-type radial basis function.

$$m_i = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(d(\mathbf{p}-\mathbf{W}_{Pi}))^2}{2\sigma^2}};\ i = 1,2,\ldots,C \quad (6)$$

where m_i is membership value, and σ is Standard Deviation in cluster i. σ will be used to show the scatter of the data in the cluster. σ value between 0.001 and 0.05 and after the patterns near the prototype are included in the same prototype, the standard deviation is updated accordingly.

Step 6) Find the winner node that has highest membership value [14].

$$\text{winner} = \arg max_i\,(m_i);\ i = 1, 2, \ldots, C \tag{7}$$

$$J = winner\ index = \arg\ max_i(m_i) \tag{8}$$

Step 7) Update the Winning node [14]

If winner $> m_{th}$, the instance is similarity to the winning node, then update weight of the winning node.

$$\mathbf{W}_{\mathrm{P_{J,new}}} = \frac{\mathbf{W}_{\mathrm{P_{J,old}}}(\mathrm{C_{J,new}}-1) + \mathbf{p}}{\mathrm{C_{J,new}}} \tag{9}$$

$$\boldsymbol{\sigma}_{\mathrm{J,new}} = \begin{cases} \sqrt{\left(1 - \frac{1}{\mathrm{C_{J,new}}}\right)\boldsymbol{\sigma}^2_{\mathrm{J,old}} + \left(\frac{(\boldsymbol{\sigma}_{\mathrm{J,old}} - \mathbf{p})^2}{\mathrm{C_{J,new}}}\right)} & \text{if } \mathrm{C_{J,new}} > 1, \\ \boldsymbol{\sigma}_0 & \text{otherwise;} \end{cases} \tag{10}$$

where $\mathbf{W}_{\mathrm{P_{J,new}}}$ is new weight, $\mathbf{W}_{\mathrm{P_{J,old}}}$ is original weight, $\mathbf{p}$ is latest input data , and $\mathrm{C_J}$ is number of members in the cluster.

If winner $< m_{th}$, the instance is considered node a member of the winning node; then a new node $\mathbf{W}_{\mathrm{T}}$ is created.

Step8) If in prediction mode, i.e. no target, assign the predicted class to the unseen pattern.

$$y = \mathbf{W}_{\mathrm{T_J}} \tag{11}$$

where y is predicted class, $\mathbf{W}_{\mathrm{T_J}}$ is target of the Winner.

If in training mode, i.e. there is target of the input $\mathbf{p}$, compute error.

$$e = t - y \tag{12}$$

where e is error, t is target, and y is output.

Step 9) Continue process to step 3, until stop condition is met.

4. Experimental Results

4.1. Experimental Data

The data set used in the experiment is the gas sensor data set obtained from the UCI Machine Learning Repository [22]. This data set includes measurements using 16 chemical sensors to measure the gas response at different concentrations. The data set consists of 16 attributes with a total of 13,910 records. This data set has been collected for 36 months from January 2008 to February 2011. The data set was collected from 6 types of pure gas: 1) Ethanol, 2) Ethylene, 3) Ammonia, 4) Acetaldehyde, 5) Acetone, and 6) Toluene. Each gas has added characteristics extracted with different values, making each sensor have 8 features. Therefore, the data set consists of 128 features and is divided into 10 batches in time sequence. The details are shown in Table 1.

The response of the said sensor is read in the form of resistance. Each measurement creates a 16-channel time series data set that responds to the chemicals being measured. In the experiment, the steady-state feature (DR), was selected which means the maximum resistance change compared to the base line.

The signals from 16 sensors showed the characteristics of data consisting of Multivariate Time-series. Since the data in each batch were collected several times, and each batch has a different number of classes (Imbalance class). Therefore, the data obtained are at different concentrations as shown in Figure 3.

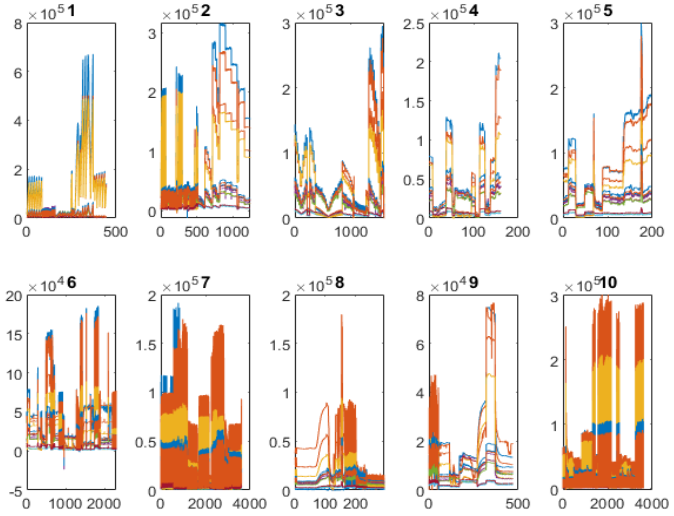


Figure 3: Gas Sensor Array Drift at Different Concentrations Data Set

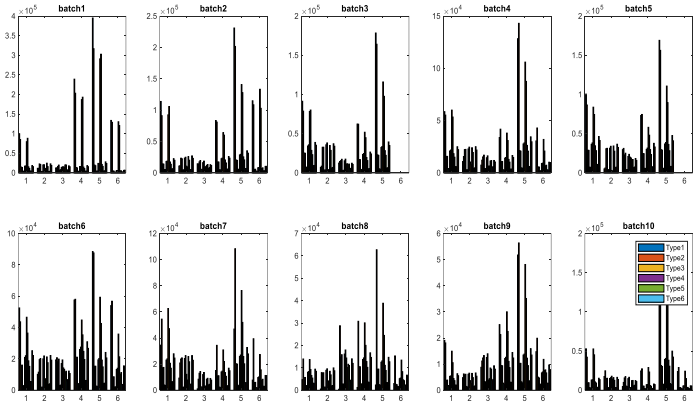


Figure 4: The Odor's Fingerprint of Gas Sensor Drift Compensation Data Set

Table 1: Experimental data

Batch ID	Month	Acetone	Acetaldehyde	Ethanol	Ethylene	Ammonia	Toluene	Total
Batch 1	1, 2	90	98	83	30	70	74	445
Batch 2	3-10	164	334	100	109	532	5	1,244
Batch 3	11, 12, 13	365	490	216	240	275	0	1,586
Batch 4	14, 15	64	43	12	30	12	0	161
Batch 5	16	28	40	20	46	63	0	1997
Batch 6	17, 18, 19, 20	514	574	110	29	606	467	2,300
Batch 7	21	649	662	360	744	630	568	3,613
Batch 8	22, 23	30	30	40	33	143	18	294
Batch 9	24, 30	61	55	100	75	78	101	470
Batch 10	36	600	600	600	600	600	600	3,600

The characteristic response spectrum is called odor's fingerprint. Therefore, according to the characteristics of response spectrum, different odors can be distinguished showing the fingerprint of each gas type at various times of each batch, as shown in Figure 4.

4.2. Experimental Setting

To evaluate the effectiveness of the proposed CILN, we selected only 16 steady-state (DR) features that have not been extracted from a total of 128 features. Therefore, the data set has 13,910 records, divided into 6 classes: 1) Ethanol, 2) Ethylene, 3) Ammonia, 4) Acetaldehyde, 5) Acetone, and 6) Toluene. The data were normalized and split into training sets and test sets. The training sets were randomly selected for 10%, 20%, 30%, 40%, and 50%; the rest of the data were used for test sets. The proposed CILN algorithm, which uses Correlation distance measurement, was compared with Euclidean distance and Mahalanobis distance measurement, as well as other well-known classifiers, including NaiveBayes, BayesNet, RBF, SVM, MLP, and Simple Logistics. Comparison of experimental results by Accuracy, Precision, Recall, F-Measure. The experiment steps are shown in Figure 5.

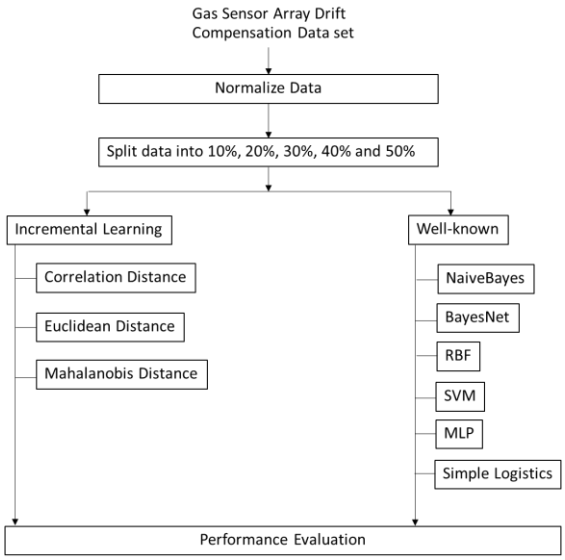


Figure 5: Experimental Steps

4.3. Experimental Results

The experimental results under certain settings show the comparison of the efficiency of the proposed CILN algorithm which uses Correlation distance measurement method and the incremental learning algorithm which uses Euclidean distance and Mahalanobis distance measurement methods. According to Table 2, CILN yielded top 5 accuracy scores: 98.96%, 98.74%, 98.51%, 97.87% and 96.08% by using 50%, 40%, 30%, 20%, and 10% of the training sets, respectively. Considering all the 3 measurement methods, Correlation distance method gave the highest mean of 98.03%, followed by Euclidean distance method of 93.16% and Mahalanobis distance method of 93.10%. Euclidean distance and Mahalanobis distance methods supply similar accuracy, while Correlation distance method gives a high accuracy. Therefore, the similarity measurement method affects accuracy, and the proposed method is suitable for the gas sensor array drift at different concentrations data set which contains continuous data.

Table 2: Comparison the performance of incremental learning using different similarity measurement methods.

Similarity Measure	Training (%)	Accuracy (%)	Precision	Recall	F-Measure
Correlation distance	10	96.08	0.85	0.96	0.90
	20	97.87	0.90	0.99	0.94
	30	98.51	0.94	0.98	0.96
	40	98.74	0.94	0.99	0.97
	50	98.96	0.95	0.99	0.97
	Average	98.03	0.92	0.98	0.95
Euclidean distance	10	90.13	0.69	0.84	0.76
	20	92.44	0.74	0.91	0.82
	30	93.75	0.77	0.93	0.85
	40	93.82	0.79	0.91	0.84
	50	95.64	0.84	0.94	0.89
	Average	93.16	0.77	0.91	0.83
Mahalanobis distance	10	87.86	0.63	0.85	0.72
	20	93.22	0.75	0.95	0.84

Similarity Measure	Training (%)	Accuracy (%)	Precision	Recall	F-Measure
	30	94.38	0.80	0.94	0.86
	40	94.36	0.79	0.94	0.86
	50	95.64	0.85	0.94	0.89
	Average	93.10	0.76	0.92	0.83

Table 3: Comparison of Performance with well-known classifiers

Algorithm	Training (%)	Accuracy (%)	Precision	Recall	F-Measure
CILN	10	96.08	0.85	0.96	0.90
	20	97.87	0.90	0.99	0.94
	30	98.51	0.94	0.98	0.96
	40	98.74	0.94	0.99	0.97
	50	98.96	0.95	0.99	0.97
Simple Logistics	10	93.52	0.94	0.94	0.94
	20	93.66	0.94	0.94	0.94
	30	95.48	0.96	0.96	0.96
	40	95.70	0.96	0.96	0.96
	50	96.05	0.96	0.96	0.96
MLP	10	91.37	0.92	0.91	0.91
	20	93.37	0.94	0.93	0.94
	30	94.64	0.95	0.95	0.95
	40	94.42	0.95	0.94	0.95
	50	95.64	0.96	0.96	0.96
SVM	10	70.71	0.80	0.71	0.71
	20	76.23	0.85	0.76	0.78
	30	79.48	0.87	0.80	0.81
	40	82.19	0.88	0.82	0.84
	50	83.89	0.89	0.84	0.85
BayesNet	10	59.41	0.65	0.59	0.60
	20	65.05	0.69	0.65	0.66
	30	65.27	0.69	0.65	0.66
	40	66.30	0.70	0.66	0.67
	50	66.48	0.70	0.67	0.67
RBF	10	64.29	0.67	0.64	0.64
	20	61.94	0.65	0.62	0.60
	30	63.09	0.67	0.63	0.61
	40	62.73	0.66	0.63	0.61
	50	62.78	0.66	0.63	0.61
NaiveBayes	10	51.56	0.60	0.52	0.52
	20	51.24	0.60	0.51	0.52
	30	51.43	0.61	0.51	0.52
	40	51.72	0.61	0.52	0.53
	50	51.45	0.60	0.52	0.52

Table 3 shows the comparison of the performance of the proposed CILN algorithm with the well-known classifiers, such as NaiveBayes, BayesNet, RBF, SVM, MLP, and Simple Logistics. Overall, CILN still yielded the highest accuracy of 98.96%, Simple Logistics of 96.05%, MLP of 95.64%, SVM of 83.89%, BayesNet of 66.48%, RBF of 64.29%, and NaiveBayes of 51.72% by using 50%, 50%, 50%, 50%, 50%, 10%, and 40% of the training sets, respectively. It was found that the number of training data sets can affect the accuracy. However, new data entered the classifiers may change according to the environment. Without incremental learning, it will also affect accuracy. Therefore, the algorithm should have incremental learning ability for effective classification.

5. Conclusions and Future Work

Machine learning is a useful tool for analysis and learning. However, data imported into the system are both static data and dynamic data which fluctuate and change according to the environment. Conventional algorithms still lacks incremental learning new data. When new data are added, the algorithm will not be able to learn and adapt, resulting in reduced efficiency. Moreover, all the trained data sets are no longer available, and a new model needs to be created for new data. This process leads to a phenomenon known as catastrophic forgetting, resulting in poor classification performance. Therefore, to solve this problem, this research has proposed Correlation-Based Incremental Learning algorithm that allows the model to learn and improve automatically while maintaining old knowledge using the Correlation distance to measure similarities and membership functions using Gaussian-type Radial Basis Function to determine membership of each node.

Gas sensor data from the UCI machine learning repositories used to evaluate the performance of the proposed algorithm. This data set holds 13,910 measurements from 16 chemical sensors exposed to 6 gases at different concentration levels. This data set has been collected for 36 months from January 2008 to February 2011. The data were collected from 6 types of pure gas: 1) Ethanol, 2) Ethylene, 3) Ammonia, 4) Acetaldehyde, 5) Acetone, and 6) Toluene. In the experiment, only 16 steady-state features (DR) were chosen. The data were normalized and split into training sets and test sets. The training sets were randomly selected for 10%, 20%, 30%, 40%, and 50%; the rest of the data were used for test sets.

The results show that CILN allows the system to learn new patterns while maintaining the old knowledge. The proposed CILN algorithm supplies an initial accuracy of 96.08% by using only the 10% of the training data set which is higher than all classifiers. In addition, CILN yields the highest accuracy of 98.96% when 50% of the training data set was used. It shows that CILN can learn from a small sample size and can adapt and learn new data automatically while keeping the existing knowledge. Therefore, CILN can increase the accuracy of classification and support the time series data which are dynamic data and can be used for environmental or other inspections. Moreover, it was found that using only 16 steady-state features (DR) was sufficient for gas classification without additional feature extraction. In the future work, we will consider reducing the dimensions of data by selecting features, removing noise, and selecting the proper signal range.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to gratefully acknowledge the King Mongkut's University of Technology North Bangkok and Rajamangala University of Technology Isan.

References

[1] A. Loutfi, S. Coradeschi, G. K. Mani, P. Shankar, J.B.B. Rayappan, "Electronic noses for food quality: A review" Journal of Food Engineering, **144**, 103-111, 2015. http://dx.doi.org/10.1016/j.jfoodeng.2014.07.019

[2] L. Huang, J. Zhao, Q. Chen, Y. Zhang, "Nondestructive measurement of total volatile basic nitrogen (TVB-N) in pork meat by integrating near infrared spectroscopy" computer vision and electronic nose techniques Food Chemistry, **145**, 228-236, 2014. http://dx.doi.org/10.1016/j.foodchem.2013.06.073

[3] L. Pan, W. Zhang, N. Zhu, S. Mao, K. Tu, "Early detection and classification of pathogenic fungal disease in post-harvest strawberry fruit by electronic nose and gas chromatography–mass spectrometry" Food Research International, **62**, 162-168, 2014. http://dx.doi.org/10.1016/j.foodres.2014.02.020

[4] B. Tudu, A. Jana, A. Metla, D. Glosh, N. Bhattacharyya, "Electronic nose for black tea quality evaluation by an incremental RBF network" Sensors and Actuators B: Chemical, **138**(1), 90-95, 2009. http://dx.doi.org/10.1016/j.snb.2009.02.025

[5] Y. Yang, H. Liu, Y. Gu, "A Model Transfer Learning Framework With Back-Propagation Neural Network for Wine and Chinese Liquor Detection by Electronic Nose" IEEE Access, **8**, 105278-105285, 2020. https//doi.org/10.1109/ACCESS.2020.2999591

[6] M. Aleixandre, J. P. Santos, I. Sayago, M. C. Horrillo, J. M. Cabellos, T. Arroyo, "Use of an electronic nose as a tool to differentiate winemaking techniques" in 2015 10th Spanish Conference on Electron Devices (CDE), Madrid, Spain, 2015. https//doi.org/10.1109/CDE.2015.7087490

[7] B. Schmekel, F. Winquist, A. Vikström, "Analysis of breath samples for lung cancer survival" Analytica Chimica Acta, **840**, 82-86, 2014. https://doi.org/10.1016/j.aca.2014.05.034

[8] E. Westenbrink, R.P. Arasaradnam, N. O'Connell, C. Bailey, C. Nwokolo, K.D. Bardhan, J.A. Covington, "Development and application of a new electronic nose instrument for the detection of colorectal cancer" Biosensors and Bioelectronics, **67**, 733-738. 2015. http://dx.doi.org/10.1016/j.bios.2014.10.044

[9] Q. Wu, D. Snell, "An Introduction to Incremental Learning" Predictive Analytics and Futurism, **13**, 18-19, 2016.

[10] R. M. French, "Catastrophic forgetting in connectionist networks" Trends in Cognitive Sciences, **3**(4), 128-135, 1999. http://dx.doi.org/10.1016/S1364-6613(99)01294-2

[11] Z. Zheng, X. Lin, "Study on Application of Medical Diagnosis by Electronic Nose" World Science and Technology, 14(6), 2115-2119. 2012. http://dx.doi.org/10.1016/S1876-3553(13)60016-2

[12] V. Losing, B. Hammer, H. Wersing, "Incremental on-line learning: A review and comparison of state of the art algorithms" Neurocomputing, **275**, 1261-1274, 2018. https://doi.org/10.1016/j.neucom.2017.06.084

[13] S. Wan, and L.E. Banta, "Parameter Incremental Learning Algorithm for Neural Networks" IEEE Transactions on Neural Networks, 17(6), 1424-1438, 2006. https://doi.org/10.1016/j.neucom.2017.06.084

[14] G. G. Yen, P. Meesad, "An effective neuro-fuzzy paradigm for machinery condition health monitoring" IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **31**(4), 523-536. 2001. https://doi.org/10.1109/3477.938258

[15] H. O. Myint, P. Meesad, "Incremental Learning Algorithm based on Support Vector Machine with Mahalanobis distance (ISVMM) for intrusion prevention" in 2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. Pattaya, Chonburi, Thailand, 2009. https://doi.org/10.1109/ECTICON.2009.5137129

[16] Y. Pang, J. Deng, Y. Yuan, "Incremental threshold learning for classifier selection" Neurocomputing, **89**, 89-95, 2012. https://doi.org/10.1016/j.neucom.2012.01.012

[17] K. Diaz-Chito, K. Georgouli, A. Koidis, J. M. del Rincon, "Incremental model learning for spectroscopy-based food analysis" Chemometrics and Intelligent Laboratory Systems. **167**, 123-131, 2017. https://doi.org/10.1016/j.chemolab.2017.06.002

[18] M. Režnáková, L. Tencer, M. Cheriet, "Incremental Similarity for real-time on-line incremental learning systems" Pattern Recognition Letters **74**, 61-67, 2016. https://doi.org/10.1016/j.patrec.2016.01.010

[19] L. Li, Z. Jun, J. Fei, S. Li, "An incremental face recognition system based on deep learning" in 2017 15th IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 2017. https://doi.org/10.23919/MVA.2017.7986845

[20] C. L. P. Chen and Z. Liu, "Broad Learning System: An Effective and Efficient Incremental Learning System Without the Need for Deep Architecture" IEEE Transactions on Neural Networks and Learning Systems, **29**(1), 10-24, 2018. https://doi.org/10.1109/TNNLS.2017.2716952

[21] L. Tian and W. Liu, "Incremental intrusion detecting method based on SOM/RBF" in 2010 International Conference on Machine Learning and Cybernetics. Qingdao, China, 2010. https://doi.org/10.1109/ICMLC.2010.5580770

[22] A. Vergaraa, S. Vembu, T. Ayhan, M.A. Ryan, M.L. Homer, R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles" Sensors and Actuators B: Chemical, **166-167**, 320-329, 2012. https://doi.org/10.1016/j.snb.2012.01.074

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 664-675 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Improved System Based on ANFIS for Determining the Degree of Polymerization

Marcel Nicola[1,*], Marian Duță[1], Maria-Cristina Nițu[1], Ancuța-Mihaela Aciu[1], Claudiu-Ionel Nicola[1,2]

[1]*Research and Development Department, National Institute for Research, Development and Testing in Electrical Engineering – ICMET, Craiova, 200746, Romania*

[2]*Department of Automatic Control and Electronics, University of Craiova, Craiova, 200585, Romania*

ARTICLE INFO

Article history:
Received: 20 August, 2020
Accepted: 06 November, 2020
Online: 24 November, 2020

Keywords:
Degree of polymerization
2-furfuraldehyde
Fuzzy logic
ANFIS
Neural network

ABSTRACT

Transformer health assessment techniques, based on applicable standards, such as the dissolved gas analysis (DGA), through laboratory testing or online monitoring are used to analyze the symptoms of a failure which develops in transformers from an early stage. The DGA from a sample of dielectric liquid taken from the main tank generates information on the state of degradation of the active part. It was found that of all the furan derivatives which result from the degradation of insulation, 2-furfuraldehyde (2-FAL) is the only derivative which dissipates in large quantities in oil. Because of this, and due to its thermal stability as compared to other derivatives, 2-FAL is the best unit of measurement to determine and monitor the degree of polymerization (DP) of insulation. The poor accessibility of paper samples has led to difficulties in testing the ageing state of paper directly by measuring the tensile strength and the DP. It was developed methods to indirectly assess the ageing state of paper, by means of the chemical markers in oil which are associated with paper ageing. This article presents a method to determine the DP of solid insulation in transformers, which provides a faster and more accurate interpretation, as compared to the classical ones. The 2-FAL data resulting from the lab are recorded in a MySQL database, which is embedded in an intelligent system for diagnosis of DP (ISDDP) based on Adaptive Neuro-Fuzzy Inference System (ANFIS), generating at the output the report with the interpretation of the faults as word files. The automation of the DP diagnostic process will be achieved, allowing the operator to make timely decisions, to avoid any possible damages.

1. Introduction

Transformers are expensive pieces of equipment that play a significant role in electricity transmission and distribution systems. Although transformers are generally robust pieces of equipment, they also undergo failures, and there are numerous degradation mechanisms operating in components and subsystems that will limit the useful life. It is important that transformer users and distribution system administrators are properly trained to assess the condition of a group of transformers in operation, in order to take the best decisions at critical moments of possible damages, for the operations of minor faults repairing, their rehabilitation or replacement [1-4].

The assessment of the condition of the operating transformer consists in identifying the indicators that can provide accurate data on the extent of degradation of transformer components and subcomponents. The detailed health assessment of a power transformer requires extensive testing (electrical, as well as physico-chemical tests of the dielectric fluid), inspection of dataset from acquisition of other relevant data (operational history, as well as maintenance, including data of putting into service). These data are then analyzed by an expert based on the relevant standards and expert knowledge [5-8].

Most common transformer health assessment techniques apply only to the active part of the transformer. The DGA from a sample of dielectric liquid taken from the main tank generates information on the state of degradation of the active part. However, all components and subcomponents of a transformer should be considered for a full assessment. Unexpected failures of the

* Corresponding Author: Marcel Nicola, ICMET Craiova, Romania, marcel_nicola@yahoo.com

bushings and the on-load tap-changer (OLTC) can often lead to a state of transformer failure or even to its removal from the system [9-17].

The assessment of the active part of the transformer (magnetic core and windings) can be performed as follows [6], [10-16]:

- The thermal assessment concerns the faster-than-anticipated aging due to excessive temperatures resulting from the hot spot of the winding, the hot spot in the core and the defects of paper-insulated hot metals;
- The dielectric assessment tracks the process of degradation of the paper-oil insulation system due to high thermal or electrical stress, oil-paper moisture, manufacturing defects and contamination with (semi) conducting particles (solid particles);
- The mechanical assessment concerns the disintegration of the structure in the core, the windings and the solid insulation, due to one of the many reasons, such as strong electromagnetic forces resulting from the flow of the short-circuit current through the windings; the mechanical deformation of the windings due to shocks during transport, installation or during system failures, etc; and the degradation of solid insulating materials.

The oil plays a key role as an insulating medium that acts as a barrier between the windings and the tank and acts as a "cooling agent" that dissipates the heat generated in the windings and the core. It is a conservation medium which protects the windings and the solid insulation against moisture and other physical contamination.

The molecular structure of oil is subjected to electrical, mechanical and thermal stress during transformer operation, which leads to a process of decomposition, and, as a result hydrocarbon gases are generated. These gases are found in the mass of the dielectric fluid and can be quantized.

Transformer health assessment techniques, based on applicable standards, such as the DGA, through laboratory testing or online monitoring are used to analyze the symptoms of a failure which develops in transformers from an early stage [12-14].

The complexity of real world problems, the need to develop complex software systems that adapt to a dynamic environment such as the real world lead to growing interest in the use of computational intelligence techniques to develop specific applications. Computational intelligence is a subdomain of the artificial intelligence that deals with the design, application and development of computational paradigms inspired for example by biological, social and linguistic areas. The techniques covered by the computational intelligence are the fuzzy logic, the artificial neural networks and the evolutionary computation. These are mainly fuzzy systems, neural networks, cognitive systems, nature-inspired systems and hybrid intelligent systems containing these paradigms. The most common of these techniques used for the diagnostic of power transformers are the fuzzy logic and the artificial neural networks [18-27].

This article presents a method to determine the DP of solid insulation in transformers, which provides a faster and more accurate interpretation, as compared to the classical methods. The 2-FAL data resulting from the lab are recorded in a MySQL database through a client software application for writing to the database and then read through a database read client, which is embedded in the diagnostic system, generating at the output the report with the interpretation of the faults as word files. Thus, the automation of the DP diagnostic process will be achieved, allowing the operator to make timely decisions, to avoid any possible damages.

The proposed ISDDP to determine the DP of the insulation of power transformer windings based on the content of the 2-FAL derivative uses the fuzzy logic, and, for the improvement of the system, the fault learning facility is used, based on the artificial neural networks.

This article is an extension of [28] and is structured as follows: Section 2 presents the chemical markers used to determine the DP. The classical methods for the determination of the DP are presented in Section 3. Section 4 presents the implementation as fuzzy logic of the diagnostic system in the LabVIEW graphical programming environment, as well as its improvement by using artificial neural networks through an adaptive neuro-fuzzy inference system. Section 5 presents conclusions and new pursuits in future work.

2. Chemical markers used for the determination of the DP

When paper insulation ages, its electrical properties do not change significantly, but the mechanical properties are considerably diminished, due to cellulose degradation (depolymerization). During normal operation, paper cellulose will split according to the general condition of the transformer (temperature, moisture and acids). As a result, the tensile strength of the paper is reduced over time, and therefore it is important to monitor the condition of the insulation [5-8].

The poor accessibility of paper samples has led to difficulties in testing the ageing state of paper directly by measuring the tensile strength and the DP. Thus, the researchers developed methods to indirectly assess the ageing state of paper, by means of the chemical markers in oil which are associated with paper ageing.

2.1. Carbon oxides (CO and CO2)

The first studies covered the relation between the amount of carbon oxides, carbon monoxide (CO) and carbon dioxides (CO2), in oil and the DP of the insulating paper. Carbon oxides, CO2 and CO, are determined based on the DGA by applying the gas chromatography technique. As a result of thermal stresses, these gases are generated by the degradation of both paper and other parts of the transformer. In free breathing transformers, the generation of carbon oxides is influenced by free air ingress [12].

In [21], it is presented a relationship between the amount of carbon oxides, CO and CO2, dissolved in transformer oil and the DP of the insulating paper. In [17] is demonstrated that water and CO2 are the main byproducts of the thermal degradation of cellulose. Therefore, the CO2/CO ratio is normally utilized as an indicator of the thermal decomposition of cellulose [8, 9].

According to [18], the CO2/CO ratio is normally when its value is above seven, while the respective values of CO2 and CO must be greater than 5000 ppm and 500 ppm to improve the certainty factor.

In accordance with [19], the CO2/CO ratio below 3 is expected to determine a severe degradation of the paper. When the ratio

exceeds 10, it may indicate a slight overheat of 160°C. It was noted that in order to obtain safe results, correction of CO2 absorption from the atmosphere and correction of CO2 and CO background are necessary.

According to [17], high paper degradation rates are reported when the CO2/CO ratio drops below 6. However, the application of this ratio in determining the condition of the cellulose insulation is not suitable due to the long-term oxidation of the oil or as a result of free air ingress due to leakage [24]. The following additional tests are proposed, as solutions to solve this problem: the analysis of furan compounds; the determination of the DP of paper using the paper samples, to analyze the cellulose insulation.

2.2. Furan compounds

Oxidation, hydrolysis and pyrolysis of cellulose generate a group of chemical compounds which are used as a tool to analyze the condition of the cellulose insulation. Due to their solubility in oil, these chemical compounds (2-furfural; 2-acetylfuran; 5-methyl-2-furfural; 2-furfuryl alcohol; 5-hydroxymethyl-2-furfural) provide key information necessary for the diagnosis of the health of solid insulation. The limitations and applicability of the furan compounds as tools for diagnosing the state of the degradation of cellulose insulation in oil-filled transformers were constant in [6] and [7]. Studies have shown that of the five furan compounds specified in [9], 2-FAL is the most stable, by providing relevant information on the state of degradation of the solid insulation.

Their main conclusion indicates that, of the five furan compounds identified in [9], 2-FAL provides more relevant information on the paper degradation. 2-FAL has been shown to have a high degree of generation and better stability than the other furan compounds and can be extracted from oil and used to characterize the thermal decomposition of the electroinsulating paper. A high concentration of 2-FAL indicates a high level of paper degradation. The constructive differences between transformers, the thermal and electrical stresses limit the setting of typical threshold values for 2-FAL, thus hindering the diagnosis of the condition of the cellulose insulation [16], [17], [20].

Some noted disadvantages of 2-FAL, such as its low detection threshold, when the paper is thermally upgraded (TUK). TUK paper is frequently used for new transformers, has a high rate of 2-FAL generation from hemicellulose, which shows an exponential increase, so the applicability of 2-FAL as a chemical marker for paper degradation must be considered prospectively.

3. Classical methods for the determination of the DP

Cellulose insulating paper can be damaged due to the effects of hydrolysis, pyrolysis and oxidation that occur in the transformer during operation. Due to the deterioration of the cellulose insulation, the group of furan compounds appears [6].

The cellulose insulation in transformers consists of cellulose, hemicellulose, lignin and mineral substances, and their degradation results in furans (see Figure 1) [9]. The generation of furans can also be caused by pentoses. As a common point of the degradation processes, the final result is the 2-FAL [17], [20].

Following the degradation processes of solid insulation that take place in the transformer, 2-FAL has the ability to dissipate in oil and due to its stability is an analysis tool for determining and monitoring DP of cellulose insulation. The other furan compounds can be used as markers to identify the process that causes the degradation of solid insulation [17], [22].

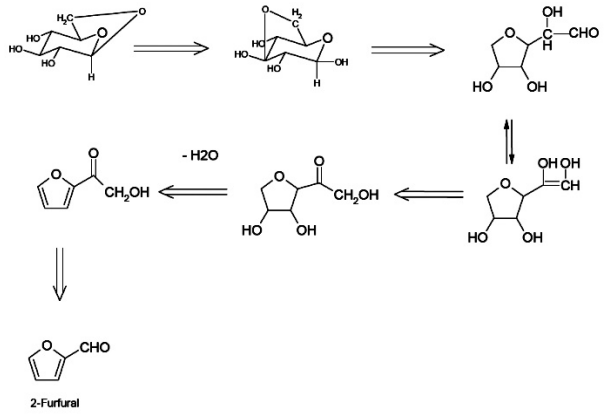


Figure 1: Determination of 2-FAL from lignin

The determination of the content of furan compounds, especially 2-FAL, was given a certain use for the diagnostic of transformers. It is believed that there is a certain correlation between the degree of polymerization of paper and the content of 2-FAL (or total furan compounds) of oil. Several researchers [25] have developed various relations for the determination of the DP, depending on 2-FAL concentration, but Pahlanvanpour's relation [15] is the most commonly used, because, when developing the formula, he relied on the states of the transformer and an assumption that 20% of the winding paper and the inner layer of the paper degrades twice as fast as the rest of the paper.

In [17] is builted a degradation model based on experimental data and field measurements in a research program conducted for CIGRE (International Council on Large Electric Systems) and found that a furfural molecule will result for every three cellulose chain splits [9], [24]. The system proposed in this paper was designed to determine the DP based on the 2-FAL ranges proposed in [6, 7] as is presented in Table 1.

Table 1: The state of the transformer according to the furan content

2-FAL content	DDP
0-0.1	1200-700
0.1-1	700-450
1-10	450-250
>10	< 250

To determine the DP, the following mathematical equations were used, which relate the DP to the content of 2-FAL:

- Chendong equation [16], [26]:

$$DP = \frac{1.51 - \log_{10} F}{0.0035} \quad (1)$$

- Scholinik et al. equation [18] , [26]:

$$DP = \frac{1.17 - \log_{10} F}{0.00288} \quad (2)$$

- Pahlavanpour equation [15] , [26]:

$$DP = \frac{800}{(0.186 \cdot F) + 1} \quad (3)$$

- De Pablo equation [17] , [26]:

$$DP = \frac{7100}{8.88 + F} \quad (4)$$

where: F represents 2-FAL concentration.

4. The architecture of ISDDP

The fuzzy logic-based approach for solving the control problems is highly recommended in very complex, nonlinear systems with uncertainties for internal or input parameters. The fuzzy logic is an approximate method by which vague knowledge, stored in a base of rules can be formally modeled. The practical implementation of the fuzzy systems is based on the advantages which allow the human experience to be transposed for the creation of the inference rules by using linguistic variables. The efficiency of the fuzzy systems or the artificial neural networks work is demonstrated by the widespread use in various fields in recent years [12], [26-30].

The fuzzy control provides a method of designing control algorithms in a user-friendly manner and provides the ability to capture human non-linear control behavior that has been shown to be suitable in many complex adjustment problems. A controller design method close to human thinking and perception reduces the time of application development and requires less trained staff in the field of regulator design. The advantage of the fuzzy tuning is the ability to program the knowledge of process operators and engineers, consisting in process operation experience, in a manner which is easy to understand and friendly. The fuzzy systems mainly have the advantage of explaining behaviors based on a set of rules and therefore their performance and development can be adjusted by modifying them [28], [30].

The neural networks are utilized to adapt the membership functions of the fuzzy systems in the applications of automatic control of the dynamic processes. Although the fuzzy logic allows the coding of expert knowledge through linguistic terms, usually it is a difficult endeavor to design and adjust the membership functions which quantitatively describe these terms. The learning techniques of the neural networks can automate the process and can therefore substantially reduce the cost and the time for the application development, leading to increased performance. To prevent the problem of knowledge acquiring, the neural networks can be designed to automatically extract the set of fuzzy rules from the disponible numerical data. Other approaches involve the use of neural networks to optimize a number of certain parameters of the fuzzy systems or to preprocess input data into the fuzzy systems [29], [30].

The general scheme for the proposed system for diagnosis of DP is shown in Figure 2. The system is achieved by using the facilities provided by LabVIEW and Matlab/Simulink graphic development environments for the implementation based on the fuzzy logic and neural networks.

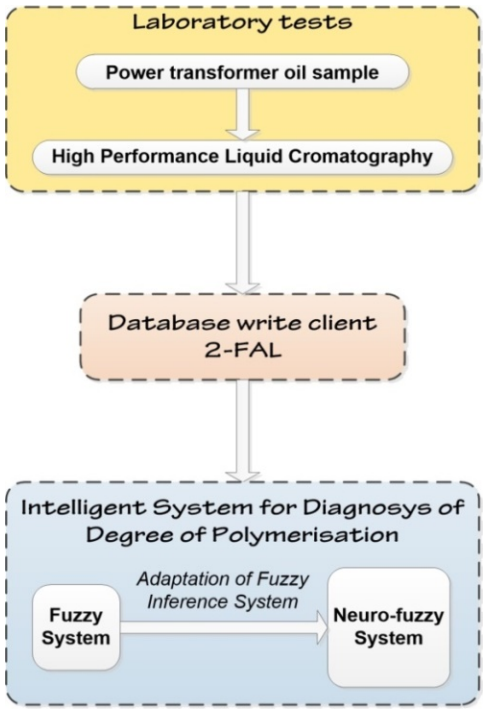


Figure 2: General scheme for the proposed system for diagnosis of DP

4.1. Fuzzy system for the determination of the DP

The system for the diagnostic of the DP of power transformers based on the fuzzy logic generates flexibility in interpreting the fault condition based on the 2-FAL data sets. The fuzzy logic diagnostic system is implemented in the Fuzzy System Designer, which is a software tool from LabVIEW graphical programming environment and basic model of the proposed fuzzy logic controller is shown in Figure 3. To overcome the difficulties in determining the boundary conditions for furan derivatives, the fuzzy logic technique is used, which allows the based rules to be implemented in a type of structure with natural language appropiate and widely used.

In the fuzzy logic, the domains correspond to the concepts of linguistic variables, so that such a variable corresponds to a domain of possible values. The input to the diagnostic system with fuzzy controller is the value of the 2-FAL derivative concentration, and the output is the result of DP value. For the 2-FAL input, the range of possible values (the set grading the membership) is between 0 and 20 and has as linguistic values the next form: {low; medium; high; very_high}. For the output, the DP is [0, 1200] and has as linguistic variables {healthy; mod_det; ext_det; end}, which represent: healthy, moderate degradation, extreme degradation, and end of life.

In order to achieve the best admissible response of the system, two types of fuzzy controllers are proposed to be used, and two types of membership functions are defined for these fuzzy controllers, namely the triangular and the trapezoidal membership functions, which correspond to the input and output of each of the two controllers. The result of the average of the two controllers of the system will be compared to the result of the fault values which is also implemented and calculated based on the classical methods presented in the equations (1) to (4).

The implementation of the two types of membership functions (triangular and trapezoidal) for the inputs and outputs of the two

logic fuzzy type controllers is shown in Figure 4 and Figure 5 respectively.

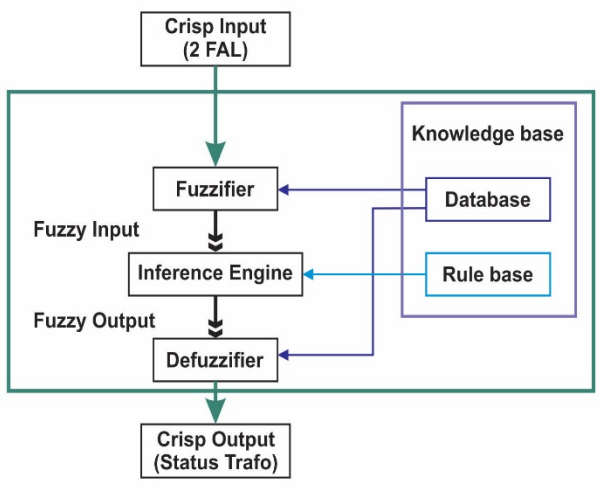


Figure 3: The proposed model of the fuzzy logic controller

Based on the 2-FAL ranges presented in Table 1, the implementation of the four triangular membership functions is carried out, and it is defined by the use of three parameters {a, b, c} covered by these ranges; the functions are described as follows:

$$triangle(x;a,b,c)=\begin{cases}0; & x\le a\\ \dfrac{x-a}{b-x}; & a\le x\le b\\ \dfrac{c-x}{c-b}; & b\le x\le c\\ 0; & c\le x\end{cases} \quad (5)$$

or, by using min and max functions:

$$triangle(x;a,b,c)=\max\left\{\min\left\{\frac{x-a}{b-a};\frac{c-x}{c-b}\right\};0\right\} \quad (6)$$

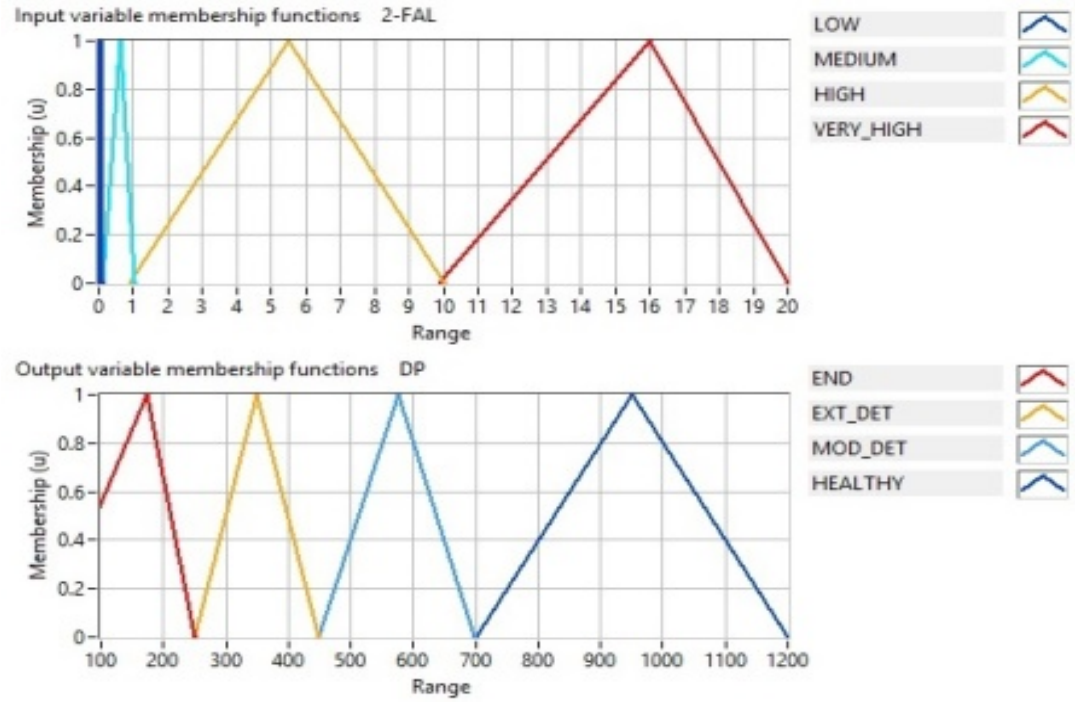
The trapezoidal membership function used for the second fuzzy controller is defined by using four parameters {a, b, c, d} for each of the ranges described in Table 1 and is expressed in the next form:

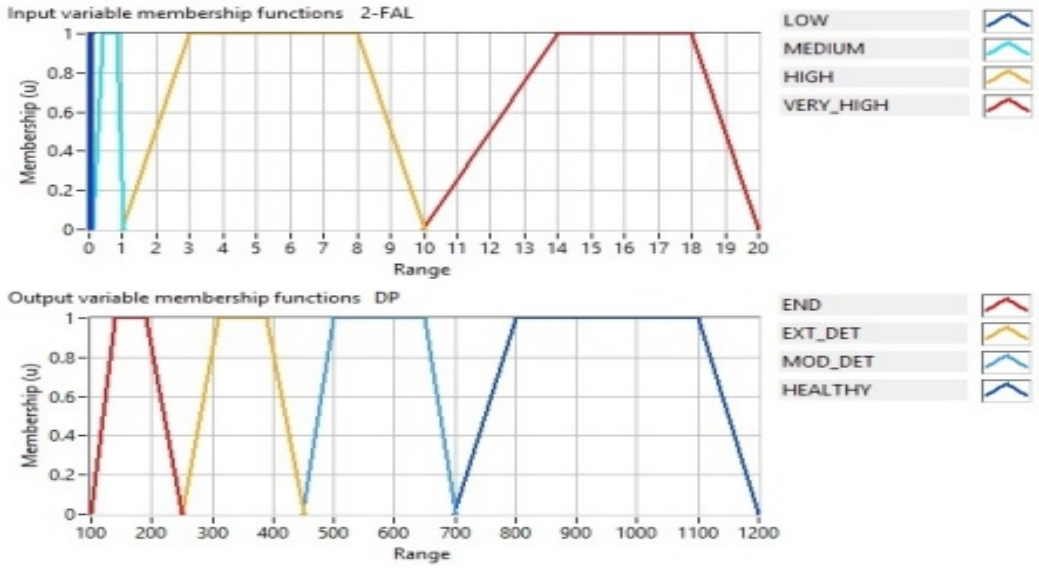
$$trapezoid(x;a,b,c)=\begin{cases}0; & x\le a\\ \dfrac{x-b}{b-a}; & a\le x\le b\\ 1; & b\le x\le c\\ \dfrac{d-x}{d-c}; & c\le x\le d\\ 0; & d\le x\end{cases} \quad (7)$$

or, by using the min and max functions:

$$trapezoid(x;a,b,c,d)=\max\left\{\min\left\{\frac{x-a}{b-a};1;\frac{d-x}{d-c}\right\};0\right\} \quad (8)$$

The defuzzification is required to transform the fuzzy output of the fuzzy controllers achieved based on the triangular and trapezoidal membership functions and averaged into a numerical representation (see Figure 5 and Figure 6). Theoretically, the fuzzy output is a fuzzy relation. The center of gravity method is the most applied practical defuzzification method and is used for the implementation of the two fuzzy controllers. The inference method used for the fuzzy logic implementation of the DP diagnostic system is the Mamadani method.





The fuzzy reasoning is an inference procedure which provides conclusions according to a set of fuzzy "if-then" rules and a set of known data and is based on the ranges specified in Table 1. The method consists in determining the crisp value for the output (DP), taking into consideration, in a balanced manner, all the influences obtained from the rules activated by the particular state of the input (2-FAL) at a given time.

Figure 6 and Figure 7 show the testing of the diagnostic system by entering a 2-FAL value for the two types of fuzzy controllers. The rule that applies for this value of the input and the obtained result of the DP output can be noted.

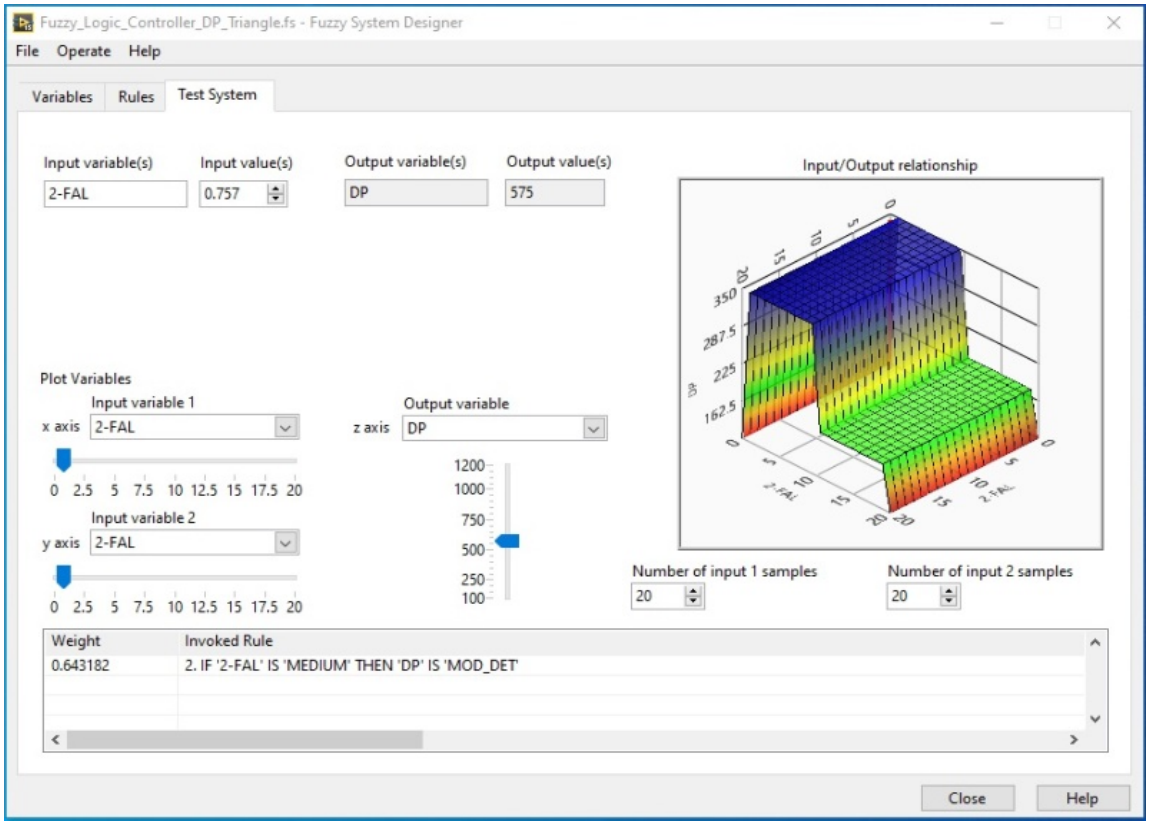


Figure 6: Surface viewer of DP result based on triangular fuzzy logic controller

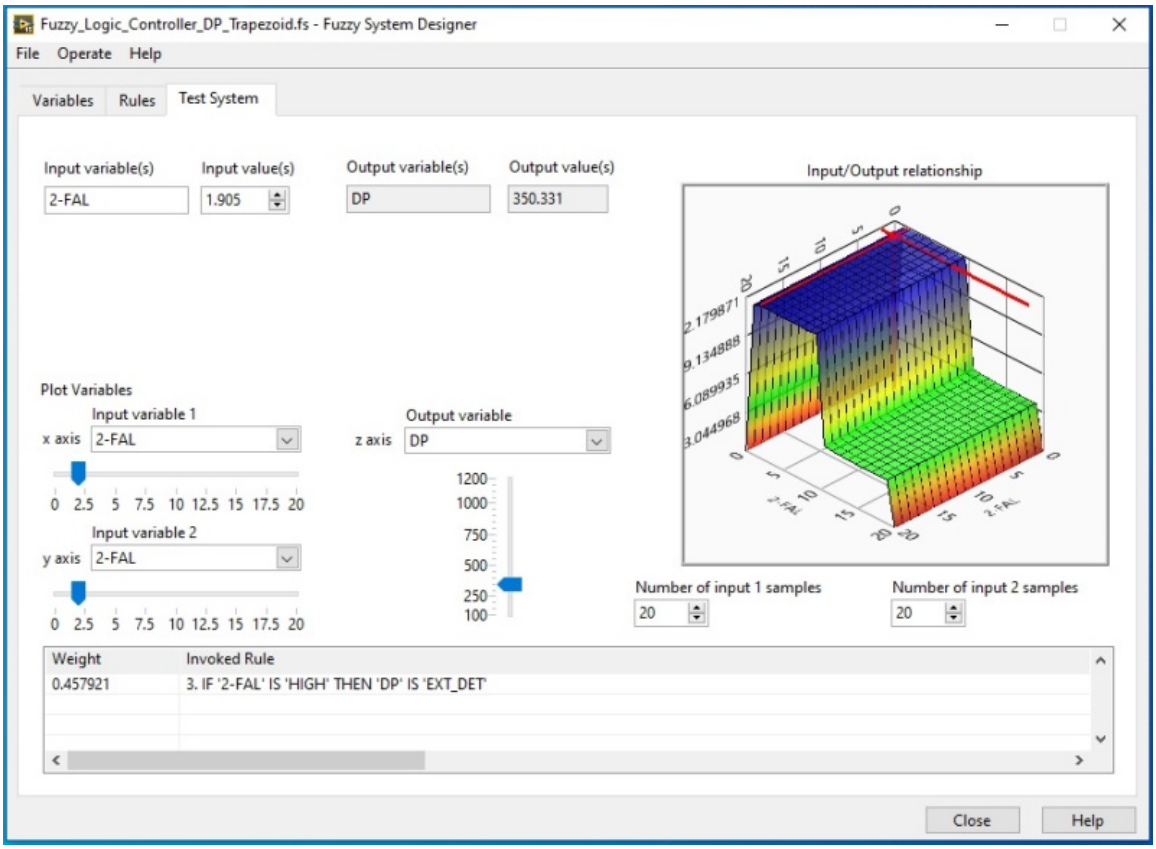
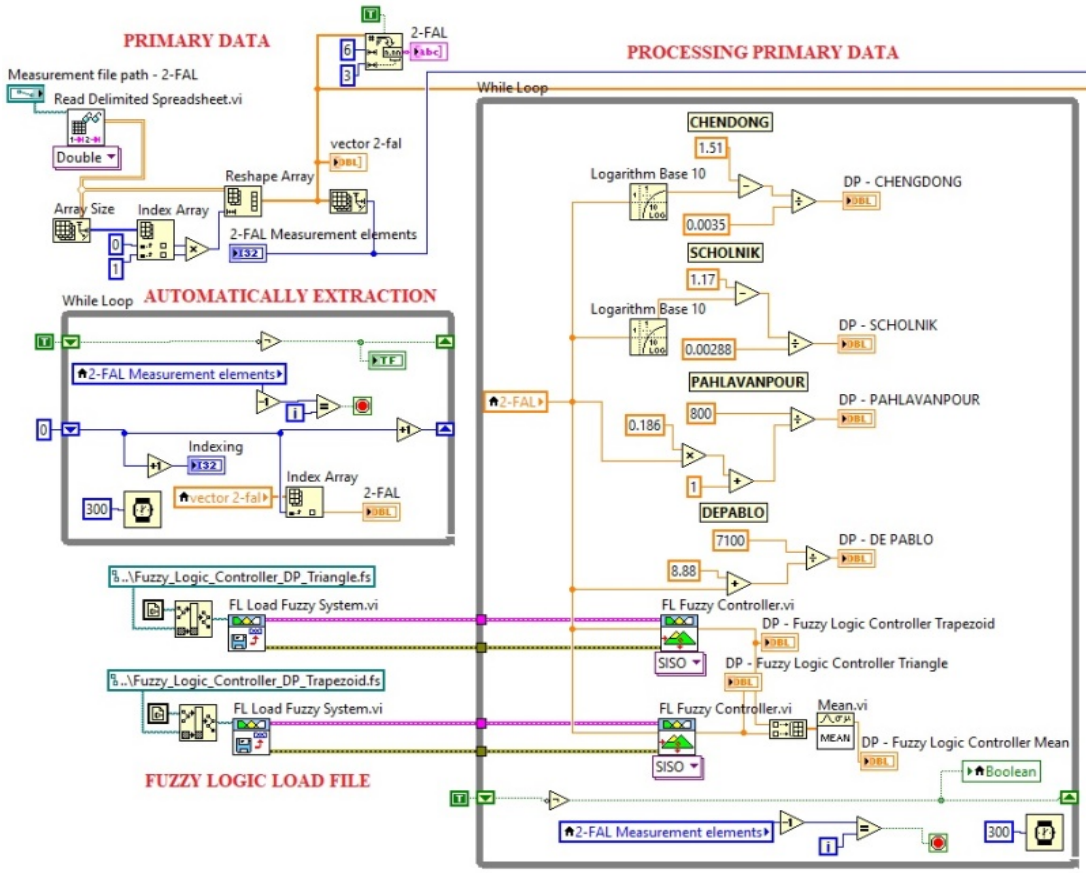


Figure 7: Surface viewer of DP result based on trapezoidal fuzzy logic controller



The software block diagram of the software application developed in LabVIEW for the main stages of the DP diagnostic software application is presented in Figure 8.

4.2. Neuro-fuzzy system for the determination of the DP

The neuro-fuzzy hybrid systems are a combination of neural networks and fuzzy logic where both techniques play a key role: the fuzzy logic provides the structure in which the learning capability of neural networks can be exploited [23], [27]. The ISDDP for power transformers is developed in the Matlab environment in the ANFIS toolbox, which is based on the adaptive neural network. The inference system used in ANFIS toolbox is of Sugeno type, also called fuzzy inference Takagi-Sugeno; it uses singleton output membership functions which for the presented application are linear functions of the input values [26, 27]. Unlike the fuzzy system described above, the proposed neuro-fuzzy diagnostic system has the capability to adapt during a learning process. Thus, by applying the hybrid optimization method, both the membership functions of the fuzzy sets which occur in the premise part of the rules and the parameters of the consequence parts of the fuzzy rules are matched, and the structure of the ANFIS model is presented in Figure 9. The dark circle (in black) on the left side represents the system input (2-FAL), and the one on the right side represents the system output (DP). The white circles on the left and right sides represents the membership functions, and the blue circles represents the rules developed by means of the logical operator "and".

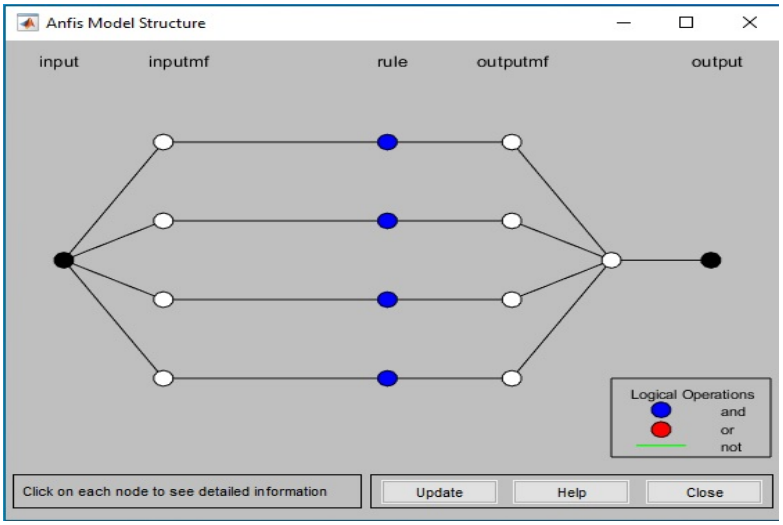


Figure 9: The model structure of the diagnostic system

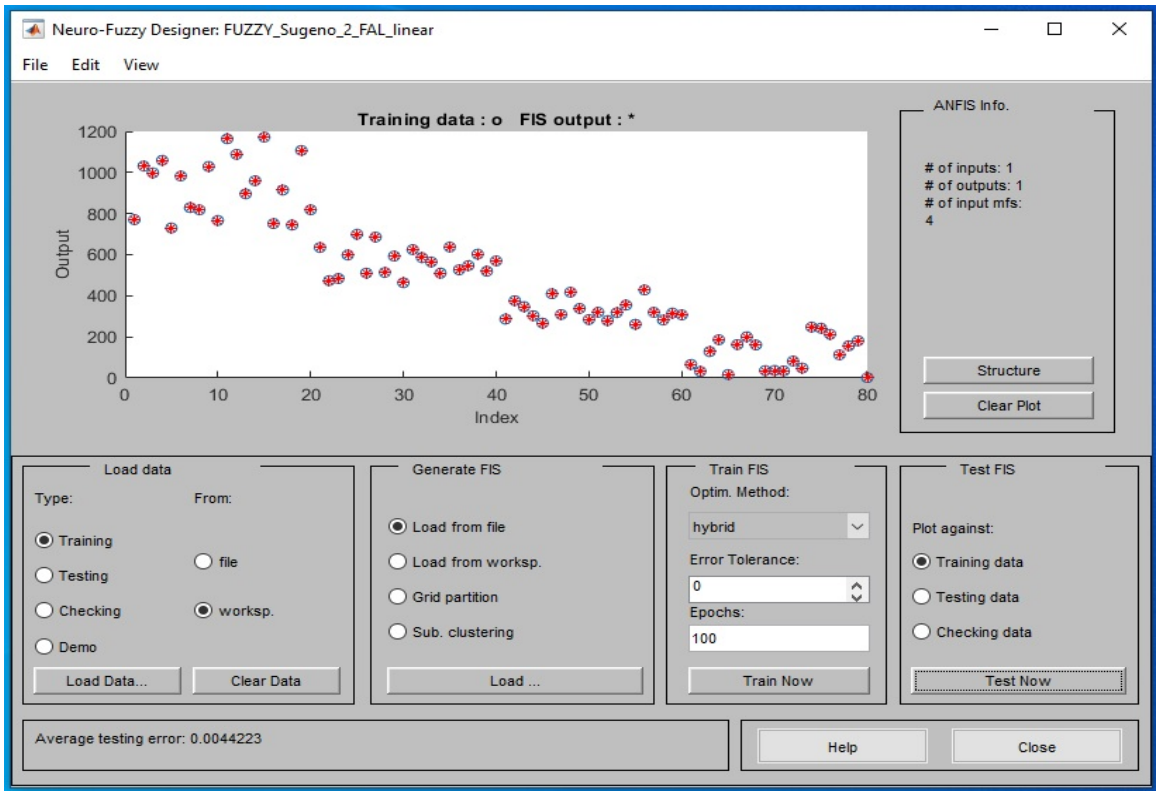


Figure 10: Training data of the ANFIS diagnostic system

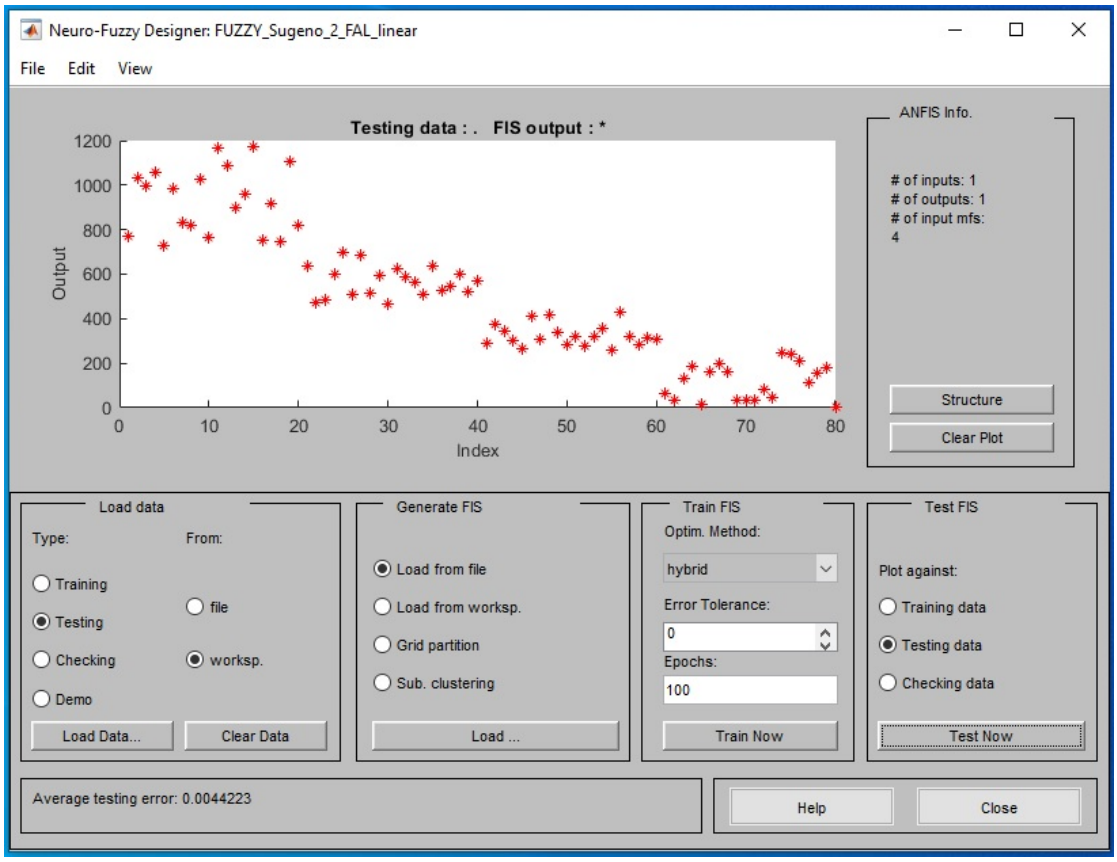


Figure 11: Testing data of the ANFIS diagnosis system

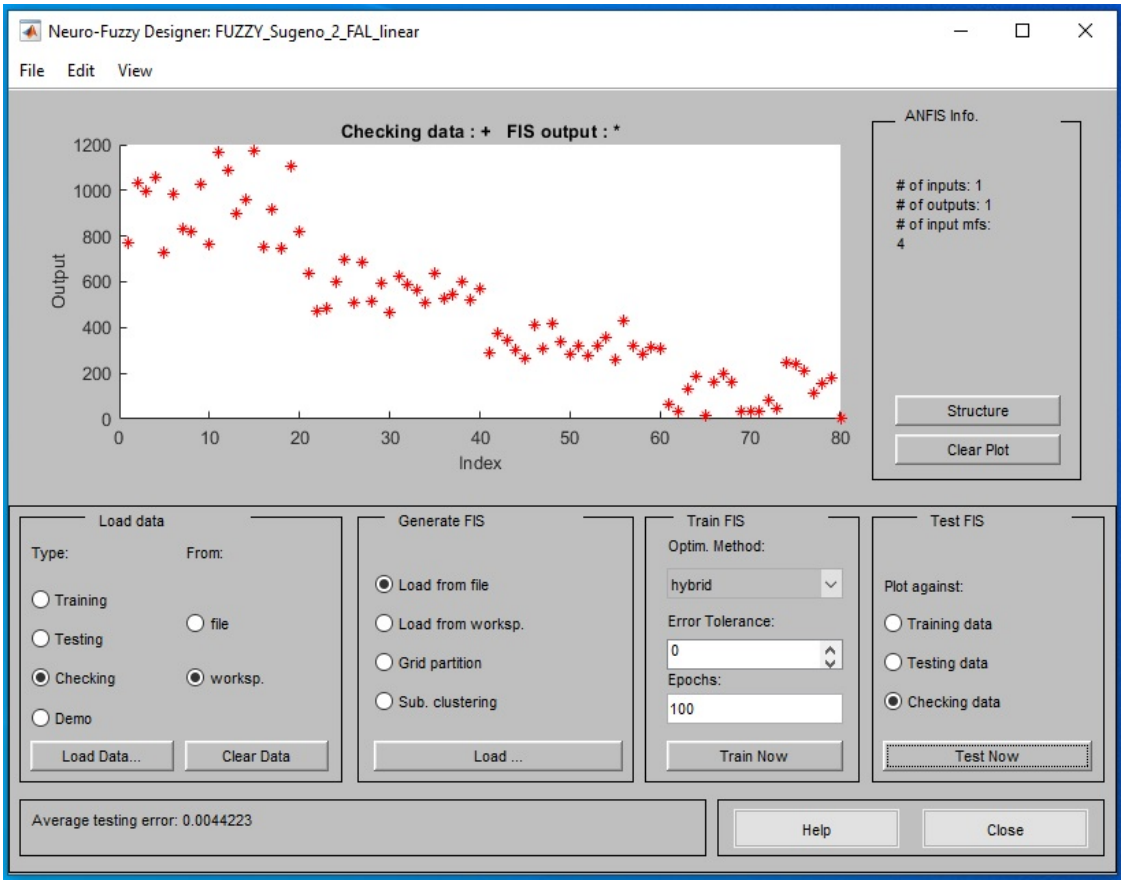


Figure 12: Checking data of the ANFIS diagnosis system

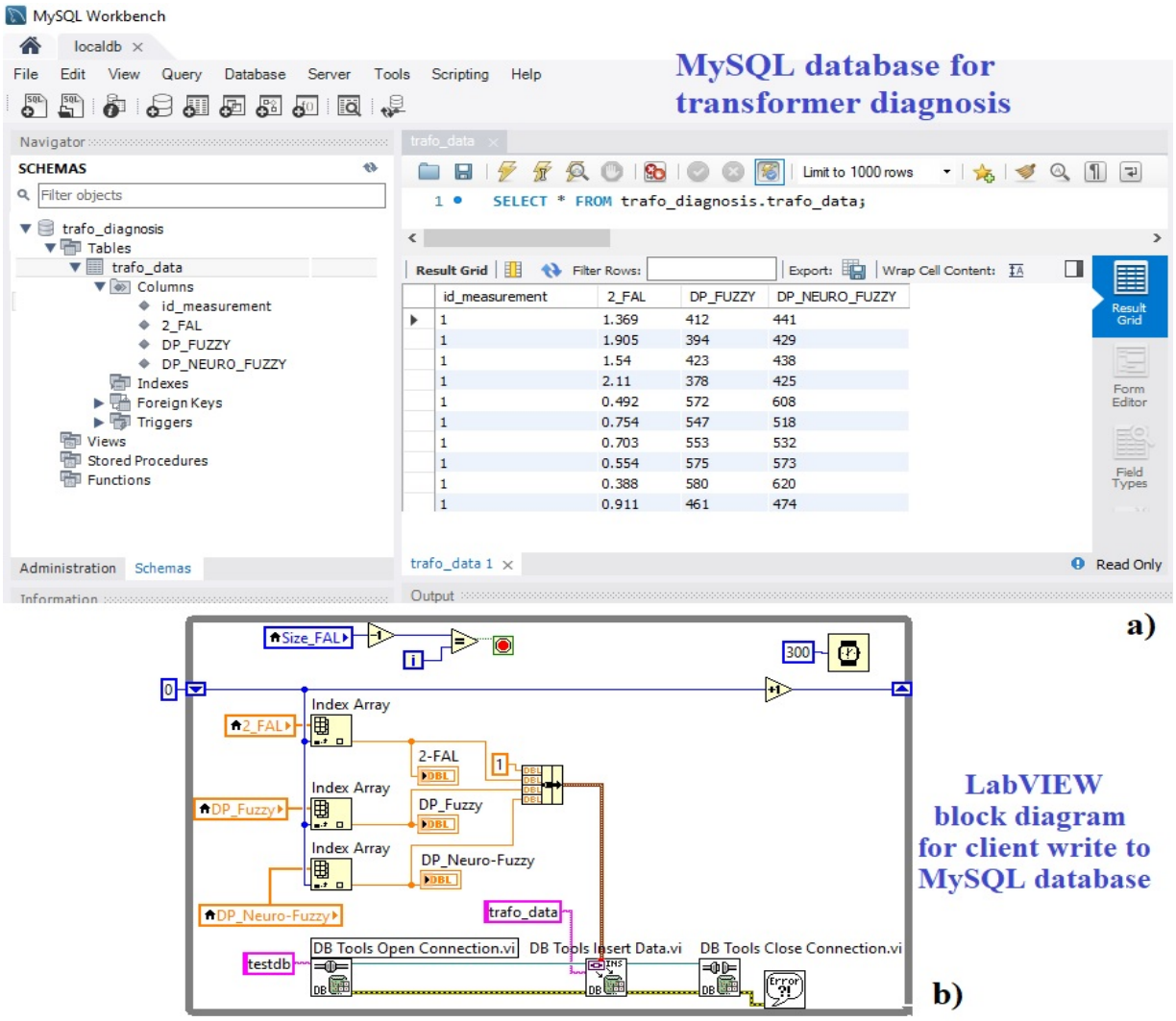


Figure 13: MySQL database connection: a) Table of results in MySQL database, b) block diagram of the client write software module into database

The data set for training, testing and checking comes from the application of classical Pahlavanpour and Chendong methods. The input/output data sets (2-FAL/DP) are entered in the ANFIS based diagnostic system, and they are trained (see Figure 10), verified (see Figure 11), and tested (see Figure 12) in 100 epochs, and the accuracy is given by the mean error obtained with a value of 0.00444223.

The data of the 2-FAL values resulting from the laboratory tests are recorded in the MySQL database by means of a client write software module into the database, and these are taken through the client write software module from the database to be provided to the ISDDP. In the Figure 13 a) is presented the MySQL database interface with the recorded 2-FAL table, and in Figure 13 b) is presented client write software module into the database.

5. Analysis and results

To demonstrate the efficiency of the proposed method and the manner in which data summarization is carried out, we present the analysis of the condition of the insulation of 10 transformers, based on 2-FAL concentration. The results obtained by using the proposed system in items 2 and 4 of Table 2 show a high DP of cellulose insulation, thus, the whole set of physical-chemical measurements of the oils of the two transformers was used. These analyses showed highly degraded wet oil, with temperature spots of >700°C in the transformer in item 2, and high-energy arc discharges, in addition to the wet oil, generated in the transformer in item 4. These discharges may be the result of oil breakdown by arcing between the coils, between terminals and earth, or arcing in the on-load tap changer along the contacts during the switching accompanied by oil leakage in the main tank.

2FAL	DP Chendong	DP Pahlavanpour	DP FUZZY	NEURO-FUZZY
1.369	442	684	412	441
1.905	392	590	394	429
1.540	377	621	423	438
2.110	338	574	378	425
0.492	519	732	572	608
0.754	466	701	547	518
0.703	475	707	553	532
0.554	504	725	575	573
0.388	548	746	580	620
0.911	442	684	461	474

Taking into account these possible problems, the two transformers were un-tanked. In the case of the first transformer, the low-voltage terminals a, and b had loose nuts and showed signs of overheating (see Figure 14), and all connections to the on-load tap-changer selector switch were also loose.

Following the untaking of the second transformer, it was found that the insulation of the flexible connections of the on-load tap-changer selector switch is completely charred, as a result of overheating caused by the arcing during switching as is shown Figure 15.



Figure 14: Low-voltage terminals a and b

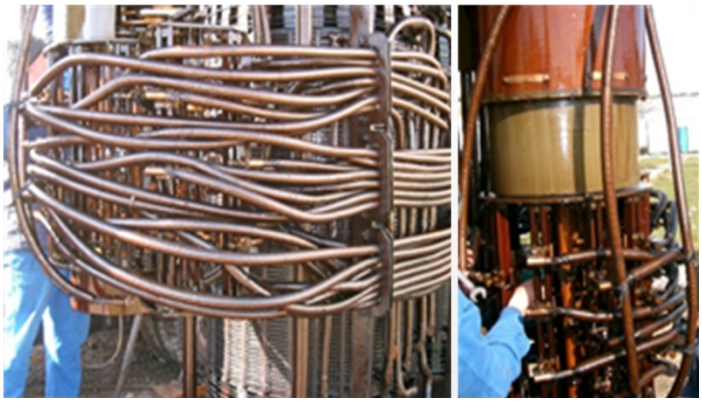


Figure 15: On-load tap-changer selector switch

The faults found and the condition of the insulation of the two transformers confirm the result of the analysis performed using the proposed system for diagnosing the degree of polymerization of solid insulation.

The faults in both transformers were remedied by replacing the damaged insulation; cleaning the terminals, nuts and flexible connection; achieving the tightening to the appropriate torque. The new cotton tape insulation was restored and all degraded insulation residue was removed from the pressboard supports. The on-load tap-changer selector switch was cleaned and the tightening of the connections on the plots of the selector switch was achieved.

The results obtained for the DP by means of the neuro-fuzzy system have a higher precision than the results obtained using mathematical equations and comply with the thresholds proposed by CIGRE according to the 2-FAL content.

The validation of the proposed system is performed by statistical means [31]. To verify the precision and the accuracy, the proposed system was applied to 80 data sets, establishing a precision of approximately 94% and an accuracy of 96% .

6. Conclusions

This article presents a method to determine the DP of solid insulation in transformers, which provides a faster and more accurate interpretation, as compared to the classical methods. The main stages in DP determination are presented, starting with the data provided by the laboratory, the use of classic methods of DP determination, ANFIS training based on a fuzzy controller, data exchange for storage and archiving in MySQL database, as well as the presentation of the validation of this system on a large number of samples, together with the comparison with the results provided by ISDDP with those obtained through un-tanked of transformer.

Thus, the automation of the DP diagnostic process will be achieved, with better results than those obtained by previous methods [26] of precision and accuracy, allowing the operator to make timely decisions, to avoid any possible damages. The main factors of comparison are the accuracy and precision. The values obtained for the improved ISDDP system are around the values presented in the current literature. In the future we propose the improvement of this system by using high-performance neural algorithms for all stages of training, testing and verification in order to improve the accuracy and precision of the system.

Acknowledgment

The paper was developed with funds from the Ministry of Education and Scientific Research - Romania as part of the NUCLEU Program: PN 19 38 01 03 and PN 19 38 02 01.

References

[1] M. Saad, E. teNyenhuis, “On-line gas monitoring for increased transformer protection,” in 2017 IEEE Electrical Power and Energy Conference (EPEC), Saskatoon, SK, Canada, 2017. doi: 10.1109/EPEC.2017.8286169

[2] G. S. Kuleshova, S. A. Yanchenko, “Online Identification of Transformer Internal Faults Based on the Parameters of the Operating Mode” in 2020 V International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russia, 2020. doi: 10.1109/Inforino48376.2020.9111735

[3] N. T. Waugh, D. D. Muir, “Improving the life cycle management of power transformers transforming data to life” in SoutheastCon 2015, Fort Lauderdale, FL, USA, 2015. doi: 10.1109/SECON.2015.7132977

[4] J. Haema, R. Phadungthin, “Development of condition evaluation for power transformer maintenance” in 4th International Conference on Power Engineering, Energy and Electrical Drives, Istanbul, Turkey, 2013. doi: 10.1109/PowerEng.2013.6635680

[5] M. Wang, A. J. Vandermaar, K. D. Srivastava, “Review of condition assessment of power transformers in service” IEEE Electrical Insulation Magazine, **18**(6), 12–25, 2002. doi: 10.1109/MEI.2002.1161455

[6] Condition assessment of power transformers, in *Cigré brochure 761*, 1–163, March 2019

[7] Mineral insulating oils. Methods for the determination of 2-furfural and related compunds, in *CEI/IEC 61198*, 1–18, June 1994

[8] Ageing of cellulose in mineral-oil insulated transformers, in *Cigré brochure 323*, 1–87, Oct. 2007

[9] J. Scheirs, G. Camino, M. Avidano, W. Tumiatti, “Origin of furanic compounds in thermal degradation of cellulosic insulating paper” Journal of Applied Polymer Science, **69**, 2541–2547, 1998

[10] C. Perrier, M. Coulibaly, M. Marugan, “Efficiency of ageing markers for different transformer insulation systems” in 2016 IEEE International Conference on Dielectrics (ICD), Montpellier, France, 2016. doi: 10.1109/ICD.2016.7547743

[11] T. V. Oommen, T. A. Prevost, "Cellulose insulation in oil-filled power transformers: part II maintaining insulation integrity and life" IEEE Electrical Insulation Magazine, **22(**2), 5–14, 2006, doi: 10.1109/MEI.2006.1618996
[12] C. I. Nicola, M. Nicola, M. C. Nițu, A. M. Aciu, "Power Transformer Fault Diagnosis Using Fuzzy Logic Technique Based on Dissolved Gas Analysis and Furan Analysis" UPB Scientific Bulletin, Series C: Electrical Engineering and Computer Science, **81**(3), 183–196, 2019
[13] S. Apte, R. Somalwar, A. Wajirabadkar, "Incipient Fault Diagnosis of Transformer by DGA Using Fuzzy Logic" in 2018 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Chennai, India, 2018. doi: 10.1109/PEDES.2018.8707928
[14] N. Mahmoudi, M. H. Samimi, H. Mohseni, "Experiences with transformer diagnosis by DGA: case studies" IET Generation, Transmission & Distribution, **13**(23), 5431–5439, 2019. doi: 10.1049/iet-gtd.2019.1056
[15] B. Pahlavanpour, G. Duffy, "Development of a rapid spectrophotometry method for analysis of furfuraldehyde in transformer oil as an indication of paper ageing" in Proceedings of IEEE Conference on Electrical Insulation and Dielectric Phenomena - (CEIDP '93), Pocono Manor, PA, USA, 1993. doi: 10.1109/CEIDP.1993.378924
[16] Chendong X., "Monitoring paper insulation ageing by measuring ageing by measuring furfural contents oil" in Seventh International Symposium on High Voltage Engineering, Dresden, Germany, 1991.
[17] M. Duval, A. de Pablo, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," IEEE Electrical Insulation Magazine, **17**(2), 31–41, 2001. doi: 10.1109/57.917529
[18] Guide for the interpretation of gases generated in oil-immersed transformers, in *IEEE C57.104*, 1–36, February 2009
[19] Guidance on the interpretation of dissolved and free gases analysis, in *IEC 60599*, pp. 1–15, September 2015
[20] A. B. Shkolnik, R. T. Rasor, "Statistical insights into furan interpretation using a large dielectric fluid testing database" in 2012 IEEE PES Transmission and Distribution Conference and Exposition (PES T&D), Orlando, FL, USA, 2012.
[21] M. Tamura, K. Ito, M. Honda, Y. Nakagawa, H. Sugimoto, K. Tomishig, "Direct Copolymerization of CO2 and Diols" Scientific Reports 6:24038, 1–9, 2016. doi: 10.1038/srep24038
[22] A. Abu-Siada, S. Hmood, S. Islam, "A new fuzzy logic approach for consistent interpretation of dissolved gas-in-oil analysis" IEEE Transactions on Dielectrics and Electrical Insulation, **20**(6), 2343–2349, 2013. doi: 10.1109/TDEI.2013.6678888
[23] R. A. Prasojo, K. Diwyacitta, Suwarno, H. Gumilang, "Transformer Paper Expected Life Estimation Using ANFIS Based on Oil Characteristics and DissolvedGases (Case Study: Indonesian Transformers)" Energies, **10**(8), 1–18, 2017. doi:10.3390/en10081135
[24] M. M. Nezami, S. A. Wani, S. A. Khan, N. Khera, M. A. Khan, "Fuzzy approach for residual life assessment of paper insulation" in 2017 3rd International Conference on Condition Assessment Techniques in Electrical Systems (CATCON), Rupnagar, India, 2017. doi: 10.1109/CATCON.2017.8280178
[25] D. Urquiza, B. Garcia, J. C. Burgos, "Statistical study on the reference values of furanic compounds in power transformers" IEEE Electrical Insulation Magazine, **31**(4), 15–23, 2015. doi: 10.1109/MEI.2015.7126070
[26] M. C. Nițu, A. Aciu, C. Nicola, M. Nicola, "Fuzzy Controller vs. Classical Methods for Determining the Degree of Polymerization" in 2019 International Conference on Electromechanical and Energy Systems (SIELMEN), Craiova, Romania, 2019. doi: 10.1109/SIELMEN.2019.8905847
[27] R. A. Prasojo, E. F. Nasution, Suwarno, "Implementation of ANFIS and FIS-ID3 in Transformer Oil-Paper Assessment" in 2018 Conference on Power Engineering and Renewable Energy (ICPERE), Solo, Indonesia, 2018. doi: 10.1109/ICPERE.2018.8739422
[28] R. A. Prasojo, K. Diwyacitta, Suwarno, H. Gumilang, "Transformer paper condition assessment using Adaptive Neuro-Fuzzy Inference System model" in 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, Indonesia, 2017, doi: 10.1109/ICECOS.2017.8167141
[29] S. Forouhari, A. Abu-Siada, "Application of adaptive neuro fuzzy inference system to support power transformer life estimation and asset management decision" IEEE Transactions on Dielectrics and Electrical Insulation, **25**(3), 845–852, 2018, doi: 10.1109/TDEI.2018.006392
[30] J. Travis, J. Kring, LabVIEW for Everyone: Graphical Programming Made Easy and Fun 3rd Edition, Prentice Hall, 1996.
[31] J. P. Marques de Sá, Applied Statistics Using SPSS, STATISTICA, MATLAB and R, Springer, Berlin, Heidelberg, 2007.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 669-673 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Investigation of Dielectric Properties of Indigenous Blended Ester oil for Electric System Applications

D.M. Srinivasa[1,*], Usha Surendra[2], V.V. Pattanshetti[3]

[1]*Department of E&E Engineering, PES College of Engineering, Mandya, 571401, India*

[2]*Department of E&E Engineering, School of Engineering & Technology, Christ (Deemed to be University), Bangalore - 560074, India*

[3]*Central Power Research Institute, Bangalore, 560074, India*

ARTICLE INFO

Article history:
Received: 12 August, 2020
Accepted: 19 September, 2020
Online: 05 October, 2020

Keywords:
Antioxidants
Degasification
Indigenous oil

ABSTRACT

The insulation condition of a transformer decides the longevity of the equipment. The unpredicted failure of power transformer will lead to major disaster in the distribution network and it affects both environment and public safety. Nowadays synthetic oil and natural esters are alternatives to transformer oil because of the biodegradable nature. In this paper, investigations were carried out to study the performance of the blended ester. The different properties investigated were viscosity, breakdown voltage, flash point, dielectric dissipation factor and moisture content. Comparisons of the properties were made between mineral oil, vegetable oil without additives and with additives. Further Investigation was carried out to study the impact of antioxidants and degasification. The results indicated that the addition of antioxidants and degasification of the vegetable oil improve significantly its voltage withstanding capacity. The Indigenous oil is code named as DM; Indigenous oil with DBPC is codenamed as DM1, Indigenous oil with BHA is codenamed as DM2. The results have been tabulated and found to be satisfactory.

1. Introduction

The life of insulating medium is investigated by comparing the Breakdown strength and Viscosity of different oil in its pure form with that of the contaminated oil and to find the alternative for mineral oil. The presence of moisture content in the mineral oil compared to vegetable oil would be the main cause for failure of insulation. Vegetable oils which are reliable, cost-effective and environmentally friendly [1]. The evaluation of dielectric properties of Nomex-910 and thermally upgraded Kraft (TUK) paper is being investigated in soya-based natural ester oil by performing accelerated thermal ageing (at 1200C, 1500C and 1800C as per modified ASTM D1934) [2].

The waste cooking oil (WCO) acts as a potential alternative to the existing transformer insulating oil. The use of WCO promotes the optimal consumption of plant-based resources and more efficient waste management. Trans esterification method is performed to eliminate the free fatty acids in the WCO and to reduce viscosity. The trans esterification process is based on the chemical modification reaction between WCO, methyl alcohol (methanol) and sodium hydroxide (NaOH) catalyst lye that produces waste cooking oil methyl ester (WCOME). Chemical and electrical properties of the developed WCOME are compared with the existing WCO [3]. The analysis has been carried out to investigate the critical parameters in different vegetable oil before and after the inclusion of an aging derivative of 2–furfuraldehyde (2-FAL) [4]. The survey has been made to present recent research progress and also for future research. The research scenarios related to the performance of ester fluids versus mineral oils, miscibility, and retro filling of insulating fluids are discussed along with challenges and future aspects on the investigation to enhance the knowledge of ester fluids [5].

The impact of electrical and physicochemical properties of different vegetable oils with different ageing has been studied [6]. Due to lack of oxidative stability, the vegetable oils are less exposed as dielectric in transformer application [7]. The dielectric properties and partial discharge of mineral oil, natural ester and FR3 have been investigated. The suitability in practical use of oils is verified with impregnation of pressboard insulation [8]. The dielectric property of pongamia pinnata oil (PPO) under thermally aged condition has been investigated. [9]. Comparative study of five different types of natural ester liquid under a variation of temperature has been made with respect to different properties. [10]. A study has been carried out to explore the performance of coconut oil as an insulator [11]. Investigation of different properties of Palm Oil (PO) and Coconut Oil (CO) has been carried out under open thermal ageing condition [12]. The analysis of breakdown strength and

*Corresponding Author D. M. Srinivasa, dmsrinivasa.pesce@gmail.com, srinivasa.dm@res.christuniversity.in

physical characteristics of extra virgin olive oil and castor oil under unaged and thermally aged conditions is performed and compared the results with mineral oil with respect to same aging process [13]. The natural ester-based oils are investigated to analyse the suitability of it as better insulating medium with respect to their properties. Further analysis is carried out to on the fatty acid content and cost of vegetable oil [14]. Different vegetable oils and their blends are considered to investigate their dielectric properties with thermal ageing for about120 hours [15]. Neem oil, sunflower oil and mahua oil are blended with antioxidants plus nano powders to measure various properties. To minimize the oxidation stability, Synthetic antioxidants such as Beta Carotene and TBHQ (Tert-Butyl Hydro Quinone) and selenium are preferred as natural antioxidants [16].

Many a times vegetable oils develop bad odour due to oxidation and also rendered unfit for edible applications. These oils can be used effectively reconditioned for applications such as dielectrics by removal of the undesirable products.

In this research paper experiments have been carried out to determine the physical and electrical properties of Indigenous oil which is initially rancid in nature. This oil has chemically treated to bring to status of edible oil and then investigated for different properties. In this paper, section I explained the introduction and its related works. Section II, indicated the methodology and experimental procedure. Results and discussions have been given in section III and Section IV reserved for conclusion.

2. Experimental Procedure

The different physical and electrical characteristics of vegetable oil require special considerations. Due to the presence of oxygen, the degradation of the fluid takes place. So, the dielectric properties are investigated for the smooth operation of transformers.

2.1. Experimental Procedure

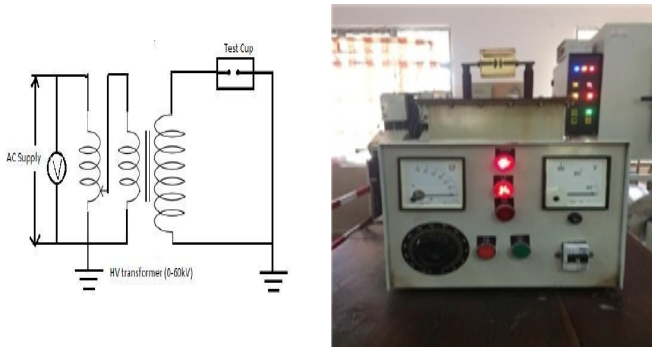
The process of refinement adopted in the present work has been given below.

- The vegetable oil was mixed with hot water (70˚C), shaken for 10 minutes and allowed to cool for 30 minutes. The separated supernatant was again mixed with hot water (60-70˚C) and stirred for 30 minutes. This process helped to remove the gums, waxes and other water soluble components from the oil.
- Removal of free acids: Oil has been mixed with equal volume of 5% Sodium Hydroxide solution and stirred for 20 minutes by treating with dilute caustic soda (5% solution). Free fatty acids reacted with caustic soda result in sodium soap. These acids can be coagulated by heating. Allow the soaps to settle and separate them from filtration.
- Removal of bad odour: Vegetable oil has bubbled with steam in the oil for 1 hour to remove the low volatile oxidised products. Supernatant oil layer has been taken out, leaving behind the condensed water.
- Removal of Pigments and Polar contaminants: Oil has been subsequently passed through an adsorbent column of korvi earth wherein pigments and other polar contaminants will be adsorbed leaving the clean oil ready for use as dielectric fluid.
- Removal of dissolved oxygen and removal of trace moisture and other polar components: Oil has been bubbled with hot nitrogen gas in a controlled manner to strip any dissolved oxygen, carbon dioxide, traces of moisture for almost 24 hours. The process has been used to reduce the moisture.
- Sample preparation: Mineral oil (Electrol-IS335), Indigenous oil and a quantity of 1108 ml of Indigenous oil mixed with 9 grams of DBPC (2,6-di-tert-butyl-p-cresol) and also 9 grams of BHA(Butylated hydroxyl Anisole) is mixed with 1108ml of Indigenous oil is magnetically stirred to ensure complete dissolution are the samples considered for investigation of physical and electrical properties. Samples prepared with indigenous oil without antioxidant has been named as DM, the indigenous oil with DBPC has named as DM-1and the indigenous oil with BHA has named as DM-2.

The following test has been conducted to investigate both electrical and physical properties.

2.2. Breakdown Voltage Conduction Test (BDV)

As per IEC standard, electrodes of spherical in shape has been placed at a distance of 2.5 mm and the liquid specimen filled test cup were used to carry out the test as per Figure 1. Repeatedly conducted the test for 6 times and noted the values for analysis purpose. AC high voltage supply varied gradually till its breakdown at a step of 2kV/sec. BDV test has been conducted as per IEC 60156 standards.



(a) (b)

Figure 1: (a). Circuit diagram for breakdown voltage measurement (b) Breakdown voltage test.

2.3. Kinematic Viscosity

Cannon-Fensken viscometer (Figure 2 and Figure 3) is used for the viscosity test. The test specimen has been filled in the tube up to the level indicated in the bulb. By maintaining a constant temperature of 40 degree centigrade, recorded the time taken by the test specimen to flow in a vertical direction from higher point to lower point of the bulb. Viscosity has been calculated by considering the calibration constant.

Kinematic Viscosity (centistokes-cSt) = (time in seconds) × (viscometer constant)

2.4. Flash Point

The ground level temperature at which flame causes the fire in the form of a flash under specified conditions is called the flash point. Flash point is measured using PENSKY MARTEN'S closed setup as shown in Figure 4. The procedure to identify the flash point has given as follows:

- To heat the oil, the liquid has been filled up to the specified level given in the flash point apparatus.
- External fire has been injected at regular time interval until the flash over occurs. Note the temperature at which flash occurs. Flash point=Recorded temperature at which flash occurs.



Figure 2: Cannon –Fensken viscometer



Figure 3: Viscometer tube.



Figure 4: Pensky martens.

2.5. Dielectric Dissipation factor

Dielectric dissipation factor is measured as per IEC-60247 standards using Eltel model-ADTR-2K Capacitance - Tan δ bridge.

2.6. Degasification Method

Oxidation stability is important to measure the quality and life time of liquid dielectrics. Moisture in oil will degrade the insulation. Transformer failure occurs due to the presence of moisture as well as heat, oxidation, and electrical stresses. Moisture has a significant effect on the dielectric strength. Improvement in the oxidation stability and reduction of moisture content level is achieved by adding antioxidants and also by degasification.

Indigenous oil with DBPC (DM-1) and BHA (DM-2) is considered for degasification with nitrogen gas to remove oxygen and also to reduce moisture content present in the Indigenous oil to investigate improvement in breakdown voltages.

3. Result and Discussion

Experiments have been conducted on DM, DM-1 and DM-2 to identify their breakdown voltage, viscosity, flash points, dielectric dissipation factor and moisture content.

3.1. Standard Values

The ranges of the various characteristics of liquid dielectrics as per IEC62770 standards, has been indicated in the table 1. After experimentation, results of MO, DM, DM-1and DM-2 has been compared with standard values and found out that, the values comes under with in the acceptance limits of IEC62770-2013.

Table 1: IEC62770-2013 standard ranges of liquid dielectrics

Characteristics	Acceptable limits	Mineral oil (MO)	Indigenous oil without Antioxidants(DM)	Indigenous oil with DBPC(DM-1)	Indigenous oil with BHA (DM-2)
Breakdown voltage, kV	30(min)	32	29	37	43
Dielectric dissipation factor	0.05(max)	0.00064	0.038	0.034	0.0511
Kinematic Viscosity, cSt @40 ˚C	50(max)	13.74	46	44	42
Water content, ppm	200(max)	11	501	454	340
Flash point, ˚C	250(min)	152	256	251	245

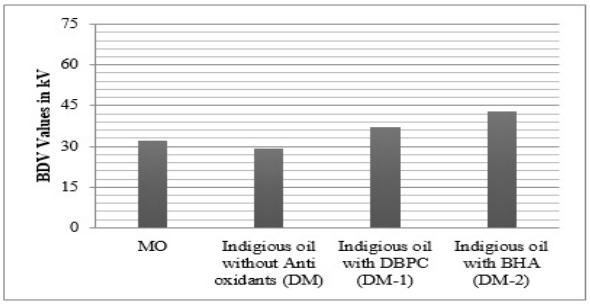


Figure 5: BDV Values of oils.

From Fig 5. BDV of DM-1 is higher than MO and DM. This indicates breakdown voltage of DM-2 is more than MO and DM and DM-1. As the BDV of Indigenous oil is high, this results in higher breakdown strength of Indigenous oil than MO.

Figure 10 indicated the Dielectric dissipation factor of MO, DM and DM-1 oils and the value has been observed and found out that it has well within the 0.05 of standard limits. But the DDF of DM-2 is slightly more than the standard limits. The desired low dissipation factor of oil has been achieved and also it indicated the low power factor.

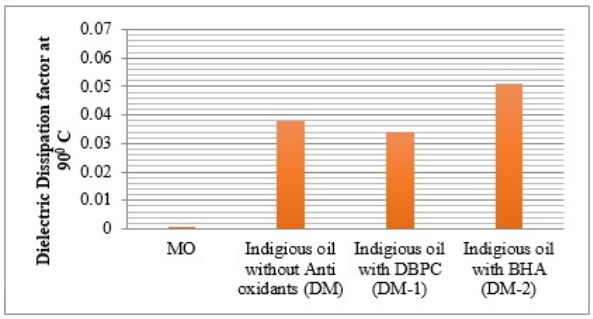


Figure 6: Comparative chart of dielectric dissipation factor of oils.

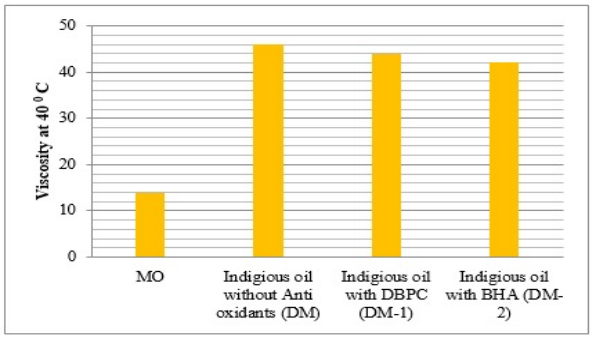


Figure 7: Viscosity of oils

From Figure 7 Viscosity of both DM, DM-1and DM-2 is higher than MO. For better cooling capability, the insulating liquid should have low viscosity. Still the value of viscosity of DM, DM-1and DM-2 will be reduced by reducing the moisture content.

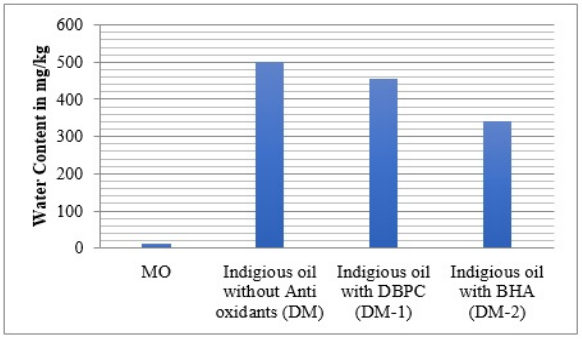


Figure 8: Water content of oils

From Figure 8 Water content is more in DM, DM-1 and DM-2 compared to MO. Reducing the water content in the oil enhances the properties, can be reached by adding antioxidants and also by degasification.

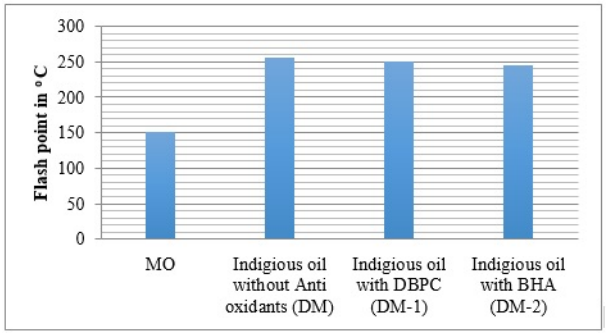


Figure 9: Flash point of oils.

From Figure 9. Flash points of DM, DM-1and DM-2 is very much higher than MO. This indicates higher withstanding capacity of Fire.

3.2. Improvement in BDV Values

Figure 10 indicates that there is a reduction in the moisture content of DM-2 after nitrogen bubbling.

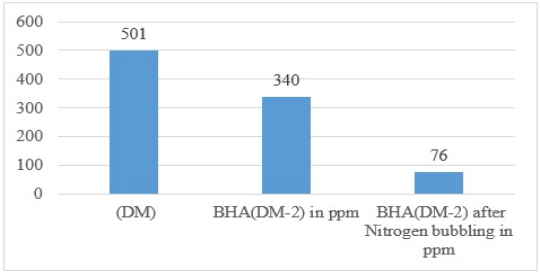


Figure 10: Reduction of Moisture content in DM-2(ppm)

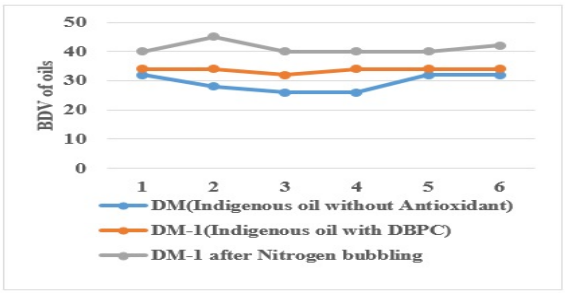


Figure 11: Comparison of breakdown voltage of oils

Figure 11 and 12 indicates that average breakdown voltage is increased in DM-1 after nitrogen bubbling compared with DM (Indigenous oil without Antioxidant) and DM-1(Indigenous oil with DBPC)

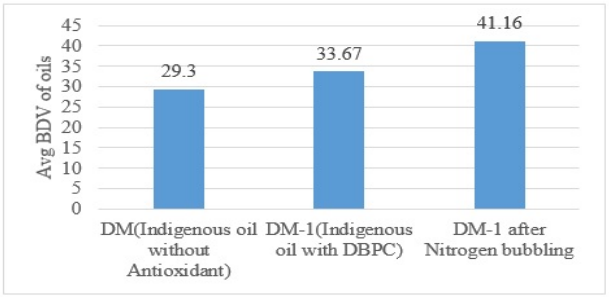


Figure 12: Average breakdown voltage of oils

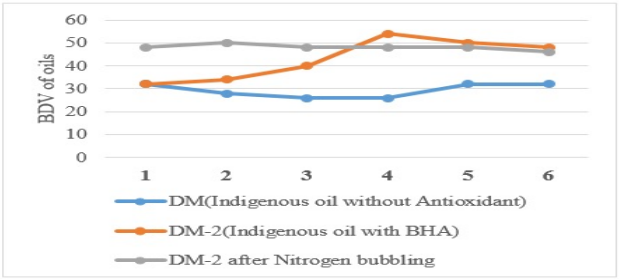


Figure 13: Comparison of breakdown voltage of oils

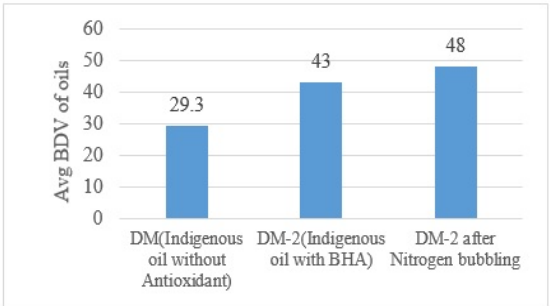


Figure 14: Average breakdown voltage of oils

Figure 13 &14 indicates that average breakdown voltage is increased in DM-2 after nitrogen bubbling compared with DM (Indigenous oil without Antioxidant) and DM-2(Indigenous oil with BHA).

4. Conclusion

Insulation property will be acted as the prime active portion of the high voltage equipment's. The cooling of the transformer depends on the viscosity of the insulating oil and governed by convection. Higher the flash point which in turn reduces the risk of transformer explosions and fire catching criteria and protect the environment.

After adding antioxidants, voltage withstanding capacity of Indigenous oil has been enhanced. It was observed that flash point of the DM, DM-1 and DM-2 is higher than that of mineral oil. In the similar manner, break down voltage (BDV), Dielectric Dissipation factor and Viscosity of DM, DM-1and DM-2 has been observed and found out that their values were well within the range of IEC62770-2013 standard values.

Further Breakdown voltage is enhanced due to addition of Anti-oxidants and also due to degasification.

In case of equipment failure or spillage, vegetable oil decomposition is very slow. The above experimental results justified that, the antioxidants blended indigenous oil can be treated as alternate oil for transformer applications in power system and also it is environmental supportive by means of biodegradation. As the flash point of Indigenous oil is better compared to Mineral oil and results in high fire withstanding capacity. Further this work can be extended by considering Indigenous oil with different antioxidants. The other properties can also be investigated by reducing the moisture content by adopting nitrogen bubbling method and also by other method.

References

[1] D M Srinivasa and Usha Surendra **"**Comparative study of Breakdown Phenomena and Viscosity in Liquid Dielectrics**"** 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), 2019 Vellore, India https**:**//doi.org/: 10.1109/i-PACT44901.2019.8960134.

[2] Ankit Chouhan, Raj Kumar Jarial, and U. Mohan Rao "Thermal Performance of Nomex-910 and TUK Insulating Papers in Soya based Natural Ester Oil" International Journal on Electrical Engineering and Informatics, **12**(1), March 2020. https://doi.org/: 10.15676/ijeei.2020.12.1.5.

[3] Muhammad Nazori Deraman, Norazhar Abu Bakar, Nur Hakimah Ab Aziz, Imran Sutan Chairul,Sharin Ab Ghani "The Experimental Study on the Potential of Waste Cooking Oil as a New Transformer Insulating Oil" Journal of Advanced Research in Fluid Mechanics and Thermal Sciences **69**(1), 74-84, 2020, https://doi.org/10.37934/arfmts.69.1.7484.

[4] C. Subalaksmi, M. Bakrutheen, M. Willjuice Iruthayarajan "Influence of Aging Derivatives on Properties of Natural Ester Oil" International Journal of Innovative Technology and Exploring Engineering (IJITEE), **9**(4), February 2020, https://doi.org/: 10.35940/ijitee.D2031.029420.

[5] M. Rao, I. Fofana, T.Jaya, Esperanza Mariela Rodriguez-Celis, Jocelynalbert and Patrick Picher "Alternative Dielectric Fluids for Transformer Insulation System: Progress, Challenges, and Future Prospects" IEEE Access, **7**, 2019, 10.1109/ACCESS.2019.2960020.

[6] S.Senthil Kumar, A. Arul Marcel Moshi, S.R.Sundara Bharathi, K.Karthik Kumar "Optimization of Various Natural Ester Oils Impregnated Nomex Paper Performance in Power Transformer Applications under Different Ageing Conditions", International Journal of Recent Technology and Engineering (IJRTE), **8**(3), September 2019, https://doi.org/: 10.35940/ijrte.C5832.098319.

[7] K.Sindhuja, M.Srinivasan, N.Niveditha "Natural Esters as an Alternative to Mineral oil in Transformer Applications", International Journal of Pure and Applied Mathematics, **118**(20), 723-732,2018, http://www.ijpam.eu.

[8] K. Jariyanurat, S. Maneerot, P. Nimsanong, P. Kitcharoen, P. Chaisiri, N. Pattanadech "Dielectric Properties of Mineral oil compared with Natural Ester", 19th IEEE International Conference on Dielectric Liquids (ICDL), Manchester, United Kingdom, 25 – 29 June, 2017, https://doi.org/: 978-1-5090-4877-9/17.

[9] T. Mariprasath and V. Kirubakaran "Thermal degradation analysis of pongamia pinnata oil as alternative liquid dielectric for distribution transformer", Indian Academy of Sciences, **41**(9), 933–938 , September 2016, https://doi.org/: 10.1007/s12046-016-0529-0.

[10] M. H. A. Hamid, M. T. Ishak, M. F. Md. Din, N. S. Suhaimi, N. I. A. Katim, "Dielectric Properties of Natural Ester Oils Used for Transformer Application Under Temperature Variation", IEEE 6th International Conference on Power and Energy (PECON 2016), 2016, Melaka, Malaysia, https://doi.org/: 10.1109/PECON.2016.7951472.

[11] N. A. Muhamad and S. H. M. Razali, Electrical and Chemical Properties of New Insulating Oil forTransformer Using PureCoconut Oil", Journal of Advanced Research in Materials Science, **25**(1). 1-9, 2016, ISSN (online): 2289-7992.

[12] Nur Aqilah Mohamad, Norhafiz Azis, Jasronita Jasni, Mohd Zainal Abidin Ab Kadir, Robiah Yunus, Mohd Taufiq Ishak and Zaini Yaakub, "Investigation on the Dielectric, Physical and Chemical Properties of Palm Oil and Coconut Oil under Open Thermal Ageing Condition", J Electr Eng Technol; **11**(3): 690-698,2016, https://doi.org/: 10.5370/JEET.2016.11.3.690.

[13] S.Banumathi, S.Chandrasekar "Analysis of Breakdown Strength and Physical Characteristics of Vegetable Oils for High Voltage Insulation Applications" Journal of Advances in Chemistry, **12**(16), 4902-4912, 2016.

[14] S. Senthil Kumar, Dr. M. Willjuice Iruthayarajan, M. Bakrutheen "Analysis of Vegetable Liquid Insulating Medium for Applications in High Voltage Transformers", In 2014 International Conference on Science Engineering and Management Research (ICSEMR), 1-5. IEEE, 2014.

[15] M. S. Ahmad Kamal,N. Bashir, N. A. Muhamad "Insulating Properties ofVegetableOilsandTheirBlends", 2013 IEEE 7th International Power Engineering and Optimization Conference (PEOCO2013), Langkawi, Malaysia, 3-4 June, 2013. https://doi.org/: 978-1-4673-5074-7/13.

[16] Uma Devi Sankarasubbu and Senthil Kumar Suburaj Examining the Properties of Neem Oil, Sunflower Oil and Mahua Oil with Antioxidants and Nano Powders for Power Transformer" American Journal of Electrical and Computer Engineering, **3**(1), 20-29, 2019. doi: 10.11648/j.ajece.20190301.13

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 686-699 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Fuzzy Recognition by Logic-Predicate Network

Tatiana Kosovskaya*

Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, 199034, Russia

ARTICLE INFO

Article history:
Received: 28 May, 2020
Accepted: 04 August, 2020
Online: 28 August, 2020

Keywords:
Predicate calculus
logic-predicate approach to AI
predicate network
fuzzy predicate network

ABSTRACT

The paper presents a description and justification of the correctness of fuzzy recognition by a logic-predicate network. Such a network is designed to recognize complex structured objects that can be described by predicate formulas. The NP-hardness of such an object recognition requires to separate the learning process, leaving it exponentially hard, and the recognition process itself. The learning process consists in extraction of groups of features (properties of elements of an object and the relations between these elements) that are common for objects of the same class. The main result of a paper is a reconstruction of a logic-predicate recognition cell. Such a reconstruction allows to recognize objects with descriptions not isomorphic to that from a training set and to calculate a degree of coincidence between the recognized object features and the features inherent to objects from the extracted group.

1 Introduction

This paper is an extension of work originally presented in conference CSIT-2019 [1].

The term "logical approach to solving Artificial Intelligence (AI) problems" is usually understood as data notation in the form of a binary string that defines the values of some object properties under study. In this case, if the object is represented as a set of its elements that have some properties and are connected by given relations, then when describing it as a binary string, firstly, the structure of the object itself is lost, and secondly, you have to store a large number of "unnecessary" information that the element a_j does not have the properties $p_{i_1}, \ldots, p_{i_k}$, that the elements a_{j_1}, a_{j_2} are not in the relations $q_{i_1}, \ldots, q_{i_l}$, etc.

Consider objects composed of smaller elements with preassigned properties, and some smaller object having preassigned relations. In such a case predicate formulas are an adequate description language.

The use of predicate calculus and automatic proof of a theorem for AI problems solving was offered by many authors [2]-[4] in the 70-th years of the XX century. At the same time the notion of NP-completeness was introduced and began to be actively used by the scientific community [5].

In 2003 the use of predicate calculus and automatic proof of a theorem for AI problems solving (without complexity bounds) was described in [6].

In particular, in that book it is shown that if a binary string simulates a description of an economic problem in which interaction agents have given properties and are in given relations then the notation length of such a string is exponential of the length of the description of the same problem input data by setting properties of these agents and relations between them, i.e. in fact, using the language of predicate calculus.

In 2007 the author of this paper has proved NP-completeness of recognition problems described in the terms of predicate formulas and upper bounds of number of steps for two solving algorithms [7]. When such problems are solved by an exhaustive search algorithm, their computational complexity coincides with the length of their input data encoding using a binary string [6].

A hierarchical level description of classes was suggested in [8] to decrease the computational complexity of these problems. Formulas which are isomorphic to "frequently appeared" sub-formulas of "small complexity" are extracted from class descriptions in order to construct such a level description. This allows to decompose the main problem into a series of similar problems with input data with the smaller length. An instrument for this extraction is partial deduction of a predicate formula [9],[10].

Later, the author noticed that the level description is actually a recognition network, which, after training (having exponential computational complexity), can quickly solve the recognition problem.

The concept of "network" is increasingly used both in theoretical and in practical research, particularly, when solving pattern recognition problems. Research and practical applications of neural networks [11],[12], Bayesian networks [13]-[15], technical networks [16] are widespread. The inputs of such networks are usually signals characterizing the properties of the studied objects or pro-

*Corresponding Author: Tatiana Kosovskaya, St. Petersburg State University, St. Petersburg, 199034, Russia, +7 921 323 23 07 & kosovtm@gmail.com

cesses, and expressed, as a rule, by numerical characteristics, which can be encoded by binary strings. The processing of binary strings in the most cases is carried out in linear or polynomial (usually quadratic) number of steps under the length of these strings. A convenient model for processing such strings is propositional logic (Boolean functions).

As noted above, the use of binary strings in the recognition of complex structured objects has disadvantages. To recognize such objects, the author offered the notion of a logic-predicate network using level description of classes [17]. Such a network contains two blocks: a training block and a recognition block. The training block run is based on the extraction from a class of objects description such fragments, which are inherent in many objects of the class. Recognition block run is reduced to the sequential solution of problems of the same type with a smaller length of the input data.

The disadvantage of such a recognition network is that it is able to recognize only objects that differ from those included in the training set (TS) by renaming the elements of the object (objects with isomorphic descriptions). But such a network can be retrained by adding unrecognized objects to the TS. This was the reason for modification of the logic-predicate network permitting to recognize objects, that not coincide but only similar to the objects from a training set. Such a modification was offered in [1, 18]. The degree of coincidence for description of an object part and the formula, satisfiability of which is checking in the cell, is calculated. Such a fuzzy network allows not only quickly enough to recognize objects isomorphic to those represented in the TS, but also to calculate the degree of coincidence that a "new" object belongs to one or another class. If necessary, it can be retrained using such a "new" object.

The structure of the paper is as follows.

Section 2 "Logic-predicate approach to AI problems" contains 2 subsections: "Problem Setting" and "Important Definitions". This section describes the results previously obtained by the author and defines the terminology previously introduced by the author, which are necessary for understanding the further presentation.

Section 3 "Level description" presents the basic idea of constructing a level description of classes. It contains two sub-sections "Construction of a level description of classes" and "The use of level description of a class". They give algorithms for constructing a level description of classes and the use of such a description for object recognition. These algorithms are further used in constructing the logic-predicate network. It is shown that the computational complexity of an object recognition decreases while using a level description of a class.

Section 4 "Logic-predicate network" describes the structure of such a network.

Section 5 "Example of logic-predicate network" describes an example illustrating the formation of a network, recognizing of an object with the use of a network, and its retraining. It consists of 4 sub-sections "Training Block Run: Extraction of Sub-Formulas", "Training Block Run: Forming a Level Description", "Recognition Block Run" and "Retraining the Network".

Section 6 "Fuzzy recognition by a logic-predicate network" describes the changes that need to be made to the logic-predicate network in order to recognize objects that are not isomorphic to those presented in the TS. A description of the contents of a fuzzy network cell is given. A cell of this type replaces each cell in the logic-predicate network in which the logical sequence of the target formula from the description of an object is checked. In this case, the degree of coincidence is calculated that the fragment being tested partially (and to what extent) satisfies the target formula.

Section 7 "Model example of fuzzy recognition" presents an example of fuzzy recognition of a "new" object.

In comparison with [1], the setting of problems that can be solved in the framework of logic-predicate approach are described in the presented paper in more details. A justification for the introduction of a level description is given. Algorithms for a level description construction and a level description use are described. Scheme 1 of a common up to the names of arguments sub-formulas extraction is added. A section 5 has been added with an example of constructing and modifying a logic-predicate network. It describes in details how a network is formed by the training set, a new object is recognized, and the network is rebuilt. The scheme of a fuzzy logic-predicate network cell is presented. An example of fuzzy recognition is described in more details.

2 Logic-Predicate Approach to AI Problems

A detailed presentation of the logic-predicate approach to solving AI problems is available in [19]. In this paper only a general setting of problems and main methods for their solving are formulated.

2.1 Problem Setting

As it was mentioned in the Introduction, if an object to be recognized is a complex structured one, then predicate formulas are a convenient language for its description and recognition. Let an investigated object ω be represented as a set of its elements $\omega = \{\omega_1, \ldots, \omega_t\}$ and be characterized by predicates $p_1, \ldots, p_n$ which define some properties of its elements or relations between them. The description $S(\omega_1, \ldots, \omega_t)$ of the object ω is a set of all constant literals (atomic formulas or their negations) with predicates $p_1, \ldots, p_n$ which are valid on ω.

Let the set of all investigated objects Ω be is divided into classes $\Omega_1, \ldots, \Omega_K$ such that $\Omega = \cup_{k=1}^{K} \Omega_k$. Logical description of the class Ω_k is such a formula $A_k(\overline{x})$[1] that if the formula $A_k(\overline{\omega})$ is true then $\omega \in \Omega_k$. The class description may be represented as a disjunction of elementary conjunctions of atomic formulas.

Many AI problems may be formulated as follows with the use of such descriptions.

Identification problem. To pick out all parts of the object ω which belong to the class Ω_k

$$S(\omega) \Rightarrow \exists \overline{x}_{k\neq} A_k(\overline{x}_k)^2. \quad (1)$$

Classification problem. To find all such class numbers k that $\omega \in \Omega_k$

$$S(\omega) \Rightarrow \vee_{k=1}^{K} A_k(\overline{x}_k). \quad (2)$$

[1] Here and below the notation $\overline{x}$ is used for an ordered list of the set x.

[2] To denote that there exist distinct values for variables from the list $\overline{x}$ the notation $\exists \overline{x}_{\neq} A_k(\overline{x})$ is used.

Analysis problem. To find and classify all parts τ of the object ω which may be classified

$$S(\omega) \Rightarrow \vee_{k=1}^{K} \exists \overline{x}_{k \neq} A_k(\overline{x}_k). \quad (3)$$

In fact, instead of the quantifier $\exists$ (*whether exists*) a symbol ? (*for what*) must be written. But if one uses a constructive method of its proof (an exhaustive search algorithm or logical methods such as derivation in a sequential predicate calculus or resolution method for predicate calculus) then not only existence of such arguments would be proved but the values of them would be found.

It was proved that the problems (1) and (3) are NP-complete [7] and the problem (2) is GI-complete [20], i.e. it is polynomially equivalent to the "open" problem of Graph Isomorphism, for which it is not proved its NP-completeness and a polynomial in time algorithm is not found.

As the formulas $A_k(\overline{x}_k)$ may be represented as a disjunction of elementary conjunctions of atomic formulas, the checking of each of the problems (1), (2) and (3) may be reduced to sequential checking of the formula

$$S(\omega) \Rightarrow \exists \overline{x}_{\neq} A(\overline{x}), \quad (4)$$

where $A(\overline{x})$ is an elementary conjunction. Note that in the case of the problem (2) the number of variables in $\overline{x}$ equals to the number of constants in ω.

The formula (4) may be checked, for example, by an exhaustive algorithm and an algorithm based on the derivation in sequential calculus or on the use of resolution method.

The upper bound of number of steps for an exhaustive algorithm is $O(t^m)$[3], where t is the number of elements in ω, m is the number of variables in $A(\overline{x})$ [7]. This upper bound coincides with the upper bound of a binary string simulating input data in the form of $S(\omega)$ and $A(\overline{x})$.

The upper bound of the number of steps for a logical algorithm is $O(s^a)$ where s is the maximal number of atomic formulas in $S(\omega)$ with the same predicate symbol having occurrences in $A(\overline{x})$, a is the number of atomic formulas in the elementary conjunction $A(\overline{x})$ [21].

Obviously, these estimates depend exponentially on the parameters of the formula $A(\overline{x})$. That's why it will be useful to break the solution of the problem into a series of sub-problems of the type (4) with a shorter right-hand side.

2.2 Important Definitions

The objects and notions satisfying the following definitions will be used later in the text.

Definition 1: Elementary conjunctions $P(a_1, \ldots, a_m)$ and $Q(b_1, \ldots, b_m)$ are called **isomorphic**

$$P(a_1, \ldots, a_m) \sim Q(b_1, \ldots, b_m),$$

if there are an elementary conjunction $R(x_1, \ldots, x_m)$ and substitutions of arguments $a_{i_1}, \ldots, a_{i_m}$ and $b_{j_1}, \ldots, b_{j_m}$ instead of the variables $x_1, \ldots, x_m$ such that the results of these substitutions $R(a_{i_1}, \ldots, a_{i_m})$ and $R(b_{j_1}, \ldots, b_{j_m})$ coincide with formulas $P(a_1, \ldots, a_m)$ and $Q(b_1, \ldots, b_m)$, respectively, up to the order of literals.

The substitutions $(a_{i_1} \rightarrow x_1, \ldots, a_{i_m} \rightarrow x_m)$ and $(b_{j_1} \rightarrow x_1, \ldots, b_{j_m} \rightarrow x_m)$ are called **unifies** of $R(x_1, \ldots, x_m)$ with $P(a_1, \ldots, a_m)$ and $Q(b_1, \ldots, b_m)$ and are denoted as $\lambda_{R,P}$ and $\lambda_{R,Q}$, respectively. [1], [20]

Note that concept of isomorphism of elementary conjunctions of atomic predicate formulas differs from the concept of equivalence of these formulas, because they can have significantly different arguments. In fact, for isomorphic formulas there are permutations of their arguments such that they define the same relationship between their arguments.

Definition 2: Elementary conjunction $C(x_1, \ldots, x_n)$ is called a **common up to the names of arguments sub-formula** of two elementary conjunctions $A(a_1, \ldots, a_m)$ and $B(b_1, \ldots, b_k)$ if it is isomorphic to some sub-formulas $A'(a'_1, \ldots, a'_n)$ and $B'(b'_1, \ldots, b'_n)$ of $A(a_1, \ldots, a_m)$ and $B(b_1, \ldots, b_k)$, respectively.

The unifiers of $C(x_1, \ldots, x_n)$ with $A'(a'_1, \ldots, a'_n)$ and $B'(b'_1, \ldots, b'_n)$ will be denoted as $\lambda_{C,A}$ and $\lambda_{C,B}$, respectively [19].

For example, let

$$A(x, y, z) = p_1(x) \,\&\, p_1(y) \,\&\, p_1(z) \,\&\, p_2(x, y) \,\&\, p_3(x, z),$$

$$B(x, y, z) = p_1(x) \,\&\, p_1(y) \,\&\, p_1(z) \,\&\, p_2(x, z) \,\&\, p_3(x, z).$$

The formula

$$P(u, v) = p_1(u) \,\&\, p_1(v) \,\&\, p_2(u, v)$$

is their common up to the names of variables sub-formula with the unifiers $\lambda_{P,A} = (x \rightarrow u, y \rightarrow v)$ and $\lambda_{P,B} = (x \rightarrow u, z \rightarrow v)$ because

$$P(x, y) = p_1(x) \& p_1(y) \& p_2(x, y)$$

is a sub-formula of $A(x, y, z)$ and

$$P(x, z) = p_1(x) \& p_1(z) \& p_2(x, z)$$

is a sub-formula of $B(x, y, z)$.

Definition 3: Elementary conjunction $C(x_1, \ldots, x_n)$ is called a **maximal common up to the names of arguments sub-formula** of two elementary conjunctions $A(a_1, \ldots, a_m)$ and $B(b_1, \ldots, b_k)$ if it is their common up to the names of arguments sub-formula and after adding any literal to it, it ceases to be one [19].

For further presentation, the concept of partial sequence [9] is important.

The problem of checking if the formula $A(\overline{x})$ or some its sub-formula $A'(\overline{y})$ is a consequence of the set of formulas $S(\omega)$ is under consideration in [9]. Here the list of arguments $\overline{y}'$ is a sub-list of the list of arguments $\overline{y}$.

Definition 4: Let $A(\overline{x})$ and $B(\overline{y})$ be elementary conjunctions. If $A(\overline{x}) \Rightarrow \exists \overline{y}_{\neq} B(\overline{y})$ is not valid but for some sub-formula $B'(\overline{y}')$ of $B(\overline{y})$ the sequence $A(\overline{x}) \Rightarrow \exists \overline{y}'_{\neq} B'(\overline{y}')$ is true, we will say that $B(\overline{y})$ is a **partial sequence** from $A(\overline{x})$ and denote this by $A(\overline{x}) \Rightarrow_P \exists \overline{y}_{\neq} B(\overline{y})$ [1].

A constructive algorithm for a proof of partial sequence is in [22]. While using a constructive algorithm to prove $A(\overline{x}) \Rightarrow_P \exists \overline{y}_{\neq} B(\overline{y})$ we can find the maximal sub-formula $B'(\overline{y}')$ such that $A(\overline{x}) \Rightarrow \exists \overline{y}'_{\neq} B'(\overline{y}')$ and such values $\overline{x}'$ ($\overline{x}'$ is a permutation of a sub-string of $\overline{x}$) for $\overline{y}'$ that $B'(\overline{x}')$ is a sub-formula of $A(\overline{x})$. It means that we find a maximal common up to the names of arguments

[3] $f(x) = O(g(x))$ means that there is such a constant C that for every x the inequality $f(x) \leq C \cdot g(x)$ is valid.

sub-formula $B'(\overline{y}')$ of two elementary conjunctions $A(\overline{x})$ and $B(\overline{y})$ and its unifier with $B'(\overline{x}')$ (a sub-formula of $A(\overline{x})$).

Further, for fuzzy recognition of an object we need an extension of the partial sequence concept.

Every sub-formula $A'(\overline{x}')$ of the formula $A(\overline{x})$ is called its **fragment** [1].

Let a and a' be the numbers of atomic formulas in $A(\overline{x})$ and $A'(\overline{x}')$, respectively, m and m' be the numbers of objective variables in $\overline{x}$ and $\overline{x}'$, respectively.

Numbers q and r are calculated by the formulas $q = \frac{a'}{a}$, $r = \frac{m'}{m}$ and characterize the degree of coincidence between $A(\overline{x})$ and $A'(\overline{x}')$. For every elementary conjunction $A(\overline{x})$ and its fragment $A'(\overline{x}')$ it is true that $0 < q \leq 1$, $0 < r \leq 1$. Besides, $q = r = 1$ if and only if $A'(\overline{x}')$ coincides with $A(\overline{x})$.

Under these notations, the formula $A'(\overline{x}')$ will be called a (q, r)-**fragment of the formula** $A(\overline{x})$ [1].

If $S(\omega) \Rightarrow \exists \overline{x}_{\neq} A(\overline{x})$ is not valid but for some (q, r)-fragment $A'(\overline{x}')$ $(q \neq 1)$ of $A(\overline{x})$ the sequence $S(\omega) \Rightarrow \exists \overline{x}'_{\neq} A'(\overline{x}')$ is true, we will say that $S(\omega) \Rightarrow_P \exists \overline{x}_{\neq} A(\overline{x})$ is a **partial** (q, r)**-sequence** for description $S(\omega)$ [1].

A (q, r)-fragment $A'(\overline{x}')$ of the formula $A(\overline{x})$ with maximal value of q satisfying $S(\omega) \Rightarrow \exists \overline{x}'_{\neq} A'(\overline{x}')$ will be called a **maximal fragment of the formula** $A(\overline{x})$ for description $S(\omega)$ [1].

As for a maximal fragment of the formula $A(\overline{x})$ the checking whether $S(\omega) \Rightarrow \exists \overline{x}'_{\neq} A'(\overline{x}')$ may be done by some constructive method then such values $\overline{\tau}$ $(\tau \subseteq \omega)$ for the list of variables $\overline{x}'$ that $S(\omega) \Rightarrow A'(\overline{\tau})$ will be found.

Definition 5: Conjunction of literals from $A(\overline{x})$ which are not in $A'(\overline{x}')$ is called a **complement** of $A'(\overline{x}')$ up to $A(\overline{x})$ [1].

A complement of $A'(\overline{x}')$ up to $A(\overline{x})$ will be denoted by $C_{A(\overline{x})}A'(\overline{x}')$.

Definition 6: A (q, r)-fragment $A'(\overline{x}')$ of the formula $A(\overline{x})$ is called **contradictory** to the description $S(\omega)$ on the list of constants $\overline{\tau}$ if $S(\omega)$ and $C_{[A(\overline{x})]^{\overline{x}'}_{\overline{\tau}}}A'(\overline{\tau})$ lead to the contradiction, i.e., $S(\omega) \Rightarrow \neg C_{[A(\overline{x})]^{\overline{x}'}_{\overline{\tau}}}A'(\overline{\tau})$ [1].

Here the denotation $[A(\overline{x})]^{\overline{x}'}_{\overline{\tau}}$ is used for the result of substitution of the constants from the list $\overline{\tau}$ instead of the corresponding variables from the list $\overline{x}'$.

3 Level Description

As noted above, estimates of the number of steps of the proof of (4) exponentially depends on the number of variables or on the number of literals of the formula $A(\overline{x})$. Moreover, the proof of formulas (1), (2) and (3) in which $A_k(\overline{x}_k) = A_{k,1}(\overline{x}_{k,1}) \vee \cdots \vee A_{k,m_k}(\overline{x}_{k,m_k})$, with the available algorithms for solving (4), can be reduced to a sequential proof of (4) with $A_{k,1}(\overline{x}_{k,1}), \ldots, A_{k,m_k}(\overline{x}_{k,m_k})$ on the right side.

Due to the unprovable, but repeatedly confirmed in practice, statement of Einstein "God is subtle, but he is not malicious" elementary conjunctions included in the description of one class must have common up to the names of arguments sub-formulas. Therefore, it is natural to break the proof of this formula into a series of formulas of the same kind, but with lower values of the essential parameters. This may be done by means of the following procedure.

– Extract "frequently occurred" common up to the names of arguments sub-formulas $P^1_i(\overline{y}^1_i)$ $(i = 1, \ldots, n_1)$ of goal formulas $A_{k,1}(\overline{x}_{k,1}), \ldots, A_{k,m_k}(\overline{x}_{k,m_k})$ with "small complexity". Simultaneously we find unifiers of $P^1_i(\overline{y}^1_i)$ and sub-formulas of $A_k(x_1, \ldots, x_m)$.

– For every sub-formula $P^1_i(\overline{y}^1_i)$ a new 1st level predicate p^1_i with one 1st level variable y^1_i, defined by the equivalence $p^1_i(y^1_i) \Leftrightarrow P^1_i(\overline{y}^1_i)$, is introduced. This 1st-level variable y^1_i is a variable for a string of initial variables.

– Replace each occurrence of a formula, isomorphic to $P^1_i(\overline{y}^1_i)$ $(i = 1, \ldots, n_1)$, into $A_{k,j}(\overline{x}_{k,j})$ $(j = 1, \ldots, m_k)$ with a literal $p^1_i(y^1_i)$. The earlier found unifier gives us the particular string of initial variables which are in this occurrence of y^1_i. Denote the received from $A_{k,j}(\overline{x}_{k,j})$ formula by means of $A^1_{k,j}(\overline{x}^1_{k,j})$.

For $l = 1, \ldots, L-1$ repeat this procedure for $A^l_{k,1}(\overline{x}^l_{k,1}), \ldots, A^l_{k,m_k}(\overline{x}^l_{k,m_k})$ and receive formulas $A^{l+1}_{k,1}(\overline{x}^{l+1}_{k,1}), \ldots, A^{l+1}_{k,m_k}(\overline{x}^{l+1}_{k,m_k})$. The process will end because the number of literals in $A^l_{k,j}(\overline{x}^l_{k,j})$ decreases with increasing l.

Level description of formulas $A_{k,1}(\overline{x}_{k,1}), \ldots, A_{k,m_k}(\overline{x}_{k,m_k})$ has the following form [8].

$$\begin{cases} & A^L_{k,j}(\overline{x}^L_{k,j}) & (j = 1, \ldots, m_k) \\ p^1_1(y^1_1) & \Leftrightarrow & P^1_1(\overline{y}^1_1) \\ & \vdots & \\ p^1_{n_1}(y^1_{n_1}) & \Leftrightarrow & P^1_{n_1}(\overline{y}^1_{n_1}) \\ & \vdots & \\ p^l_i(y^l_i) & \Leftrightarrow & P^l_i(\overline{y}^l_i) \\ & \vdots & \\ p^L_{n_L}(y^L_{n_L}) & \Leftrightarrow & P^L_{n_L}(\overline{y}^L_{n_L}) \end{cases} . \quad (5)$$

The described procedure deals with such intuitive notions as "frequently occurred" and having "small complexity" sub-formula. In the following algorithms it is supposed that "frequently occurred" means that a sub-formula appears at least twice and "small complexity" means that the number of a sub-formula literals is less then such a number of at least one of the initial formulas.

3.1 Construction of a Level Description of Classes

An algorithm for construction of a level description of classes is presented in [19], [23]. This algorithm contains two parts: 1) extraction of common up to the names of arguments sub-formula from the class description, 2) forming a level description. Here it is presented with more accuracy.

1) Extraction of common up to the names of arguments sub-formulas from the class description

Let the description of a class be $A(\overline{x}) = A_1(\overline{x}_1) \vee \cdots \vee A_K(\overline{x}_K)$, where $A_1(\overline{x}_1), \ldots, A_K(\overline{x}_K)$ are elementary conjunctions of literals.

• For every i and j $(i < j)$ check whether $A_i(\overline{x}_i) \Rightarrow_P \exists \overline{x}_{j_{\neq}} A_j(\overline{x}_j)$.

Using the notion of partial sequence we can obtain a maximal common up to the names of arguments sub-formula $Q^1_{i,j}(\overline{z}^1_{i,j})$ of two elementary conjunctions $A_i(\overline{x}_i)$ and $A_j(\overline{x}_j)$ and unifiers $\lambda_{Q^1_{i,j},A_i}$ and $\lambda_{Q^1_{i,j},A_j}$. Let s_i be the lists of indices of $Q^l_{s_i}(\overline{z}^l_{s_i})$.

• For $l = 1, \ldots, L-1$ do.

• For every s_i and s_j $(s_i \neq s_j)$ do.

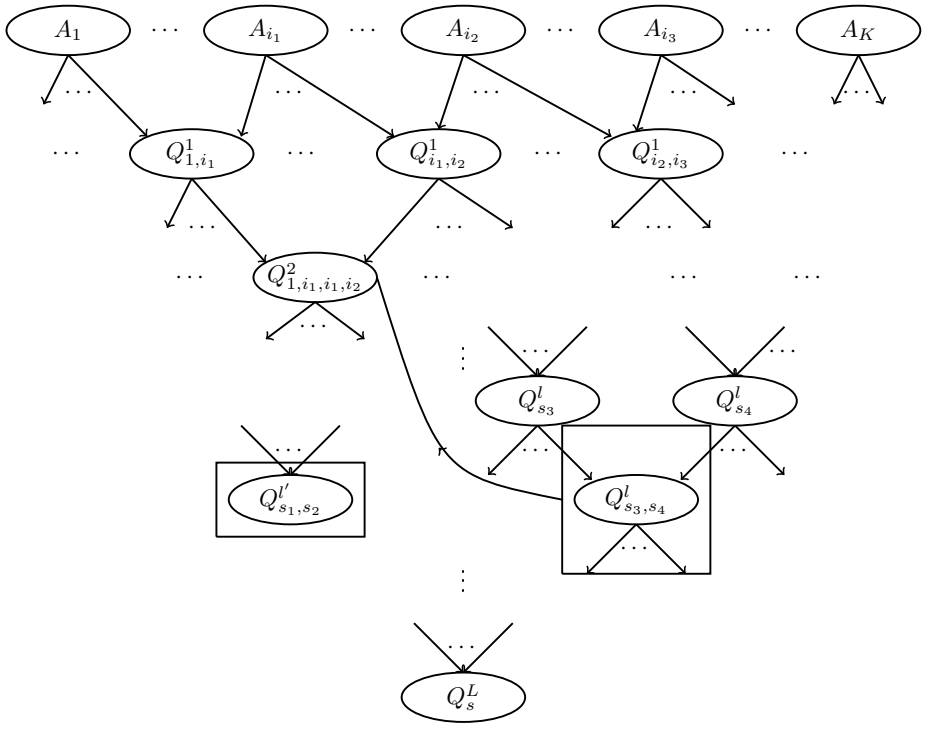


Figure 1: Scheme of common up to the names of arguments sub-formulas extraction.

– In the similar way find a maximal common up to the names of arguments sub-formula $Q^{l+1}_{s_i,s_j}(\overline{z}^{l+1}_{s_i,s_j})$ of the formulas $Q^l_{s_i}(\overline{z}^l_{s_i})$ and $Q^l_{s_j}(\overline{z}^l_{s_j})$ and unifiers $\lambda_{Q^{l+1}_{s_i,s_j},Q^l_{s_i}}$ and $\lambda_{Q^{l+1}_{s_i,s_j},Q^l_{s_j}}$.

– If a formula $Q^{l+1}_{s_i,s_j}(\overline{z}^{l+1}_{s_i,s_j})$ is isomorphic to some previously obtained $Q^{l'}_s(\overline{z}^{l'}_s)$ for some s and $l' \le l+1$ then it is deleted from the set of the $(l+1)$th level formulas and we wright down unifiers $\lambda_{Q^{l'}_s,Q^l_{s_i}}$ and $\lambda_{Q^{l'}_s,Q^l_{s_j}}$.

Note that the length of $Q^{l+1}_{s_i,s_j}(\overline{z}^{l+1}_{s_i,s_j})$ decreases with increasing l. That is why the process would stop.

A scheme of the extraction is presented in Figure 1. In order not to overload the scheme, the arguments of the formulas are not written on it and instead of $Q^l_{s_i}(\overline{z}^l_{s_i})$ it is written $Q^l_{s_i}$.

In this scheme it is supposed that highlighted together with its connections formula $Q^l_{s_3,s_4}$ is isomorphic to some previously obtained formula, for example, to Q^2_{1,i_1,i_1,i_2} and that's why it is deleted from the set of the lth level formulas and all its connections are rewrited to Q^2_{1,i_1,i_1,i_2}.

A formula $Q^{l'}_{s_1,s_2}$ that does not have a common sub-formula with any other formulas, is highlighted on it.

2) Forming a level description.

Further, we assume that $Q^0_i(\overline{z}^0_i)$ is $A_i(\overline{x}_{k,i})$.

• Every formula of the type $Q^l_s(\overline{z}^l_s)$ which has no sub-formula of the type $Q^{l+1}_{s,s_j}(\overline{z}^{l+1}_{s,s_j})$ for any s_j is declared as $P^1_i(\overline{y}^1_i)$ $(i = 1, \dots, n_1)$.[4]

A new 1st level predicate defined as $p^1_i(y^1_i) \Leftrightarrow P^1_i(\overline{y}^1_i)$ is introduced and $p^1_i(y^1_i)$ is substituted (using the correspondent unifies) instead of $P^1_i(\overline{y}^1_i)$ into all formulas of the type $Q^l_s(\overline{z}^l_s)$ $(l \ge 0)$ such that there is an oriented path from $Q^l_s(\overline{z}^l_s)$ to $P^1_i(\overline{y}^1_i)$.

• For $l = 1, \dots, L-1$ do.

For every $i = 1, \dots, n_l$ if $P^l_i(\overline{y}^l_i)$ is a maximal common up to the names of arguments sub-formula of formulas $Q^{l'}_{s_{i'}}(\overline{z}^{l'}_{s_{i'}})$ and $Q^{l'}_{s_{j'}}(\overline{z}^{l'}_{s_{j'}})$ then these formulas are declared as $P^{l+1}_{i''}(\overline{y}^{l+1}_{i''})$ and $P^{l+1}_{i''+1}(\overline{y}^{l+1}_{i''+1})$ $(i'' = 1, \dots, n_{l+1} - 1)$.

New $(l+1)$th level predicates $p^{l+1}_{i''}(y^{l+1}_{i''})$ and $p^{l+1}_{i''+1}(y^{l+1}_{i''})$ defined as $p^{l+1}_{i''}(y^{l+1}_{i''}) \Leftrightarrow P^{l+1}_{i''}(\overline{y}^{l+1}_{i''})$ and $p^{l+1}_{i''+1}(y^{l+1}_{i''+1}) \Leftrightarrow P^{l+1}_{i''+1}(\overline{y}^{l+1}_{i''+1})$ are introduced and every of them is substituted (using the correspondent unifies) instead of $P^{l+1}_{i''}(\overline{y}^{l+1}_{i''})$ and $P^{l+1}_{i''+1}(\overline{y}^{l+1}_{i''+1})$, respectively, into all formulas of the type $Q^{l'}_s$ such that there is an oriented path from Q^l_s to at least one of them.

If $P^l_i(\overline{y}^l_i)$ corresponds to the cell marked as Q^2_{1,i_1,i_1,i_2} in Figure 1 then $p^{l+1}_{i''}(y^{l+1}_{i''})$ is substituted instead of $P^{l+1}_{i''}(\overline{y}^{l+1}_{i''})$ into the formulas corresponding the cells marked as Q^1_{1,i_1}, A_1 and A_{i_1}. The literal $p^{l+1}_{i''+1}(y^{l+1}_{i''+1})$ is substituted instead of $P^{l+1}_{i''+1}(\overline{y}^{l+1}_{i''+1})$ into the formulas corresponding the cells marked as $Q^1_{i_1,i_2}$, A_{i_1} and A_{i_2}.

The words "using the correspondent unifies" may be explained in such a way. If $P^l_i(\overline{y}^l_i)$ corresponds to the cell marked as Q^2_{1,i_1,i_1,i_2} in Figure 1 then $p^{l+1}_{i''}(y^{l+1}_{i''})$ is substituted instead of $P^{l+1}_{i''}(\overline{y}^{l+1}_{i''})$ into the formulas corresponding the cells marked as Q^1_{1,i_1}, A_1 and A_{i_1}. The literal $p^{l+1}_{i''+1}(y^{l+1}_{i''+1})$ is substituted instead of $P^{l+1}_{i''+1}(\overline{y}^{l+1}_{i''+1})$ into the formulas corresponding the cells marked as $Q^1_{i_1,i_2}$, A_{i_1} and A_{i_2}.

It may be easily proved that the problem of level description

[4] These are formulas Q^L_s and $Q^{l'}_{s_1,s_2}$ in Figure 1.

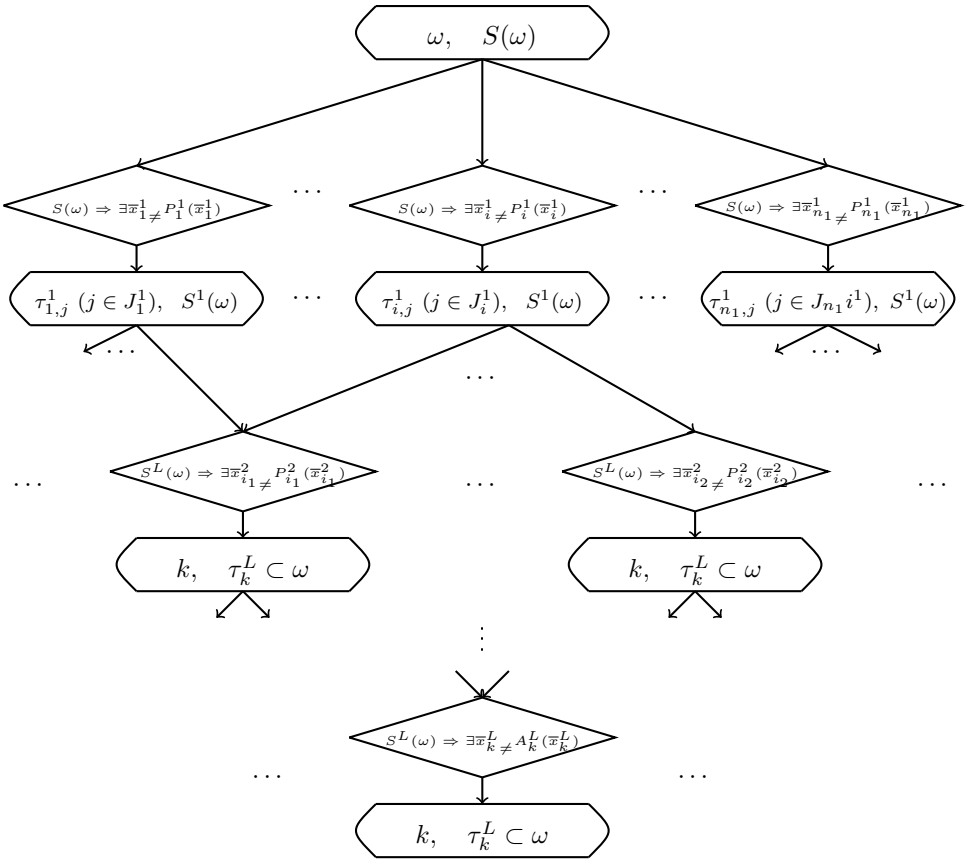


Figure 2: Scheme of level recognition.

construction is NP-hard. In particular, the problem of checking if there exists a level description is NP-complete and the upper bounds of number of steps for traditional algorithms exponentially depends on parameters of $A_1(\overline{x}_1), \ldots, A_K(\overline{x}_K)$. At the same time such a description must be constructed only once and then may be used as many times as you like.

3.2 The Use of Level Description of a Class

Checking for what k the sequence $S(\omega) \Rightarrow \exists \overline{x}_{k \neq} A_k(\overline{x}_k)$ is valid and if "yes" then finding for what value of $\overline{x}_k$ it is valid with the use of level description of a class, may be done according to the following algorithm.

• For $i = 1, \ldots, n_1$ check $S(\omega) \Rightarrow \exists \overline{y}^1_{i \neq} P^1_i(\overline{y}^1_i)$ and find all lists of values $\overline{\tau}^1_{i,j}$ for the variables $\overline{y}^1_i$ such that $S(\omega) \Rightarrow P^1_i(\overline{\tau}^1_{i,j})$.[5]

Let $\tau^1_{i,j}$ be a notation for the list $\overline{\tau}^1_{i,j}$. All atomic formulas of the form $p^1_i(\tau^1_{i,j})$ with $\tau^1_{i,j}$ a notation for the list $\overline{\tau}^1_{i,j}$ are added to $S(\omega)$ to obtain $S^1(\omega)$ – a 1st level description of ω.

• For $l = 1, \ldots, L-1$ do.

– For $i = 1, \ldots, n_l$ check $S^l(\omega) \Rightarrow \exists \overline{y}^l_{i \neq} P^l_i(\overline{y}^l_i)$ and find all lists of values $\overline{\tau}^l_{i,j}$ for the variables $\overline{y}^l_i$ such that $S^l(\omega) \Rightarrow P^l_i(\overline{\tau}^l_{i,j})$.

– All atomic formulas of the form $p^l_i(\tau^l_{i,j})$ with $\tau^l_{i,j}$ a notation for the list $\overline{\tau}^l_{i,j}$ are added to $S^l(\omega)$ to obtain $S^{1+1}(\omega)$ – a lth level description of ω.

• – For $k = 1, \ldots, K$ check $S^L(\omega) \Rightarrow \exists \overline{y}^L_{k \neq} A^L_k(\overline{y}^L_k)$ and find all lists of values $\overline{\tau}^L_{k,j}$ for the variables $\overline{y}^L_k$ such that $S^L(\omega) \Rightarrow A^L_k(\overline{\tau}^L_{k,j})$.

The object ω satisfies such formulas $A_k(\overline{x}_k)$ for which the corresponding sequence $S^L(\omega) \Rightarrow \exists \overline{y}^L_{k \neq} A^L_k(\overline{y}^L_k)$ is valid.

The scheme of level recognition is presented in Figure 2.

It is proved in [8] that for a two-level description ($L = 1$) the number of steps with the use of 2-level description decreases. For an exhaustive algorithm the number of steps decreases from $O(t^{m_k})$ to $O(n_1 \cdot t^r + t^{\delta^1_k + n_1})$, where r is a maximal number of arguments in the 1st level predicates, n_1 is the number of such predicates, δ^1_k is the number of initial variables which are presented in $A_k(\overline{x}_k)$ and are absent in $A^1_k(\overline{x}^1_k)$. Both parameters r and $\delta^1_k + n_1$ are less than m_k.

Number of steps for an algorithm based on the derivation in predicate calculus decreases from $O(s_k{}^{a_k})$ to $O(s^{1 a^1_k} + \sum_{j=1}^{n_1} s^{\rho^1_j})$, where a_k and a^1_k are maximal numbers of literals in $A_k(\overline{x}_k)$ and $A^1_k(\overline{x}^1_k)$, respectively, s and s^1 are the numbers of literals in $S(\omega)$ and $S^1(\omega)$, respectively, ρ^1_j is the numbers of literals in $P^1_i(\overline{y}^1_i)$. Both parameters a^1_k and ρ^1_j are less than a_k.

[5]Note that for every predicate P^1_i there may be many lists of values from ω satisfying it.

So, the recognition of an object with the use of a level description requires less time than that with the use of initial description.

4 Logic-Predicate Network

The inputs of a traditional neuron network are binary or many-valued characteristics of an object. The neuron network configuration is fixed that does not correspond to biological networks.

Below a logic-predicate network is described. Such a network inputs are atomic formulas setting properties of the elements composing an investigated object and relations between them [17], [19]. Further it will be shown that its configuration may be changed during its retraining.

A level description of a class of objects permits quickly enough to recognize an object which description is isomorphic to the description of an element of a training set.

A logic-predicate network consists of two blocks: a training block and a recognition block. Training block forms a level description of a class of objects according to a training set (or some formalized description of the class). Its content is presented in Figure 1.

Recognition block implements the algorithm of the use of level description of a class. It must be marked that the level description may be represented as an oriented graph presented in Figure 2.

Scheme of logic-predicate network is presented in Figure 3. Here inputs and outputs are situated in ovals and checking the conditions are situated in rhombus.

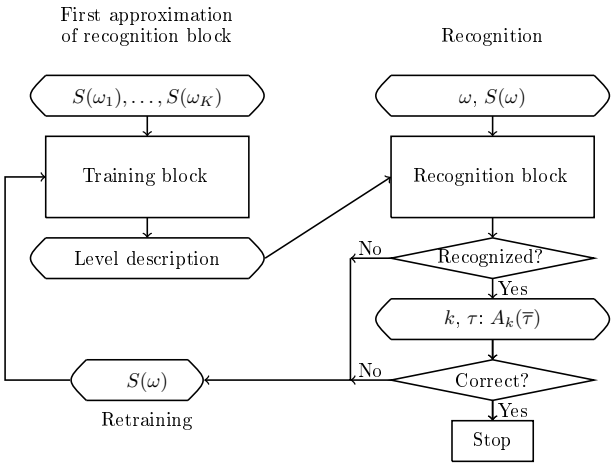


Figure 3: Scheme of logic-predicate network.

Let a training set of objects $\omega^1, \ldots, \omega^K$ be given. An initial variant of the class description may be formed by replacement of every constant ω_j^k in $S(\omega^k)$ by a variable x_j^k ($k = 1, \ldots, K$, $j = 1, \ldots, t^k$) and substitution of the sign & between the atomic formulas. Construct a level description for these goal formulas. The first approximation to the recognition block is formed.

If after the recognition block run an object is not recognized or has wrong classification, then it is possible to retrain the network. The description of the "wrong" object must be added to the input set of the training block. The training block extracts common up to the names sub-formulas of this description and previously received formulas forming a new recognition block. Some sub-formulas, the number of layers and the number of formulas in every layer may be changed in the level description. Then the recognition block is reconstructed.

5 Example of Logic-Predicate Network Construction, its Running and Retraining

Let a set of contour images of "boxes" be to recognize. Every image may be described in the terms of two predicates V and L defined as it is presented in Figure 4.

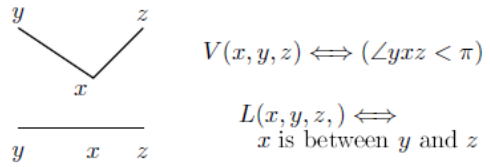


Figure 4: Initial predicates.

A training set $\{\omega_1, \omega_2, \omega_3\}$ of such images is presented in Figure 5. Here $\omega_1 = \{a_1, \ldots, a_{10}\}$, $\omega_2 = \{b_1, \ldots, b_8\}$, $\omega_3 = \{c_1, \ldots, c_{10}\}$.

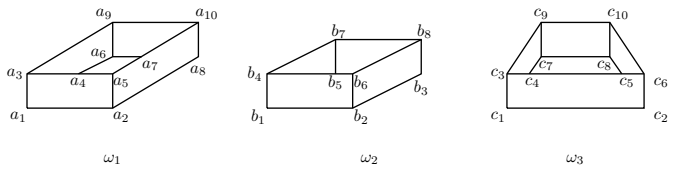


Figure 5: Training set.

In order not to overload the text, we write down only the formula $A_2(y_1, \ldots, y_8)$, which is obtained from the description $S(\omega_2)$ by replacing the names b_i with the variables y_i and placing the sign & between the literals.

$A_2(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8) =$
$V(y_1, y_4, y_2)$ & $V(y_2, y_1, y_6)$ & $V(y_2, y_1, y_3)$ & $V(y_2, y_6, y_3)$ &
$V(y_3, y_2, y_8)$ & $V(y_4, y_7, y_6)$ & $V(y_4, y_7, y_5)$ & $V(y_4, y_7, y_1)$ &
$V(y_4, y_5, y_1)$ & $V(y_4, y_6, y_1)$ & $V(y_5, y_4, y_7)$ & $V(y_5, y_4, y_6)$ &
$V(y_5, y_7, y_6)$ & $L(y_5, y_4, y_6)$ & $V(y_6, y_4, y_8)$ & $V(y_6, y_5, y_8)$ &
$V(y_6, y_8, y_2)$ & $V(y_6, y_2, y_5)$ & $V(y_6, y_2, y_4)$ & $V(y_7, y_4, y_5)$ &
$V(y_7, y_5, y_8)$ & $V(y_7, y_4, y_8)$ & $V(y_8, y_3, y_6)$ & $V(y_8, y_6, y_7)$ &
$V(y_8, y_3, y_7)$

Further, for clarity, we will write down not formulas, but drawings.

5.1 Training Block Run: Extraction of Sub-Formulas

For constructing a level description, find pairwise maximal common up to the names of arguments sub-formulas $Q_i^1(\overline{y}_i^1)$ of formulas $A_1(\overline{x}), A_2(\overline{y}), A_3(\overline{z})$ and their unifiers. See Figure 6.

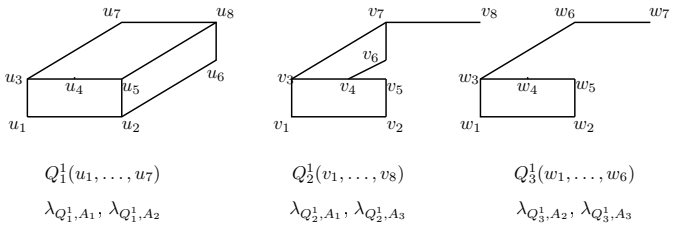


Figure 6: First extraction of pairwise maximal common up to the names of arguments sub-formulas.

Here the following unifier are formed.

Let $u^1 = (u_1, \ldots, u_8)$, $v^1 = (v_1, \ldots, v_8)$, $w^1 = (w_1, \ldots, w_7)$.

$\lambda_{Q_1^1,A_1} = (u_1 \to x_1, u_2 \to x_2, u_3 \to x_3, u_4 \to x_4, u_5 \to x_5, u_6 \to x_8, u_7 \to x_9, u_8 \to x_{10}) = (u^1 \to x_1^1)$,

$\lambda_{Q_1^1,A_2} = (u_1 \to y_1, u_2 \to y_2, u_3 \to y_4, u_4 \to y_5, u_5 \to y_6, u_6 \to y_3, u_7 \to y_7, u_8 \to y_8) = (u^1 \to y_1^1)$,

$\lambda_{Q_2^1,A_1} = (v_1 \to x_1, v_2 \to x_2, v_3 \to x_3, v_4 \to x_4, v_5 \to x_5, v_6 \to x_6, v_7 \to x_9, v_8 \to x_{10}) = (v^1 \to x_2^1)$,

$\lambda_{Q_2^1,A_3} = (v_1 \to z_1, v_2 \to z_2, v_3 \to z_3, v_4 \to z_4, v_5 \to z_6, v_6 \to z_7, v_7 \to z_9, v_8 \to z_{10}) = (v^1 \to z_1^1)$,

$\lambda_{Q_3^1,A_2} = (w_1 \to y_1, w_2 \to y_2, w_3 \to y_4, w_4 \to y_5, w_5 \to y_6, w_6 \to y_7, w_7 \to y_8) = (w^1 \to y_2^1)$,

$\lambda_{Q_3^1,A_3} = (w_1 \to z_1, w_2 \to z_2, w_3 \to z_3, w_4 \to z_4, w_5 \to z_6, w_6 \to z_9, w_7 \to z_{10}) = (w^1 \to z_2^1)$.

Pairwise extractions of maximal common up to the names of arguments sub-formulas of $Q_1^1(u_1, \ldots, u_8)$, $Q_2^1(v_1, \ldots, v_8)$ and $Q_3^1(w_1, \ldots, w_6)$ give only just obtained formula $Q_3^1(w_1, \ldots, w_6)$. That's why we have a situation presented in Figure 7.

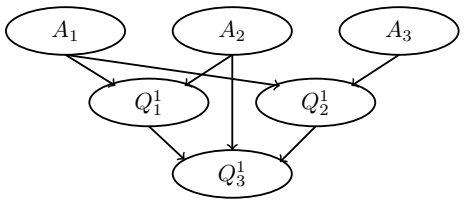


Figure 7: Connections between the extracted formulas.

5.2 Training Block Run: Forming a Level Description

The formula $Q_3^1(w_1, \ldots, w_7)$ has no common sub-formulas with the other ones. Therefore, it is the 1st level predicate $P_1^1(w_1, \ldots, w_7)$ of the network. A new 1st level predicate $p_1^1(w^1)$ defining as $p_1^1(w^1) \Leftrightarrow P_1^1(w_1, \ldots, w_7)$ is introduced. Remind that w^1 is a new 1st level variable for the string of initial variables $(w_1, w_2, w_3, w_4, w_5, w_6, w_7)$.

Formula $P_1^1(w_1, \ldots, w_7)$ (which corresponds to $Q_3^1(w_1, \ldots, w_7)$) is on the end of edges from $A_2(y_1, \ldots, y_8)$, $Q_1^1(u_1, \ldots, u_8)$ and $Q_2^1(v_1, \ldots, v_8)$ in Figure 7. The unifier of $P_1^1(w_1, \ldots, w_7)$ with $A_2(y_1, \ldots, y_8)$ coincides with $\lambda_{Q_3^1,A_2}$ obtained above.

The unifiers of $P_1^1(w_1, \ldots, w_7)$ with $Q_1^1(u_1, \ldots, u_8)$ and $Q_2^1(v_1, \ldots, v_8)$ are the following.

$\lambda_{P_1^1,Q_1^1} = (w^1 \to u^2)$, where $u^2 = (u_1, u_2, u_3, u_4, u_5, u_7, u_8)$.

$\lambda_{P_1^1,Q_2^1} = (w^1 \to v^2)$, where $v^2 = (v_1, v_2, v_3, v_4, v_5, v_7, v_8)$.

The formulas $Q_1^1(u_1, \ldots, u_8)$ and $Q_2^1(v_1, \ldots, v_8)$ with the substitutions of $p_1^1(u^1)$ and $p_1^1(v^1)$, respectively, instead of sub-formulas isomorphic to $P_1^1(w_1, \ldots, w_7)$ will form the 2nd level of the network and are denoted by $P_1^2(u_1, \ldots, u_8; u^1)$ and $P_2^2(v_1, \ldots, v_8; v^1)$.

It must be marked that if we know the value of the 1st level variable u^1 then we know the values of $(u_1, u_2, u_3, u_4, u_5, u_7, u_8)$. That's why the essential variables of $P_1^2(u_1, \ldots, u_8; u^1)$ are $(u_6; u^1)$. Similarly the essential variables of $P_2^2(v_1, \ldots, v_8; v^1)$ are $(v_6; v^1)$.

New 2nd level predicates $p_1^2(u^2)$ and $p_2^2(v^2)$ defining as $p_1^2(u^2) \Leftrightarrow P_1^2(u_1, \ldots, u_8; u^1)$ and $p_2^2(v^2) \Leftrightarrow P_2^2(v_1, \ldots, v_8; v^1)$ are introduced.

Taking into account only essential variables and earlier received unifiers we have the last layer formulas.

$A_{1,1}^2(x_7; x^1, x_1^2)$, where

$x^1 = (x_1, x_2, x_3, x_4, x_5, x_8, x_9, x_{10})$, $x_1^2 = (x_6; x^1)$.

$A_{1,2}^2(x_7; x^1, x_2^2)$, where

$x^1 = (x_1, x_2, x_3, x_4, x_5, x_8, x_9, x_{10})$, $x_2^2 = (x_6; x^1)$.

$A_{2,1}^2(; y^1, y_1^2)$, where

$y^1 = (y_1, y_2, y_4, y_5, y_6, y_7, y_8)$, $y_1^2 = (y_3, y_8; y^1)$.

$A_{2,2}^2(y_3 y_7; y^1)$, where

$y^1 = (y_1, y_2, y_4, y_5, y_6, y_7, y_8)$.

$A_3^2(z_5, z_8; z^1, z^2)$, where

$z^1 = (z_1, z_2, z_3, z_4, z_6, z_9, z_{10})$, $z^2 = (z_4, z_7; z^1)$.

The contents of the last layer cells are the following:

$A_{1,1}^2(\overline{x}; w^1, u^2) \vee A_{1,2}^2(\overline{x}; w^1, v^2)$,

$A_{2,1}^2(\overline{y}; w^1, u^2) \vee A_{1,2}^2(\overline{y}; w^1)$,

$A_3^2(\overline{z}; w^1, v^2)$.

5.3 Recognition Block Run

The recognition block is formed according to the level description constructed in the previous subsection. The scheme of recognition block is presented in Figure 8. In this scheme only names of a checking formula and results of checking are written down.

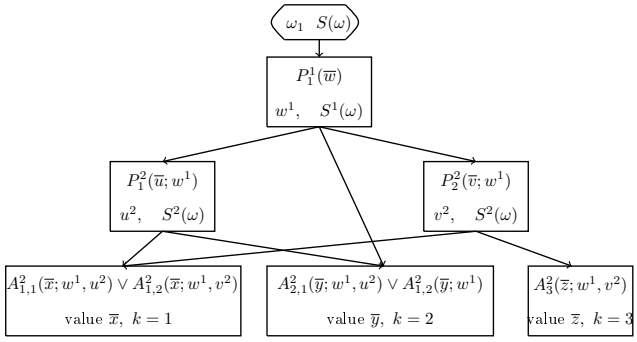


Figure 8: Predicate network.

It is obvious that if an input object ω is isomorphic to one from the training set, then it will be recognized. It means that the number k and such a permutation $\overline{\omega}$ of ω that $S(\omega) \Rightarrow A_k(\overline{\omega})$ will be found.

Let a control object with the description not isomorphic to anyone from the training set is to be recognized. Such a control object $\{d_1, \ldots, d_8\}$ is presented in Figure 9.

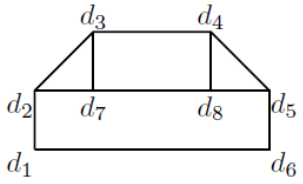


Figure 9: Control object.

Checking $S(\omega) \Rightarrow \exists \overline{w}_{\neq} P_1^1(\overline{w})$ in the first layer (see Figure 8) gives us $\overline{w} = (d_1, d_6, d_2, d_7, d_5, d_3, d_4)$. A new 1st level variable w^1 takes this list of constants as a value. Add new atomic formula to $S(\omega)$ and obtain $S^1(\omega) = S(\omega) \cup \{p_1^1(w^1)\}$.

Neither $P_1^2(\overline{u}, w^1)$ or $P_2^2(\overline{v}, w^1)$ are a consequence of $S^1(\omega)$ because, in fact, 7 variables in $\overline{u}$ and $\overline{v}$ are changed by constants from w^1.

This control object is not recognized.

5.4 Retraining the Network

Let's retrain the network with the object presented in Figure 9.

Replace every constant d_i $(i = 1, \ldots, 8)$ in its description by a variable x_i and substitute the sign & between the atomic formulas. The formula $A_4(\overline{x})$ is obtained.

Pairwise checking of partial sequence between $A_4(\overline{x})$ and $A_j(\overline{x}_j)$ $(j = 1, 2, 3)$ permits to obtain maximal common up to the names of arguments formulas Q_4^1, Q_5^1 and Q_6^1. The images of thees fragments are shown in Figure 10.

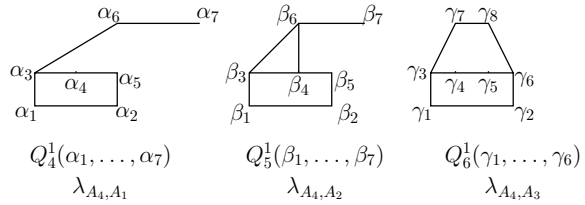


Figure 10: Pairwise extraction of maximal common up to the names of arguments sub-formulas between $A_4(\overline{x}_4)$ and $A_j(\overline{x}_j)$ $(j = 1, 2, 3)$.

One of the obtained formulas (formula $Q_4^1(\alpha_1, \ldots, \alpha_7)$) is isomorphic to the earlier obtained formula $Q_3^1(w_1, \ldots, w_7)$. That's why it would not be added to the 2nd layer of the retrained network. But all its connections with the cells of another layers would be saved.

Pairwise checking of partial sequence between Q_5^1, Q_6^1 and $Q_j^1(\overline{x}_j^1)$ $(j = 1, 2, 3)$ does not give any new formulas.

So, the 1st layer would have only the formula $Q_3^1(w_1, \ldots, w_7)$ renamed as $P_1^1(w_1, \ldots, w_7)$.

On the 2nd layer the cell with formula $P_4^2(\overline{\beta}; w^1)$ and output t^2 and the cell with formula $P_5^2(\overline{\gamma}; w^1)$ and output r^2 would be added on the end of the edge from the cell with $P_1^1(\overline{w})$. Edges from these cells to the cells with $A_j^2(\overline{x}_j; x_j^1, x_j^2)$ $(j = 1, 2, 3)$ would be added as it is shown in Figure 11.

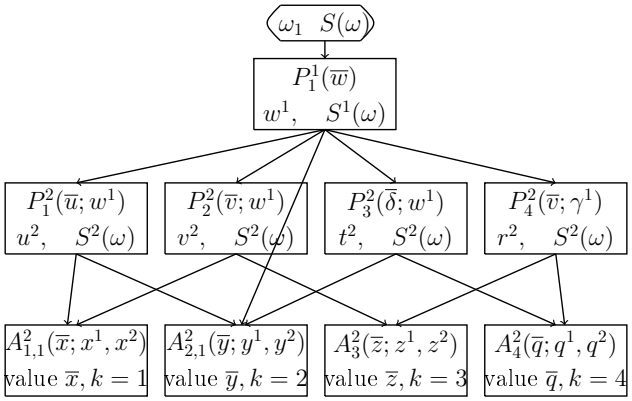


Figure 11: Pairwise extraction of maximal common up to the names of arguments sub-formulas between $A_4(\overline{x}_4)$ and $A_j(\overline{x}_j)$ $(j = 1, 2, 3)$.

The network is retrained.

It should be noted that the examples of images discussed in the article are taken from [3]. The original predicates W, Y and T in that book for describing the images differ from the predicates V and L used here. This is due to the fact that W, Y and T are expressed in terms of the predicates V and L and do not significantly reduce the computational complexity of recognition when constructing a level description.

Both $\{W, Y, T\}$ and $\{V, L\}$ are not enough to describe real images of "boxes". So, for example, the description of the image in Figure 12 is isomorphic to the description of $A_2(y_1, \ldots, y_8)$, but you need to have a rich imagination to guess the box in it. The choice

of initial predicates in each specific problem should be chosen very carefully.

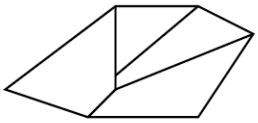


Figure 12: Is it a box?

Here we do not consider an example when, while retraining the network, not only additional cells in a layer appear, but the number of layers also changes. This could be done by retraining the network using the example of a mirror (relative to the vertical axis) image of an object from the training set. In this case, rectangle recognition will appear in the 1st layer. The contents of the remaining cells will also be changed.

6 Fuzzy Recognition by a Logic-Predicate Network

The disadvantage is that the proposed logic-predicate network does not recognize objects with descriptions that are not isomorphic to the descriptions that participate in its formation. This can be eliminated by slight modifications of the network.

Further the *i*th cell of the *l*th level will be denoted as $cell_i^l$.

Let's change the contents of the network cells presented in Figure 2 by replacing the checking of $S^{l-1}(\omega) \Rightarrow \exists\overline{x}_{i\neq}^l P_i^l(\overline{x}_i^l)$ in the $cell_i^l$ with the **partial** sequence checking $S^{l-1}(\omega) \Rightarrow_P \exists\overline{x}_{i\neq}^l P_i^l(\overline{x}_i^l)$. The degree of coincidence of a recognized object with the description of a class and the degree of recognition correctness certainty is calculated during partial sequence checking.

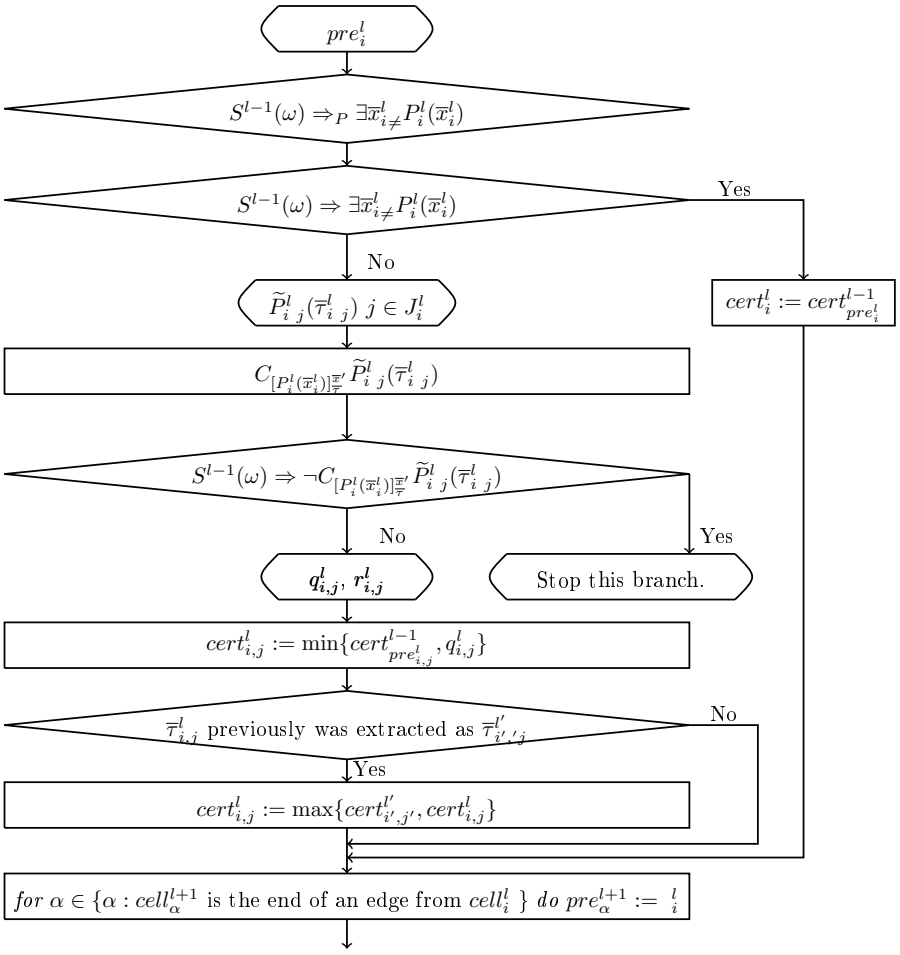


Figure 13: The fuzzy network *i*th cell of the *l*th level contents.

For a fully identified object, both of these degrees equal to 1. For an object from a given class for which these degrees are essentially small (this is determined by an expert), you can still retrain the network.

A fuzzy network has two additional parameters:

$cert_i^l$ ($l = 0, \dots, L$, $i = 1, \dots, n_l$) – a **degree of certainty in the correctness of recognition** by the $cell_i^l$;

pre_i^l ($l = 1, \dots, L$, $i = 1, \dots, n_l$) – the **cell number of a cell preceding** the $cell_i^l$ in the current traversal of a network. The cell number has the form $\frac{l'}{j}$, where l' is the number of a layer and j is the number of a cell in this layer

Initially, $cert_1^0 := 1$, $pre_i^1 := \frac{0}{1}$ ($i = 1, \dots, n_1$).

The scheme of $cell_i^l$ contents of a fuzzy network is presented in Figure 13.

First of all, a partial sequence $S^{l-1}(\omega) \Rightarrow_P \exists \overline{x}_{i\neq}^l P_i^l(\overline{x}_i^l)$ is checked in this cell. If this sequence is not partial but total, then the parameter $cert_i^l$ is not changed and the calculation goes to the end of the cell.

Otherwise, all lists of constants $\overline{\tau}_{i\,j}^l$ and maximal not contradictory on $\overline{\tau}_{i\,j}^l$ with $S^{l-1}(\omega)$ sub-formulas $\widetilde{P}_{i\,j}^l(\overline{x}_{i\,j}^l)$ of the formula $P_i^l(\overline{x}_i^l)$ are found while partial sequence checking.

If the sub-formulas $\widetilde{P}_{i\,j}^l(\overline{x}_{i\,j}^l)$ is contradictory to the description $S(\omega)$ on the list of the found constants, then this recognition branch stops.

Remind that $S(\omega)$ and $C_{[P_i^l(\overline{x}_i^l)]_{\overline{\tau}}^{\overline{\tau}'}} \widetilde{P}_{i\,j}^l(\overline{\tau}_{i\,j}^l)$ are in the contradiction means that $S(\omega) \Rightarrow \neg C_{[P_i^l(\overline{x}_i^l)]_{\overline{\tau}}^{\overline{\tau}'}} \widetilde{P}_{i\,j}^l(\overline{\tau}_{i\,j}^l)$, where $C_{[P_i^l(\overline{x}_i^l)]_{\overline{\tau}}^{\overline{\tau}'}} \widetilde{P}_{i\,j}^l(\overline{\tau}_{i\,j}^l)$ is a complement (conjunction of literals from $P_i^l(\overline{x}_i^l)$ which are not in $\widetilde{P}_{i\,j}^l(\overline{x}_{i\,j}^l)$) of $\widetilde{P}_{i\,j}^l(\overline{x}_{i\,j}^l)$ up to $P_i^l(\overline{x}_i^l)$.[6]

Parameters $q_{i\,j}^l$ and $r_{i\,j}^l$ [7] are calculated with the use of the full form of formulas $P_i^l(\overline{x}_i^l)$ and $\widetilde{P}_i^l(\overline{y}_i^l)$, i.e., with the replacement in them of each atomic formula of the levels l' ($l' < l$) by the defining elementary conjunction.

Except this, a degree of certainty $cert_i^l$ that the recognition would be valid is calculated in every cell. While first visit to the $cell_i^l$ the value of $cert_i^l$ will be $cert_i^l := \min\{cert_{pre_i^l}^{1-1}, q_i^l\}$. This corresponds to conjunction of degrees of certainty while successive passage along one branch of network traversal.

While next visit of the $cell_i^l$ the value of $cert_i^l$ will be $cert_i^l := \max\{cert_i^l, \min\{cert_{pre_i^l}^{1-1}, q_i^l\}\}$. This corresponds to disjunction of degrees of certainty while parallel passage along different branches of network traversal.

Why can we visit a cell not only once? For example, in Figure 11 the 3rd layer cell with A_1^2 may be visited by a path including the cells with P_1^1, P_1^2 and A_1^2 and by a path including the cells with P_1^1, P_2^2 and A_1^2.

In the last layer of the network, objects τ_k^L would be received and for each object the degree of certainty q_k^L, that it is the r_k^Lth part of an object satisfying the description $A_k(\overline{x}_k)$, would be calculated.

A flowchart of fuzzy recognition network can not be presented in the paper because it is very large. It consists of the "Scheme of logic-predicate network" shown in Figure 3, in which instead of the "Training block" the "Scheme of common up to the names of arguments sub-formulas extraction" (Figure 1) is inserted, and the "Scheme of level recognition" (Figure 2) is inserted instead of the "Recognition block". At the same time, in Figure 2, each element enclosed in a rhombus is replaced by "The fuzzy network *i*th cell of the *l*th level contents" (Figure 13).

7 Model Example of Fuzzy Recognition

Let's look how a fuzzy network (corresponding to that constructed in the section 6 and presented in Figure 8) recognizes the control object presented in Figure 9.

To recognize it, a partial sequence $S(d_1, d_2, \dots, d_8) \Rightarrow_P \exists w_1 \dots w_6 P_1^1(w_1, \dots, w_6)$ is checked in the **1st layer**. There is a total sequence and the 1st level variable $w^1 = (w_1, w_2, w_3, w_4, w_5, w_6, w_7)$ takes the value $(d_1, d_6, d_2, d_7, d_5, d_3, d_4)$.

$cert_1^1 := 1$, $pre_1^2 := \frac{1}{1}$, $pre_2^2 := \frac{1}{1}$, $pre_2^3 := \frac{1}{1}$.

The result of the 1st layer fuzzy cell run is presented in Figure 14.[8]

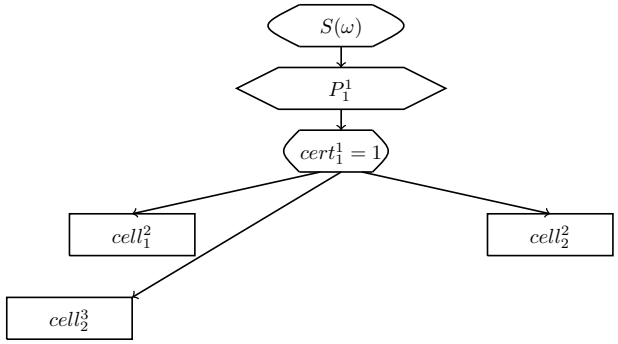


Figure 14: The result of the 1st layer fuzzy cell run.

In the **2nd layer**, partial sequence $S^1(d_1, d_2, \dots, d_8, w^1) \Rightarrow_P \exists u_1 \dots u_{7\neq} P_1^2(u_1 \dots u_8; w^1)$ is checked. It must be noted that in fact $S^1(d_1, d_2, \dots, d_8, w^1) \Rightarrow_P \exists u_6 P_1^2(d_1, d_6, d_2, d_7, d_5, u_6, d_3, d_4; w^1)$ is checked. That is, there is only one variable at the right-hand side of the formula.

Since we need to check formulas for consistency, we note that $V(x, y, z) \Rightarrow \neg V(x, z, y)$.

There is only a partial sequence. A formula $\tilde{P}_1^2(d_1, d_6, d_2, d_7, d_5, d_3, d_4; w^1)$ is the maximal sub-formula such that $S^1(d_1, d_2, \dots, d_8, w^1) \Rightarrow \tilde{P}_1^2(d_1, d_6, d_2, d_7, d_5, d_3, d_4; w^1)$. Its complement up to $P_1^2(d_1, d_6, d_2, d_7, d_5, d_3, d_4, u_6; w^1)$ contains three literals with predicate V, one of them is $V(u_6, d_4, d_6)$. This contradicts $S^1(d_1, d_2, \dots, d_8, w^1)$ containing $V(u_6, d_6, d_4)$. This branch of the network stops.

There is another cell in the 2nd layer on the end of an edge from $cell_1^1$. This is $cell_2^2$.

Check partial sequence $S^1(d_1, d_2, \dots, d_8, w^1) \Rightarrow_P \exists v_1 \dots v_8 P_2^2(v_1 \dots v_8; w^1)$. In fact, $S^1(d_1, d_2, \dots, d_8, w^1) \Rightarrow_P \exists v_6 P_2^2(d_1, d_6, d_2, d_7, d_5, v_6, d_3, d_4; w^1)$ is checked.

[6] See the definitions of a complement and contradictory formulas in the section 2.

[7] Here $q_{i\,j}^l$ and $r_{i\,j}^l$ are ratios of the number of literals and variables, respectively, in $\overline{P}_{i\,j}^l(\overline{x}_{i\,j}^l)$ and $P_i^l(\overline{x}_i^l)$. See definition of (q, r) fragment in the section 2.

[8] Not to overload the following schemes, only the name of the formula is written in the cell instead of the formula of corresponding partial seguence.

There is only partial sequence. A formula $\tilde{P}_2^2(d_1, d_6, d_2, d_7, d_5, v_6, d_3, d_4; w^1)$ is the maximal sub-formula such that $S^1(d_1, d_2, \ldots, d_8, w^1) \Rightarrow \tilde{P}_2^2(d_1, d_6, d_2, d_7, d_5, v_6, d_3, d_4; w^1)$. Its complement up to $P_2^2(d_1, d_6, d_2, d_7, d_5, v_6, d_3, d_4; w^1)$ contains 5 literals with predicate V

$V(v_6, d_7, d_3)$, $V(d_3, v_6, d_2)$, $V(d_7, v_6, d_2)$, $V(d_7, d_8, v_6)$, $V(d_7, d_5, v_6)$

and 3 literals with predicate L

$L(d_7, d_2, d_8)$, $L(d_8, d_2, d_5)$, $L(d_8, d_7, d_5)$, but no one of them is in contradiction with $S^1(d_1, d_2, \ldots, d_8, w^1)$. The value of v^2 is $(v_6; w^1)$. It means that one 1st level variable v_6 does not take value.

The number of literals in $P_2^2(v_1 \ldots v_8; w^1)$ (taking into account its full form in which the 1st level predicate is changed by the defining formula only with initial variables) is $a_2^2 = 16$. The number of literals in $\tilde{P}_2^2(d_1, d_6, d_2, d_7, d_5, v_6, d_3, d_4; w^1)$ is $\tilde{a}_2^2 = 8$.

That's why $q_2^2 = \frac{8}{16} = \frac{1}{2}$ and $cert_2^2 := \min\{1, \frac{1}{2}\} = \frac{1}{2}$.

There are two edges from the cell with P_2^2 to the 3rd layer cells with $A_{1,2}^2$ and with A_3^2. That's why $pre_1^3 := \frac{2}{2}$, $pre_3^3 := \frac{2}{2}$.

The result of the 2nd layer fuzzy cells run is presented in Figure 15.

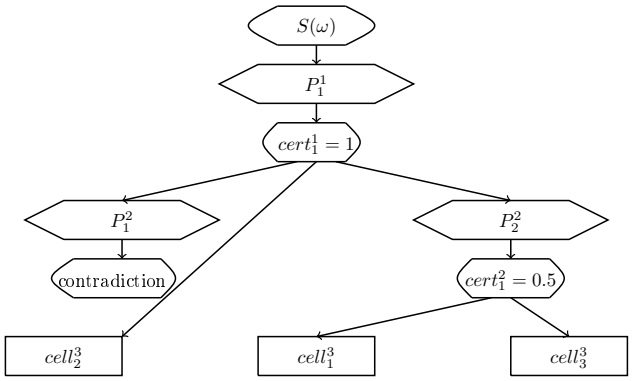


Figure 15: The result of the 2nd layer fuzzy cell run.

In the **3rd layer** while going from the cell with P_1^1 to the $cell_2^3$ with $A_{2,1}^2$ partial sequence $S^1(d_1, \ldots, d_8; w^1) \Rightarrow_P \exists y_1 \ldots y_{8\neq} A_{2,1}^2(y_1 \ldots y_8; v^1)$ is checked. In fact, a partial sequence $S^1(d_1, \ldots, d_8; w^1) \Rightarrow_P \exists y_3 A_{2,1}^2(d_1, d_6, y_3, d_2, d_7, d_5, d_3, d_4; v^1)$ with the only one initial variable y_3 is checked.

There is only a partial sequence. The same string of values for variables in v^2, as it was in the second layer, partially satisfies $A_{1,2}^2(d_1, d_6, y_3, d_2, d_7, d_5, d_3, d_4; v^1)$. But the description of this fragment has two additional literals $V(d_7, d_2, d_3)$ and $V(d_7, d_3, d_5)$.

The number of literals in $A_2(y_1 \ldots y_8; v^1)$ is $a_2 = 22$. The number of literals in $\tilde{A}_{1,2}^2(y_1 \ldots y_8; v^1, v^2)$ is $\tilde{a}_{1,2}^2 = 10$. That's why $q_2^3 = \frac{10}{22} \approx 0.455$ and $cert_2^3 := \min\{1, \frac{10}{22}\} \approx 0.455$.

While going from the cell with P_2^2 to the $cell_1^3$ with $A_{1,2}^2$ partial sequence $S^2(d_1, d_2, \ldots, d_8; v^1, v^2) \Rightarrow_P \exists v_1 \ldots v_{10\neq} A_{1,2}^2(v_1 \ldots v_{10}; v^1, v^2)$ is checked. In fact $S^2(d_1, d_2, \ldots, d_8; v^1, v^2) \Rightarrow_P \exists v_6, v_7, v_{8\neq} A_{1,2}^2(d_1, d_6, d_2, d_7, d_5, v_6, v_7, v_8, d_3, d_4; v^1, v^2)$ with 3 initial variables v_6, v_7, v_8 is checked.

There is only a partial sequence. The same string of values for variables in v^2, as it was in the second layer, partially satisfies $A_{1,2}^2(v_1 \ldots v_{10}; v^1, v^2)$.

The number of literals in $A_1(x_1, \ldots, x_{10})$ is $a_1 = 34$. The number of literals in $\tilde{A}_{1,2}^2(v_1 \ldots v_{10}; v^1, v^2)$ is $\tilde{a}_{1,2}^2 = 8$. That's why $q_1^3 = \frac{8}{34} \approx 0.235$ and $cert_1^3 := \min\{1, \frac{8}{34}\} \approx 0.235$.

While going from the cell with P_2^2 to the $cell_2^3$ with A_3^2 partial sequence $S^2(d_1, d_2, \ldots, d_8; v^1, v^2) \Rightarrow_P \exists z_1 \ldots z_{10\neq} A_3^2(z_1 \ldots z_{10}; v^1, v^2)$ is checked. In fact $S^2(d_1, d_2, \ldots, d_8; v^1, v^2) \Rightarrow_P \exists z_7, z_{8\neq} A_3^2(d_1, d_6, d_2, d_7, d_8, d_5, z_7, z_8, d_3, d_4; v^1, v^2)$ with 2 initial variables z_7, z_8 is checked.

There is only a partial sequence. The variables z_7, z_8 have not taken values. All literals with these variables are absent in $\tilde{A}_3^2(d_1, d_6, d_2, d_7, d_8, d_5, z_7, z_8, d_3, d_4; v^1, v^2)$. The number of such literals is 14.

The number of literals in $A_3^2(z_1 \ldots z_{10}$ is $a_3 = 32$. The number of literals in $\tilde{A}_3^2(z_1 \ldots z_{10}; v^1, v^2)$ is $\tilde{a}_3^2 = 32 - 14 = 18$. That's why $q_2^3 = \frac{18}{32} \approx 0.563$ and $cert_2^3 := \min\{1, \frac{18}{32}\} \approx 0.563$.

The full degree of certainty that it is a representative of the class of "boxes" that is given as a control image is $cert = \max\{\frac{8}{34}, \frac{10}{22}, \frac{18}{32}\} = \frac{18}{32} \approx 0.563$.

Moreover, $\frac{7}{8}$ elements of this image coincide with $\frac{8}{10}$ elements of the image ω_3 from the training set.

The result of the fuzzy network run is presented in Figure 16.

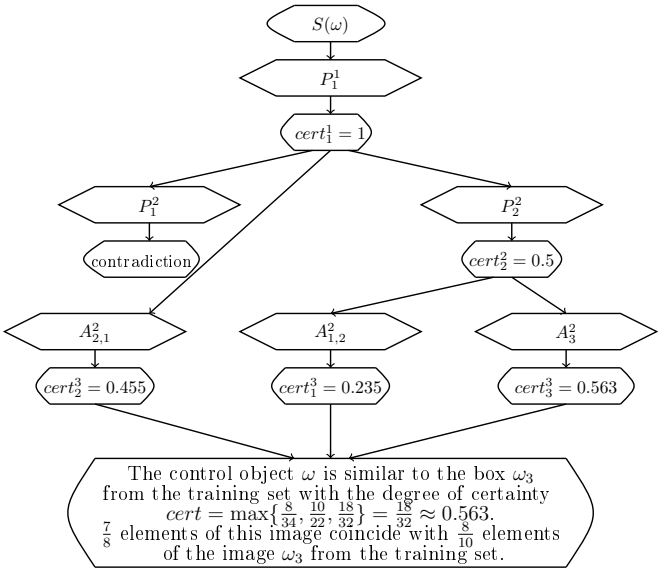


Figure 16: The result of the fuzzy network run.

8 Results and Discussion

The main result of this article is the proposal of fuzzy recognition of complex structured objects by logic-predicate network. For a detailed presentation of this result, it was necessary to describe in detail a logic-predicate approach to the recognition of complex structured objects.

A significant disadvantage of this approach is that the problems arising are NP-hard, which to a large extent "scares off" some researchers. On the other hand, the use of predicate logic does

not always seem legitimate, since the propositional logic is much simpler for understanding by non-mathematicians.

The use of level descriptions is not something essentially original to decrease the computational complexity of the main problem. It is a well-known technique in which sufficiently complex problems are divided into a sequence of problems of the same type of lower dimension. The main problem was HOW to build such a description. To this end, the mathematical apparatus for extracting the general sub-formulas of two elementary conjunctions of predicate formulas was developed and briefly presented in this paper.

Based on the level description of classes, it was possible to formulate the concept of a logic-predicate network. Unlike existing networks, for example, neural or Bayesian, when retraining such a network, a logic-predicate network can (and, as a rule, this happens) change its configuration. This corresponds to the training of a person in whose brain connections between one neural cells break and connections between others arise.

This is one of the essential features of the proposed approach, which distinguishes it from the use of various other networks. Its other essential feature is to calculate the degree of coincidence of the object being recognized with those ones for which the network was trained.

A trained logic-predicate network quickly recognizes objects with descriptions that are isomorphic to the ones on which it was trained. But its training and retraining processes remain to be NP-hard. This corresponds to the process of teaching a person. But in some cases, it is not necessary to accurately recognize an object or process, but to find out what it looks like and to what extent.

To answer such a question, the concept of a fuzzy logic-predicate network was proposed. The structure of its cell, shown in Figure 13, is quite complex and contains a partial sequence check. Such verification is NP-hard, but after repeated retraining of the network, the formulas $P_i^l(\overline{x}_i^l)$, setting intermediate goal formulas, are quite short and have a small number of essential arguments, as it was shown by the example in section 7. This means that an exponent of the computational complexity of the partial sequence check is a fairly small value.

Special importance has the choice of "good" initial features. As already mentioned in Subsection 5.4, the predicates W, Y and T, which were used to describe contour images in [3], do not give a good decreasing of computational complexity for the logic-predicate network, constructed from them. Apparently, this is due to the fact that they reflect the perception of a person and are essentially composite.

9 Conclusion

In development of the proposed approach, many-valued predicates may be considered. Especially this relates to the properties of an object elements. Such multi-valued predicates $p(x) = a$ with $a \in D$ can be interpreted as binary ones with a numerical argument: $q(x; a) \Leftrightarrow p(x) = a$. The inclusion of such predicates in the formulation of the problem may serve as the subject of further research.

It is also planned to consider the possibility of introducing cell weights in a logic-predicate network. Such weights can be determined depending on the frequency with which parts of an object description satisfy the formula checked in the corresponding cell. Other weighting strategies are possible, similar to strategies for determining the cell weights of a classical artificial neural network.

In addition, if in the problem setting global features of the object that characterize the entire object are considered, then such an object is usually described by a binary or multi-valued string. Class descriptions can be presented as a propositional formula in disjunctive normal form. Extraction the "frequently encountered" sub-formulas of such propositional formulas [24] allows to form a level description of classes. Using this description in a manner similar to that described in this article, it is possible to build a logic-algebraic network, which changes its configuration in the process of learning. This corresponds to a change of connections in biological neural networks.

Acknowledgments The author is grateful to Vakhtang Kvaratskhelia from Muskhelishvili Institute of Computational Mathematics of the Georgian Technical University, whose question at the conference in 2017 allowed this paper to be born. The author thanks the anonymous referees for their useful suggestions.

References

[1] T. Kosovskaya . "Implementation of Formula Partial Sequence for Rough Solution of AI Problems in the Framework of the Logic-Predicate Approach" in 2019 Computer Science and Information Technologies (CSIT),Yerevan, Armenia, 2019. https://doi.org/10.1109/CSITechnol.2019.8895153

[2] N. J. Nilson, Problem-solving methods in Artificial Intelligence, McGraw-Hill Book Company, 1971.

[3] R. Duda, P. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, 1973.

[4] E.B. Hunt, Artificial Intelligence, Academic Press, 1975.

[5] M.R. Garey, D.S Jonson, Computers and Intractability: A Gide to the Theory of NP-Completeness, W.H. Freeman and Company, 1979.

[6] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Third edition, Prentice Hall Press Upper Saddle River, 2009.

[7] T.M. Kosovskaya. "Estimating the number of steps that it takes to solve some problems of pattern recognition which admit logical description" in VESTNIK ST PETERSBURG UNIVERSITY-MATHEMATICS, **40**(4), 82–90, 2007. (In Russian) https://doi.org/10.3103/S1063454107040061

[8] T.M. Kosovskaya, "Multilevel descriptions of classes for decreasing of step number of solving of a pattern recognition problem described by predicate calculus formulas" VESTNIK SANKT-PETERBURGSKOGO UNIVERSITETA SERIYA 10 PRIKLADNAYA MATEMATIKA INFORMATIKA PROTSESSY, **1**, 64–72 (2008). (In Russian)

[9] T.M. Kosovskaya, "Partial deduction of predicate formula as an instrument for recognition of an object with incomplete description" VESTNIK SANKT-PETERBURGSKOGO UNIVERSITETA SERIYA 10 PRIKLADNAYA MATEMATIKA INFORMATIKA PROTSESSY, **1**, 74–84, 2009. (In Russian) https://doi.org/10.24412/FgeTsEutG6w

[10] T.M. Kosovskaya, "Partial Deduction in Predicate Calculus as a Tool for Artificial Intelligence Problem Complexity Decreasing" in 2015 IEEE 7th International Conference on Intelligent Computing and Information Systems, ICICIS 2015, 73–76, 2016. https://doi.org/10.1109/IntelCIS.2015.7397199

[11] J. Schmidhuber, "Deep learning in neural networks: An overview" Neural Networks, **61**, 85–117, 2015. https://doi.org/10.1016/j.neunet.2014.09.003

[12] T. Chunwei, X. Yong, Z. Wangmeng, "Image denoising using deep CNN with batch renormalization" Neural Networks, **121**, 461–473, 2020. https://doi.org/10.1016/j.neunet.2019.08.022

[13] E.H. Dawn, C. J. Lakhmi (Eds), Innovations in Bayesian Networks. Theory and Applications, Springer-Verlag, 2008.https://doi.org/10.1007/978-3-540-85066-3

[14] R. Agrahari, A. Foroushani, T.R. Docking et al. “Applications of Bayesian network models in predicting types of hematological malignancies”, Sci Rep 8, 6951 (2018). https://doi.org/10.1038/s41598-018-24758-5

[15] A. Tulupiev, C. Nikolenko, A. Sirotkin, Bayesian Networks: logic-probabilistic approach, Nauka, 2006. (in Russian)

[16] A. Shahzad, R. Asif, A. Zeeshan, N. Hamad, A. Tauqeer, “Underwater Resurrection Routing Synergy using Astucious Energy Pods” Journal of Robotics and Control (JRC), **1**(5), 173–184, 2020. https://doi.org/10.18196/jrc.1535

[17] T. Kosovskaya, “Self-modificated predicate networks” International Journal on Information Theory and Applications, **22**(3), 245–257, 2015.

[18] T. Kosovskaya, “Fuzzy logic-predicate network” in Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019), Part of series ASUM (Atlanties Studies in Uncertainty Modeling), **1**, 9–13, 2019. https://doi.org/10.2991/eusflat-19.2019.2

[19] T. Kosovskaya, “Predicate Calculus as a Tool for AI Problems Solution: Algorithms and Their Complexity”, 1–20. 2018. Chapter 3 in “Intelligent System”, open access peer-reviewed Edited volume. Edited by Chatchawal Wongchoosuk Kasetsart University https://doi.org/10.5772/intechopen.72765

[20] T. M. Kosovskaya, N. N. Kosovskii, “Polynomial Equivalence of the Problems Predicate Formulas Isomorphism and Graph Isomorphism” VESTNIK ST PETERSBURG UNIVERSITY-MATHEMATICS, **52**(3), 286–292, 2019. https://doi.org/10.1134/S1063454119030105

[21] T.M. Kosovskaya, “Some artificial intelligence problems permitting formalization by means of predicate calculus language and upper bounds of their solution steps” SPIIRAS Proceedings, **14**, 58–75, 2010. (In Russian) https://doi.org/10.15622/sp.14.4

[22] T. M. Kosovskaya, D. A. Petrov, “Extraction of a maximal common sub-formula of predicate formulas for the solving of some Artificial Intelligence problems” VESTNIK SANKT-PETERBURGSKOGO UNIVERSITETA SERIYA 10 PRIKLADNAYA MATEMATIKA INFORMATIKA PROTSESSY, **3**, 250–263, 2017. (In Russian) https://doi.org/10.21638/11701/spbu10.2017.303

[23] T. Kosovskaya, “Construction of Class Level Description for Efficient Recognition of a Complex Object” International Journal “Information Content and Processing”, **1**(1), 92–99, 2014.

[24] D.A. Koba, “Algorithm for constructing a multi-level description in pattern recognition problems and estimating the number of its run steps” in Proceedings of the XVII International School-Seminar “Synthesis and Complexity of Control Systems” named after Acad. O.B. Lupanov (Novosibirsk, October 27 - November 1, 2008), Publishing House of the Institute of Mathematics, Novosibirsk, 55–59, 2008. (In Russian)

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 689-698 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Hand-Based Biometric Recognition Technique - Survey

Katerina Prihodova*, Miloslav Hub

Faculty of Economics and Administration, University of Pardubice, 532 10, Czech Republic

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 01 November, 2020
Online: 24 November, 2020

Keywords:
Hand geometry
Biometric recognition
Hand-base authentication

ABSTRACT

In today's modern world, reliable automatic personal recognition is a crucial area of discussion, primarily because of the increased security risks. A large number of systems first require the recognition of a person before they can access their services. Biometric recognition can be used, which can be understood as automatic identification or automatic verification of persons based on physiological or behavioural characteristics. Examples of biometric characteristics may be the fingerprint, face, signature, hand geometry. There are many approaches, and each has its pros and cons. However, there is currently no current extensive research and evaluation of hand-based biometric systems. Given the user-friendliness of these systems and the needs of society, this article aims to compare different methods of biometric recognition based on hand, so that the article reduces the entry barrier into this area of research. Furthermore, the article aims to determine the research gap and suggest possible directions for research in the future.

1. Introduction

This paper is an extension of work originally presented in International Conference on Information and Digital Technologies [1]. In today's modern society, there is an increasing emphasis on working with information. Information is one of the most valuable assets we can own, so it needs to be protected. Among the basic requirements for their security are their confidentiality, integrity and availability. The requirement of confidentiality information requires that only authorized entities have access to it. One way to ensure the confidentiality of information is through authentication, during which the identity of the entity to which the relevant rights are subsequently assigned is verified [2].

Authentication can be defined as: “A method of verifying a user's identity on a system to control access to the system or resources.” [3] Authentication can be performed using knowledge (password, pin), an authentication object (token, mobile phone), or using what we are physically and mentally (biometric characteristics). The last option is used by biometric recognition systems, which have come to the fore, especially in recent years, thanks to scientific and technical progress.

Biometric systems are applications of biometric technologies that allow automatic identification or verification of a specific natural person, using physicavl (e.g. face, fingerprint) or behavioural (e.g. voice, gait) characteristics. Biometric characteristics must be universal (the characteristics must exist for all persons), they must have high external variability (high diversity for different persons) and low internal variability (low diversity for one person).

Identification ("One-To-Many Matching") is understood in this article as the process of comparing a set of biometric features obtained from a scanned biometric sample with all reference templates stored in a database. If there is a reference template in the database that sufficiently corresponds to the set of biometric features, the biometric system identifies the person by the identity associated with this reference template, otherwise the person is not identified [4].

Verification ("One-To-One Matching") is understood in this article as the process of verifying that the person being examined is who he or she claims to be. The set of biometric features obtained from the scanned sample is compared with one own biometric template, which is stored in a database. If the set of biometric features sufficiently matches this reference template, the identity is verified, otherwise it is not verified [4].

Biometric recognition of people is based on their biometric characteristics, and it is their unique physiological and behavioural characteristics. Physiological characteristics are those characteristics with which we were born and include, e.g. iris [5, 6], retina [7], face [8], shape of the outer ear [9], fingerprints [10], palmprints [11] and footprints [12], fingers geometry [13], hand geometry, [14], topography (layout and shape) of wrist veins

*Corresponding Author: Katerina Prihodova, st30351@student.upce.cz

[15], DNA [16]. Behavioural characteristics include, e.g. voice [17], writing [18], signature [19], and gait [20].

From the biometric characteristics, distinguishing repeatable biometric features can be obtained for biometric recognition. Biometric features are numbers or designations obtained from biometric samples that are used for subsequent comparison. Biometric samples are an analogue or digital representation of biometric characteristics before the actual extraction of biometric features [21].

The article aims to provide a comprehensive survey with the inclusion of most recent research papers up to 2020, covering a total of 117 journal publications, conference proceeding, patents, and papers. The aim is also lower the entry barrier to this research area by providing a comprehensive reference for novices. Offer a wide range of comparisons in diverse research angles and perspectives in hand geometry recognition. Identify current trends in this area and the need for future research.

2. Development of Hand-Based Biometric Systems

The first modern biometric systems used the hand geometry as a biometric characteristic, and then it was an authentication system in the verification mode (the user submitted an identity and it was subsequently confirmed or rejected). Robert Miller dealt with the issue of hand-based biometric systems in the 1960s and based on research, and he created a device at the Stanford Research Institute, which he patented in 1971. This device performed the pattern recognition by purely mechanical means, where four spring bars slid towards the user's fingers and subsequently the length of the user's fingers was measured. If the finger lengths matched the pattern embossed on the user's identity card, a simple switch was activated, which was used to operate the electric lock [4, 22].

The first commercial hand scanner was the Identimat, based on the patent Robert Miller's as mentioned above [22]. The measurement of the hand was performed by scanning the photocells located under grooves on which the user placed the hand (fingers). The light source was light with a power of 1000 watts. A reading head with a magnetic stripe was attached to this scanning mechanism, to which a user's magnetic card was attached during verification. If the pulses on the card matched the signal times from the photocell sensors within the threshold value, the user's identity was verified. This system was successfully used in various applications. The first use was of a military nature, and it was the security of nuclear weapons. Later, Identimat was used to access US nuclear power plants and also to control employee attendance at an investment company on Wall Street. The second development line of Identimat was already two-dimensionally oriented, measuring not only the length but also the width of all fingers on one hand [4, 23].

The history of the practical implementation of hand geometry recognition is further associated with the name David Sidlauskas and his company Recognition System Inc., which he founded in 1986. This company was the first in the world to start developing and selling modern hand scanners. In the mid-1980s, David Sidlauskas patented the first biometric system to use a three-dimensional hand image. The patented system consists of a camera and an optical measuring board. The camera has a plan view and a side view of the hand. The identification code is entered using the keypad [24]. Based on this patent, the first commercially successful biometric system using the hand geometry of the Handkey ID3D scanner was created.

BioMet Partners, Inc. developed a two-finger version of the Digi-2 hand scanner and was launched in 1995. The scanner uses a CCD camera to scan only the index finger and middle finger and obtain their 3D image [4].

In the 1990s, biometric hand scanners were used in the INSPASS program, a program at selected US airports that accelerated the check-in of frequent travellers. These persons could apply for inclusion in a program in which their hand geometry was scanned and then obtained an identification card with a reference template. These persons were able to complete check-in quickly at selected airports with the help of installed hand scanners [25].

Another hand-scanner recognition project was an Israeli project that automated motion control on the Israeli-Palestinian border. During the creation of the biometric system, suitable biometric characteristics were selected. Biometric characteristics face, hand geometry and fingerprints were considered. A multi-biometric system was chosen that combined face recognition and hand geometry, which was chosen mainly because most people crossing the border are workers whose fingerprints are not reliable for biometric verification [26].

In terms of security, hand scanners were also used at the 1996 Atlanta Olympics, which controlled access to the Olympic Village. Hand scanners were also used by the Colombian legislature, which voted safely through them [27]. Until 1998, only two scientific articles in the field of hand-based biometric systems were published, namely: "Vital Signs of Identity" a "A Performance Evaluation of Biometric Identification Devices." [2, 23, 28] From 1998 to 2006, there was an almost exponential increase in the number of publications in this area [29].

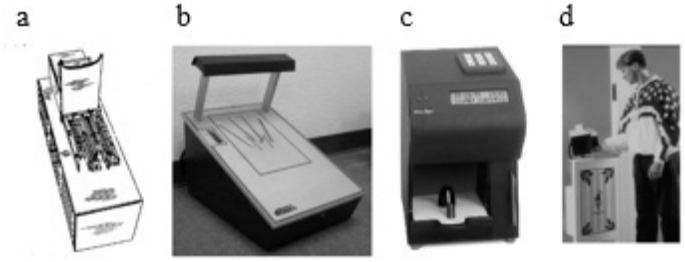


Figure 1: a) Drawing of mechanical hand reader [22] b) Identimat [30] c) two-finger scanner from Biomet [31] d) utilization hand geometry scanner at the Atlanta Olympics [32]

3. Current methods in hand-based biometric systems

Current hand-based biometric systems can be divided into the following groups according to biometric characteristics [33]:

- hand geometry and handshape,
- palmprint,
- fingerprint,
- topography (finger, hand).

At the beginning of the research of hand-based biometric systems, hand geometry was used, e.g. widths, lengths, angles [35-37], later the silhouette of the hand was added as another biometric characteristic [38-40].

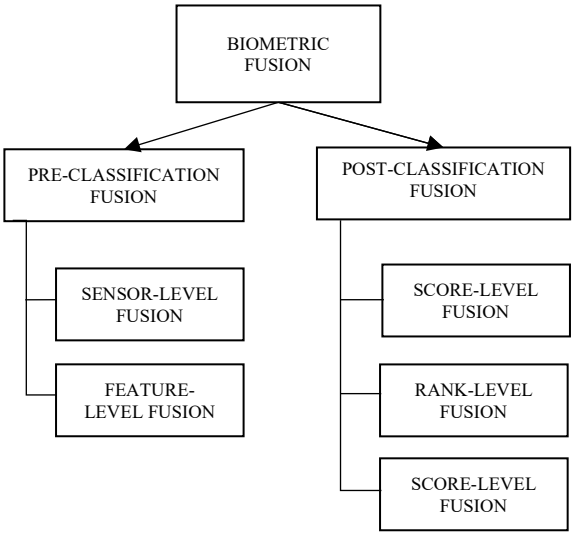


Figure 2: Fusion at various levels of multi-biometric systems (source: [34])

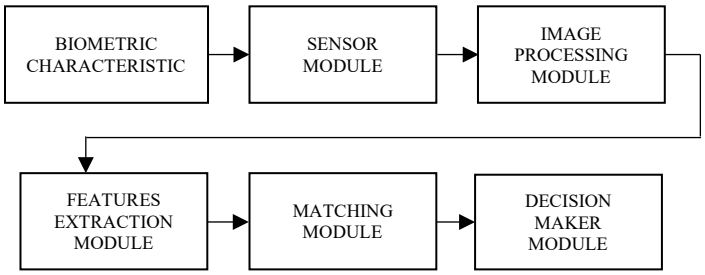


Figure 3: General model of biometric system

Currently, there are also multi-biometric systems that use a combination of multiple biometric characteristics. A palmprint [41-45] or bloodstream topography [46-48] is used in combination with hand geometry or silhouette. The topography of bloodstream is measured either on the palm or back of the hand when the whole hand is scanned. Systems have also been developed that combine more than two characteristics, such as hand geometry, palmprint, and fingerprint [49, 50]. Multi-biometric systems achieve higher accuracy [51] than biometric systems using only one biometric characteristic. Newly emerging biometric systems also use as a characteristic the heartbeat captured on the finger [52], pores on the hands [53], or hand gestures [44].

Combining multiple biometric characteristics in one biometric system requires fusion, which is the interconnection of information. Fusion can be achieved at different levels of the biometric system. Fusion levels can be divided into two large groups: pre-classification fusion and post-classification fusion. Pre-classification fusion include sensor-level fusion and feature-level fusion while post-classification fusion includes score-level fusion, rank-level fusion rank and decision-level fusion [34]. The individual fusion levels are shown in Figure 2.

The biometric recognition system has several modules: the sensor captures the biometric characteristic, followed by image processing, extraction of identification features, comparison with the biometric template and decision. The individual modules of the biometric system are shown in its general model in Figure 3.

3.1. Sensor

CCD chip cameras, a light-sensitive image detector used by digital cameras, camcorders, and webcams, e.g. are most commonly used to capture biometric characteristics [24, 53]. The scanner can also be a biometric sensor [42, 49, 54]. The sensor captures either a 2D [55-57] or a 3D [48, 58, 59] image of the hand. 3D images are obtained using two cameras, with the help of mirrors, or they can be obtained using a 3D digitizer [60].

Lighting also plays an essential role in obtaining a biometric sample. Some systems have their own light source, mainly commercial biometric scanners. Most sensors acquire images of the hand in visible light, but some also use infrared radiation [47, 61], which achieves a higher accuracy of biometric recognition [62].

Today, commercially available hand readers use spacer pins that firmly determine the position of the hand when scanning [48, 58, 63]. As spacer pins can deform the silhouette of the hand [64], systems without spacer pins have come to the fore in scientific circles [53, 65, 66]. The latest possibilities are contactless systems [38, 61, 67, 68], where, in addition to the spacer pins, direct contact with the reader is not required. These systems are considered to be the most user-friendly, mainly due to the elimination of the need for physical contact.

3.2. Image processing

After obtaining the image of the hand, it is necessary to process the acquired image. The first step in image processing is image pre-processing. It is a set of methods that are applied to an image because only a minimal percentage of hand images are taken under optimal conditions. Almost no captured image is completely noise-free, not entirely focused, obtained in optimal lighting conditions. For this reason, filters are used to sharpen the image or to remove noise. It is also possible to change the brightness using geometric transformation methods. [69]

The second step of image processing is image segmentation, i.e. dividing the image into segments with common properties. The aim of segmentation in hand-based biometric systems is to separate the hand from the rest of the image automatically. The more hand the image is obtained in more limited conditions, the easier the segmentation and can be done by faster methods. Segmentation methods can be divided into groups [69]:

- statistic methods,
- methods based on edge detection,
- hybrid methods (morphological),
- knowledge methods.

If the hand image is acquired on a scanner, the background is constant, and segmentation is relatively simple, then the segmentation method of image thresholding can be used. Image thresholding is the simplest statistical segmentation method based on pixel brightness evaluation. During thresholding, problems arise with rings, bracelets, or dirty skin [70-72].

Segmentation of the hand image can also be performed using edge detection methods. Edges are detected by edge detectors based on significant differences in the values of neighbouring pixels [37, 73].

There are also multi-stage segmentation methods where multiple methods are combined. These multi-stage methods can detect a hand from a crowded background. An example of such a method is a method that first detects the skin by colour (knowledge) and then detects the shape of the hand [74].

Another method of segmentation and at the same time localization of biometric features is AAM (Active Appearance Model) [55]. The AAM method presents the hand using an appearance model, which is created from a shape model and a hand texture model based on a statistical method - principal component analysis (PCA). The advantage of the method is the complexity of the information it contains about the hand [75].

The disadvantage of some segmentation methods is that their use breaks the silhouette of the hand. Mathematical morphology is a mathematical tool that can solve this problem and connect disconnected fingers. Mathematical morphology is also used to remove rings, bracelets or watches [56, 76, 77].

If the position of the hand is not fixed when acquiring the image, it is necessary to align the image of the hand after segmentation within the image processing. This can be done using landmarks, which can be extremes (fingertips, finger valleys) [36, 57, 78], of the centre of gravity, or a fixed reference point on the wrist [79].

Image processing is performed to facilitate the extraction of the region of interest. Biometric features are then extracted from the area of interest.

3.3. Extraction of biometric features

The image processing is followed by the extraction of biometric features. Extraction of biometric features is defined in [21] as*: "A process applied to a biometric sample to isolate and obtain a repeatable output of characteristic numbers or designations that can be compared with features extracted from other biometric samples."* For the biometric characteristic of hand geometry, biometric features are geometric measurements, such as finger lengths and widths, finger area, radii of circles on fingers and palms, and angles between fingers. Figure 4 shows 17 geometric measurements, but some systems use a different number of measurements, most often 13-50 measurements [35, 49, 80].

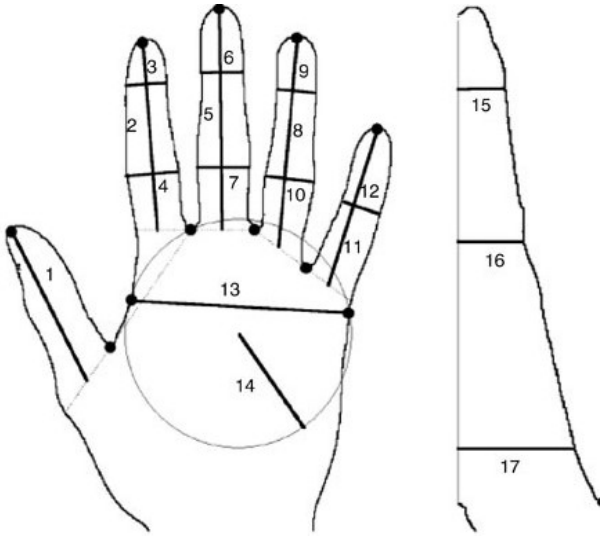


Figure 4: Geometric measurements (source: [81])

More information than hand measurements is provided by the silhouette of the hand, which is used by many systems [40,70]. These systems use biometric features to extract generic or customized feature extractors. For example, customize shape extractors [40, 82, 83] or figure features extractors algorithms Scale Invariant Feature Transform (SIFT) [82], LBP [84]. They are also used palm-based textures [85]. The silhouette is also distinguishable by its contour pixels; it represents the real shape of the hand [67].

Contours are less explored than geometric features, especially since geometric features are easier to determine from a normalized hand image. The issue of normalization of various hand positions and accurate localization of fingertips and finger valleys in unrestricted and contactless systems requires further research. [67]

Some systems extract biometric features of hand geometry and hand shape at the same time, extractors are specially adapted to the shape and geometry of the hand [86, 79].

Recent research uses deep neural networks, such as convolutional neural networks [87, 88] or Siamese networks, to extract biometric features of the hand [87].

3.4. Comparison with the reference template

In this phase, a set of biometric features obtained from a recognized person's biometric sample is compared with a reference template or templates (stored in a database). The score is then generated using the following methods:

- statistic methods,
  - minimum distance (Euclidean metric, Hamming metric, Chebyshev metric),
  - Kullback–Leibler divergence,
  - correlation coefficient,
  - Gaussian mixture models (GMM),
- machine learning,
  - support vectors machine (SVM),
  - k-nearest neighbours,
  - Bayesian networks,
  - neural networks.

Statistic methods

Statistical methods of comparing a set of biometric features with a reference template when recognizing a person based on the geometry and shape of the hand include a comparison by a minimum distance. If M reference templates are stored in the database x_{1R}, x_{2R}, …, x_{MR}, the minimum distance of the hand biometric sample x from the reference template is calculated as follows:

$$\min_{\forall s} \|x_{sR} - x\| \quad (1)$$

x_{sR}…… reference templates

s …… number of reference templates

x …… biometric features set (vector)

When recognizing a person based on hand geometry and hand shape, various metrics are used to determine the distance, such as the Euclidean metric [37, 56, 70], the Hamming metric [39], or the Chebyshev metric [80].

- Euclidean metric is defined by the relation:

$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad (2)$$

- Hamming's metric, sometimes called Manhattan, is defined by relations:

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n}|x_{1i} - x_{2i}| \quad (3)$$

- Chebyshev's metric is defined by the relation:

$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i}|x_{1i} - x_{2i}| \quad (4)$$

The Kullback – Leibler divergence can also be used to express "distance", which expresses the degree of dissimilarity between two probability distributions [36]. The correlation coefficient can also be used as a measure of similarity [81, 89].

Gaussian mixture models are located at the interface of statistical and machine methods (GMM). It is a pattern recognition method that is based on the GMM's ability to approximate any probability density using a mixture of Gaussian densities. The probability of the vector of characteristics x_n is estimated as a weighted sum of Gaussian densities. The result is, therefore, the probability with which the pattern (hand image) belongs to a given class. This method gives better results in the field of recognizing a person based on hand geometry than methods working with distance [37, 59].

The Gaussian mixture model is defined by the relation:

$$p(\vec{x}|u) = \sum_{i=1}^{M} \frac{c_i}{(2\pi)^{L/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x} - \overrightarrow{u_i})^T {\Sigma_i}^{-1}(\vec{x} - \overrightarrow{u_i})\right\} \quad (5)$$

u ……identity of the person to whom the hand image is assigned with probability p

x ……. vector of biometric features

c_i …… weight of individual Gaussian models

μ_i …… average vector model

Σ_i …… covariance matrix

M …… number of models

L ……. dimension vector of biometric features

Machine learning

Machine learning is a group of algorithms that allow a computer system to learn. Mittchell [90] d defined machine learning as follows: *"A computer program is said to learn from experience E concerning some classes of T and the measured power P, if its performance P improves with experience E."*

A large and used group of machine learning are algorithms called support vectors or support vector machines (SVM). These are classifiers that use boundaries in image space. SVM aims to find a superstructure that divides the symptom space. SVM used Vinodkumar and Srikantaswamy as a classifier in a multi-biometric system, where the biometric characteristics were hand geometry, fingerprints and palmprint [91]; however, they have also been used in many other hand-based multi-biometric systems [41, 42, 49, 92].

A frequently used machine learning method in hand-based biometric systems is the nearest neighbour (k-NN) method [54, 93]. It is one of the simplest classification methods. Bayesian networks [93, 94] or random forest [84] can also be used for comparison with the pattern.

Some types of neural networks, multilayer perceptron (MLP) can also be used as classifiers [95]. Generalized regression neural network (GRNN) [96] and convolutional neural network (CNN) were used as classifiers that were not preceded by classical features extraction, classification was performed using SOFTMAX [86].

3.5. Decision

In most systems, the final decision on a user's identity is based on a threshold that is chosen so that the biometric system meets security and throughput requirements. The score calculated in the previous system module is compared with the threshold value. If the score is higher than the threshold, the user is identified or verified, but if the score is lower than the threshold, the user is not identified or verified. The outcome of decisions for hand-based biometric systems is significantly influenced by user training, with the right habits of users, the score increases (agreement of the compared samples) [97, 98].

4. Quantitative evaluation of biometric system

Possible identification results are shown in Table 1.

Table 1: Possibilities of recognition decision

		Decision	
		Accept	Reject
Actual output	Authorized user	Correct acceptance	Incorrect rejection
	Unauthorized intruder	Incorrect acceptance	Correct rejection

Source: [2]

Based on the decision, the performance of biometric systems can be quantified. Selected quantified identifiers are [99].

False Rejection Rate – FRR

The probability that the biometric system does not recognize the authorized user will lead to an erroneous rejection, i.e. a type 1 error. If an error of this type occurs, the user can try to prove his identity again. The FRR is calculated according to the formula:

$$FRR = P(FR/OU) = \lim_{N_{OU}\to\infty} \frac{N_{FR}}{N_{OU}} \quad (6)$$

FR… occurrence of a rejection error

OU … recognition (verification or identification) by an authorized user

N_{FR}… number of false rejections

N_{OU}… number of attempts by authorized users to recognize

False Acceptance Rate – FAR

The probability that the biometric system will accept an unauthorized intruder will lead to incorrect acceptance, i.e. a type 2 error.

The FAR calculation is considered according to the formula:

$$FAR = P(FA/NN) = \lim_{N_{NN}\to\infty} \frac{N_{FA}}{N_{NN}} \quad (7)$$

FA… occurrence of an acceptance error

NN… recognition by an unauthorized intruder

N_{FA}… number of incorrect acceptances

N_{NN}… the number of attempts by an unauthorized intruder to recognize

The FAR and FRR rates depend on the threshold and are also inversely proportional. The higher the threshold, the more secure the biometric system, but the more common the type 1 error occurs and the user-friendliness decreases. On the contrary, the lower the threshold value, the more user-friendly the biometric system, but more often an unauthorized intruder, i.e. a type 2 error, is admitted, and the security of the system is reduced. The threshold value must always be set according to the requirements that are placed on the biometric system.

Equal Error Rate – EER

Value where the error rates of both FAR and FRR are equal, (i.e. FAR = FRR). With such a setting, the system incorrectly rejects and incorrectly identifies the same number of people. The lower the EER value, the higher the accuracy of the system. It is the most widely used indicator for evaluating and comparing biometric systems.

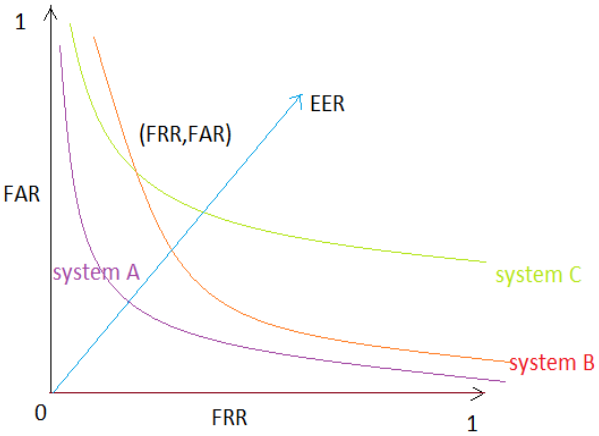


Figure 5: ROC Curves [100]

Receiver Operating Characteristic – ROC

The probability's curve that is given by the relationship between the probability of false acceptance (FAR) on the x-axis and the corresponding probability of correct acceptance on the y-axis for all allowable threshold values. Using ROC curves, multiple models can be compared with each other. Comparison and evaluation of multiple ROC curves are performed using the Area Under Curve (AUC) areas, which are located below the individual ROC curves [101]. Figure 5 shows the ROCs of the three systems, system A being the best system under the circumstances.

EER, FRR, FAR, ROC was selected for this work. They are a good indicator and the most frequently used tool for evaluating and managing biometric systems.

In articles dealing with biometric systems, we also encounter the concept of accuracy. The accuracy is calculated according to the formula:

$$\text{accuracy} = \frac{CD}{NR} \quad (8)$$

CD … correct decision (correct rejection and correct acceptance)

NR… number of recognition attempts

5. Comparison

The most commonly used physiological characteristics in biometric are fingerprint, face, hand geometry, and iris [102]. Table 2 compares biometric systems that use commonly biometric characteristics based on EER. Of these biometric characteristics, most often the lowest EER is achieved, biometric systems working with the iris, but as can be seen in Table 2, other systems, such as fingerprints, also achieve very low EER. However, in addition to error rates compilation speed, acqurability, privacy, cost and ease of use, should also play a role in choosing a biometric system [102].

Table 2: Comparison of selected hand other biometric systems

Sources	Biometric Characteristic of Recognition	Size of the Database	Performance
Chen et al. [103]	Iris	3 500	EER = 0.43
Yingzi et al. [104]	Iris	610	EER = 0.0295
Wijaya [105]	Face	2 268	EER = 0.0457
Marvadi [106]	Face	427	EER = 0.1525
Liu et al. [107]	Fingerprint	350	EER = 0.0042
Jain & Feng [108]	Fingerprint	449	EER = 0.081
Varchol & Levický [109]	Hand geometry	408	EER = 0.0462
Jain et al. [110]	Hand geometry	500	EER= 0.06

Table 3 summarizes the essential characteristics of some systems and methods that appeared in the journal and conferences. Some systems have been tested in identification mode, some in verification mode. Only some papers report EER, FAR / FRR or Accuracy values, some authors evaluate systems using ROC curves. The selection in Table 3 seeks to compare a representative sample of hand-base systems by characteristics (Hand-base Biometric Characteristic, Mode, Techniques Applied for Features Extraction, Techniques Applied for Recognition, Size of the Database, Performance).

Older systems use geometric measurements as features and are mostly in verification mode [110, 115]. Newer systems [111, 117] use a combination of multiple biometric characteristics such as hand shape and hand texture and use deep learning methods. Today's systems are mainly in identification mode, and FRR or EER has less than 1%.

Table 3: Comparison of selected hand based biometric system

Sources	Biometric characteristic / Mode	Techniques Applied for Features Extraction	Techniques Applied for Recognition	Size of the Database	Performance
Afifi [111]	Hand texture and shape / identification	CNNs + LBP	Support vector machine (SVM)	11 076	EER = 0.009
Kumar & Zang [112]	Hand geometry / identification	discretization of hand-geometry features	SVM	1 000	Accuracy = 0.94
Shanmukhappa & Sanjeevakumar [113]	Hand geometry /identification	Feature vector construct by hand image graph	SVM (Radial basic function)	1 440	FRR = 0.0205
Jain et al. [110]	Hand geometry/ verification	Geometric measurements	Euclidean distance	500	EER= 0.06
Varchol & Levický [109]	Hand geometry/ verification	Geometric measurements	Hamming distance	408	EER = 0.0973
Villegas et al. [114]	Hand geometry /identification	Wavelet Features	Nearest Neighbour	120	FAR = 0.11 FRR = 0.1
Varchol & Levický [109]	Hand geometry/ verification	Geometric measurements	Gaussian Mixture Model	408	EER = 0.0462
Burques et al. [115]	Hand geometry/ verification	Geometric measurements	Distance measure	12 800	EER = 0.0016
Charfi et al. [116]	Hand shape/identification	Scale Invariant Feature Transform (SIFT)	SIFT Matching	1 170	EER = 0.0586
Yoruk et al. [29]	Hand shape /identification	Independent Component Analysis (ICA)	L1/L2 norm of the difference of the feature vectors	458	Accuracy = 0.9731
Yoruk et al. [29]	Hand texture /identification	Principal Component Analysis (PCA)	L1/L2 norm of the difference of the feature vectors	458	Accuracy = 0.9791
Yoruk et al. [29]	Hand texture /identification	Angular radial transform (ART)	L1/L2 norm of the difference of the feature vectors	458	Accuracy = 0.976
Prihodova & Hub [117]	Hand texture and shape / identification	CCN (GoogLeNet)	CCN (Softmax)	456	FRR= 0.001

6. Research gap

It was not possible to cover all the literature dealing with hand biometric systems, but we included a representative sample of modern methods. However, we were able to identify some research gaps. To make hand-based biometric systems more efficient in real-world applications, the following challenges need to be overcome:

A large number of users – In the real world, it is often necessary to recognize millions of identities. However, the variability between some individuals may be minimal.

Quality of input data – Thanks to the progress of sensing sensors, the input image obtained in a real application under standard lighting conditions is correctly exposed. However, the image quality deteriorates in poor lighting conditions. Other problems we face despite the significant progress of the sensors are accurate colour capture and noise.

The permanence of biometric characteristics – Hand-based biometric characteristics have high stability over time but are not 100 %. For the hand geometry and hand shape, the topography of the bloodstream is primarily a problem of biological ageing and injury.

Attacks on systems – Hand-based biometric systems must face attacks at various levels in real-world applications. One of them is an attack on an input device, where the attacker forged a biometric sample, as a defence against this type of attack is the detection of liveliness.

Datasets – Small datasets are also a problem in this area of research, most current systems are tested on small datasets.

Costs – Costs should always be considered, and efforts should be made to make systems more accessible.

7. Discussion

Given the current state of hand-based biometric recognition, which was described in previous chapters of the article and due to the unresolved challenges in the field, one of the possibilities for further research is the use of a multi-biometric system, which uses hand geometry and thermal properties of hand. The thermal characteristic can be used in the form of blending two images (thermogram and image from the visible part of the electromagnetic spectrum), this blending creates a multispectral image. The multispectral image provides more significant variability between certain individuals than the mere geometry of the hand. A multispectral image is also advantageous in low lighting conditions, the quality of input data decreases less because the thermogram is not dependent on lighting conditions. When using a multispectral image, some types of attacks on biometric systems are also eliminated, because the thermal map of the hand, which is captured by the thermogram, is complicated to imitate and at the same time liveness is detected. At the same time, the multi-biometric system should increase recognition accuracy.

Another research option is also a multi-biometric system using hand geometry or palmprint in combination with skin colour (soft biometrics). This multi-biometric system could help to solve the problem of low variability between specific individuals.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This article was supported by grant No. 2020_018 supported by the Student Grant Competition.

References

[1] K. Prihodova, M. Hub. "Biometric Privacy through Hand Geometry - A Survey" International Conference on Information and Digital Technologies (IDT). IEEE; 395-401, 2019. https://doi.org/10.1109/DT.2019.8813660
[2] M. Hub. "Bezpečnost a ochrana informací v prostředí internetu" (book in Czech) Pardubice: University of Pardubice, 2013.
[3] T. Doseděl. "Počítačová bezpečnost a ochrana dat." (book in Czech) Brno: Computer Press, 2004. ISBN 80-251-0106-1.
[4] A. Jain, P. Flynn, A. Ross. "Handbook of biometrics" New York: Springer, c2008.
[5] A. W. K. Kong, D. Zhang, M.S. Kamel. "An Analysis of IrisCode" IEEE Transactions on Image Processing. **19**(2), 522-532, 2010. DOI: 10.1109/TIP.2009.2033427.
[6] D. Nguyen, T. Pham, Y. Lee, K. Park. "Deep Learning-Based Enhanced Presentation Attack Detection for Iris Recognition by Combining Features from Local and Global Regions Based on NIR Camera Sensor" Sensors. 2018, **18**(8). DOI: 10.3390/s18082601.
[7] S. M. Lajevardi, A. Arakala, S. A. Davis, K. J. Horadam. "Retina Verification System Based on Biometric Graph Matching" IEEE Transactions on Image Processing. **22**(9), 3625-3635, 2013. DOI: 10.1109/TIP.2013.2266257.
[8] D. Wang, H. Lu, M-H Yang. "Kernel collaborative face recognition" Pattern Recognition. **48**(10), 3025-3037, 2015. DOI: 10.1016/j.patcog.2015.01.012.
[9] Ž. Emeršič, V. Štruc, P. Peer. "Ear recognition: More than a survey" Neurocomputing., **255**, 26-39, 2017. DOI: 10.1016/j.neucom.2016.08.139.
[10] A. K. Jain, J. Feng, K. Nandakumar. "Fingerprint Matching" International Journal of Computer, **43**(2), 36-44, 2010.
[11] N. Nedjah, R. S. Wyant, L. De Macedo Mourelle. "Efficient biometric palmprint matching on smart-cards for high security and privacy" Multimedia Tools and Applications. **76**(21), 22671-22701, 2017. DOI: 10.1007/s11042-016-4271-8.
[12] V. D. Ambeth Kumar, M. Ramakrishnan. "Multiple Feature Extraction for FootPrint Image" National Journal of System and Information Technology. 2012, **5**(2), 142, 2012.
[13] L. Zhu, S. Zhang. "Multimodal biometric identification system based on finger geometry, knuckle print and palm print" Pattern Recognition Letters. **31**(12), 1641-1649, 2010. DOI: 10.1016/j.patrec.2010.05.010.
[14] B. Aghili, H. Sadjedl. "Personal Authentication Using Hand Geometry" In: 2009 International Conference on Computational Intelligence and Software Engineering. IEEE, 2009, 2009, s. 1-4. DOI: 10.1109/CISE.2009.5363947.
[15] Y. Wang, D. Zhang, Q. Ql. "Liveness detection for dorsal hand vein recognition" Personal and Ubiquitous Computing. **20**(3), 447-455, 2016. DOI: 10.1007/s00779-016-0922-z.
[16] D. Frumkin, A. Wasserstrom, A. Davidson, A. Grafit. "Authentication of forensic DNA samples" Forensic Science International: Genetics. 2010, **4**(2), 95-103. DOI: 10.1016/j.fsigen.2009.06.009.
[17] "Voice biometrics growth likely" Biometric Technology Today. 2009, **17**(6), 3-4. DOI: 10.1016/S0969-4765(09)70085-2.
[18] T. Scheidat, M. Kalbitz, C. Vielhauer. "Biometric authentication based on 2D/3D sensing of forensic handwriting traces" IET Biometrics. 2017, **6**(4), 316-324. DOI: 10.1049/iet-bmt.2016.0127.
[19] T. Hafs, L. Bennacer, M. Boughazi, A. Nait-Ali. "Empirical mode decomposition for online handwritten signature verification" IET Biometrics. 2016, **5**(3), 190-199. DOI: 10.1049/iet-bmt.2014.0041.
[20] T. K. M. Lee, M. Belkhatir, S. Sanei. "A comprehensive review of past and present vision-based techniques for gait recognition" Multimedia Tools and Applications. 2014, **72**(3), 2833-2869. DOI: 10.1007/s11042-013-1574-x.
[21] Information technology - Biometric performance testing and reporting - ISO/IEC 19795-1. City, 2006.
[22] R. Miller. "Finger dimension comparison identification system" USA. US3576538 (A). Written down 27.4.1971.
[23] B. Miller. "Vital signs of identity [biometrics]" IEEE Spectrum. 1994, **31**(2), 22 30. DOI: 10.1109/6.259484.
[24] D. Sisdlouskas. "3D hand profile identification apparatus" 1988. USA. US4736203A. Written down 5.4.1988.
[25] Immigration and naturalization service passenger accelerated service system pilot program. Office of the Inspector General. Washington, DC.
[26] Technology assessment. Edtion ed. Washington, DC (PO Box 37050, Washington 20013): US General Accounting Office, 2002.
[27] D. Maltoni. "Handbook of fingerprint recognition" 2nd ed. London: Springer, c2009.
[28] X. Fan, H. Zhang, C. Leung, C. Miao. "Comparative study of machine learning algorithms for activity recognition with data sequence in home-like environment" In: 2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). IEEE, 2016, 2016, s. 168-173. DOI: 10.1109/MFI.2016.7849484.
[29] E. Yoruk, H. Dutagaci, B. Sankur. "Hand biometrics. Image and Vision Computing" 2006, **24**(5), 483-497. DOI: 10.1016/j.imavis.2006.01.020.
[30] D. P. Sidlauskas, S. Tamer. "Hand Geometry Recognition" 2008, In: Jain A.K., Flynn P., Ross A.A. (eds) Handbook of Biometrics. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-71041-9_5
[31] R. Rak, V. Matyáš, Z. Říha. "Biometrie a Identita Člověka: Ve Forenzních a Komerčních Aplikacích" (book in Czech) 1electronic edition ed. Prague: Grada; 2008.
[32] B. Lindstrom "Hand Geometry A Biometric Technology with a Pedigree BiometriTech June 2003 Bruce Lindstrom, Regional Sales Manager" 2003. https://slideplayer.com/slide/5032579/
[33] J. A. Unar, W. C. Seng, A. Abbasi. "A review of biometric technology along with trends and prospects" Pattern Recognition. 2014, **47**(8), 2673-2688. DOI: 10.1016/j.patcog.2014.01.016.
[34] A. Ross "An introduction to multibiometrics." IEEE; 2007.
[35] B. Aghili, H. Sadjedi. "Personal Authentication Using Hand Geometry" In: 2009 International Conference on Computational Intelligence and Software Engineering. IEEE, 2009, 2009, s. 1-4. DOI: 10.1109/CISE.2009.5363947.
[36] G. Fouquier, L. Likforman, J. Darbon, B. Sankur. "The Biosecure Geometry-Based System for Hand Modality" In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. IEEE, 2007, 2007, I-801-I-804. DOI: 10.1109/ICASSP.2007.366029. ISBN 1-4244-0727-3.
[37] R. Sanchez-Reillo, C. Sanchez-Avila, A. Gonzalez-Marcos. "Biometric identification through hand geometry measurements" IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000, **22**(10), 1168-1171. DOI: 10.1109/34.879796.
[38] De Santos Sierra, A. J. G. Casanova, C. S. Avila, V. J. Vera. "Silhouette-based hand recognition on mobile devices" In: 43rd Annual 2009 International Carnahan Conference on Security Technology. IEEE, 160-166, 2009. DOI: 10.1109/CCST.2009.5335548.
[39] C. Oden, A. Ercil, B. Buke. "Combining implicit polynomials and geometric features for hand recognition" Pattern Recognition Letters. 2003, **24**(13), 2145-2152. DOI: 10.1016/S0167-8655(03)00087-4.
[40] E. Yoruk, E. Konukoglu, B. Sankur, J. Darbon. "Shape-based hand recognition" In: IEEE Transactions on Image Processing. IEEE, 2006, **15**(7), s. 1803-1815. DOI: 10.1109/TIP.2006.873439.

[41] S. A. Angadi, S. M. Hatture. "Biometric Person Identification System: A Multimodal Approach Employing Spectral Graph Characteristics of Hand Geometry and Palmprint" International Journal of Intelligent Systems and Applications. 2016, 8(3), 48-58. DOI: 10.5815/ijisa.2016.03.06.

[42] M. Ferrer, C. Travieso, J. Alonso. "Multimodal Biometric System based on Hand Geometry and Palm Print Texture" In: Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology. IEEE, 2006, 2006, s. 92-97. DOI: 10.1109/CCST.2006.313435.

[43] A. Kirthika, S. Arumugam. "Texture and colour intensive biometric multimodal security using hand geometry and palm print. "International Journal of Advances in Engineering & Technology, **4**(1), 226-235, 2012.

[44] T. O. Oladele, K. Adeniyi, T. O. Aro. "Framework for User Authentication at a Distance for Mobile Phones Using Contactless Hand-based Multimodal Biometric System" Journal of Computer Science and Control Systems. ProQuest Central, 2019, **12**(1), 24-27, 2019.

[45] W.-C. Wang, W.-S. Chen, S.-W. Shih. "Biometric recognition by fusing palmprint and hand-geometry based on morphology" In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009, 2009, s. 893-896. DOI: 10.1109/ICASSP.2009.4959728.

[46] P. Gupta, S. Srivastava, P. Gupta. "An accurate infrared hand geometry and vein pattern-based authentication system" Knowledge-Based Systems.**103**, 143-155, 2016. DOI: 10.1016/j.knosys.2016.04.008. ISSN 09507051.

[47] E. C. Lee, H. Jung, D. Kim. "New Finger Biometric Method Using Near Infrared Imaging" Sensors. **11**(3), 2319-2333, 2011. DOI: 10.3390/s110302319.

[48] C. Oden, A. Ercil, B. Buke. "Combining implicit polynomials and geometric features for hand recognition "Pattern Recognition Letters.**24**(13), 2145-2152, 2003. DOI: 10.1016/S0167-8655(03)00087-4.

[49] M. A. Ferrer, A. Morales, C. M. Travieso, J. B. Alonso. "Low Cost Multimodal Biometric Identification System Based on Hand Geometry, Palm and FingerPrint Texture" In: 2007 41st Annual IEEE International Carnahan Conference on Security Technology. IEEE, 2007, 2007, s. 52-58. DOI: 10.1109/CCST.2007.4373467.

[50] F. Yang, B. Ma. "A New Mixed-Mode Biometrics Information Fusion Based-on Fingerprint, Hand-geometry and Palm-print" In: Fourth International Conference on Image and Graphics (ICIG 2007). IEEE, 2007, 2007, s. 689-693. DOI: 10.1109/ICIG.2007.39.

[51] A. Ross, A. Jain. "Information fusion in biometrics. Pattern" Recognition Letters. 2003, **24**(13), 2115-2125. DOI: 10.1016/S0167-8655(03)00079-5.

[52] M. Kim, W.-Y. Kim, J. Paik. "Optimum Geometric Transformation and Bipartite Graph-Based Approach to Sweat Pore Matching for Biometric Identification" Symmetry. 2018, **10**(5), 175. DOI: 10.3390/sym10050175.

[53] M. Adán, A. Adán, A. S. Vázquez, R. Torres. "Biometric verification/identification based on hands natural layout" Image and Vision Computing. 2008, **26**(4), 451-465. DOI: 10.1016/j.imavis.2007.08.010.

[54] S. Gonzeilez, C. M. Travieso, J. B. Alonso, M. A. Ferrer. "Automatic biometric identification system by hand geometry" In: IEEE 37th Annual 2003 International Carnahan Conference onSecurity Technology, 2003. Proceedings. IEEE, 2003, s. 281-284. DOI: 10.1109/CCST.2003.1297573.

[55] R. Gross, Y. Li, L. Sweeney, X. Jiang, W. Xu, D. Yurovsky "Robust Hand Geometry Measurements for Person Identification using Active Appearance Models" In: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems. IEEE, 2007, 2007, s. 1-6. DOI: 10.1109/BTAS.2007.4401936.

[56] N. Charfi, H. Trichili, A. M. Alimi, B. Solaiman. "Personal recognition system using hand modality based on local features" In: 2015 11th International Conference on Information Assurance and Security (IAS). IEEE, 2015, 2015, s. 13-18. DOI: 10.1109/ISIAS.2015.7492764.

[57] S. Xu, M. Li, J. F. Ding, Y. Q. Cui. "Personal Identification by Fusing Hand Shape Geometry and Palmprint Features" Applied Mechanics and Materials, 278-280, 1228-1231, 2013. DOI: 10.4028/www.scientific.net/AMM.278-280.1228.

[58] A. Iula, G. Hine, A. Ramalli, F. Guidi. "An Improved Ultrasound System for Biometric Recognition Based on Hand Geometry and Palmprint" Procedia Engineering.**87**, 1338-1341, 2014. DOI: 10.1016/j.proeng.2014.11.709.

[59] R. Sanchez-Reillo. "Hand geometry pattern recognition through Gaussian mixture modelling" In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. IEEE Comput. Soc, 2000, s. 937-940. DOI: 10.1109/ICPR.2000.906228.

[60] V. Kanhangad, A. Kumar, D. Zhang. "Human hand identification with 3D hand pose variations" In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. IEEE, 2010, 2010, s. 17-21. DOI: 10.1109/CVPRW.2010.5543236.

[61] J.-M. Guo, C.-H. Hsia, Y.-F. Liu, J.-C. Yu, M.-H. Chu, T.-N. Le. "Contact-free hand geometry-based identification system. Expert Systems with Applications" 2012, **39**(14), 11728-11736. DOI: 10.1016/j.eswa.2012.04.081.

[62] A. Morales, M. A. Ferrer, J. B. Alonso, C. M. Travieso. "Comparing infrared and visible illumination for contactless hand based biometric scheme." In: 2008 42nd Annual IEEE International Carnahan Conference on Security Technology. IEEE, 2008, 2008, s. 191-197. DOI: 10.1109/CCST.2008.4751301.

[63] A. Jian, N. Duta. "Deformable matching of hand shapes for user verification" In: Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348). IEEE, 1999, s. 857-861. DOI: 10.1109/ICIP.1999.823019.

[64] M. Ferrer, A. Morales, C. M. Travieso, J. B. Alonso. "Influence of the pegs number and distribution on a biometric device based on hand geometry" In: 2008 42nd Annual IEEE International Carnahan Conference on Security Technology. IEEE 2008, 221-225, 2008. DOI: 10.1109/CCST.2008.4751305.

[65] G. Boreki, A. Zimmer. "Hand Geometry: A New Approach for Feature Extraction" In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05). IEEE, 2005, s. 149-154, 2005. DOI: 10.1109/AUTOID.2005.33.

[66] M. Faundez-Zanuy. "Biometric verification of humans by means of hand geometry" In: Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology. IEEE, 2005, 2005, s. 61-67. DOI: 10.1109/CCST.2005.1594816.

[67] A. Bera, D. Bhattacharjee, M. Nasipuri. "Finger contour profile-based hand biometric recognition" Multimedia Tools and Applications. 2017, **76**(20), 21451-21479. DOI: 10.1007/s11042-016-4075-x.

[68] J. Doublet, O. Lepetit, M. Revenu. "Contactless Hand Recognition Based on Distribution Estimation" In: 2007 Biometrics Symposium. IEEE, 2007, 2007, s. 1-6. DOI: 10.1109/BCC.2007.4430547.

[69] M. Španěl, V. Beran. "Obrazové segmentační techniky: Přehled existujících metod" (book in Czech) Brno, Brno University of Technology, 2006.

[70] G. Amayeh, G. Bebis, A. Erol, M. Nicolescu. "Peg-Free Hand Shape Verification Using High Order Zernike Moments." In: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). New York: IEEE, 2006, s. 40-40. DOI: 10.1109/CVPRW.2006.155.

[71] A. M. Badawi, M.S. Kamel. "Free style hand geometry verification system" In: 2003 46th Midwest Symposium on Circuits and Systems. IEEE, 2003, s. 324-327, 2003. DOI: 10.1109/MWSCAS.2003.1562284.

[72] M. Golfarelli, D. Maio, D. Malton. "On the error-reject trade-off in biometric verification systems" IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997, **19**(7), 786-796. DOI: 10.1109/34.598237.

[73] J. Mintalvao, L. Molina, J. Canuto. "Robust hand image processing for biometric application" Pattern Analysis and Applications. 2010, **13**(4), 397 407, 2010. DOI: 10.1007/s10044-010-0185-7.

[74] A. Bapat, V. Kanhangad. "Segmentation of hand from cluttered backgrounds for hand geometry biometrics" In: 2017 IEEE Region 10 Symposium (TENSYMP). IEEE, 2017, 2017, s. 1-4. DOI: 10.1109/TENCONSpring.2017.8070016.

[75] J. Král, "Rozpoznávání výrazu práce" Brno, 2010. Thesis. Technical University Brno, 2010.

[76] N. Covavisaruch, P. Prateepamornkul, P. Ruchikachorn, P. Taksaphan. "Personal verification and identification using hand geometry" ECTI Transactions on Computer and Information Technology (ECTI-CIT), **1**(2), 134-140, 2005.

[77] V. Kanhangad, A. Kumar, D. Zhang. "Combining 2D and 3D hand geometry features for biometric verification" In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2009, 39-44, 2009. DOI: 10.1109/CVPRW.2009.5204306.

[78] X. Jiang, W. Xu, L. Sweeney, Y. Li, R. Gross, D. Yurovsky. "New Directions in Contact Free Hand Recognition" In: 2007 IEEE International Conference on Image Processing. IEEE, 389-392, 2007. DOI: 10.1109/ICIP.2007.4379174.

[79] S. Sharma, S. R. Dubey, S. K. Singh, R. Saxena, R. K. Singh. "Identity verification using shape and geometry of human hands" Expert Systems with Applications. Elsevier, 2015, **42**(2), 821-832, 2015. DOI: 10.1016/j.eswa.2014.08.052.

[80] T. Fotak, P. Koruga, M Baca. "Trends in Hand Geometry Biometrics" In: Central European Conference on Information and Intelligent Systems. Varazdin: Faculty of Organization and Informatics Varazdin, 2012, 319-324, 2012.

[81] D. L. Woodard, P. J. Flynn. "Finger surface as a biometric identifier. Computer Vision and Image Understanding" Elsevier, 2005, **100**(3), 357-384, 2005. DOI: 10.1016/j.cviu.2005.06.003.

[82] N. Charfi, H. Trichili, A. M. Alimi, B. Solaiman. "Novel hand biometric system using invariant descriptors" In: 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR). IEEE, 2014, 2014, 261-266. DOI: 10.1109/SOCPAR.2014.7008016.

[83] W. Kang, Q. Wu. "Pose-invariant hand shape recognition based on finger geometry" IEEE Transactions on Systems, Man, and Cybernetics: Systems **44**(11), 1510–1521. 2014.

[84] A. Ungureanu, S. Salahuddin, P. Corcoran. "Towards unconstrained palmprint recognition on consumer devices: A literature review", 2020. DOI:10.1109/ACCESS.2020.2992219

[85] R. S. Choraś, M. Choraś. "Hand shape geometry and palmprint features for the personal identification" 2006. IEEE Proc. of 6th Intl Conf. on Intelligent Systems Design and Applications, 1085–1090, 2006.
[86] A. Bera, D. Bhattacharjee. "Human Identification Using Selected Features from Finger Geometric Profiles" IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2020, **50**(3), 747-761. DOI: 10.1109/TSMC.2017.2744669. ISSN 2168-2216.
[87] A. Ungureanu, S. Salahuddin, P. Corcoran. "Towards unconstrained palmprint recognition on consumer devices: A literature review" 2020. DOI:10.1109/ACCESS.2020.2992219.
[88] K. Příhodová. "Convolutional neural networks in hand-based recognition system" In: 34th International Business Information Management Association Conference. 2019, 4744-4750.
[89] A. Kumar, D. C. M. Wong, H. C. Shen, A. K. Jain. "Personal Verification Using Palmprint and Hand Geometry Biometric" In: Audio- and Video-Based Biometric Person Authentication. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, 2003-6-24, s. 668-678. Lecture Notes in Computer Science. DOI: 10.1007/3-540-44887-X_78.
[90] T. Mitchell, "Machine learning" Boston: McGraw-Hill, 1997.
[91] Vinodkumar, R. Srikantaswamy. "Fusion of fingerprint, palmprint and hand geometry for an efficient multimodal person authentication system" In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE, 2016.
[92] J. J. Fuertes, C. M. Travieso, M. A. Ferrer, J. B. Alonso. "Intra-modal biometric system using hand-geometry and palmprint texture" In: 44th Annual 2010 IEEE International Carnahan Conference on Security Technology. IEEE, 2010, 2010, 318-322. DOI: 10.1109/CCST.2010.5678697.
[93] A. Kumar, D. Zhang. "Hand-Geometry Recognition Using Entropy-Based Discretization. IEEE Transactions on Information Forensics and Security" 2007, **2**(2), 181-187. DOI: 10.1109/TIFS.2007.896915.
[94] X.Fan, X. H. Zhang, C. Leung, C. Miao. "Comparative study of machine learning algorithms for activity recognition with data sequence in home-like environment" In: 2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). IEEE, 2016, 2016, s. 168-173. DOI: 10.1109/MFI.2016.7849484.
[95] M. Ferrer, J. Fabregas, M. Faundez, J. B. Alonso, C. Travieso. "Hand geometry identification system performance" In: 43rd Annual 2009 International Carnahan Conference on Security Technology. IEEE, 2009, 2009, s. 167-171. DOI: 10.1109/CCST.2009.5335545.
[96] O. Polat, T. Yildirim. "Hand geometry identification without feature extraction by general regression neural network" Expert Systems with Applications. 2008, **34**(2), 845-849. DOI: 10.1016/j.eswa.2006.10.032.
[97] S. J. Elliott, B. Senjaya, E. P. Kukula J. M. Werner, M. Wade. "An evaluation of the Human Biometric Sensor Interaction using hand geometry" In: 44th Annual 2010 IEEE International Carnahan Conference on Security Technology. IEEE, 2010, 2010, 259-265. DOI: 10.1109/CCST.2010.5678710.
[98] E. P. Kukula, B. P. Gresock, S. J. Elliott, N. W. Dunning. "Defining Habituation using Hand Geometry" In: 2007 IEEE Workshop on Automatic Identification Advanced Technologies. IEEE, 2007, 2007, 242-246. DOI: 10.1109/AUTOID.2007.380627.
[99] ISO/IEC 19795-1. Information technology: Biometric performance testing and reporting. 2016.
[100] M. R. Rajput and G. S. Sable, "IRIS biometrics survey 2010–2015," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2016, 2028-2033, DOI: 10.1109/RTEICT.2016.7808195.
[101] J. Továrek, "Využití umělé inteligence pro vícenásobnou bezkontaktní biometrickou autentizaci" (in Czech) Ostrava, 2017. Ph. D Thesis. VSB - Technical University of Ostrava, 2017.
[102] R.M. Bolle, J.H. Connell, Pankanti S., Ratha N.K. and Senior A.W. The Common Biometrics. In: Guide to Biometrics. Springer Professional Computing. Springer, New York, NY. 2004.
[103] R. Chen, X.R. Lin, T.H. Ding "Iris segmentation for non-cooperative recognition systems" IET Image Process., **5** (2011), 448-456, 2011.
[104] Y. Du, E. Arslanturk, Z. Zhou, C. Belcher. "Video-Based Noncooperative Iris Image Segmentation" IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **41**(1):64-74, 2011.
[105] I G. P. Suta Wijaya. "Face Recognition Using Map Discriminant on YCbCr Color Space." 2018.
[106] D. Marvadi, C. Paunwala, M. Joshi, A. Vora. "Comparative analysis of 3D face recognition using 2D-PCA and 2D-LDA approaches" IEEE; 2015;2016.
[107] F. Liu, G. Liu, Q. Zhao, L. Shen. "Robust and high-security fingerprint recognition system using optical coherence tomography" Neurocomputing (Amsterdam). 2020. 402:14-28, 2020.
[108] F. Liu, J. Liang, L. Shen, M. Yang, D. Zhang, Z. Lai. "Case study of 3D fingerprints applications" PLoS ONE. 2017; **12**(4) 0175261.
[109] P. Varchol and D. Levický, "Using of hand geometry in biometric security systems", Radio Eng., **16**(4), 82-87, 2007.
[110] A.K. Jain, A. Ross and S. Pankanti, "A prototype hand geometry-based verification system", Proc. Second Int. Conf. Audio and Video based Personal Authentication, 166-171, 1999.
[111] M. Afifi. "11K Hands: Gender recognition and biometric identification using a large dataset of hand images" Multimedia tools and applications, **78**(15), 20835-54, 2019.
[112] A. Kumar, D. Zhang "Hand-Geometry Recognition Using Entropy-Based Discretization" IEEE Transactions on Information Forensics and Security. 2007, **2**(2),181-7, 2007.
[113] S. Angadi, S. Hatture "Hand geometry-based user identification using minimal edge connected hand image graph" IET computer vision, **12**(5),744-52, 2018.
[114] O. Vergara Villegas, H. J. Ochoa Dominguez, V. G. Cruz Sanchez, L. Ortega and H. Madero. "Biometric Human Identification of Hand Geometry Features Using Discrete Wavelet Transform" Discrete Wavelet Transforms - Biomedical Applications. InTech, 09-12, 251-266. 2011.
[115] J. Burgues, J. Fierrez, D. Ramos, M. Puertas and J. Ortega-Garcia, „Detection Invalid Samples in Hand Geometry Verification via Geometric Measurements", 2010 International Workshop on Emerging Techniques and Challenges for Hand-Based Biometrics, Istanbul, 2010, 1-6, DOI: 10,1109 / ETCHB.2010.5559296.
[116] N. Charfi, H. Trichili, A. M. Alimi and B. Solaiman, "Novel hand biometric system using invariant descriptors," 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, 2014, 261-266, doi: 10.1109/SOCPAR.2014.7008016.
[117] K. Prihodova and M. Hub. "Hand-Based Biometric System Using Convolutional Neural Networks". Acta Informatica Pragensia, **9**(1), 48-57. 2020. doi: 10.18267/j.aip.131.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 692-697 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Optimal Hydrokinetic Turbine Array Placement in Asymmetric Quasigeostrophic Flows

Victoria Monica Miglietta*, Manhar Dhanak

Department of Ocean and Mechanical Engineering, Florida Atlantic University, Boca Raton, Florida, 33431, USA

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 10 January, 2021
Online: 05 February, 2021

Keywords:
Tidal power
Array design
Coriolis
Rossby number
Blocking parameter

ABSTRACT

The Coriolis force in the ocean at mid to high latitudes can cause significant deviation of flow over bottom topography, including formation of Taylor columns. Structures in a tidal zone will experience zero inertial current between every tidal change. Around periods of directional change, the Coriolis force may be tapped into for energy. Factors like timescales and other environmental factors like local currents could influence the flow characteristics in an undesirable way and are outside of the scope of this study. The focus of this study is to assess how the design of a structure influences the asymmetric flow patterns produced around it by an incident quasigeostrophic flow. Analytical solutions existing for inviscid quasigeostrophic flow over isolated elongated elliptical topography are used for flows with small Rossby numbers. These solutions are used to predict and explore the characteristics of the flows expected during a change in the tidal cycle. Results show that a linear array placed perpendicular to a quasi-geostrophic flow will experience flow acceleration on the left-hand side when looking downstream. On the other hand, a linear array placed parallel to the quasi-geostrophic flow will experience a sharp velocity gradient over the array. This suggests that an array placed perpendicular to the quasi-geostrophic flow will provide for a more robust design when compared to a linear array placed parallel to the flow.

1. Introduction

Special considerations for large underwater turbine farms or arrays at mid and high latitudes will be necessary because at such latitudes the Coriolis force can measurably impact the dynamics of flow in asymmetric patterns. A thoughtful and robust array design can position turbines to harness the Coriolis force[1]. Array design requires knowledge of the local dynamics and an understanding of all forces including the Coriolis force. Tools for evaluating such an ocean engineering endeavor includes numerical solutions for flows moving past extended obstacles in instances when rotational forces are required to describe the flow [2].

The Coriolis force is a pseudo force caused by Earth spinning on its axis. Typical ocean engineering problems ignore Coriolis effects because they are minimal on the typical scale of ocean engineering problems. Also, local inertial forces often dominate engineering problems where a good approximation allows for the Coriolis force to drop out of consideration. But for tidal currents, the local inertial speeds will drop to zero at least twice daily, and it is at these periods where the Coriolis force may impact an array. A renewable energy system's array which produces electricity for a municipal power grid will be sensitive to electrical and therefore environmental energy fluctuations. For locations with higher values of latitude the Coriolis force should be considered at periods surrounding slack tide.

Often tidal current power sites are limited to channels and bays [3-7], however tides are not limited to these locations. Tides occur throughout the Earth's oceans. Here the design of a turbine array in open water is analyzed. Vertical boundaries are far from the array and the domain's distinct boundaries are the surface of the water and the seafloor. A very long array, capable of being economically viable, composed of a single line of turbines is considered. The turbine array is taken to be 10,000m long. The large dimension is comparable to a topographic feature and certainly would allow for the array to meet a production goal on the order of megawatts of electricity. It is hypothesized that a very large array in the ocean would experience the same amplification of flow rate around it that a seamount experiences [8, 9].

Studies of seamounts, and underwater ridges indicate flow behavior around large obstructions in the ocean are influenced in a measurable way by the Coriolis force [2, 8, 10]. The largest dimension of 10,000m would put a turbine array on the scale of a

*Corresponding Author: Victoria Monica Miglietta, vmiglietta2013@fau.edu

https://dx.doi.org/10.25046/aj060175

small seamount and on par with the scale of a seamount [10]. It has been observed that flows past/around seamounts are amplified and that the flow above them is best described as stagnant [10]. These are rotational effects and have been of interest to marine biologists and oceanographers [9, 11].

The flow patterns and Coriolis effects on an underwater turbine array differ from those of wind turbine arrays because of the huge density difference between air and water. The nature of air and water differ in a way where the rotational phenomenon of Taylor columns are not found in the atmosphere for quasigeostrophic flow conditions [12, 13]. These Coriolis effects are not seen around wind farms or wind turbines because they disappear in a density gradient.

2. Background

2.1. Not a channel

Studies like [4-6] look at tidal channels being utilized for tidal current power. However, taking over an entire tidal channel or even a good portion of one for power generation would have serious consequences for the established use of a channel. Many channels have daily uses which include maritime traffic for shipping, fishing and recreation. Obtaining permits and backing of local communities to re-purpose heavily used tidal channels for power production would likely be a challenge. If the tidal power project was for a remote community or some remote operations then challenges could be different. However, for large coastal communities looking to harness tidal power across a tidal channel such challenges could be expected. For example, the offshore wind development off the coast of Massachusetts, Vineyard Wind, has experienced a lot of push back by the fishing industry in the area. The project has been on hold since 2019 while the Interior Department conducts a review. Harnessing tidal current power in more open areas, could be less oppressive to current users of waterways. In this study the ideal deployment area is far from boundaries like shorelines, piers, or breakwaters.

2.2. Static Obstruction in a Rotational Fluid

Here the effects of the static physical presence of an obstruction to flow in a changing tide were studied. The review of studies of the effects of seamounts on the flow around them shows that large obstructions, like that which could be expected from a tidal turbine array, produce asymmetric patterns seen in rotational fluids [14]. These rotational effects are caused by the spinning of the Earth. The fundamental physical dynamics for an obstruction to flow in a rotational fluid with a small inertial current is similar for a bottom seated obstruction and an obstruction suspended in the water column [14]. An array would have similar effects for both bottom seated and moored or suspended turbines in the water column.

Furthermore, the research presented here would be applicable to the tidal array design engineering as well as the deployment site assessment. The physical phenomenon of Taylor columns and their associated flow patterns could develop locally at the site of the engineered structure itself, and it could also develop on/around a bathymetric feature like a ridge or seamount where an array or turbine is deployed. The University of Strathclyde encountered what was described as a spatial-temporal eddy with a 'calm' center when it was testing a tidal current turbine [15]. This could have been caused by a Taylor column which has a stagnant interior. The latitude for their deployment was approximately 55N.

2.3. Unchanging Coriolis Parameter in the f-plane

This study evaluates an array so large it can be viewed at the scale of topography, i.e. the array will define the 'landscape' in the area it is deployed. The *f*-plane approximation is made, where the Coriolis parameter does not change. If the array was to extend over several degrees of latitude the Coriolis parameter would change and a structure so great in size would need to be analyzed in the β-plane. For the *f*-plane approximation, the absolute velocity of the tidal current will be the parameter that changes the blocking parameter cyclically over time i.e. with the tidal cycle.

2.4. Blocking Parameter

The blocking parameter correlates the size of the obstruction with environmental parameters to indicate the degree of influence the structure has on the local flow patterns. A critical blocking parameter predicts the formation of a Taylor column. A Taylor column is a hydrodynamic phenomenon, parallel to the axis of rotation, which appears like an extension of an object, obstructing flow when the blocking parameter is at a critical value. The critical value of the blocking parameter S, is $O(1)$. There is no consensus on a more exact value for the blocking parameter. For the [10] observations the blocking parameter appeared to be approximately 0.7. The exact blocking parameter is likely to be site specific and will vary on a range of environmental parameters. Important parameters include the height of the array, h; the depth of the water, d, and the latitude which changes the Coriolis parameter, f.

This study uses the definition of the blocking parameter by [16],

$$S = h/(d \bullet R_0)$$

$$S = fhL/(d \bullet U) \quad (1)$$

There is no consensus on the calculation of the blocking parameter or its critical value. The formation of a Taylor column was initially linked to the Rossby number dropping below the value of $1/\pi$ [12]. Effectively this original critical blocking parameter was when the Rossby number, R_0,was $1/\pi$ or smaller.

Further research [16] on rotational fluids added a coefficient to the inverse of the Rossby number, the ratio of the height of the obstruction to the depth of the water, h/d. Decades later [17] used the ratio of the greatest horizontal length of the obstruction to the water depth as the coefficient to the inverse of the Rossby number. The similar thread in all these blocking parameters is the Rossby number. The Rossby number is the ratio of inertial forces to rotational forces. The rotational forces are represented in the Rossby number by the terms fL. Where L is the characteristic length, the longest horizontal dimension of the object. The Coriolis parameter is larger for higher latitudes. The Rossby number will

decrease for higher latitudes because at these locations rotational forces are stronger. The Rossby number is inversely proportional to the Coriolis parameter

$$R_0 \propto 1/f \quad (1)$$

and the blocking parameter is proportional to the Coriolis parameter

$$S \propto f \quad (2)$$

Consequently, a structure located at a higher latitude will experience a higher blocking parameter; with velocity, and structure size, and water depth all remaining constant.

The Rossby number includes *L*, the characteristic length of an object. The [16] blocking parameter is the most complete because it not only includes the characteristic length of an object, but it also includes its height. The [17] blocking parameter is redundant in that it uses the length of the obstruction in its coefficient and the length also appears in the Rossby number. So effectively the [17] blocking parameter uses the square of the characteristic length. Returning to the perspective of a tidal current, the velocity will change and with it the blocking parameter will change. Recall that the blocking parameter is inversely proportional to the velocity

$$S \propto 1/U \quad (3)$$

Looking at the critical blocking parameter for various latitudes, faster speeds can bring about a critical blocking parameter at higher latitudes. This means that the period of the tidal cycle where the blocking parameter is critical will be longer at higher latitudes.

3. Problem Statement

Since it is during the slack tide that the Coriolis force could impact the dynamics of a turbine array, this period cannot be passed over simply as the period where tidal current turbines will not produce electricity. At slack tide, the water is in quasigeostrophic flow; meaning that the inertial current is minimal and the Coriolis force is needed to describe the flow. In quasigeostrophic flow, the consequence of a small inertial current is that it prevents the equilibrium of the pressure force and Coriolis force (this equilibrium is described as geostrophic flow).

During each tidal change the blocking parameter becomes critical. The critical value of the blocking parameter is associated with the formation of a Taylor column. When this happens the array (the obstruction) will behave like a taller structure, or it will have a sort of physical shadow above it (and possibly below it). Also, for a critical blocking parameter the flow can be described well by a 2-D approximation, where the velocity does not vary in the water column. This flow described by the 2-D approximation, i.e. a Taylor column, can appear almost instantaneously in a laboratory tank. The question addressed is what are the characteristics of the flow influenced by the Coriolis force in the immediate vicinity of a linear array? The equipment in a tidal zone is likely to experience rotational effects during periods when currents change direction. Asymmetric flow for quasigeostrophic conditions have been observed around seamounts and in this study modelling was used to predict how a large linear array would experience such conditions.

3.1. Method

The blocking parameter, *S*, will have a spiked value at every tidal change because it is inversely proportional to the velocity. Numerical [2] solutions were used to plot the velocity around a linear array when the blocking parameter was critical or *O(1)*. The solutions depend on the angle of the incident flow, α, and also depend on the ratio, γ, of the largest dimension, *L*, to the smallest dimension, *l*, in the X-Y plane.

$$\gamma = L/l \quad (4)$$

The solution is insensitive to the details of the shape of the obstacle [2]. Solutions model flow patterns closely to those in the oceans where the bathymetry is more complex than a simple ellipse [2]. The solution is for a homogeneous, inviscid, incompressible flow, with a constant depth, in a rotating frame, and it satisfies the conservation of potential vorticity. The stream function used [2] in elliptical coordinates (μ, θ) is:

$$\Psi = \begin{cases} \frac{1}{32}a^2[(1-e^{-2\mu_0}\cosh(2\mu)]\cos(2\theta) \\ \qquad + \frac{1}{32}a^2[\cosh(2\mu)-\cosh(2\mu_0)], \quad \mu < \mu_0 \\ \frac{1}{32}a^2\sinh(2\mu_0)\left(e^{-2\mu}\cos(2\theta)+2(\mu-\mu_0)\right), \mu > \mu_0 \end{cases}$$

where $0.5a\sinh(\mu_0) = 1$, and $0.5a\cosh(\mu_0) = \gamma$.

4. Results

An idealized smooth diurnal tide, with equal peaks is considered for analysis. The far field velocity, *U*, changes the blocking parameter, *S*, as the tidal velocity changes. The height, *h*, is taken to be 10m, and the water depth, *d,* to be 35m. Aspect ratios between 1,000 and 250 do not vary the results remarkably. For steady quasigeostrophic flow the streamlines do not vary in the *Z* direction as long as there is no density gradient. The results show that directly at, or adjacent to, the structure is where the system could be impacted by the unique dynamics of periods dominated by quasigeostrophic flow. The character of the rotational effects is asymmetric and for the linear array placed in quasigeostrophic incident flow, the flow around it is distinctly asymmetrical when modelled with [2] analytical solutions. The results for different blocking parameters were considered as "snapshots" of what may occur over time because the time variability of these processes is not understood. The velocity fields at both ends of an array placed perpendicular to the quasi-geostrophic flow were examined for a number of velocities/blocking parameters and plotted against time. The plots reveal that the accelerations on the left-hand side looking downstream were greater than the decelerations on the opposite side, Figure 1. The corresponding effect on power over the period of quasi-geostrophic flow may be estimated from a plot of velocity cubed against time, Figure 2. The flow adjacent to the structure is most affected by the Coriolis force.

For such a perpendicular array (Figures 3 and 4) the flow is accelerated by a factor of 2 on one end, while on the opposite end it the flow rate is 1/2 to 1/5 of the far field velocity. The velocity coefficients at the end points do not change significantly as the blocking parameter increases. For blocking parameters larger than one, the location of the minimal flow (i.e. smallest velocity coefficient) moves away from the structure. The location of the most accelerated flow does not move away from the structure in the same manner. When the blocking parameter increases, flow adjacent to a greater area of the structure experiences acceleration. Furthermore, there is a steady decrease in velocity across the perpendicular array for critical blocking parameters of $O(1)$.

For an array placed parallel to the quasi-geostrophic flow (Figures 5 and 6) the velocity gradient is very sharp over the array. It is remarkable that for such a parallel array, the location of greatest flow deceleration moves significantly farther from the structure when the blocking parameter increases. The sharp velocity gradient appears unrealistic and indicates occurrence of flow separation and high turbulence.

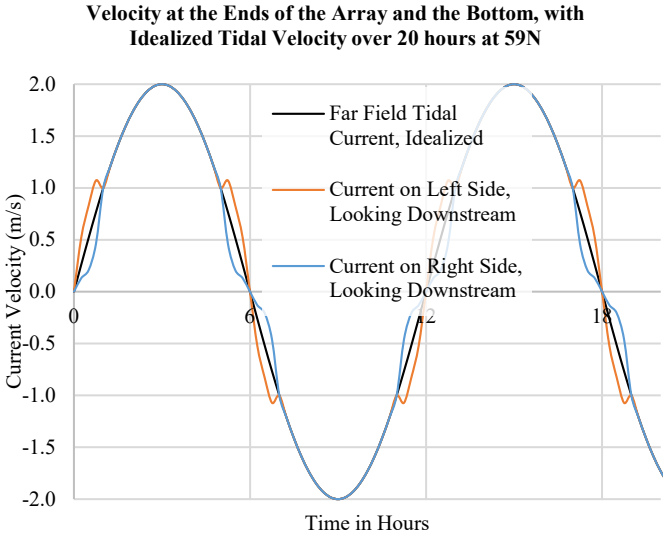


Figure 1: Graph of the idealized far field tidal velocity U and the different velocities at both ends of a linear array aligned perpendicular to the tidal velocity. Solutions for perpendicular flow and $\gamma = 1{,}000$

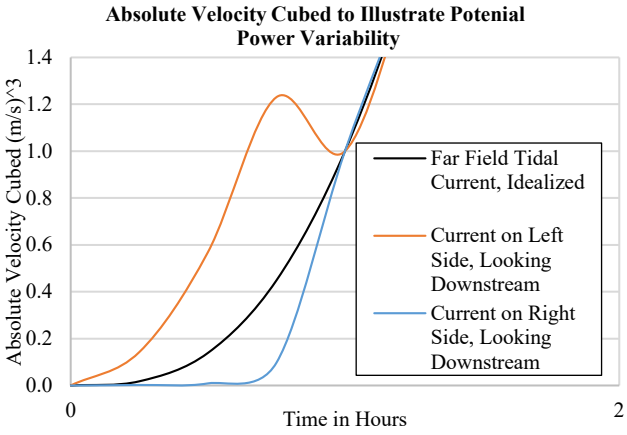


Figure 2: The power produced from an ocean current is proportional to the velocity cubed. A change in the flow around the structure during a tidal change may be experienced by sensitive power generation systems.

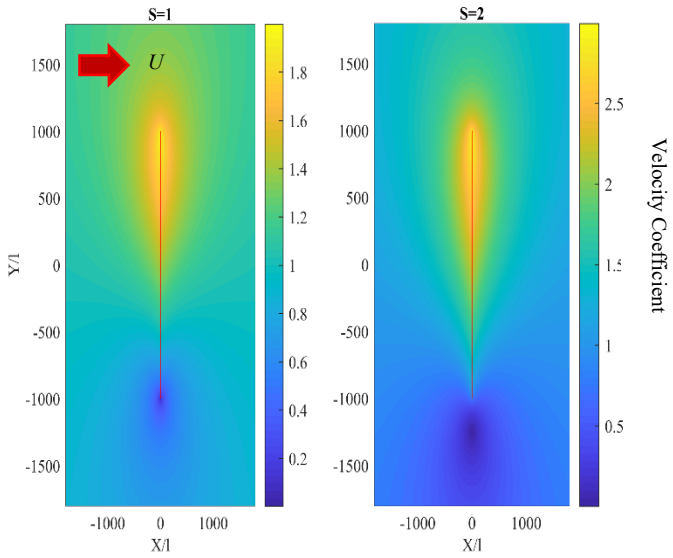


Figure 3: Modeled flow velocity change factors are plotted in color for flow at the depth of the array during quasigeostrophic flow with a blocking parameter, S, equal to 1 and 2. Flow direction is from the left to the right. The coordinate system is normalized using l, the minor axis of the elliptical body, which is 10m for the proposed array. The major axis L is 10,000m. A velocity coefficient of 1 corresponds to the far field velocity U.

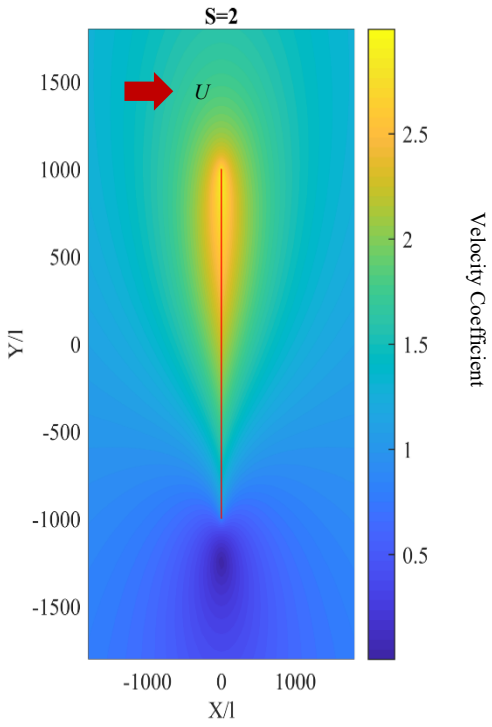


Figure 4: Modeled flow velocity coefficients are plotted in color for flow at the depth of the array during quasigeostrophic flow with a blocking parameter equal to 2. A reduced velocity will increase the blocking parameter. The decelerated flow has moved away from the obstruction, as compared with S=1. More of the flow adjacent to the obstruction has a velocity coefficient >1 when compared with S=1. Flow direction is from the left to the right. The coordinate system is normalized using l, the minor axis of the elliptical body.

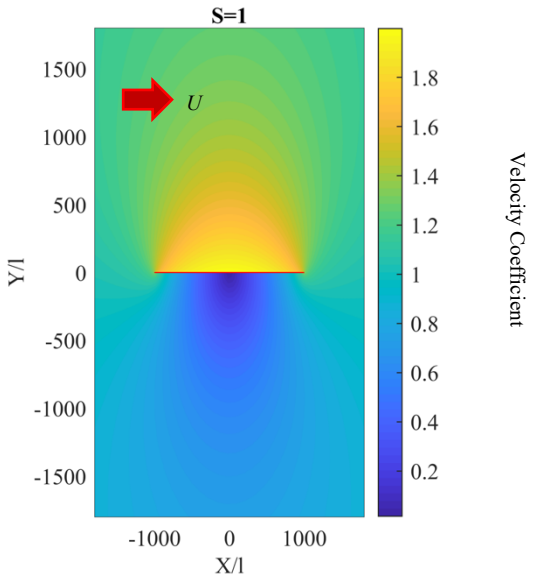


Figure 5: The asymmetry of the flow is striking for an array oriented perpendicular to the current. The area adjacent to the array has the greatest change in flow velocity. The array is modeled as static. Flow past the array has a blocking parameter, S = 1. The far field velocity has a velocity coefficient of one. The coordinate system is normalized using l, the minor axis of the elliptical body.

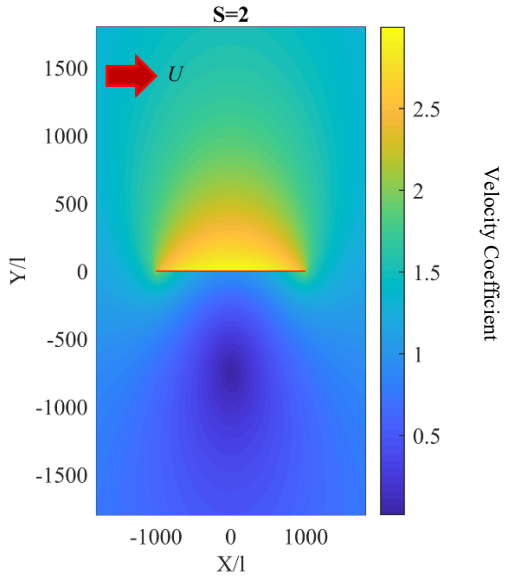


Figure 6: Flow past the array during quasigeostrophic flow with a blocking parameter, S = 2. The darkest blue point is the minimum velocity in the vicinity of the structure and this point has moved a significant distance, approximately half the characteristic length, L, away from the structure.

4.1. Design Plan

For a line of tidal current turbines at mid to high latitudes rotational flow effects caused by the Coriolis force can impact local dynamics at periods with small inertial currents. The strongest changes in flow dynamics can occur directly adjacent to a structure obstructing a quasigeostrophic flow. Flow acceleration and deceleration can occur around the structure asymmetrically. It appears prudent to make the following design considerations:

1) ensure that the linear array is placed perpendicular to the flow;
2) review local tidal current data and decide if both ebb and flood currents, or just one direction, will be used for power production;
3) prepare your system for an asymmetric velocity gradient, down the line of the array, at periods of tidal change.

One design recommendation for a site where both ebb and tide are utilized for energy production would be to separate the array into a minor and a major field with the two fields separated by supporting equipment. The major field placed on the left side of the array (looking downstream), for the strongest tidal direction. Dummy turbines could be included in the central field of supporting equipment to elongate the array if the asymmetric flow pattern is especially desirable for a production location.

An array parallel to the incident flow is not recommended because of the sharp velocity gradient modeled over it for quasi-geostrophic flow. The modeled gradient is not realistic and suggests occurrence of flow separation and high turbulence.

5. Final Comments

Site specific investigations are crucial for effective array design and deployment. All sites will have unique bathymetry and unique local currents. The variables affecting the layout of the array will include the angle of incident flow; the aspect ratio of the array; the water depth and height of the array; and the latitude at which the array is placed. Placement at higher latitudes will produce higher blocking parameters. The Taylor column phenomenon can be produced by an engineered structure, bottom seated or suspended in the ocean, or by a larger bathymetrical feature on which an engineered structure is placed.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

A special thanks to the Department of Ocean and Mechanical Engineering at Florida Atlantic University.

References

[1] V. Miglietta, M. Dhanak, "Current turbine array placement in quasigeosterophic flows over bottom topography," in Oceans Conference, Seattle, WA, USA, 2019.
[2] E. Johnson, "Quasigeostrophic flow over isolated elongated topography," Deep-Sea Research, **29**(9A), 1085-1097, Pergamon Press Ltd., Great Britain, 1982.
[3] C. Garrett, P Cummins, "The efficiency of a turbine in a tidal channel," Journal of Fluid Mechanics, **588**, 243–251, 2007, doi:10.1017/S0022112007007781.
[4] R. Vennell, "Tuning turbines in a tidal channel," Journal of Fluid Mechanics, **663**, 253-267, 2010, doi:10.1017/S0022112010003502
[5] R. Vennell, "Realizing the potential of tidal currents and the efficiency of turbine farms in a channel," Renewable Energy, **47**, 95-102, Elsevier, 2012, doi: https://doi.org/10.1016/j.renene.2012.03.036
[6] C. Christian, R. Vennel, "Efficiency of tidal turbine farms," Coastal Engineering Proceedings, **1**(33), structures.4, 2012, doi:https://doi.org/10.9753/icce.v33.structures.4
[7] S. Draper, T. Nishino, "Centered and staggered arrangements of tidal turbines," Journal of Fluid Mechanics, 739, 72-93, Cambridge University Press, 2014, doi:10.1017/jfm.2013.593

[8] G. Carter and M. Gregg and M. Merrifield. "Flow and mixing around a small seamount on Kaena Ridge, Hawaii," Journal of Physical Oceanography, **36**, 1036-1052, 2006.
[9] R.E. Brainard, "Fisheries aspects of seamounts and Taylor columns," Master's Thesis, Naval Postgraduate School. Monterey, California, 1986.
[10] W. Owens and N. Hogg. "Oceanic observations of stratified Taylor columns near a bump," Deep-Sea Research, **27A**, 1029-1045, Pergamon Press Ltd., Great Britain, 1979.
[11] A. Rogers, The biology of seamounts:25 years on., **79**, Advances in marine biology, 2018. Doi 10.1016/bs.amb.2018.06.001
[12] G.I. Taylor, "Motion of solids in fluids when the flow is not irrotational," Proc. Royal Society of London, Series A, Containing papers of a mathematical and physical character, **93**, 99-113, 1917.
[13] M. Buckley, "Taylor Columns," Course website for weather and climate laboratory at the Massachusetts Institute of Technology, 2004, Accessed June 2015. http://paoc.mit.edu/12307/reports/tcolumns.pdf
[14] J.W.M. Bush, H.A. Stone, J.P. Tanzosh, "Particle motion in rotating viscous fluids: Historical survey and recent developments," Current Topics in The Physics of Fluids, **1**, 337-355, 1994.
[15] J. Clarke et al. "Development and in sea performance testing of a single point mooring supported contra-rotating tidal turbine," in ASME 28th International Conference on Ocean Offshore and Arctic Engineering, 2009, doi:10.1115/OMAE2009-79995.
[16] R. Hide, "Origin of Jupiter's Great Red Spot." Nature, **190**, 895-896, 1961.
[17] P.J. Mason, R.I. Sykes, "A numerical study of rapidly rotating flow over surface-mounted obstacles," Journal of Fluid Mechanics, **111**, 175-195. Great Britain, 1981.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 698-709 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

An Innovative Angle of Attack Virtual Sensor for Physical-Analytical Redundant Measurement System Applicable to Commercial Aircraft

Antonio Vitale*, Federico Corraro, Nicola Genito, Luca Garbarino, Leopoldo Verde

CIRA – Italian Aerospace Research Centre, On-Board Systems and ATM, Capua (CE), 81043, Italy

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 12 December, 2020
Online: 05 February, 2021

Keywords:
Angle of attack
Estimation
Fault detection
Kalman filter
Loss of control in flight
Monte Carlo simulation
Virtual sensor

ABSTRACT

The angle of attack is a critical flight parameter for commercial aviation aircraft, because automatic envelope protection systems rely on it to keep the aircraft within its safe flight envelope. Faulty measurements of the angle of attack could have catastrophic effects, leading to aircraft loss of control in flight and fatalities, as demonstrated by the recent accidents involving the Boeing 737-MAX. This paper presents a novel approach to the measurement of the angle of attack, which uses one virtual sensor and two physical sensors to implement a physical-analytical redundant system that is robust to a single fault of the physical sensors. The virtual sensor is based on an innovative and reliable estimator of the angle of attack. It was originally developed to provide General Aviation pilot with an accurate indication of trend toward stall, and has been suitably customized to fit its application to commercial aviation. One of the peculiarities of the redundant measurement system is that its implementation on-board several existing commercial aviation aircraft only needs the integration of a software code and does not require any installation of additional physical sensors. The proposed approach demonstrated very interesting performance, assessed in simulation through several Monte Carlo analyses. Its exploitation could contribute to reduce the angle of attack related accidents, improving the safety of the air transport system.

1. Introduction

This paper is an extension of work originally presented in SYSTOL'19 [1]. The original work defined an innovative method for the estimation of the angle of attack (AoA), which does not require any physical sensor dedicated to the measurement of the AoA. The present paper proposes a novel measurement system. It exploits the innovative estimator described in [1] and two physical sensors of the AoA, usually available on commercial aircraft, to compute a more reliable measurement of the angle of attack, which is robust to a single fault of the physical sensors. The paper defines the architecture and the algorithms of the proposed measurement system and discusses in depth its performance, which has been assessed through Monte Carlo simulations.

Loss of Control – In Flight (LOC-I) defines a condition in which the flight crew is unable to maintain the control of an aircraft while it is flying, resulting in an unrecoverable deviation from the intended flight path. LOC-I is the most significant cause of fatal accidents in General Aviation (GA): there are approximatively 37 fatal LOC-I accidents per year in Europe involving GA aircraft, leading to 67 persons on average losing their lives every year (for fixed-wing aircraft only) [2]. The LOC-I accidents can result from several contributing factors, which can act individually or in combination. They often result from failure to prevent or recover from a stall. An aircraft stalls if it exceeds the critical angle of attack and it may occur at any indicated airspeed. Hence, the only use of an airspeed indicator is unreliable to detect the proximity to stall. The measurement of the angle of attack could complement the airspeed information, prevent to enter an upset condition and allow performing safer manoeuvers. General Aviation aircraft often are not equipped with AoA sensors. The availability of suitable methods or affordable and reliable instruments to estimate or measure the angle of attack and to display its current value to the pilot would be dramatically beneficial for this aircraft category. Indeed, it could facilitate earlier detection of danger conditions and could permit adjusting power and configuration to avoid displacing the aircraft from the intended path. Additionally, the knowledge of the angle of attack could reduce the pilot workload [3], [4].

*Corresponding Author: Antonio Vitale, Italian Aerospace Research Centre, via Maiorise, 81043, Capua (CE), Italy, Email: a.vitale@cira.it

Based on these motivations, several methods for the estimation of the angle of attack are described in the literature. Many methods use the Kalman Filter [5]-[9], which exploits nonlinear kinematics equations for the propagation of the AoA estimation, and measurement models, coupled with suitable measurements, to correct the estimation. Usually, this approach is model-based and requires the knowledge of the aircraft aerodynamic model [5]-[7]. The estimator proposed in [8] does not need the aerodynamics of the aircraft neither other aircraft parameters; but the method works properly only if the aircraft dynamics are excited by continuously changing pitch and yaw angles. Model-based estimation methods that do not use the Kalman Filter are proposed in [10] and [11]; both these approaches require as input a detailed model of the aircraft aerodynamics. In [10] the data fusion of GPS (Global Positioning System) and IMU (Inertial Measurements Unit) sensors together with the aerodynamic information allow computing the angle of attack, whereas in [11] a Bayesian estimator provides it. In [12] the angle of attack calculation only relies on inertial data, but obtained estimation accuracy is not good in windy and turbulent conditions. The authors of [13] estimate the AoA by using a complementary filter. The steady state low frequency component of the estimation is obtained by solving algebraically the angle of attack kinematic equation, in which the AoA derivative is set equal to the pitch rate. The angle of attack high frequency content is considered coincident with the pitch rate. However, this simple method provides estimations which are characterized by high mean error and standard deviation. A Virtual Sensor is presented in [14]. It uses a Functional Pooling Non-linear Auto-Regressive with eXogenous excitation (FP-NARX) methodology that provides very good performance. However, this method requires the availability of a wide set of flight data, which shall cover the whole flight envelope of the aircraft in order to identify the NARX model that reconstructs the AoA behavior. Model-free virtual sensors for the angle of attack measurement are available in the literature, too. They mainly use neural networks [15], [16], which has the drawback to need huge amount of data for the training of the network.

This paper is an extension of the work presented in [1], which describes an innovative approach to estimate the angle of attack, developed and patented [17] by the Italian Aerospace Research Centre (CIRA) and ASPEN Avionics. The new AoA estimator aimed at dramatically improving the safety of GA aircraft, by providing an immediate, accurate, and clear visual display of trend toward stall and stall margin. It does not need the installation of external dedicated sensors, which is the key factor in limiting the adoption of angle of attack instruments by General Aviation pilots. The AoA computation exploits an Extended Kalman Filter and the measurements provided by navigation sensors that are more reliable than physical angle of attack sensors. The main innovative feature of the proposed method with respect to what already proposed in the literature is that it does not need detailed information about the aerodynamics of the aircraft neither huge amount of flight data for calibration/training. It only requires the knowledge of few aircraft parameters, easily gathered from the Pilot Operating Handbook (POH), and the execution of a short calibration flight, to carry out just once. This calibration flight is aimed at identifying the relevant aircraft aerodynamic parameters needed to tune the algorithm for the estimation of the angle of attack. The proposed estimator is operative on-board several hundreds of private General Aviation aircraft since July 2015 without experiencing any problem.

The availability of reliable information about the angle of attack is very valuable to commercial aircraft, too. Higher levels of automation support the piloting tasks of commercial aircraft: the envelope protection systems exploit the angle of attack measurement to implement automatic corrective actions if the aircraft is approaching its flight envelope boundaries. These systems should improve the flight safety. However, in case of wrong angle of attack measurement, they could lead to LOC-I and fatalities, especially if the aircraft is performing critical flight phases such as take-off or landing. In fact, LOC-I is the leading cause of fatal accidents in commercial aviation and it is one of the accident categories with the lowest survivability ratio [18].

These statistics are confirmed by the recent flight accidents in March 2019 [19] and in October 2018 [20], which involved the Boeing 737-8 (MAX) operated by the Ethiopian Airlines and by the Lion Air, respectively. The investigation reports state that both the aircraft had a failure of one angle of attack sensor that dramatically affected the behavior of the envelope protection system. The Boeing 737 MAX has two independent physical sensors to measure the angle of attack, one sensor on each side of the forward fuselage. These sensors consist of an external vane, which rotates to align with the local airflow, connected to an internal resolver that measures the rotation angle. The aircraft is also equipped with a stability augmentation function, called the Maneuvering Characteristics Augmentation System (MCAS), which aims to improve the aircraft handling characteristics. Specifically, the MCAS was introduced to counteract the aircraft tendency to pitch-up when it flies at elevated angles of attack. This unwanted behavior is an effect of the new engines that were installed in the 737 MAX with respect to the previous 737 models. During the Ethiopian Airlines accident flight, soon after the lift-off, the left and right vanes provided different measurements of the AoA: the left measurement reached 74.5 degrees, while the right sensor showed a maximum value of 15.3 degrees. The difference between left and right AoA measurements (about 59 degrees) remained constant until the end of the recording [21]. The erroneous measurement of the angle of attack was not detected and it affected the pitch command. Indeed, the MCAS considered valid the left sensor's measurement and pushed down the nose of the aircraft several times. The crew tried to keep the aircraft along its planned flight path, but lost the control and the aircraft crashed. Figure 1, that is an excerpt of [21], shows the angle of attack measurements before the crash, pointing out the failure of the left sensor. The Flight Data Recorder of the Lion Air accident also registered differences between the measurements of the angle of attack (about 20 degrees) provided by the left and right sensors, as shown in Figure 2, taken from [22]. The erroneous angle of attack measurement produced multiple alerts and repetitive MCAS activations, which combined with pilot distractions due to numerous ATC communications contributed to the flight crew difficulties to control the aircraft and leaded to the crash after a steep dive.

Faulty angle of attack measurement was the cause of other incidents in the past, such as the crash of the Airbus A320-232 operated by XL Airways Germany on 27th November 2008 [23] and the incident of the Airbus A321-231 operated by Lufthansa

(flight 1829) on 5th November 2014. In this last case, the fault of the angle of attack sensor caused the anti-stall system to push down the nose of the aircraft, but since the aircraft was flying at high altitude, the crew was able to react, avoiding catastrophic results [24]. All these incidents remark that the availability of a fault tolerant system for the measurement of the angle of attack is very valuable. Triplex redundancy of physical sensors could be a solution to be robust with respect to a single sensor fault; however, triplex hardware redundancy is not always implemented for angle of attack sensor (the Boeing 737 MAX is an example).

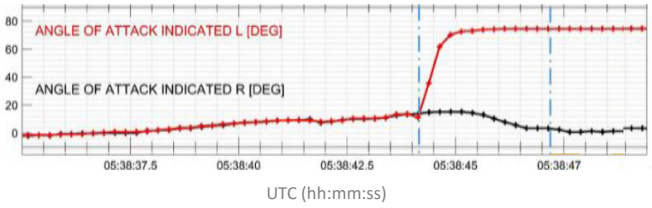


Figure 1: AoA measurements for the Ethiopian Airlines accident [21]

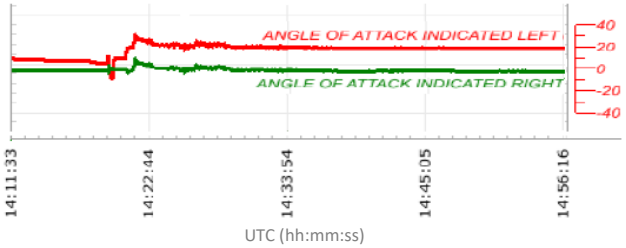


Figure 2: AoA measurements for the Lion Air accident [22]

The authors of the present paper deem that a suitable customization of the AoA estimator developed for GA aircraft and presented in [1] and [17] could help addressing the issues discussed for Commercial Aviation aircraft. Indeed, it could be exploited as an additional virtual sensor that, coupled with existing physical sensors and suitable voting algorithms [25]-[27], allows to implement a physical-analytical redundant measurement system that is robust to a single fault of the physical sensors. This measurement system could significantly improve the safety of the aircraft, reducing the occurrence of LOC-I related accidents, saving human lives, and avoiding the economic impact on the airliners and the aircraft manufacturers consequent to those accidents. The main benefit of this research to the industry consists in the possibility to achieve this result without the need for hardware modification to the existing aircraft, installation of additional physical sensors or change to the avionics hardware architecture. In fact, the proposed approach just requires the installation of the software code for the computation of the robust angle of attack measurement, which exploits as input the measurements provided only by the sensors already available on-board.

The next sections of the paper detail the proposed approach. The innovative AoA estimator is first described in section 2. Next, a possible application to commercial aviation aircraft is proposed, defining the architecture and the algorithms to implement a robust fault tolerant measurement system. Finally, section 4 shows the performance of the proposed method, assessed in simulation through Monte Carlo analysis. A section of conclusion ends the paper.

2. Angle of Attack Virtual Sensor

The design of the proposed angle of attack estimator was originally based on the following requirements:

- to be applicable to a wide set of aircraft;
- to exploit the information commonly available for GA aircraft;
- the installation of specific additional sensors shall not be required.

These requirements imply that the estimation method shall not require the knowledge of the detailed aerodynamic model, that is specific of each type of aircraft.

The proposed approach works in two steps. First, the parameters of the estimation algorithm are tuned through a calibration procedure, which consists in performing a flight test that shall be executed just once. Next, normal operations are carried out, that is, the estimator could compute the angle of attack during the flight.

The proposed method uses the measurements provided by sensors which are usually exploited for primary navigation and does not need a dedicated angle of attack sensor. It guarantees that the measurement's accuracy is independent upon the installation of the AoA sensor. It is worthy to note that the same set of sensors is used both for the calibration and in the normal operations. Figure 3 shows the basic concept of the virtual sensor. The following subsections details the calibration procedure and the angle of attack computation.

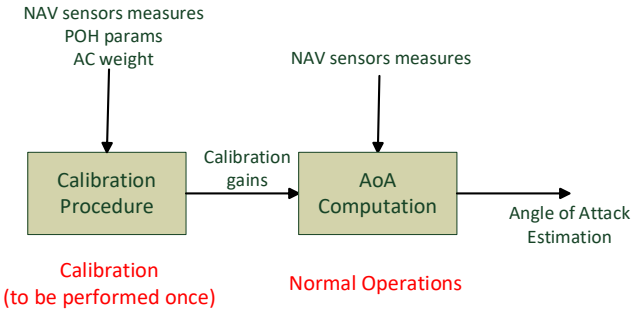


Figure 3: AoA virtual sensor basic concept

2.1. Calibration Procedure

The calibration procedure aims at identifying the lift curve of the aircraft and consists in flying steady and wing leveled at two different airspeed values (set points), which are close to the boundaries of the aircraft flight envelope. The test is repeated for each aircraft configuration that has significantly different aerodynamic characteristics. The calibration requires the knowledge of:

- few parameters available in the POH: MTOW (Maximum Take-off Weight), SEW (Standard Empty Weight), Stall Speed and Maximum Speed in clean and flapped configurations
- aircraft weight (it is worthy to remark that the weight shall be known only during the calibration flight and not during normal operations of the virtual sensor).

The lift (L) is assumed linear with respect to the angle of attack (α) and to the deflection of the longitudinal aerodynamic surfaces (δ_i):

$$L = QSC_L = QS\left(C_{L0} + C_{L\alpha}\alpha + \sum_i C_{L\delta}^i \delta_i\right) \quad (1)$$

where Q is the dynamic pressure, C_L is the lift coefficient, S is the aerodynamic reference surface, C_{L0}, $C_{L\alpha}$ and $C_{L\delta}^i$ are the lift aerodynamic derivatives. The contribution to the lift of all the longitudinal surfaces deflection is assumed negligible, except for the flaps (δ_F). Two flap configurations are considered for GA aircraft, clean and full flap. Each configuration has a different lift curve, which is still assumed linear and is uniquely identified by its slope ($C_{L\alpha}$) and intercept (C_{L0}). The calibration procedure is performed for both the configurations, the corresponding aerodynamic derivatives C_{L0}, $C_{L\alpha}$ are computed, and a linear variation for these derivatives is assumed to compute their values for intermediate configurations of the flaps. This approach implies that the values of C_{L0} and $C_{L\alpha}$ depend on the deflection of the flaps, whereas the explicit dependence in equation (1) of L on δ_F could be delated; therefore, the equation (1) could be reformulated as:

$$L = QS(C_{L0}(\delta_F) + C_{L\alpha}(\delta_F)\alpha) = QW\left(\hat{k}_0(\delta_F) + \hat{k}_1(\delta_F)\alpha\right) \quad (2)$$

where W is the aircraft weight and the parameters $\hat{k}_0, \hat{k}_1$, named calibration gains, are defined as follows (the dependence of C_{L0}, $C_{L\alpha}$, $\hat{k}_0$ and $\hat{k}_1$ on δ_F is not shown for the sake of conciseness):

$$\hat{k}_1 = \frac{S}{W} \cdot C_{L\alpha} \quad (3)$$

$$\hat{k}_0 = \frac{S}{W} \cdot C_{L0} \quad (4)$$

During the calibration flight, the aircraft weight is denoted as W_{cal} and the calibration gains (a set for each configuration) are computed as defined in [17]:

$$\hat{k}_{1cal} = \frac{S}{W_{cal}} \cdot C_{L\alpha} = \left(\frac{1}{Q_2} - \frac{1}{Q_1}\right) / (\vartheta_2 - \vartheta_1) \quad (5)$$

$$\hat{k}_{0cal} = \frac{S}{W_{cal}} \cdot C_{L0} = \left(\frac{1}{Q_1} - \frac{{}^1/_{Q_2} - {}^1/_{Q_1}}{\vartheta_2 - \vartheta_1} \cdot \vartheta_1\right) \quad (6)$$

where ϑ is the pitch angle. All the variables on the right-hand sides of the above equations represent measurements gathered during the calibration flight and the subscripts 1 and 2 refer to the set point index (two set points for each aircraft configuration). Therefore, the calibration gains could be easily evaluated.

Replacing (5) and (6) in (2) and rearranging, we get the lift equation applicable during normal operations [17]:

$$L = QW_{conv}(k_0 + k_1\alpha) \quad (7)$$

$$\hat{k}_{0cal} = \frac{S}{W_{cal}} \cdot C_{L0} = \left(\frac{1}{Q_1} - \frac{{}^1/_{Q_2} - {}^1/_{Q_1}}{\vartheta_2 - \vartheta_1} \cdot \vartheta_1\right) \quad (8)$$

where W_{conv} is the actual aircraft weight if it is known, otherwise it is a conventional weight given by [17]:

$$W_{conv} = max\{W_{cal}, [C_0 + W_{SEW} + C_1(W_{MTOW} - W_{SEW})]\} \quad (9)$$

The values of the constants C_0 and C_1 are computed with the aim to optimize the angle of attack estimation error.

2.2. Angle of Attack Computation

After the completion of the calibration procedure, the virtual sensor can estimate the angle of attack. The estimation algorithm exploits an Extended Kalman Filter (EKF), which combines the knowledge of the lift curve, identified during the calibration, with the measurements get from ADS (air-data system), GPS and IMU. The EKF state vector includes five variables, that is, the angles of attack (α) and sideslip (β) and the wind velocity components (W_N, W_E, W_D), expressed in North-East-Down (NED) reference frame. Classical kinematic equations in polar form [28] describe the dynamics of the aerodynamic angles, whereas stochastic zero-order Gauss-Markov processes [29] are used to represent the wind velocity behavior. These equations are applicable to any type of aircraft [17]:

$$\begin{aligned}\dot{\alpha} = \frac{1}{TAS\cos\beta}&(a_z\cos\alpha - a_x\sin\alpha) + q \\ &- \tan\beta\,(p\cos\alpha + r\sin\alpha) \\ &+ \frac{1}{TAS\cos\beta}[u_w q\cos\alpha \\ &- v_w(r\sin\alpha + p\cos\alpha) \\ &+ w_w q\sin\alpha] + \eta_\alpha\end{aligned}$$

$$\begin{aligned}\dot{\beta} = \frac{1}{TAS}&(-a_x\cos\alpha\sin\beta + a_y\cos\beta \\ &- a_z\sin\alpha\sin\beta) + p\sin\alpha \\ &- r\cos\alpha \\ &+ \frac{1}{TAS}[-u_w(r\cos\beta \\ &+ q\sin\alpha\sin\beta) \\ &+ v_w(p\sin\alpha\sin\beta \\ &- r\cos\alpha\sin\beta) \\ &+ w_w(p\cos\beta \\ &+ q\sin\beta\cos\alpha)] + \eta_\beta\end{aligned}$$

$$\dot{W}_i = \eta_{W_i} \qquad with\ i = N, E, D \quad (12)$$

where:

- a_x, a_y and a_z are the inertial acceleration components in the body reference frame (provided by the IMU);
- p, q and r are the angular rates components (roll, pitch and yaw rate) in the body reference frame (provided by the IMU);
- TAS is the module of the true air speed (provided by the ADS);
- $\eta_\alpha, \eta_\beta, \eta_{W_N}, \eta_{W_E}$ and η_{W_D} are process noises, which are assumed zero mean multivariate Gaussian with covariance matrix Q_{noise};
- u_w, v_w and w_w are the wind velocity components expressed in the body reference frame and computed as:

$$\begin{bmatrix} u_w \\ v_w \\ w_w \end{bmatrix} = \bar{\bar{R}}_{BI}^{-1} \times \begin{bmatrix} W_N \\ W_E \\ W_D \end{bmatrix} \quad (13)$$

- $\bar{\bar{R}}_{BI}$ is the Transformation matrix from body axes to NED axes.

The measurement vector is given by the inertial velocity components (V_N, V_E, V_D) in the NED reference frame (provided by the GPS) and the normal load factor (N_z) in body axes (provided by the IMU). The following equations are used [17]:

$$\begin{bmatrix} V_N \\ V_E \\ V_D \end{bmatrix} = \bar{\bar{R}}_{BI} \times \begin{bmatrix} TAS \cdot \cos\alpha \sin\beta \\ TAS \cdot \sin\beta \\ TAS \cdot \cos\beta \sin\alpha \end{bmatrix} + \begin{bmatrix} W_N \\ W_E \\ W_D \end{bmatrix} + \begin{bmatrix} v_{V_N} \\ v_{V_E} \\ v_{V_D} \end{bmatrix} \quad (14)$$

$$N_z = Q \cdot (k_1 \cdot \alpha + k_0) + tg(\alpha) \cdot N_x \quad (15)$$

where:

- N_x is the axial load factor;
- v_{V_N}, v_{V_E} and v_{V_D} are measurement noises, which are assumed zero mean multivariate Gaussian with covariance matrix R_{noise};
- k_1 and k_0 are the calibration gains.

The last measurement equation holds in the hypotheses of thrust aligned to the X-body axis (that is, with negligible contribution to the normal load factor) and small angle of attack approximation. If a measurement of the current thrust is available, then the contribution along the normal load factor could be added in the equation to remove the approximation. Finally, it is also worthy to note that if the calibration gains k_1 and k_0 are computed using the conventional weight instead of the actual one, then an approximation is introduced that affects the computation of the right-hand side of the equation (15) and consequently of the angle of attack. The availability of the measurement of the actual weight allows removing this approximation and improving the estimation results.

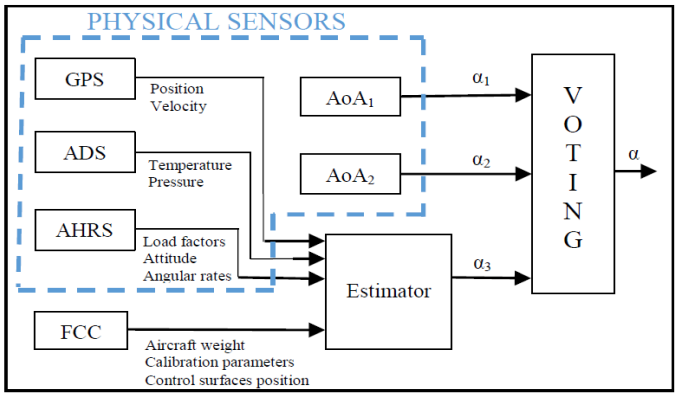


Figure 4: Physical-analytical redundant angle of attack architecture

3. Fault Tolerant Angle of Attack Measurement System for Commercial Aircraft

The proposed angle of attack estimator is applicable to commercial aviation aircraft after a suitable customization. Commercial aircraft are already equipped with one or more (but sometimes less than three [25]) physical angle of attack sensors. Indeed, the angle of attack measurement is nowadays a critical flight parameter as well as other primary navigation measurements. The introduction of a virtual sensor provides an additional and independent measurement, which in conjunction with the physical sensors, allows implementing a single fault tolerant measurement system based on physical-analytical redundancy. Figure 4 shows a possible architecture of the proposed angle of attack measurement system, which exploits two physical sensors of the AoA.

The following subsections detail the customization of the estimator with respect to the one developed for GA aircraft and the voting algorithms applied to compute the consolidated AoA measurement out of the three AoA inputs.

3.1. Virtual Sensor Customization to Commercial Aviation

Some of the assumptions introduced in the design of the estimator for GA application can be removed if the estimator is applied to commercial aircraft.

First, in commercial flight an estimation of the aircraft weight is entered in the aircraft Flight Control Computer (FCC) before each flight, therefore it is a known parameter (although roughly) that can be provided as input to the AoA estimator. The use of a fuel consumption model allows computing the weight variation during the flight execution. Consequently, in the equations (7) and (8) the conventional weight can be replaced by the actual one and equation (9) is not used. Moreover, it is worthy to note that the availability of two physical sensors of the angle of attack allows improving the knowledge of the aircraft weight online while the aircraft is flying. Indeed, when the physical measurements coincide (that means supposedly the physical sensors are working properly), the value of the aircraft weight in the AoA estimation algorithm could be tuned (within a reasonable range) to force the virtual sensor measurement to converge to the physical ones. A suitable strategy can be defined to perform periodically this check and possibly to update the aircraft weight.

Second, commercial aircraft have several control aerodynamic surfaces and their configuration affects the lift curve of the aircraft. The deflection of each of these control surfaces is available to the FCC and it shall be used as an additional input to the AoA estimation algorithm. This consideration implies a modification to the aerodynamic modelling. The lift equation is more complex than the one defined by equation (7). It shall consider the effect of all the relevant aerodynamic surfaces, which in first approximation are assumed having a linear effect (however, quadratic, cubic or in general polynomial effects could be also used). Accordingly, the lift is given by:

$$L = QSC_L = QS\left(C_{L0} + C_{L\alpha}\alpha + \sum_i C^i_{L\delta}\delta_i\right) = QW\left(k_0 + k_1\alpha + \sum_i k^i_\delta \delta_i\right)$$

where δ_i is the deflection of the i-th aerodynamic surface and k^i_δ is an additional calibration gain for each surface (more than one gain could be introduced for each surface if nonlinear effects are considered). Concerning the calibration process, the main differences with respect to the General Aviation aircraft are:

- an increased number of set points during the calibration flight shall be gathered, due to the increased number of calibration gains that shall be computed;

- the measurement of the deflection of each aerodynamic control surface is available and could be used as input to the calibration computation;
- the angle of attack measurement is available during the calibration (provided by the physical sensors), therefore it shall not be replaced by the pitch angle measurement.

The calibration test is performed first gathering two set points at different speeds (close to the aircraft's envelope limits) in clean configuration. The measurements in these set points allow the computation of the gains k_0 and k_1 by using equations (5) and (6), in which the angle of attack measurements replace the pitch angle ones. Next, two additional set points are gathered for each aerodynamic surface. These set points are collected again at two different speeds; only the aerodynamic surface related to the gain to be computed is deflected, whereas the deflections of all the other surfaces are null. Each gain is computed through:

$$k_\delta^i = \frac{S}{W_{cal}} C_{L\delta_i} = \left(\frac{Q_{i1}-Q_{i2}}{Q_{i2}\delta_{i2}-Q_{i1}\delta_{i1}}\right) k_0 + \left(\frac{Q_{i1}\alpha_{i1}-Q_{i2}\alpha_{i2}}{Q_{i2}\delta_{i2}-Q_{i1}\delta_{i1}}\right) k_1 \quad (17)$$

where the subscripts i_1 and i_2 identify the two set points related to the i-th surface; α, δ and Q are measured and k_0 and k_1 are already computed at this stage.

The availability of measurements of the actual weight of the aircraft and of the aerodynamic control surfaces deflection allows removing the approximations introduced in the estimator algorithm designed for GA application, leading to a significant improvement of the estimation accuracy.

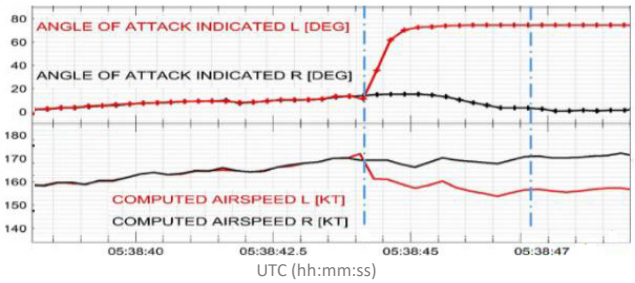


Figure 5: Effect of faulty AoA measurement on the airspeed computation [21]

Finally, it is worthy to remark that, during normal operations, the angle of attack computed by the estimator shall be completely independent from the AoA measurements provided by the physical sensors, in order to implement correctly the analytical redundancy. The estimator, as shown in Figure 4, exploits the measurements from GPS, AHRS and ADS sensors. GPS and AHRS are completely independent from the vane sensor. On the other hand, the AoA measured by the vane is usually exploited to correct the raw value of the static pressure measured by the ADS, in order to compensate for errors due to sensor position and aerodynamic effect. The corrected static pressure is then used by the air data computer to calculate the true airspeed, which therefore is influenced by the AoA measured by the vane. The effect of the AoA on the computation of the airspeed is pointed out by the data of the Ethiopian Airlines accident flight. Figure 5, excerpt of [21], shows the computed airspeed based on the static pressure corrected with left and right angle of attack measurements. The erroneous AoA values, provided by the left sensor, produced corrected static pressure values greater than the true ones and consequently the computed airspeed values are lower than the true ones.

The row static pressure values (before applying the correction based on the AoA measurement) from both sides were identical, as shown in Figure 6 [21]. The different static pressure measurements highlighted in Figure 5 were therefore only due to the AoA correction. This effect should be considered in the selection of the inputs to the angle of attack estimation algorithm. Specifically, the uncorrected values of all the air data sensors shall be provided to the estimator. Possibly corrections of these inputs depending on the angle of attack shall be based on the AoA value computed by the estimator itself, at the previous computation time step. This approach guarantees the independence of the measurements provided by the analytical and physical sensors and used as input to the voting algorithm.

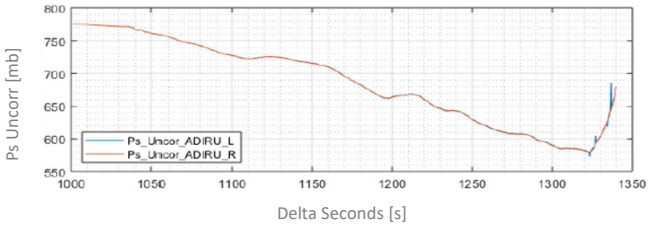


Figure 6: Static pressure reconstruction without AoA correction [21]

3.2. Voting and Signal Validity Check Algorithms

The voting algorithm consists in generating one single angle of attack measurement (consolidated value) out of three input signals of a triplex redundant architecture. Different voting techniques are available in the literature. The proposed approach applies a weighted voting, in which the consolidated signal is computed through a weighted mean of the valid inputs:

$$\alpha = \sum_{i=1}^{3} p_i \alpha_i \quad (18)$$

$$\sum_{i=1}^{3} p_i = 1 \qquad with \qquad 0 \le p_i \le 1 \quad (19)$$

where α is the consolidated value, α_i is the i-th measurement, and p_i is the i-th weight. Several choices could apply to the definition of the weights. They can be all equal, or a bigger weight could be assigned to the median input [25], or the weights could be all different, based on some metrics or criteria [26]. For each signal we define a weight that is inversely proportional to the supposed precision of the sensor providing the measurement. Assuming the input measurements are affected by a Gaussian error with standard deviation σ_i, the following weights are used:

$$p_i = F_i \frac{\frac{1}{\sigma_i}}{\sum_{i=1}^{3} \frac{1}{\sigma_i}} \quad (20)$$

where F_i is the validity flag of the i-th measurement, which is null if the input measurement is invalid, otherwise it values 1. To check the validity of each input, the algorithm computes at each time step the difference between the measurements of the angle of attack provided by the available sensors, and for each couple of sensors compares the difference (D_{ij}) with a suitable threshold (T_{ij}):

$$D_{12}(t_k) = |\alpha_1(t_k) - \alpha_2(t_k)| < T_{12} \quad (21)$$

$$D_{13}(t_k) = |\alpha_1(t_k) - \alpha_3(t_k)| < T_{13} \quad (22)$$

$$D_{23}(t_k) = |\alpha_2(t_k) - \alpha_3(t_k)| < T_{23} \quad (23)$$

If both the differences related to the i-th measurement exceed the related thresholds (that is, the above relations involving the signal α_i are not satisfied) continuously for a predefined number of samples N_s, the i-th measurement is declared invalid and F_i is set equal to zero. Consequently, the computation of the consolidated value does not use the i-th measurement, because related weight is null. The thresholds are given by:

$$T_{ij} = C_{ij}(\sigma_i + \sigma_j) \quad (24)$$

The values of both the constant C_{ij} and the number of samples N_s used to check conditions (21) to (23) are selected as a trade-off, in order to maximize the capability to detect invalid signals while minimizing the false alarm rate.

4. Angle of Attack Measurement System Performance

The performance of the proposed measurement system has been assessed in simulation through Monte Carlo analyses. The Monte Carlo technique is a procedure for numerically obtaining an estimation of the statistical characterization (mean, variance, cumulative probability, etc.) of a function's output. To this end, the input parameters of the function vary independently according to their statistical characterization, and the output of the function is computed for each realization of these inputs. The statistics of the output are then evaluated. Results presented in this section concern the angle of attack estimator (also including the calibration procedure) and the triplex redundant measurement system. A simulation tool was implemented as test harness to execute the analyses.

4.1. Simulation Tool

The simulation tool is implemented in Matlab/Simulink environment and is composed of:

- The main routine, which allows:
  - setting up the flight test manoeuver,
  - varying the inputs of the Monte Carlo analysis according to their statistical characterization
  - computing the performance metrics.
- A simulation model, presented in Figure 7, which includes:
  - AC Model – the detailed parametric models for the simulation of the aircraft, the atmosphere (winds and turbulence) and the on-board sensors.
  - AP – an autopilot that is able to perform automatically the requested test manoeuver.
  - Algorithms Feature – a module implementing the algorithms for the angle of attack estimation.
  - Save Results – a module that saves the simulated inputs and computed outputs.

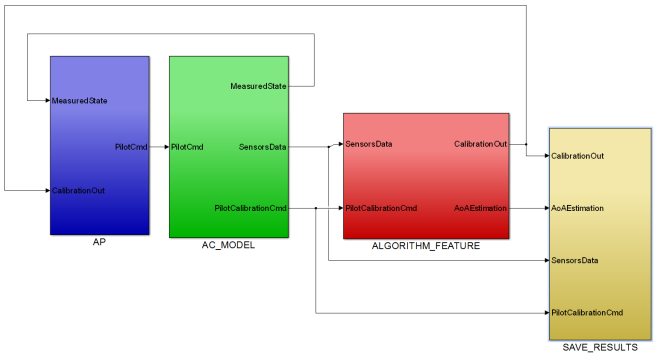


Figure 7: Simulation model for Monte Carlo analyses

- A routine that implements the triplex redundant AoA measurement system and computes the consolidated angle of attack.

The aircraft is modelled using the flight mechanics equations of a six degrees of freedom rigid body [30] and comprises a detailed aerodynamic model, also reproducing the stall phenomenon. The aircraft model has been validated in flight through the CIRA experimental aircraft [31]. The simulation of the relevant atmospheric phenomena is based on the ISA standard atmosphere [30] and the Dryden turbulence model [32]. The atmosphere model includes horizontal constant winds (variable with altitude), three-dimensional turbulence disturbances, and vertical gust (represented through "1-cosine" model) [32]. The three-dimensional turbulence is modelled by applying appropriate forming filters to band-limited white noise; the longitudinal, lateral, and vertical components (Φ_u, Φ_v, Φ_w) of the spectra of these forming filters are given by:

$$\Phi_u(\omega) = \frac{2\sigma_u^2 L_u}{\pi TAS} \frac{1}{1+\left(\frac{L_u\omega}{TAS}\right)^2} \quad (25)$$

$$\Phi_v(\omega) = \frac{2\sigma_v^2 L_v}{\pi TAS} \frac{1+12\left(\frac{L_v\omega}{TAS}\right)^2}{\left[1+4\left(\frac{L_v\omega}{TAS}\right)^2\right]^2} \quad (26)$$

$$\Phi_w(\omega) = \frac{2\sigma_w^2 L_w}{\pi TAS} \frac{1+12\left(\frac{L_w\omega}{TAS}\right)^2}{\left[1+4\left(\frac{L_w\omega}{TAS}\right)^2\right]^2} \quad (27)$$

where (L_u, L_v, L_w) and (σ_u, σ_v, σ_w) represent the turbulence scale lengths and intensities, respectively. The vertical gust velocity (V_{vg}) is computed through:

$$V_{vg}(x) = \begin{cases} 0 & x < 0 \\ \frac{V_m}{2}\left[1 - cos\left(\frac{\pi(x)}{d_m}\right)\right] & 0 \le x \le 2d_m \\ 0 & x > 2d_m \end{cases} \quad (28)$$

where V_m is the maximum gust magnitude and d_m is the gust half-width.

The sensors model simulates actual sensors (ADS, AHRS, GPS) that measure the signals needed by the algorithms under test. The models of the ADS and AHRS compute the measurements of the generic signal (s_m) at time t_k by adding bias (b_s) and Gaussian noise (ν_s) to the true signal value (s_t):

$$s_m(t_k) = s_t(t_k) + b_s + \nu_s(t_k) \quad (29)$$

A low pass filter is then applied to the signal s_m, in order to emulate the sensors dynamics (frequency band limitation). The GPS model also introduces a data latency (δt) in the measurements, which is assumed negligible for the other sensors [33]:

$$s_{mGPS}(t_k) = s_t(t_k - \delta t) + b_s + \nu_s(t_k - \delta t) \quad (30)$$

The data sheets of commercial off the shelf sensors are used to define bias, noise, latency and low pass filter characteristics. The sensors model allows injecting a failure in the measurement, such as additional bias, drift, or data freezing. In the Monte Carlo analyses, this feature was used only to simulate the failure of the angle of attack physical sensors.

4.2. Monte Carlo Analyses

Three different Monte Carlo analyses were performed to assess the performance of the following functionalities:

- calibration of the AoA estimation algorithm,
- AoA estimation,
- fault tolerant AoA measurement system.

The first two analyses were carried out considering three different classes of GA aircraft (light, medium and heavy airplanes). Indeed, the AoA estimation algorithm and its calibration procedure were originally designed for this category of aircraft. It is worthy to remark that this choice is conservative, because, as explained in the previous sections, improved results are obtained if the AoA estimator is applied to commercial aircraft. In the considered analyses the following parameters of the simulation model varied according to their statistical characterization, in order to obtain an accurate statistical assessment of the algorithms under investigation:

- aircraft weight, centre of gravity and inertia matrix;
- sensors' measurement error (bias and white noise);
- initial flight conditions;
- atmospheric conditions: horizontal wind (direction between 0 and 360 degrees, intensity up to 20Kts), turbulence (three levels of turbulence intensity), vertical gusts (cosine shape).

Uniform distribution is assumed for all these parameters, in order to get conservative results [34].

The last Monte Carlo analysis assumes the availability of a triplex AoA measurement (two physical sensors plus the AoA estimator) and assesses the performance of the proposed angle of attack measurement system, only considering the angle of attack trajectories and regardless of the aircraft being a GA or a commercial aircraft.

4.3. Performance Assessment Results

The first Monte Carlo analysis allowed getting a statistical characterization of the calibration gains and then of the aircraft lift model. The analysis comprised 1000 simulation runs for each of the considered aircraft classes. The actual aircraft weight, which is required as input to the procedure, was corrupted with a measurement error

$$W_m = W_t(1 + \nu_W) \quad (31)$$

where W_m and W_t are measured and true weights, respectively, and ν_W is a random noise that varies from one simulation run to the other with uniform distribution in the range [-0.03, 0.03] (that is, the measurement error is in the range ±3% of the true weight). Concerning the model of the atmosphere, the turbulence level varied randomly between 0 and 1. Figure 8 and Figure 9 present the results for one of the considered aircraft categories, showing the estimation of the coefficients C_{L0} and $C_{L\alpha}$ normalized with respect to their true values, which are known in simulation.

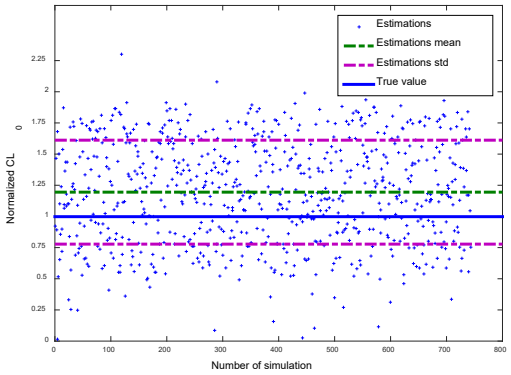


Figure 8: Normalized C_{L0} ceofficient computed in the calibration

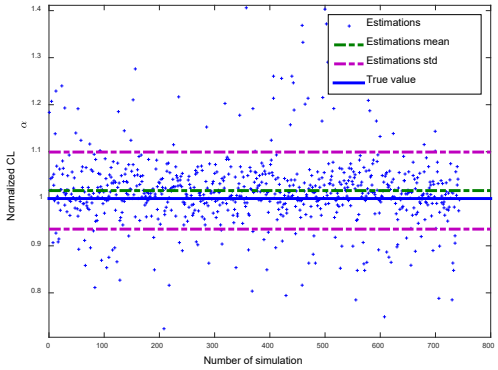


Figure 9: Normalized $C_{L\alpha}$ ceofficient computed in the calibration

Due to the turbulence and to the poor performance of the used autopilot, not all the 1000 calibrations were successful (the success rate was about 75%), that is, the aircraft was not able to keep the required stationary flight condition for a predefined period. The mean error on the C_{L0} estimation, computed on the successful tests, is about 20%, whereas the mean value of the estimated $C_{L\alpha}$ is very close to its true value. In order to assess the effect of these errors on the AoA estimation, a subset (250 samples) of the calculated couples (C_{L0}, $C_{L\alpha}$) was used to estimate the angle of attack in the same conditions flown to perform the calibration. The accuracy of the estimation, evaluated as the mean of the root mean square error (RMS), is 0.64 degree.

The second Monte Carlo focused on the AoA estimation. Several typical manoeuvers were considered, such as level flight, turn, climb, descend, stall, stall in sideslip, stall in turn. A dedicated Monte Carlo analysis was performed for each of these

manoeuvers. In each run, the calibration gains k_1 and k_0 were randomly drawn from the results obtained in the calibration Monte Carlo analysis for the same aircraft category. The number of simulations performed in each Monte Carlo Analysis was chosen in order to have a stable root mean square (RMS) error and to get the 95% confidence level on the probability to have the RMS below a given threshold. Indeed, given a condition P_X to check (such as, the RMS error is below the threshold), the confidence interval on the satisfaction of the condition can be computed as [35], [36]:

$$Low = \frac{k_{suc}}{N_{test}} + \frac{3}{4}\frac{1-\frac{2k_{suc}}{N_{test}} - \sqrt{1+4\vartheta_P k_{suc}\left(1-\frac{k_{suc}}{N_{test}}\right)}}{1+\vartheta_P N_{test}} \quad (32)$$

$$Upp = \frac{k_{suc}}{N_{test}} + \frac{3}{4}\frac{1-\frac{2k_{suc}}{N_{test}} + \sqrt{1+4\vartheta_P k_{suc}\left(1-\frac{k_{suc}}{N_{test}}\right)}}{1+\vartheta_P N_{test}} \quad (33)$$

$$\vartheta_P = \frac{9}{8log\left(\frac{2}{\varepsilon}\right)} \quad (34)$$

$$Pr\{Low < P_X < Upp\} > 1 - \varepsilon \quad (35)$$

where N_{test} is the number of tests performed in the Monte Carlo analysis, k_{suc} is the number of tests in which the condition P_X is satisfied, Pr denotes the probability, and ε is the fixed confidence parameter ($\varepsilon = 0.05$ in our case). Figure 10 and Figure 11 show the trend of RMS mean and standard deviation for one Monte Carlo, used to select the number of runs.

The metrics applied to assess the estimator performance are:

- the mean estimation error (MEE), which represents a measurement of the estimation accuracy in static conditions;
- the root mean square error (RMS), which indicates the overall accuracy of the estimation;
- the correlation between true and estimated angle of attack (CORR), which measures the capability of the estimator to track the dynamic behavior of the true value.

The assessment of the estimator's performance considers all the error's sources that are present during operative conditions, including unknown aircraft weight, sensors' error, and the effect of wind, turbulence and vertical gust. Of course, it produces a degradation of the accuracy with respect to ideal conditions. In level unaccelerated flight, under all possible aircraft configurations and turbulence levels, the MEE is lower than 1.5 degrees. The estimator is able to compute the angle of attack with RMS error varying from about 0.9 degrees for level flight under all possible environmental disturbances to less than 3 degrees for dynamic stall manoeuver, representing the worst-case condition. The average correlation factor between estimated and true angle of attack is bigger than 0.9, confirming the capability to track the dynamics. A sensitivity analysis highlighted that the main source of estimation error is the approximation introduced on the aircraft weight. Indeed, while actual aircraft weight in the simulator randomly varied with uniform distribution within the range [W_{SEW}, W_{MTOW}], the weight used by the estimator is constant in all the simulation runs and is equal to the conventional weight. The value of the conventional weight was selected to get conservative errors when approaching stall, that is, the angle of attack estimation is bigger than its actual value, also at cost of increasing maximum error. This choice derives from the need to anticipate the stall condition. It also allows using optimized algorithms (not described in the present paper) to post process the error when the AoA increases, in order to get very reliable and accurate indication about stall approaching. The knowledge (also rough) of the actual aircraft weight (usually available for commercial aircraft) could significantly improve the estimation performance. Figure 12 presents the estimation of the angle of attack in level flight condition and in presence of a vertical gust, with length equal to 150 meters and maximum amplitude equal to 1600 feet per minute.

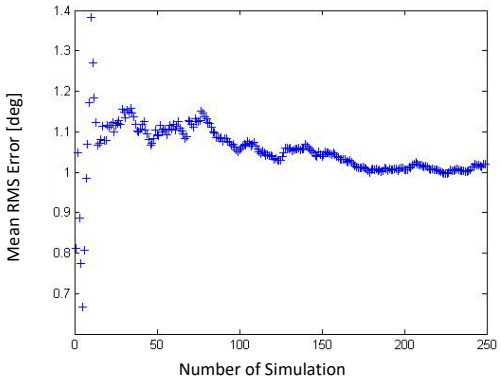


Figure 10: Mean of the RMS error versus number of simulation runs

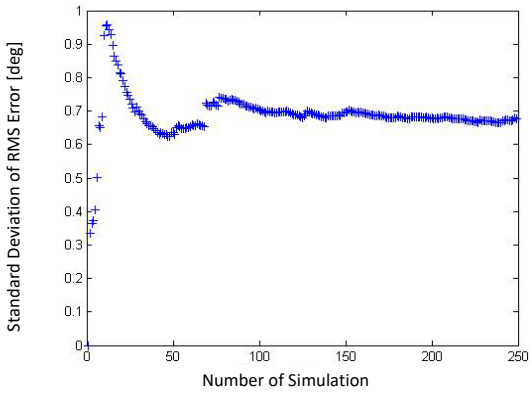


Figure 11: Standard deviation of the RMS error versus number of simulation runs

The final Monte Carlo aimed at assessing the capability of the redundant measurement system to detect a faulty sensor before its effects become significant and dangerous. To this end, one of the physical measurements of the angle of attack is corrupted by injecting one of the following faults, at a given time of the flight:

- constant bias, randomly selected in the range [5, 15] degrees;
- drift with slope randomly selected in the range [0.5, 5] degrees per second;
- signal freezing at a constant value.

The simulation lasts 90 seconds for each run (for the stall manoeuver it can end before 90 seconds if the aircraft stalls). The measurements are sampled at 10 Hz and the fault is injected 10 seconds after the start. The performance of the system is assessed by measuring the error of the consolidated angle of attack (output of the triplex redundant measurement system) with respect to the true angle of attack. Specifically, the mean and the standard deviation of the RMS error are evaluated. Moreover, the percentage of missed fault detections (that is, the percentage of faults that are not detected by the system) and of nuisance alarms

(that is, the percentage of declaration of invalid measurement when the fault is not present) are assessed. Monte Carlo analyses are carried out in four different flight conditions: climb, descent, cruise disturbed by a vertical gust, stall. In the first three conditions, the angle of attack is almost stationary; therefore, only bias and drift faults are added to the measurements of one physical AoA sensor (the freezing of the signal does not produce relevant errors). Bias, drift and signal freezing are used in the stall manoeuver, in which the angle of attack varies significantly during the test. All the Monte Carlo analyses are repeated twice. In the first set of analyses, the measurement system does not compensate for the estimation error due to the approximated aircraft weight (due to the use in the estimation of the conventional weight instead of the actual one). Table 1 shows the results for this test case. In the second set of Monte Carlo simulations, the estimated AoA is corrected by removing a constant offset, before providing it as input to the voting algorithm. The offset is computed as the mean difference, on a sufficiently long time, between the estimated AoA and the angle measured by the physical sensors, when these sensors measure about the same values (that means supposedly they are working properly). This procedure allows to compensate for, although roughly, the error due to the approximated weight. Table 2 presents the results for the second set of Monte Carlo analyses, whereas Figure 13 shows the results of one simulation run, in which the aircraft performs a stall manoeuver and one physical sensor of the angle of attack experiences a constant bias fault.

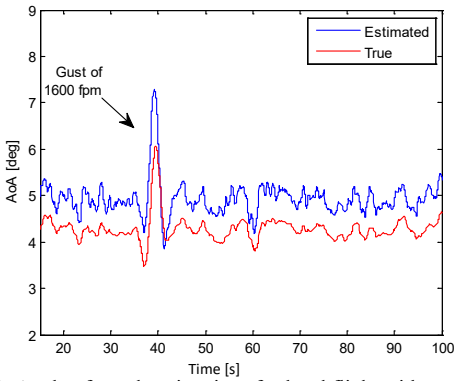


Figure 12: Angle of attack estimations for level flight with vertical gust

Table 1: Results of Monte Carlo analyses for the triplex redundant measurement system (without compensation for unknown aircraft weight)

	Fault	RMS mean [deg]	RMS std [deg]	False alarm %	Missed alarm %
Climb	Bias	0.67	0.18	6.4%	0%
	Drift	0.47	0.15	7.2%	0%
	None	0.24	0.09	1.6%	0%
Descent	Bias	0.68	0.19	0.8%	0%
	Drift	0.48	0.16	0.8%	0%
	None	0.23	0.08	0.4%	0%
Cruise	Bias	0.70	0.17	4.8%	0%
	Drift	0.49	0.16	4.0%	0%
	None	0.24	0.09	0%	0%
Stall	Bias	1.34	1.00	22.4%	5.6%
	Drift	1.08	1.11	26.4%	0%
	Freeze	1.08	1.06	28.8%	0%
	None	0.26	0.08	13.6%	0%

Table 2: Results of Monte Carlo analyses for the triplex redundant measurement system (with compensation for unknown aircraft weight)

	Fault	RMS mean [deg]	RMS std [deg]	False alarm %	Missed alarm %
Climb	Bias	0.62	0.16	0%	0%
	Drift	0.39	0.09	0%	0%
	None	0.20	0.06	0%	0%
Descent	Bias	0.63	0.16	0%	0%
	Drift	0.40	0.08	0%	0%
	None	0.20	0.06	0%	0%
Cruise	Bias	0.64	0.15	0%	0%
	Drift	0.40	0.08	0%	0%
	None	0.20	0.06	0%	0%
Stall	Bias	0.89	0.30	1.6%	0%
	Drift	0.63	0.22	2.4%	0%
	Freeze	0.73	0.23	2.8%	2.8%
	None	0.21	0.04	1.2%	0%

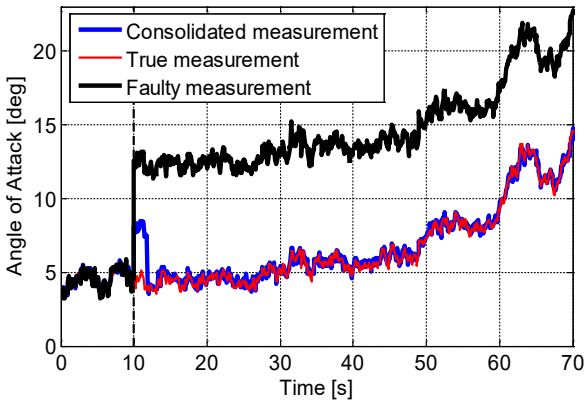


Figure 13: Angle of attack measurement for a stall manoeuver and a constant bias fault on one physical sensor

The improvements due to the weight error compensation are significant in all the examined metrics. As expected, a better estimation of the angle of attack leads to a reduction of the RMS error on the consolidated measurement and to a more effective detection of the faulty sensor. For all the considered flight phases and fault types, the mean RMS error is far below 1 degree also when the fault is present. The false alarm rate is always null, except for the stall maneuver in which however it is very low. The measurement system is able to detect all the injected faults, except for the case of measurement freezing, in which only 7 cases on 250 are missed (2.8%). Although these results are already very good, an optimization of the parameters of the voting and signal validity check algorithms could improve the result for both false and missed alarm percentage. Further investigations will be dedicated to this issue in the future.

5. Conclusion

This paper presented an original angle of attack measurement system applicable to commercial aviation aircraft. The system is based on a triplex physical-analytical redundancy, which only requires two physical sensors of the angle of attack. Indeed, the third sensor is virtual; it is an innovative angle of attack estimator, originally developed to provide General Aviation pilots with a maintenance free and low cost but accurate indication of trend toward stall and stall margin. This estimator does not require any angle of attack dedicated sensor, neither detailed information about

the aerodynamics of the specific aircraft, which is identified during a short calibration flight, carried out just once. The customization of the virtual sensor for the application to commercial aircraft, described in the present paper, allows getting better performance, due to the possibility to exploit additional information (such as, aircraft weight and deflection of the aerodynamic control surfaces), which usually are not available to General Aviation aircraft.

The physical-analytical redundant architecture defined in this paper provides a measurement system which is robust to a single fault of the angle of attack sensors. It could significantly improve the safety of the aircraft, reducing the occurrence of LOC-I related accidents, saving human lives, and avoiding the economic impact consequent to those accidents. It has the peculiarity to be applicable to the existing commercial aircraft by installing just a software code that implements the proposed algorithms, without the need to modify the on-board avionics instrumentation.

The performance of the proposed measurement system was assessed through Monte Carlo simulations. To this end, a high-fidelity simulation environment was implemented, which includes the models of aircraft, atmosphere, on-board sensors and an autopilot. The simulations highlighted very interesting results of both the angle of attack estimator and the overall measurement system. The latter is able to provide a consolidated angle of attack measurement with a mean RMS error far below 1 degree, also when a fault is present. Concerning the detection of the faults, it is generally successful; both missed and false alarm percentages are close to zero for all the examined cases. An optimized tuning of the proposed algorithms will probably further improve these results.

Future works will be focused on the validation of the system through real-time simulations and in-flight trials.

References

[1] A. Vitale, F. Corraro, N. Genito, L. Garbarino, L. Verde, "Innovative Real Time Estimator for Redundancy of Angle of Attack Sensors in Commercial Aircraft," in Proc. 4th Conference on Control and Fault Tolerant Systems SYSTOL'19, 2019, doi: 10.1109/SYSTOL.2019.8864783.

[2] General Aviation Joint Steering Committee - Loss of Control Working Groups, Final Report, 2014.

[3] M. Bromfield, B. Dillman, "The Effects of Using an Angle of Attack System on Pilot Performance and Workload during Selected Phases of Flight," Procedia Manufacturing, **3**, 3222-3229, 2015, doi: 10.1016/j.promfg.2015.07.873.

[4] ICAO, "Manual on aeroplane upset prevention and recovery training," Doc 10011, 2014.

[5] C. Ramprasadh, H. Arya, "Multi-stage fusion algorithm for estimation of aerodynamic angles in mini aerial vehicle," in Proc. 49th AIAA Aerospace Science Meeting, 2011, doi: 10.2514/1.C031322.

[6] A. Cho, Y.S. Kang, B.J. Park, and C.S. Yoo, "Airflow angle and wind estimation using GPS/INS navigation data and airspeed," in Proc. 13th Int. Conf. Control, Automation and Systems (ICCAS), 2013, doi: 10.1109/ICCAS.2013.6704159.

[7] S. Leutenegger, A. Melzer, K. Alexis, and R. Siegwart, "Robust state estimation for small unmanned airplanes," in Proc. IEEE Int. Conf. Control Applications, Antibes, France, 2014, doi: 10.1109/CCA.2014.6981466.

[8] T.A. Johansen, A. Cristofaro, K. Sorensen, J.M. Hansen, T.I. Fossen, "On estimation of wind velocity angle-of-attack and sideslip angle of small UAVs using standard sensors," in Proc. 2015 International Conference on Unmanned Aircraft Systems (ICUAS), 2015, doi: 10.1109/ICUAS.2015.7152330.

[9] A. Wenz, T.A. Johansen and A. Cristofaro, "Combining model-free and model-based angle of attack estimation for small fixed-wing UAVs using a standard sensor suite," 2016 International Conference on Unmanned Aircraft Systems (ICUAS), 2016, doi: 10.1109/ICUAS.2016.7502583.

[10] H. Long and S. Song, "Method of estimating angle-of-attack and sideslip angle based on data fusion," in Proc. International Conference on Intelligent Computation Technology and Automation, 2009.

[11] M. Shaqura, C. Claudel, "A hybrid system approach to airspeed, angle of attack and sideslip estimation in Unmanned Aerial Vehicles," in International Confonference on Unmanned Aircraft Systems ICUAS 2015, 2015, doi: 10.1109/ICUAS.2015.7152355.

[12] R.D. Colgren, "Method and system for estimation and correction of angle-of-attack and sideslip angle from acceleration measurements," U.S. Patent 6273370, November 1, 1999.

[13] Ramprasadh, S. Prem, L. Sankaralingam, Parag Deshpande, Ravi Dodamani, Suraj C S, "A Simple Method for Estimation of Angle of Attack," IFAC-PapersOnLine, **51** (1), 353-358, 2018, doi: 10.1016/j.ifacol.2018.05.048.

[14] P.A. Samara, G.N. Fouskitakis, J.S. Sakellariou, S. Fassois, "Aircraft angle-of-attack virtual sensor design via a functional pooling narx methodology," European Control Conference ECC2003, 2003, doi: 10.23919/ECC.2003.7085229.

[15] A. Lerro, M. Battipede, P. Gili, A. Brandl, "Survey on a Neural Network for Non Linear Estimation of Aerodynamic Angles," in Intelligent Systems Conference, 2017, doi: 10.1109/IntelliSys.2017.8324240.

[16] A. Lerro, M. Battipede, P. Gili, A. Brandl, "Aerodynamic angle estimation: comparison between numerical results and operative environment data," CEAS Aeronautical Journal, 11, 249–262, 2020, doi:10.1007/s13272-019-00417-x

[17] N. Genito, F. Corraro, L. Garbarino, A. Vitale, E. De Lellis, D. Bibby, S. Rieb, K. Jones, "System and method for angle of attack indication with no dedicated sensors and aircraft information," Patent WO 2016 / 164624, https://patents.google.com/patent/WO2016164624A1/en, October 13, 2016.

[18] IATA, "Loss of Control In-Flight Accident Analysis Report, 2019 edition," ISBN 978-92-9264-002-6, 2019.

[19] Aircraft Accident Investigation Bureau (AIB), "Aircraft Accident Investigation Preliminary Report - Ethiopian Airlines Group - B737-8 (MAX) Registered ET-AVJ - 28 NM South East of Addis Ababa, Bole International Airport - March 10, 2019," Report No. AI-01/19, 2019.

[20] Komite Nasional Keselamatan Transportasi, "Aircraft Accident Investigation Report - PT. Lion Mentari Airlines - Boeing 737-8 (MAX); PK-LQP - Tanjung Karawang, West Java - Republic of Indonesia - 29 October 2018," 2018.

[21] The Federal Democratic Republic of Ethiopia, Ministry of Transport, Aircraft Accident Investigation Bureau "Interim Investigation Report on Accident to the B737-8 (MAX) Registered ET-AVJ operated by Ethiopian Airlines on 10 March 2019," Report No. AI-01/19, 09 March, 2020.

[22] FINAL KNKT.18.10.35.04, "Aircraft Accident Investigation Report," PT. Lion Mentari Airlines, Boeing 737-8 (MAX), PK-LQP, Tanjung Karawang, West Java Republic of Indonesia, 29 October 2018.

[23] Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, "Report - Accident on 27 November 2008 off the coast of Canet-Plage (66) to the Airbus A320-232 registered D-AXLA operated by XL Airways Germany," 2010.

[24] German Federal Bureau of Aircraft Accident Investigation (BFU), "Interim Report," BFU 6X014-14, 2015.

[25] D. Ossmann, H.D. Joos, P. Goupil, "Enhanced Sensor Monitoring to Maintain Optimal Aircraft Handling in Case of Faults," Journal of Guidance, Control, and Dynamics, **40**(12), 3127-3137, 2017, doi: 10.2514/1.G002341.

[26] D. Berdjag, J. Cieslak and A. Zolghadri, "Fault detection and isolation of aircraft air data/inertial system," Progress in Flight Dynamics, GNC, and Avionics, **6**, 317-332, 2013, doi: 10.1051/eucass/201306317

[27] C. Seren, P. Ezerzer, G. Hardier, "Model-based techniques for virtual sensing of longitudinal flight parameters," Int. Journal of Applied Mathematics and Computer Science, **25**(1), 23-38, 2015, doi: 10.1515/amcs-2015-0002.

[28] R. E. Maine, K. W. Iliff, "Application of Parameter Estimation to Aircraft Stability and Control," NASA-RP-1168, June 1986.

[29] A. Gelb, Applied Optimal Estimation, M.I.T. Press, ISBN 0-262-20027-9, Cambridge, Massachusetts, 1989.

[30] B. Etkin, Dynamics of Atmospheric Flight, John Wiley & Sons, 1972.

[31] A. Fedele, N. Genito, A. Vitale, L. Garbarino, "Experimental aircraft system identification from flight data: Procedures and Results," CEAS 2013 Air & Space Conference, 2013.

[32] MIL-STD-1797A, "Flying Qualities of Piloted Aircraft," 1990.

[33] L. Garbarino, N. Genito, V. Baraniello, E. De Lellis, and A. Vitale, "Low Cost Air Data System For UAV In-Flight Experimentation," Society of Flight Test Engineers – 23rd European Chapter Symposium, 2012.

[34] B.R. Barmish, C.M. Lagoa, "The Uniform Distribution: A Rigorous

Justification for its use in Robustness Analysis," Mathematics of Control, Signals, and Systems, **10**, 203-222, 1997, doi: 10.1007/BF01211503.
[35] R. Tempo, G. Calafiore, F. Dabbene, Randomized Algorithms for Analysis and Control of Uncertain Systems, Springer-Verlag, New York, 2005.
[36] X. Chen, K. Zhou, J.L. Aravena, "Fast Universal Algorithms for Robustness Analysis," in Proceeding of the 42nd Conference on Decision and Control, 2003, doi: 10.1109/CDC.2003.1272897.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 709-716 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Newton-Euler Based Dynamic Modeling and Control Simulation for Dual-Axis Parallel Mechanism Solar Tracker

Sarot Srang[*,1], Sopagna Ath[2], Masaki Yamakita[2]

[1]*Institute of Technology of Cambodia, Industrial and Mechanical Engineering Department, 12150, Cambodia*

[2]*Tokyo Institute of Technology, Systems and Control Engineering Department, 152-8552, Japan*

ARTICLE INFO

Article history:
Received: 24 July, 2020
Accepted: 25 September, 2020
Online: 05 October, 2020

Keywords:
Parallel Mechanism
Kinematic Constrained
Euler Parameters
Cartesian Coordinate System
Algebraic Differential Equations

ABSTRACT

Dynamic modeling has been a crucial study in many areas of the engineering field. In this paper, we apply the Newton-Euler equation of motion to a two-DOF parallel mechanism solar tracker which is a close loop mechanism. The aim of this study is to show a simulation of the dynamical model with feedback control using a PD controller to orientate the solar panel perpendicular to the sun rays. The mechanism is modeled in the form of a system of algebraic differential equations. First, kinematic constraint equations were constructed in the form of algebraic equations to specify the dynamic interactions at joints. We use the Baumgarte stabilization method, a constraint violation method to eliminate computational error incurred by numerical approximation. Then, the dynamic equations of the system were formulated using the Newton-Euler equation of motion. To describe the translation and rotation motions, we apply Cartesian coordinates and Euler parameters. Simulation of driving the solar panel to reach the desired configuration is made, and the result shows that the PD controller provides good performance of the mechanism regardless of the complexity of the dynamic behavior of the mechanism.

1 Introduction

This paper is an extension of work originally presented in *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2019* [1]. Solar energy has been a major current research in the electricity generation field because of its unlimited resource and environmental friendly behavior. The efficiency of solar energy can be improved by implementing a tracking mechanism which keeps the solar panel perpendicular to the sun rays. There exists two main types of tracking mechanism, a single and dual axis solar tracker [2]. Energy gain from a single axis solar tracker was reported to be 20% [3] while energy gain from a dual axis solar tracker was 30-40% [4].

Dual axis solar tracker has been of an interest research topic for many researchers, [3], [5]–[7], because of its outperformance over single axis solar tracker. Many researches on energy gain from solar tracking systems compared to the tilted fixed panel had been done both theoretically and experimentally [5]. In [7], the author proposed a two-axis decoupled solar tracking system based on parallel mechanism and showed that the tracker requires less driving torque, thus less power dissipation than the conventional serial tracker does. Furthermore, the tracking system does not need reducer with large reduction ratio. Therefore, complexity and weight of the system are also reduced. The parallel mechanism solar tracker can be implemented by mounting on either a fixed or moving platform to produce electrical energy. For instance, aerial vehicles, boats, land vehicles are considered as moving platform. Assuming that the parallel mechanism solar tracker is attached to an aerial vehicle, the dynamic effect must be taken into account. The scope of this study is limited to a fixed platform.

Various methods for dynamic modeling for mechanical systems have been widely developed, and each one has its own advantages and disadvantages. In general, dynamic modeling using generalized coordinate (e.g. Lagrange's dynamic equation) yields the smallest number of differential equations and, therefore, computational efficient. However, the order of nonlinearity is high, and derivation of equation in expanded form of multibody system with loops of connected links is very tedious. For parallel mechanism, derivation can be done by virtually decomposing the system as open loop mechanism with some kinematic constraint forces by the other

[*]Corresponding Author: Sarot Srang, Department of Industrial and Mechanical Engineering, Institute of Technology of Cambodia, Russian Federation Blvd., P.O.Box 86, Sangkat Tuek L'ak 1, Khan Tuol Kouk, Phnom Penh, Cambodia. srangsarot@itc.edu.kh

parts of the system. For a system with more than one loop like the tracker we are considering, the derivation is even more difficult, and only partial reaction (or constraint) forces can be determined. With cartesian coordinates in global and body-fixed frames and Euler parameters for describing rotation, dynamic modeling using Newton's method is much simpler, and systematic generation of kinematic constraints at all joints and dynamic equations can be derived easily [8]. The method yields system equations of algebraic-differential equations. Reaction forces and the coordinates describing motion of a system are obtained from solving the equations. The reaction force has an advantage for mechanical structure design. A high-performance computer can realize the simulation modeled in detail including the reaction force effect in this case. However, the analysis of reaction forces is not reported in this work.

The remaining contents of this paper are organized as follows. In section 2, the parallel mechanism is explained on configuration, and its kinematic is described by using Cartesian coordinate. Dynamic equations for unconstrained and constrained body which are described in Cartesian coordinate and using Euler parameters are given in section 3. For constrained body, the resulted equation is in the form of system algebraic-differential equations. In section 4, the controller design is covered by using PD controller. Section 5 presents the results and discussion. We conclude our paper in section 6.

2 Kinematic Constraint

Cartesian coordinate is used to describe the system configuration, and constraint equations are obtained from individual joints. Figure 1 shows the coordinate system, where $(Oxyz)$ is global frame and $(O_i\xi_i\eta_i\zeta_i)$ is body-fixed frame attached on body i with the center of mass O_i. A point P on the body has coordinate as a vector in body-fixed frame and global frame defined by $s_i'^P$ and r_i^P respectively. Figure 2 shows a parallel mechanism which is used as dual axis solar tracker. The body numbers are labeled as seen in the figure. The mechanism consists of 7 connected rigid bodies. It has a global coordinate $(Oxyz)$, and each body has its own body-fixed frame as explained in Figure 2. Body 7 is connected with body 5 and 6 via 2 spherical joints and with body 4 via a universal joint. Body 6 is connected with body 2 via a revolute joint. Body 5 is connected to body 3 via a universal joint. Body 4 is connected to body 1 (ground) via another revolute joint. Body 3 is connected with body 1 via a translational joint. Body 2 is connected with body 1 via another translational joint. Two linear actuators are attached at the translational joints. The actuators exert forces on body 2 and 3 along vertical axes.

Denote $q_i = \left[r^T, p^T\right]_i^T = [x, y, z, e_0, e_1, e_2, e_3]_i^T$ as a coordinate vector of the body i, where r_i is position vector of the center of mass of the body as illustrated in Figure 1, and $p_i = \left[e_0, e^T\right]_i^T = [e_0, e_1, e_2, e_3]_i^T$ is Euler parameters.

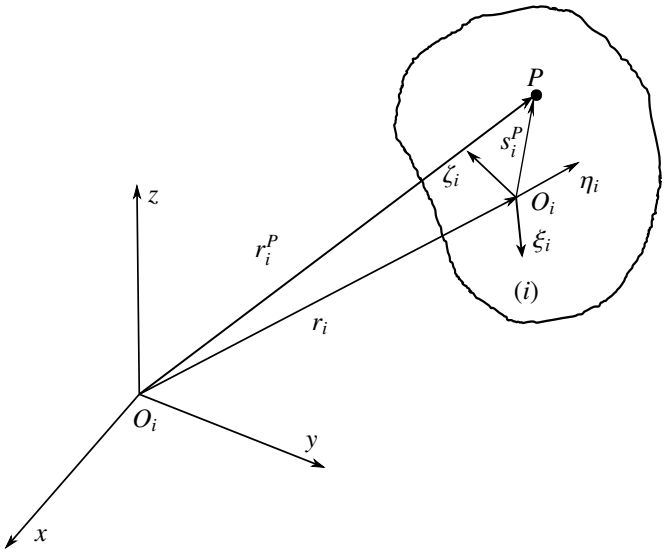


Figure 1: Cartesian Coordinate System

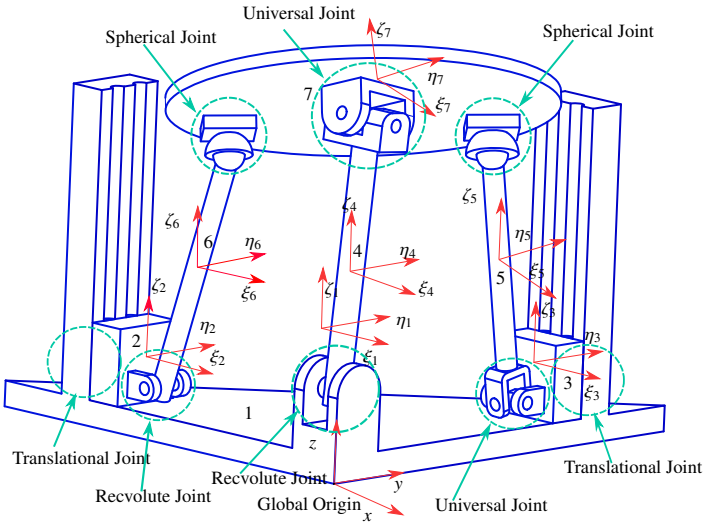


Figure 2: Global and Local coordinates of the parallel mechanism at the initial configuration

The parameter satisfies a mathematical relationship,

$$p_i^T p_i - 1 = 0. \tag{1}$$

The second time derivative of (1) is

$$p_i^T \ddot{p}_i + \dot{p}_i^T \dot{p}_i = 0. \tag{2}$$

Denote R_i as a rotational matrix of body i, and a pair of 3×4 matrices G_i and L_i defined as $G_i = [-e, \tilde{e} + e_0 I]_i$ and $L_i = [-e, -\tilde{e} + e_0 I]_i$. Then $R_i = G_i L_i^T$,[8] . The global coordinate of the point P illustrated in Figure 1 can be defined by

$$r_i^P = r_i + R_i s_i'^P = r_i + G_i L_i^T s_i'^P, \tag{3}$$

where $s_i'^P = \left[\xi^P, \eta^P, \zeta^P\right]_i^T$ is body-fixed coordinate of the point P.

Denote $q = \left[q_1^T, q_2^T, q_3^T, q_4^T, q_5^T, q_6^T, q_7^T\right]^T$ as coordinate vector for describing the configuration of the mechanism, and its respective first and second time derivative, $\dot{q}$ and $\ddot{q}$, for describing the motion of the mechanism. The number of coordinates of the system is $n = 7 \times 7 = 49$, and the number of Euler parameters relationship (or mathematical constraint) is 7.
Denote

$$\Phi \equiv \Phi(q) = 0 \tag{4}$$

as kinematic equation derived from kinematic constraints which have 34 equations. The compact form of the kinematic equation for each joint is given in Appendix A. The first and second time derivative of (4) are written as

$$\begin{aligned} \Phi_q \dot{q} &= 0, \\ \Phi_q \ddot{q} + \left(\Phi_q \dot{q}\right)_q \dot{q} &= 0 \\ \text{or } \Phi_q \ddot{q} &= \gamma, \end{aligned} \tag{5}$$

where $\gamma = -\left(\Phi_q \dot{q}\right)_q \dot{q}$ is called the right side of the kinematic acceleration equation and can be found in [8]. We adopt the constraint violation stabilization methods from [9], another similar technique can also be found in [10, 11].
Consider the following closed loop system,

$$\ddot{D} + K_P \dot{D} + K_D D = 0, \tag{6}$$

where K_P and K_D are constant and defined in Appendix B. Due to the fact that there exists numerical errors in using numerical method to solve the differential equations (5), we use the positive constraint violation stabilization method to obtain a stable response. So (5) is modified such that the kinematic constraint equations are satisfied by ensuring the shrinking of the computation error.
First time derivative of constraint equation is obtained as

$$\Phi_q \dot{q} = \dot{D} \tag{7}$$

Second time derivative of the constraint equation is given as

$$\begin{aligned} \Phi_q \ddot{q} + (\Phi_q \dot{q})_q \dot{q} &= \ddot{D} \\ &= -K_D \dot{D} - K_P D \\ &= -K_D(\Phi_q \dot{q}) - K_P \Phi(q) \end{aligned} \tag{8}$$

Thus,

$$\Phi_q \ddot{q} = \gamma - K_D(\Phi_q \dot{q}) - K_P \Phi(q). \tag{9}$$

The equation (9) is used for simulation and the performance is discussed in section 5.

3 Dynamic Modeling for Unconstrained and Constrained Bodies

Newton-Euler equations of motion is used to derive the equations of motion for multi-rigid bodies of the parallel mechanism. Prior to having the equations of motion for a set of interconnected bodies by kinematic joints which is known as constrained body, the equations of motion for body with no contact to other bodies which refers as an unconstrained body are written.

3.1 Dynamic Equations for Unconstrained Bodies

From Newton's method, an unconstrained body i with mass m_i and moment of inertia J_i' with respect to its center of mass exerted by external force f_i and moment τ_i' has dynamic equation of motion as

$$\begin{aligned} N_i \ddot{r}_i &= f_i \\ J_i' \dot{\omega}_i' + \tilde{\omega}_i' J_i' \omega_i' &= n_i', \end{aligned} \tag{10}$$

where $N_i = diag\left(\begin{bmatrix} m & m & m \end{bmatrix}\right)_i$ and ω_i' is the angular velocity defined in body-fixed frame. The equation of rotational motion in (10) is transformed by using the following relationships [8].

$$\begin{aligned} \omega' &= 2L\dot{p} \\ \dot{\omega}' &= 2L\ddot{p}. \end{aligned} \tag{11}$$

$$\widetilde{L_i \dot{p}_i} = L_i \dot{L}_i^T \tag{12}$$

Equation (10) becomes

$$\begin{aligned} 2J_i' L_i \ddot{p}_i + 4\widetilde{L_i \dot{p}_i} J_i' L_i \dot{p}_i &= n_i' \\ 2J_i' L_i \ddot{p}_i + 4L\dot{L}^T J_i' L_i \dot{p}_i &= n_i' \\ 2J_i' L_i \ddot{p}_i + L_i H_i \dot{p}_i &= n_i', \end{aligned} \tag{13}$$

where $H_i = 4\dot{L}_i^T J_i^T L_i$. The dynamic equations of motion for an unconstrained bodies are written as

$$\begin{aligned} N_i \ddot{r}_i &= f_i \\ 2J_i' L_i \ddot{p}_i + L_i H_i \dot{p}_i &= n_i' \\ p_i^T \ddot{p}_i + \dot{p}_i^T \dot{p}_i &= 0. \end{aligned} \tag{14}$$

In matrix form, (14) is rewritten as

$$\begin{bmatrix} N_i & 0 & 0 \\ 0 & 2J'L_i & 0 \\ 0 & p_i^T & 0 \end{bmatrix} \begin{bmatrix} \ddot{r}_i \\ \ddot{p}_i \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ L_i H_i \dot{p}_i \\ \dot{p}_i^T \dot{p}_i \end{bmatrix} = \begin{bmatrix} f_i \\ n_i' \\ 0 \end{bmatrix}. \tag{15}$$

3.2 Dynamic Equations for Constrained Bodies

Bodies are interconnected to form a system of constrained bodies via kinematic joints which create constraint reaction forces and moments. When applying the constraint reaction forces and moments to (14), the dynamic equations of motion for a constrained bodies are written as

$$\begin{aligned} N_i \ddot{r}_i &= f_i + f_i^{(c)} \\ 2J_i' L_i \ddot{p}_i + L_i H_i \dot{p}_i &= n_i' + n_i'^{(c)} \\ p_i^T \ddot{p}_i + \dot{p}_i^T \dot{p}_i &= 0. \end{aligned} \tag{16}$$

A constrained body i is additionally exerted by reaction forces and moments $\left[f^{(c)}, n'^{(c)}\right]_i^T$ from joints. These forces and moments can be transformed to coordinate system consistent with q denoted by $\left[f^{*(c)}, n^{*(c)}\right]_i^T$ and defined by

$$\begin{bmatrix} f^{*(c)} \\ n^{*(c)} \end{bmatrix}_i = \begin{bmatrix} \Phi_r^T \\ \Phi_p^T \end{bmatrix}_i \lambda \tag{17}$$

To be used with the formulation (16), the moment in (17) is transformed to be

$$\begin{aligned} n'^{(c)} &= \frac{1}{2} L_i n'^{*(c)} \\ &= \frac{1}{2} L_i \Phi_{p_i}^T \lambda \end{aligned} \tag{18}$$

where $\lambda = [\lambda_1, \ldots, \lambda_{34}]$ is known as Lagrange Multipliers, and $\left[\Phi_{ri}, \Phi_{pi}\right] = \Phi_{qi}$ is Jacobian Matrix of the kinematic constraint $\Phi \equiv \Phi(q)$ with respect to q_i. Therefore, the dynamic equations of motion for the mechanism are given in compact form as

$$\begin{bmatrix} N_i & 0 & \Phi_{ri}^T \\ 0 & 2J'L_i & \frac{1}{2}L_i\Phi_{pi}^T \\ 0 & p_i^T & 0 \end{bmatrix} \begin{bmatrix} \ddot{r}_i \\ \dot{p}_i \\ -\lambda \end{bmatrix} + \begin{bmatrix} 0 \\ L_i H_i \dot{p}_i \\ \dot{p}_i^T \dot{p}_i \end{bmatrix} = \begin{bmatrix} f_i \\ n'_i \\ 0 \end{bmatrix}. \tag{19}$$

$$\begin{bmatrix} M & B^T \\ P & 0 \end{bmatrix} \begin{bmatrix} \ddot{q} \\ -\lambda \end{bmatrix} + \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} g \\ 0 \end{bmatrix}, \tag{20}$$

where

$$M = \begin{bmatrix} N_1 & 0 & \ldots & 0 & 0 \\ 0 & 2J'_1 L_1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & N_7 & 0 \\ 0 & 0 & \ldots & 0 & 2J'_7 L_7 \end{bmatrix}$$

,

$$B = [\Phi_{r_1}, \frac{1}{2}\Phi_{p_1} L_1^T, \ldots, \Phi_{r_7}, \frac{1}{2}\Phi_{p_7} L_7^T]$$

,

$$P = \begin{bmatrix} 0^T & p_1^T & \ldots & 0^T & 0^T \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0^T & 0^T & \ldots & 0^T & p_7^T \end{bmatrix}, b = \begin{bmatrix} 0 \\ L_1 H_1 \dot{p}_1 \\ \vdots \\ 0 \\ L_7 H_7 \dot{p}_7 \end{bmatrix}, c = \begin{bmatrix} \dot{p}_1^T \dot{p}_1 \\ \vdots \\ \dot{p}_7^T \dot{p} \end{bmatrix},$$

$$g = \begin{bmatrix} f_1 \\ n'_1 \\ \vdots \\ f_7 \\ n'_7 \end{bmatrix}, \text{ and } \ddot{q} = \begin{bmatrix} \ddot{r}_1{}^T & \ddot{p}_1{}^T & \ldots & \ddot{r}_7{}^T & \ddot{p}_7{}^T \end{bmatrix}^T =$$

$[\ddot{x}, \ddot{y}, \ddot{z}, \ddot{e}_0, \ddot{e}_1, \ddot{e}_2, \ddot{e}_3]_i^T, \quad (i = 1, 2, ...7).$

Equation (19) consists of 49 equations with 83 variables, therefore 34 constraint equations are appended with (19) to solve for coordinate vectors q and Lagrange multiplier λ. A system of algebraic differential equation is obtained as follow

$$\begin{bmatrix} M & B^T \\ P & 0 \\ \Phi_q & 0 \end{bmatrix} \begin{bmatrix} \ddot{q} \\ -\lambda \end{bmatrix} + \begin{bmatrix} b \\ c \\ 0 \end{bmatrix} = \begin{bmatrix} g \\ 0 \\ \gamma - K_D(\Phi_q \dot{q}) - K_P \Phi(q) \end{bmatrix}. \tag{21}$$

Equation (21) were solved for the coordinate vectors q by using state equations as follow:

$$x_1 = q = [x_1, y_1, z_1, [e_0, e_1, e_2, e_3]_1, \ldots, x_7, y_7, z_7, [e_0, e_7, e_7, e_7]_7]$$
$$x_2 = \dot{q} = [\dot{x}_1, \dot{y}_1, \dot{z}_1, [\dot{e}_0, \dot{e}_1, \dot{e}_2, \dot{e}_3]_1, \ldots, \dot{x}_7, \dot{y}_7, \dot{z}_7, [\dot{e}_0, \dot{e}_1, \dot{e}_2, \dot{e}_3]_7]$$

$$\dot{x}_1 = x_2 = \dot{q}$$
$$\dot{x}_2 = \ddot{q} = \begin{bmatrix} [I] & [0] \end{bmatrix} \begin{bmatrix} M & B^T \\ P & 0 \\ \Phi_q & 0 \end{bmatrix}^{-1} \begin{bmatrix} g - b \\ -c \\ \gamma - K_D(\Phi_q \dot{q}) - K_P \Phi(q) \end{bmatrix} \tag{22}$$

4 PD Controller

The desired value of actuator's displacement is obtained from the 3D model such that the solar panel reach the position as shown in Figure 4 where the desired value of the two actuators are $z_{2d} = 0.046m$, $z_{2d} = 0.046m$. The actual displacement of the two actuators labeled as body number 2 and body number 3 are denoted as z_2 and z_3 respectively. Figure 3 illustrates a common feedback controller known as proportional derivative (PD) controller that is used to obtain the desired response, and has a form as follow:

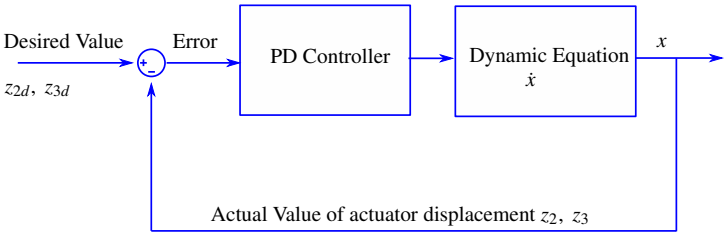


Figure 3: Block Diagram for PD Controller

$$u = K_p e_z + K_d \dot{e}_z, \tag{23}$$

where $u = \begin{bmatrix} f_{2z} \\ f_{3z} \end{bmatrix}$, $e_z = \begin{bmatrix} z_2 - z_{2d} \\ z_3 - z_{3d} \end{bmatrix}$ and $\dot{e}_z = \begin{bmatrix} \dot{z}_2 - \dot{z}_{2d} \\ \dot{z}_3 - \dot{z}_{3d} \end{bmatrix}$ are the position error and velocity error. K_p and K_d are 2 by 2 positive definite diagonal matrix. K_p plays a role as a spring that tries to keep the position error decreasing. K_d is considered as a damper which maintains the velocity error $\dot{e}$ decreasing.

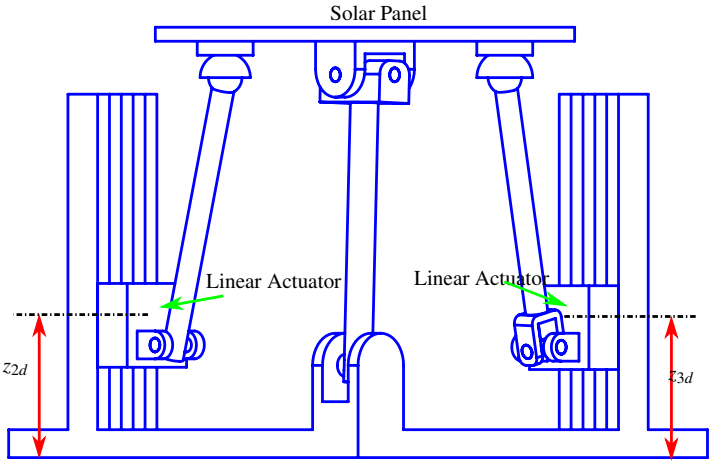


Figure 4: Desired configuration of the solar panel

5 Results and Discussion

PD control law is applied to drive the parallel mechanism from its initial configuration to the desired configuration with $K_p = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix}$, and $K_d = \begin{bmatrix} 20 & 0 \\ 0 & 25 \end{bmatrix}$ as shown in Figure 1 and Figure 4. The control input $u = \begin{bmatrix} f_{2z} & f_{3z} \end{bmatrix}^T$ is exerted to the two actuators by assigning the external forces in the vector $g = \begin{bmatrix} f_1^T & n_1'^T & \dots & f_7^T & n_7'^T \end{bmatrix}^T$ containing

$$\begin{aligned} f_2 &= \begin{bmatrix} 0 & 0 & f_{2z} \end{bmatrix}^T \\ f_3 &= \begin{bmatrix} 0 & 0 & f_{3z} \end{bmatrix}^T, \end{aligned} \tag{24}$$

while keeping the remaining components in g to be zero. Other parameters used in this paper are shown in Appendix C. As can be seen in the Figure 5, the two actuators reach the desired position quickly. The movement of the solar panel can be observed in Figure 7. It appears that the parallel mechanism can be easily controlled such that it is practically possible to move the solar panel directed to the sunlight, and thus more energy could be produced. It is evident that the desired orientation of the solar panel is with its coordinate system ($\xi_7\eta_7\zeta_7$) is parallel to the global coordinate system (xyz). In this case, the euler parameters is expected to be equal to $\begin{bmatrix} \cos\frac{\phi}{2} & 0 & 0 & \sin\frac{\phi}{2} \end{bmatrix}^T$ [8], where ϕ is the rotation angle around ζ axis, and is reported true by Figure 8. Figure 9 illustrates the mathematical relationship of euler parameters of the solar panel as described in (1) which satisfies $p_7^T p_7 = 1$. When the euler parameters are used to describe a rotational coordinate, it is important to ensure that (1) is satisfied because the four quantities of euler parameters are not independent.

6 Conclusion

Our study provides the frame work for modeling any mechanism by obtaining the algebraic differential equations of motion derived from Newton-Euler equation of motion and kinematic constraint equations. This technique can be applied to a wide range of simulation applications, especially for a close loop mechanism. Moreover, our result confirms the usefulness of the Baumgarte stabilization method that deals with numerical error when solving system of algebraic differential equations with mathematical constraint of Euler parameters. The control simulations using PD controller was to drive the solar panel to reach the desired configuration with a good performance.

Future work should focus on applying a model based controller that could provide a better control result, especially for tracking control. If the same modeling approach is applied to a system that requires fast response, an investigation on computation time should be made.

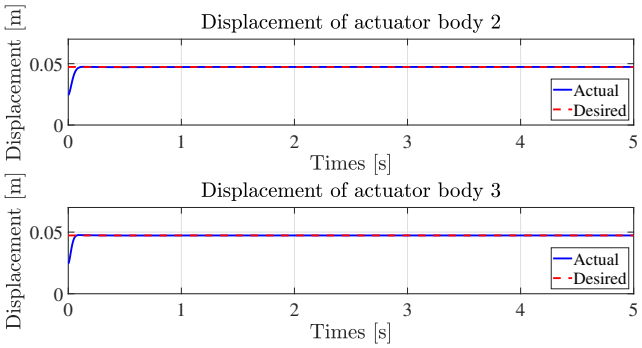


Figure 5: Response of Actuator's Displacement

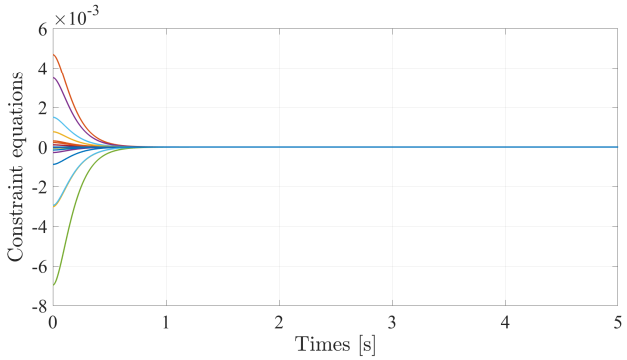


Figure 6: Kinematic constraint

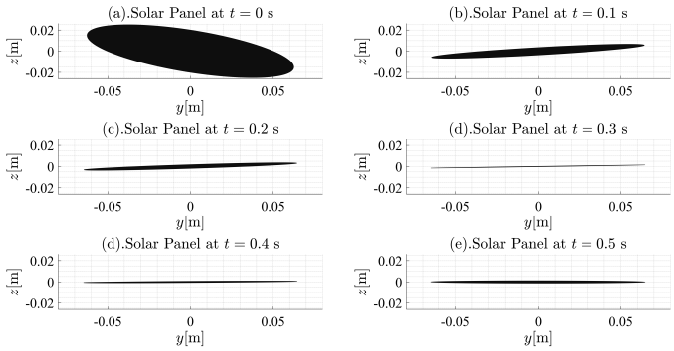


Figure 7: Orientation of Solar Panel

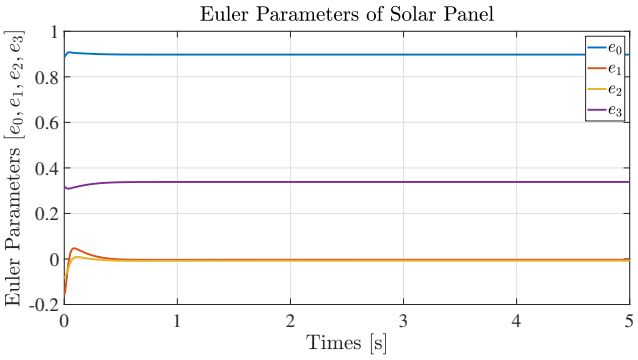


Figure 8: Euler Parameters Describing the Orientation of Solar Panel

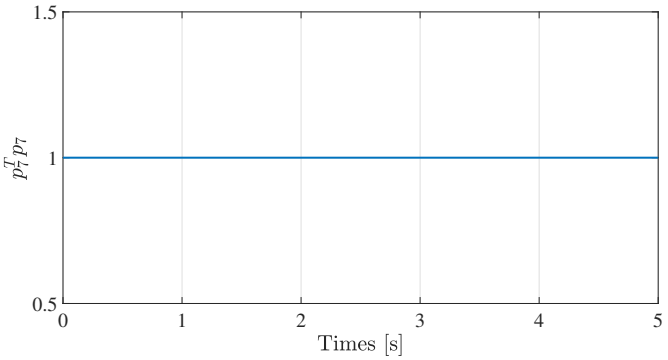


Figure 9: Mathematical Relationship of Euler Parameters of Solar Panel

APPENDIX A

For the parallel mechanism, by following the concept of relative constraints between two bodies from [8], the 34 constraint equations are obtained as follows:

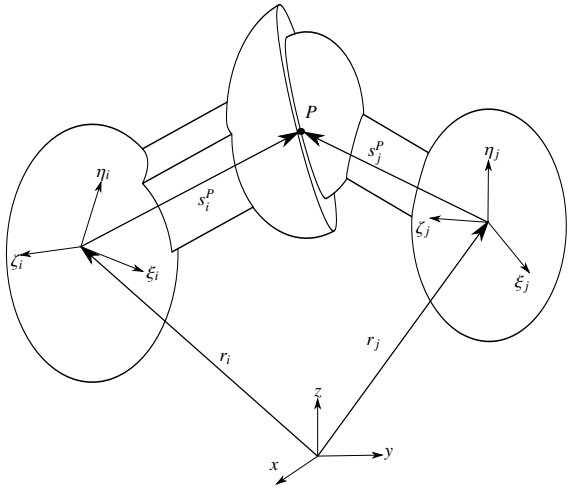


Figure 10: Spherical Joint

• Three equations for each Spherical joint between body (5,7) and (6,7).

$$\Phi^{(s,3)} \equiv r_i + A_i s_i'^P - r_j - A_j s_j'^P = 0, \tag{25}$$

where $s_i'^P = [\xi^P, \eta^P, \zeta^P]$ is body-fixed coordinate of the point P.

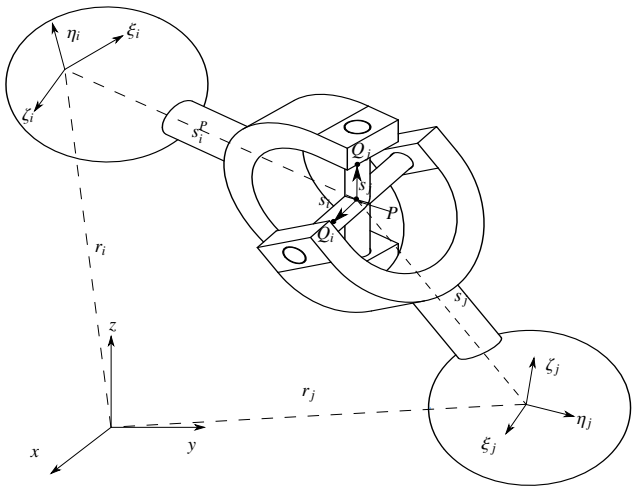


Figure 11: Universal Joint

• Four equations for each Universal joint between body (3,5) and (4,7).

$$\begin{aligned} \Phi^{(s,3)} &\equiv r_i + A_i s_i'^P - r_j - A_j s_j'^P = 0 \\ \Phi^{(n1,1)} &\equiv s_i^T s_j = 0, \end{aligned} \tag{26}$$

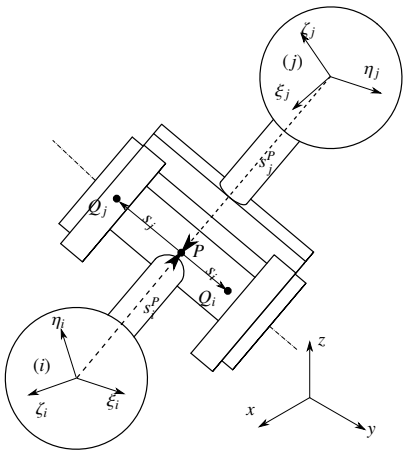


Figure 12: Revolute Joint

• Five equations for each Revolute joint between body (1,4) and (2,6).

$$\begin{aligned} \Phi^{(s,3)} &\equiv r_i + A_i s_i'^P - r_j - A_j s_j'^P = 0 \\ \Phi^{(p1,2)} &\equiv \tilde{s}_i s_j = 0, \end{aligned} \tag{27}$$

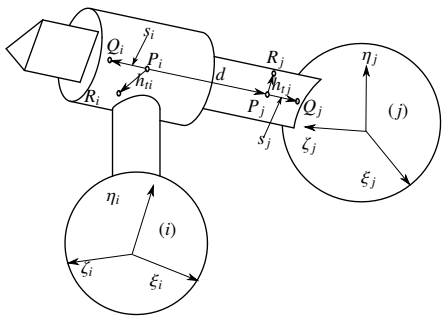


Figure 13: Translational Joint

• Five equations for each Translational joint between body (1,2) and (1,3).

$$\begin{aligned}\Phi^{(p1,2)} &\equiv \tilde{s}_i s_j = 0\\ \Phi^{(p2,2)} &\equiv \tilde{s}_i d = 0\\ \Phi^{(n1,1)} &\equiv h_{ti}^T h_{tj} = 0,\end{aligned} \tag{28}$$

APPENDIX B

Consider the following mass spring damper equation of motion

$$\ddot{D} + 2\zeta\omega_n\dot{D} + \omega_n^2 D = 0, \tag{29}$$

where ζ is a critical damping constant, and ω_n denotes a natural frequency. From (6) and (29), we can write

$$\begin{aligned}\ddot{D} + K_D\dot{D} + K_P D &= 0\\ \ddot{D} + 2\zeta\omega_n\dot{D} + \omega_n^2 D &= 0\end{aligned} \tag{30}$$

$$\begin{aligned}\zeta &= \frac{K_D}{2\omega_n} = \frac{K_D}{2\sqrt{K_P}}\\ \omega_n &= \sqrt{K_P}\end{aligned} \tag{31}$$

In the case of critically damped system, $\zeta = 1$ ensures a fast response with no overshoot. Therefore, we choose $K_D = 2\alpha$ and $K_P = \alpha^2$, where α is a positive constant and taken to be 10.

APPENDIX C

Table 1: Mass of body $i, i = 1, 2, \cdots, 7$.

Symbols	Value	Unit
m_1	475.09×10^{-3}	Kg
m_2	189.7×10^{-3}	Kg
m_3	189.7×10^{-3}	Kg
m_4	104.45×10^{-3}	Kg
m_5	53.25×10^{-3}	Kg
m_6	54.4×10^{-3}	Kg
m_7	890.5×10^{-3}	Kg

Table 2: Position vector of center of mass of body i at initial position, $i = 1, 2, \cdots, 7$

Symbols	Value	Unit
r_1	$10^{-3} \times [-63.93, 63.93, 30.41]^T$	m
r_2	$10^{-3} \times [-123.99, 15, 24.6]^T$	m
r_3	$10^{-3} \times [-15, 123.99, 24.6]^T$	m
r_4	$10^{-3} \times [-11.50, 18.37, 99.21]^T$	m
r_5	$10^{-3} \times [-12.99, 98.33, 68.84]^T$	m
r_6	$10^{-3} \times [-92.31, 15, 66.27]^T$	m
r_7	$10^{-3} \times [-34.28, 42.89, 136.56]^T$	m

Table 3: Euler parameters of body i at initial position,$i = 1, 2, \cdots, 7$

Symbols	Value
p_1	$[1, 0, 0, 0]^T$
p_2	$[1, 0, 0, 0]^T$
p_3	$[1, 0, 0, 0]^T$
p_4	$[1, 0, 0, 0]^T$
p_5	$[1, 0, 0, 0]^T$
p_6	$[1, 0, 0, 0]^T$
p_7	$[0.89043, -0.19791, -0.10481, 0.39619]^T$

Table 4: Moment of Inertia of body i,$i = 1, 2, \cdots, 7$

Symbols	Value	Unit
J_1'	$10^{-4} \times \begin{bmatrix} 20 & 8 & -1 \\ 8 & 20 & 1 \\ 1 & 1 & 26 \end{bmatrix}$	Kg m^2
J_2'	$10^{-4} \times \begin{bmatrix} 0.2774 & 0 & -0.0102 \\ 0 & 0.2015 & 0 \\ -0.0102 & 0 & 0.1987 \end{bmatrix}$	Kg m^2
J_3'	$10^{-4} \times \begin{bmatrix} 0.2015 & 0 & 0 \\ 0 & 0.2774 & 0.0102 \\ 0 & 0.0102 & 0.1987 \end{bmatrix}$	Kg m^2
J_4'	$10^{-3} \times \begin{bmatrix} 0.1217 & 0.0000 & 0 \\ 0.0000 & 0.1248 & 0 \\ 0 & 0 & 0.0054 \end{bmatrix}$	Kg m^2
J_5'	$10^{-4} \times \begin{bmatrix} 0.5827 & 0 & -0.0824 \\ 0 & 0.5940 & 0 \\ -0.0824 & 0 & 0.0172 \end{bmatrix}$	Kg m^2
J_6'	$10^{-4} \times \begin{bmatrix} 0.6242 & 0 & 0 \\ 0 & 0.6231 & -0.0000 \\ 0 & -0.0000 & 0.0058 \end{bmatrix}$	Kg m^2
J_7'	$10^{-4} \times \begin{bmatrix} 12 & 0 & 1 \\ 0 & 12 & 0 \\ 1 & 0 & 24 \end{bmatrix}$	Kg m^2

Acknowledgment This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-17-1-0148.

References

[1] S. Srang and S. Ath, "Dynamic Modeling and Simulation for 2DOF Parallel Mechanism Solar Tracker," 2019 IEEE/ASME International Conference on

Advanced Intelligent Mechatronics (AIM), Hong Kong, China, 283-288, 2019. https://doi.org/10.1109/AIM.2019.8868501

[2] C.S. Chin, A. Babu, W. McBride, "Design, modeling and testing of a standalone single axis active solar tracker using MATLAB/Simulink", Renewable Energy, **36**(11), 3075-3090, 2011. https://doi.org/10.1016/j.renene.2011.03.026

[3] Yingxue Yao, Yeguang Hu, Shengdong Gao, Gang Yang, Jinguang Du, "A multipurpose dual-axis solar tracker with two tracking strategies", Renewable Energy, **72**, 88-98, 2014. https://doi.org/10.1016/j.renene.2014.07.002

[4] P. Y. Vorobiev, J. Gonzalez-Hernandez and Y. V. Vorobiev, "Optimization of the solar energy collection in tracking and non-tracking photovoltaic solar system," (ICEEE). 1st International Conference on Electrical and Electronics Engineering, 2004., Acapulco, Mexico, 310-314, 2004. https://doi.org/10.1109/ICEEE.2004.1433900

[5] Hossein Mousazadeh, Alireza Keyhani, Arzhang Javadi, Hossein Mobli, Karen Abrinia, Ahmad Sharifi, "A review of principle and sun-tracking methods for maximizing solar systems output", Renewable and Sustainable Energy Reviews, **13**(8), 1800-1818,2009. https://doi.org/10.1016/j.rser.2009.01.022

[6] T. Zhan, W. Lin, M. Tsai and G. Wang, "Design and Implementation of the Dual-Axis Solar Tracking System," 2013 IEEE 37th Annual Computer Software and Applications Conference, Kyoto, 276-277, 2013. https://doi.org/10.1109/COMPSAC.2013.46

[7] J. Wu, X. Chen and L. Wang, "Design and Dynamics of a Novel Solar Tracker With Parallel Mechanism," in IEEE/ASME Transactions on Mechatronics, **21**(1), pp. 88-97, Feb. 2016. https://doi.org/10.1109/TMECH.2015.2446994

[8] Nikravesh, Parviz E. Computer-aided analysis of mechanical systems, Prentice-Hall, Inc., 1988.

[9] Neto, M.A., Ambrósio, J. "Stabilization Methods for the Integration of DAE in the Presence of Redundant Constraints". Multibody System Dynamics **10**, 81–105 (2003). https://doi.org/10.1023/A:1024567523268

[10] Flores, P., Machado, M., Seabra, E., and Tavares da Silva, M. (October 13, 2010). "A Parametric Study on the Baumgarte Stabilization Method for Forward Dynamics of Constrained Multibody Systems." ASME. J. Comput. Nonlinear Dynam. January 2011; **6**(1): 011019. https://doi.org/10.1115/1.4002338

[11] Flores P., Pereira R., Machado M., Seabra E. (2009) "Investigation on the Baumgarte Stabilization Method for Dynamic Analysis of Constrained Multibody Systems". In: Ceccarelli M. (eds) Proceedings of EUCOMES 08. Springer, Dordrecht.https://doi.org/10.1007/978-1-4020-8915-2_37

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 710-717 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Adaptive Identification Method of Vehicle Model for Autonomous Driving Robust to Environmental Disturbances

Yohei Yamauchi*, Mitsuyuki Saito

Systems Engineering, Graduate School of Information Sciences, Hiroshima city University, Hiroshima 731-3194, Japan

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 11 November, 2020
Online: 24 November, 2020

Keywords:
Autonomous Driving
Vehicle Model
Adaptive Identification
Modeling Error
Neural Network

ABSTRACT

Many recent studies on autonomous driving have focused on model-based control. A number of studies has addressed that simple models such as the Kinematic Bicycle Model are easier to design controls for autonomous driving systems. However, such a simple vehicle model has a weakness in that it is subject to modeling errors. This is because it does not take into account the nonlinear characteristics due to road conditions and driving conditions (environmental disturbances: road friction coefficient, large steering, acceleration, sideslip, etc.) Therefore, the purpose of this study is to identify vehicles with high accuracy and in real time, adapting to environmental disturbances.
This study propose a vehicle model based on the Kinematic Bicycle Model. The nonlinear characteristics of the vehicle are represented by the deviation of the front wheel steering angle of the Kinematic Bicycle Model. This deviation is trained and estimated online using a three-layer Neural Network. In other words, the AI is adaptive learning of modeling errors caused by nonlinear characteristics of the vehicle.
This paper presents an example of model-based control using model predictive control.

1. Introduction

In recent years, study and development of autonomous driving has been conducted in the automobile industry, IT companies, and universities in each country. In study on autonomous driving, some autonomous driving systems that combine Artificial Intelligence (AI) and model-free control methods is proposed [1,2]. However, it is considered that such the autonomous driving system is difficult to obtain system stability and reliability in unknown environments. Therefore, fusion technology of AI technology and model-based control has gained much importance in study on autonomous driving [3,4]. Model-based control is a control method in which a control target is represented by a mathematical model and optimal control input is determined based on the model. Model-based control is widely used in various industries [5]. It has problem that the control performance cannot be exhibited when the model is different from the actual dynamics. Additionally, the more complex the controlled object, the more complex the model and the more the amount of calculation. There is a limit to the number of computing units that can be equipped in an autonomous vehicle. Therefore, the model used for autonomous driving is required to be a simple model with less calculation amount. This paper proposes a simple and highly accurate method for vehicle identification (partially published in [6]).

Several studies agree that simple models, such as Kinematic Bicycle Model [7] and linear single-track model [8], are easier to design controllers for autonomous driving systems [9,10]. These vehicle models do not include nonlinear characteristics due to road conditions and driving conditions (environmental disturbances: road friction coefficient, large steering, acceleration, sideslip, etc.). Hence, the accuracy may be deteriorated due to a modeling error between the actual vehicle and the vehicle model. In order to consider the nonlinear characteristics of the vehicle, vehicle models that includes model equations such as tires and suspensions in the vehicle model has also been proposed [11,12]. However, since these vehicle models include multiple models expressions in the vehicle model, the structure of the vehicle model is complicated. It is inferred that if these are used in an autonomous driving system, it may impose calculated load on the computing unit and impair the real-time performance of the system. In other words, it is important for the vehicle model used for autonomous driving controllers to accurately model the vehicle in real time, even if there are environmental disturbances. This paper proposes a vehicle model based on the Kinematic Bicycle Model [7] in order to represent vehicle behavior simply and with high accuracy. The

*Corresponding Author: Yohei Yamauchi, Hiroshima City University, +81-82-830-1749, yamauchi-ics.info@tomatolynx4.sakura.ne.jp

Kinematic Bicycle Model does not include nonlinear characteristics due to acceleration and deceleration or large steering, etc. Therefore, an error may occur between the actual gravity center position of the vehicle and the gravity center position calculated by the vehicle model. The behavior of the actual vehicle and the behavior calculated by the model are different due to the position error of the center of gravity, and the modeling error becomes large. Therefore, this method considers the vehicle model in which the center of gravity is fixed at the center of the wheelbase of the Kinematic Bicycle Model and the modeling error is expressed by the deviation of the front wheel steering angle. In addition, this study uses Neural Network to adaptively identify vehicle model by training and estimating the deviation. This paper verifies the usefulness of the proposed method through simulation experiments using vehicle motion analysis software (CarSim: Virtual Mechanics). In this study, simulations were performed in situations closer to actual driving conditions than in [6] (Section 4). Since this method models the vehicle while determining the control input in real time, it does not exist as a modeling technology alone and must be combined with model-based control. This paper shows an example using Model Predictive Control (MPC) as an example of model-based control to show the usefulness of the method. The method requires accurate location information acquisition. Since it is expected that the measurement accuracy will improve with the development of GNSS (Global Navigation Satellite System) in the future, the simulation is performed assuming that accurate position information can be obtained.

In summary, there are two aspects of the proposed approach that are particularly unique. The first is that the structure of the model is simple and easy to identify. In conventional models, several parameters must be identified in advance, but only one parameter is required in this study in advance. This means that the controller design of autonomous driving could be simplified by relieving the task of examining cornering stiffness and tire parameters in advance. Second, by focusing on the coordinates of the center of gravity, the approach can analyze the entire vehicle as nonlinear motion. Online learning may be able to respond to changes in vehicle mass (due to the number of passengers and loads) and road surface. It is notable that the method is robust to environmental disturbances and easy to identify.

This paper sets up the issue in Section 2. Section 2.1 introduces the conventional method and Section 2.2 describes our proposed identification method in detail. This paper also presents and discuss the simulation results in Sections 3 and 4. Section 3 mainly considers the effects of acceleration, deceleration and steering on the vehicle's nonlinear characteristics, while Section 4 considers the situation with road surface changes. And Section 5 concludes this paper.

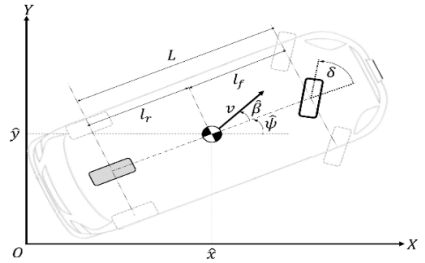


Figure 1: Kinematic Bicycle Model on the XY coordinate

2. Statements of The Issue

This section will set the issue for the proposed method. Section 2.1 introduces simple two-wheel models and accurate nonlinear models to clarify the problem. Section 2.2 details the proposed method for solving the problem.

2.1. Conventional study of vehicle models

2.1.1. Two-wheel model with simple structure

Typical vehicle models used for model-based control include simple two-wheeled models such as the Kinematic Bicycle Model [7] and linear single-track model [8]. These vehicle models are based on the assumption that the state quantities are observed instantaneously, and some conditions (e.g. constant speed, left and right tire characteristics are equal, roll and pitching motions are ignored) are set to represent vehicle dynamics in a simplified way. These vehicle models have simple structure, and thus the turning radius can be easily calculated. Therefore, they can be easily introduced to the controller design of autonomous driving systems. However, these do not take into account various nonlinear characteristics due to environmental disturbances (road friction coefficient, large steering, acceleration, etc.), which can cause modeling errors between the actual vehicle and vehicle models. This paper uses the Kinematic Bicycle Model as an example of the simple two-wheel model to test its accuracy. The model diagram of the Kinematic Bicycle Model is shown in Figure 1. Equations (1-4) show the model equations of the vehicle model. The velocity $v[k-1]$(m/s) and the front wheel steering angle $\delta[k-1]$(rad) are inputs, and the center of gravity coordinates of the vehicle model $(\hat{x}[k],\ \hat{y}[k])$, the direction of the vehicle model $\hat{\psi}[k-1]$(rad), and the sideslip angle around the center of gravity $\hat{\beta}[k-1]$(rad) are outputs. $(x[k-1], y[k-1])$ is the coordinates of the center of gravity observed one step ago.

$$\hat{x}[k] = x[k-1] + v[k-1]\Delta t \cdot \cos\left(\hat{\psi}[k-1] + \hat{\beta}[k-1]\right) \tag{1}$$

$$\hat{y}[k] = y[k-1] + v[k-1]\Delta t \cdot \sin\left(\hat{\psi}[k-1] + \hat{\beta}[k-1]\right) \tag{2}$$

$$\hat{\psi}[k-1] = \hat{\psi}[k-2] + \frac{v[k-2]\Delta t}{l_r}\sin\hat{\beta}[k-2] \tag{3}$$

$$\hat{\beta}[k-1] = \tan^{-1}\left(\frac{l_r}{l_f + l_r}\tan\delta[k-1]\right) \tag{4}$$

here k is the current time, Δt is the sampling time, $l_{f(r)}$(m) are the distance from the front (rear) wheel axle to the center of gravity, and L(m) is the wheel base.

The trajectory of the vehicle model without these nonlinear characteristics (Figure 1) is confirmed. In this case, experiments and verifications should be performed using actual vehicles, but verifications are performed by simulation experiments that are easy to analyze and verify and that can accurately acquire the

vehicle state. Specifically, this study used the Driving Simulator in Figure 2. The Driving Simulator is a Windows PC with a vehicle motion numerical analysis software (CarSim) and a game handle device (Logitech) connected. The PC used in the simulation are as follows: Windows10 64bit, CPU: Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz, and installed memory (RAM):8GB.



Figure 2: Driving Simulator

A simulation course [Slalom] was created on CarSim as shown in Figure 3. The driver drove this course between two pylons lined up in the course, gradually increasing the velocity as shown in Figure 5. The running trajectory at that time is the solid line in Figure 4 (Vehicle's running trajectory). The velocity v obtained as the vehicle data at that time is shown in Figure 5, and the front wheel steering angle δ is shown in Figure 6. The velocity v and the front wheel steering angle δ are input to the vehicle model (1-4) and the running trajectory is calculated as shown by the broken line in Figure 4 (Equation of Vehicle Model). Figure 7 shows the position error between the observed trajectory (x, y) and the trajectory calculated by the vehicle model $(\hat{x}, \hat{y})$. From Figure 4 and Figure 7, there is a maximum position error of about 0.15m in the running trajectory of the actual vehicle and the running trajectory calculated using the vehicle model of (1-4). This is thought to be due to the nonlinear characteristics (tire deformation, expansion and contraction of suspension, etc.) caused by acceleration, deceleration and steering during running. It is consider that the front wheel steering angle δ and the vehicle traveling direction do not match due to the influence of the nonlinear characteristic. This deviation affects the modeling error. In addition, this simulation is based on the assumption of asphalt surface. If it is snow or ice road, the deviation increases further. This is because the effect of the road surface is not taken into account in this model.



Figure 3: Simulation course[Slalom]

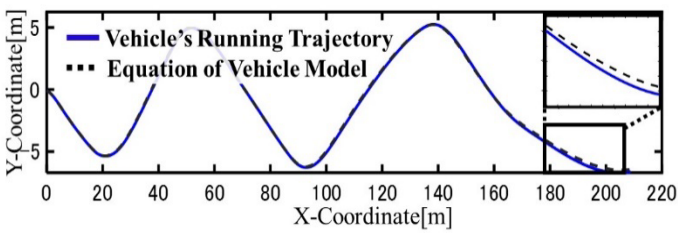


Figure 4: Driving trajectory

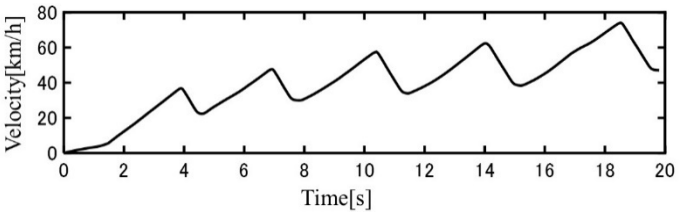


Figure 5: Velocity v

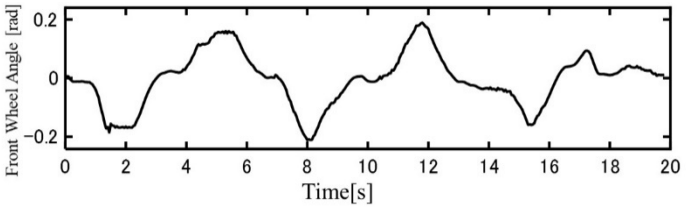


Figure 6: Front wheel steering angle δ

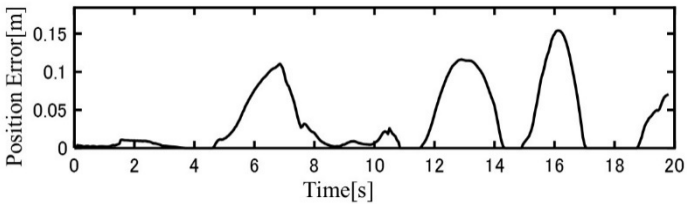


Figure 7: Position error

2.1.2. *Example of non-linear vehicle models that accurately represents vehicle behavior*

As shown in 2.1.1, simplifying the vehicle behavior may increase the modeling error. Hence, there are several conventional studies that use nonlinear vehicle models to represent the nonlinear motion of vehicles in detail. In literature [13], an autonomous driving system combined with a nonlinear vehicle model and MPC is proposed. The nonlinear model is a combination of two-wheel model and nonlinear tire model. This literature shows good results even on compacted snow surface with a low coefficient of friction. In this literature, two tire models were prepared beforehand, one for asphalt and the other for compacted snow, and were tested on each surface. In other words, the experiment is based on the assumption that the road friction coefficient is known. This means that the road friction coefficient, which changes from time to time, must be known.

To solve these problems, a combination of adaptive Model Predictive Control and tire-stiffness estimator [14] has been proposed [15]. This method estimates the tire stiffness from the tire-stiffness estimator. It is able to estimate tire stiffness in situations where the road surface changes and select the optimal road friction coefficient and tire parameters. However, the relationship between the chosen parameters and tire stiffness must be known. In order to find out the relationship between the two, it is necessary to conduct field tests or using a testbench beforehand, which may change depending on the degree of tire wear and other factors such as ageing. In addition, these literatures focused only on tire nonlinearity and did not mention nonlinear vehicle motion due to changes in vehicle mass (due to the number of passengers and loads.), suspension, body stiffness and other effects. These nonlinear motions can also lead to modeling errors. This paper proposes a vehicle model for online learning of nonlinear characteristics of the vehicle by focusing on the change of the vehicle's center of gravity position. By focusing on the change of

the center of gravity, it is possible to model not only the tires but also the nonlinear characteristics of the entire vehicle. Furthermore, online learning eliminates the hassle of pre-testing and allows you to deal with disturbances such as vehicle mass that change with each drive

2.2. Vehicle Model to Estimate Modeling Error

As described in Section 2.1.1, due to the nonlinear characteristics of automobiles, deviation occurs between the front wheel steering angle δ and the actual running direction of the vehicle. The actual direction of travel of the vehicle is defined as the front tire steering angle $\hat{\delta}$. Here, the front wheel tire steering angle $\hat{\delta}$ is an angle that includes nonlinear characteristics due to environmental disturbances and vehicle dynamics. The front wheel steering angle δ is the angle that the front wheels are facing, which can be calculated by the steering wheel angle δ_{SW}. In order to accurately represent the behavior of the vehicle, it is necessary to include in the vehicle model the deviation between the direction the front wheels are facing and the direction the vehicle is actually going, in other words, the deviation between the front wheel steering angle δ and the front wheel tire steering angle $\hat{\delta}$. However, it is difficult to directly observe and theoretically obtain the deviation. Therefore, this deviation is named the modeling error $\hat{\alpha}$ and is defined as the front tire steering angle $\hat{\delta}$ as (5).

$$\hat{\delta}[k-1] = \delta[k-1] + \hat{\alpha}[k-1] \tag{5}$$

$$\hat{\beta}[k-1] = \tan^{-1}\left(\frac{l_r}{l_f + l_r}\tan\hat{\delta}[k-1]\right) \tag{6}$$

Since the front wheel tire steering angle $\hat{\delta}$ is defined as the actual direction in which the vehicle is traveling, (4) is modified as in (6). In other words, our proposed vehicle model is (1-3,5,6). The vehicle model is as shown in the Figure 8. The vehicle model needs to identify the distance from the front (rear) wheel axle to the center of gravity $l_{f(r)}$. The position of the vehicle's center of gravity changes from moment to moment during driving. This is because acceleration, deceleration and large steering causes nonlinear motion in the vehicle, including the tires and suspensions. It is difficult to determine the exact position of the vehicle's center of gravity. Therefore, in this study, the position of the center of gravity of the vehicle is fixed at the center of the wheelbase ($l_{f(r)} = L/2$), and the identification error of the center of gravity position is corrected by $\hat{\alpha}$.

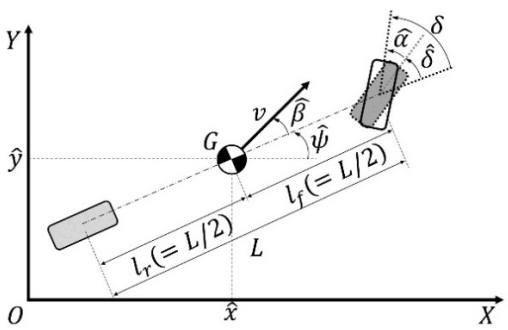


Figure 8: Vehicle model including modeling error $\hat{\alpha}$

Here, a method for estimating the modeling error $\hat{\alpha}$ is described. Since the modeling error $\hat{\alpha}$ is the parameter representing nonlinear motion due to acceleration and deceleration, steering, and road surface changes, it has nonlinearity and is expected to change from moment to moment. Therefore, this study proposes the method for estimating the model error $\hat{\alpha}$ in real time while deriving the control input by model-based control.

This paper considers the system that uses MPC to derive the front wheel steering angle δ and velocity v that are control inputs. MPC is a control law that derives the optimal control input while predicting its future behavior using a predictive model representing the dynamics of a control object. MPC solves an open-loop optimal control problem from the current time to finite horizon for each control period. MPC is an attractive method for controlling autonomous vehicles because it can consider the dynamics and constraints of the controlled object and the ability to adapt to driving scenarios [16-18]. Figure 9 shows the system configuration.

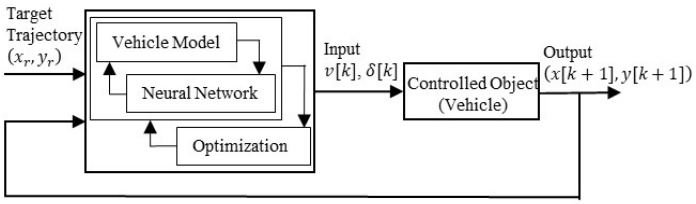


Figure 9: Block diagram of the proposed system

This system uses a Neural Network to train online the nonlinear characteristics of a vehicle that cannot be considered in the vehicle model. The trained Neural Network is used to control the vehicle while estimating the unknown parameters of the vehicle model. Figure 10 shows the flowchart of this system.

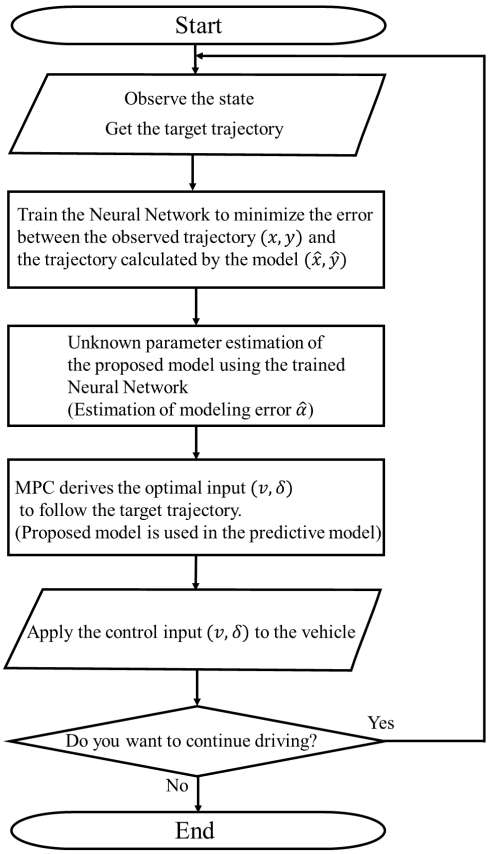


Figure 10: Flowchart of the adaptive identification method for vehicle

The control goal of MPC is to derive the control input that matches the output with the target trajectory. The control objective is to derive the control input ($\delta[k]$ and $v[k]$) that matches the output ($x[k+1]$, $y[k+1]$) with the target trajectory ($x_r[k+1], y_r[k+1]$). This system derives control inputs that minimize the cost function (7).

$$J(\boldsymbol{\delta}(k), \boldsymbol{v}(k)) = \sum_{i=1}^{H} [(\hat{x}[k+i] - x_r[k+i])^2 + (\hat{y}[k+i] - y_r[k+i])^2] \quad (7)$$

H is the prediction horizon. The prediction model (Vehicle Model and Neural Network in the Figure 9) predicts the vehicle trajectory ($\hat{x}[k+i]$, $\hat{y}[k+i]$), $1 \le i \le H$. Since the trajectory depends on the future input, the input sequence $v[k+i], \delta[k+i], 0 \le i \le H-1$ are derived so that the predicted trajectory ($\hat{x}[k+i]$, $\hat{y}[k+i]$), $1 \le i \le H$ approaches the target trajectory ($x_r[k+i], y_r[k+i]$), $1 \le i \le H$ of the obtained input sequence, only the first $v[k], \delta[k]$ are used as the actual inputs.

From here, the Neural Network that trains and estimates the modeling error $\hat{\alpha}$ is described. From Figure 4 to Figure 6, it can be confirmed that the position error (modeling error) increases as the velocity v increases and the front wheel steering angle δ increases. In other words, the position error is considered to depend on the velocity v and the front wheel steering angle δ. The parameter required to correct this position error is the modeling error $\hat{\alpha}$. This modeling error $\hat{\alpha}$ is considered to include nonlinearity. This study uses a 3-layer Neural Network with 2 inputs and 1 output for estimation. This is because the nonlinear system is modeled with high accuracy and the load on the computer memory is reduced as much as possible. The relationship between the input and output of the Neural Network is shown in (8-10). In this paper, I_1 and I_2 are inputs, c_1-c_4 are thresholds, and $w1$ and $w2$ are weighting factors. The input value of the hidden layers are s_1-s_4, and the sigmoid function is used for the output value h_1-h_4 of the hidden layers. Akaike's Information Criterion (AIC) is used to determine the number of hidden layers. The input I is the front wheel steering angle δ and the velocity v, and the output O represents the modeling error $\hat{\alpha}$.

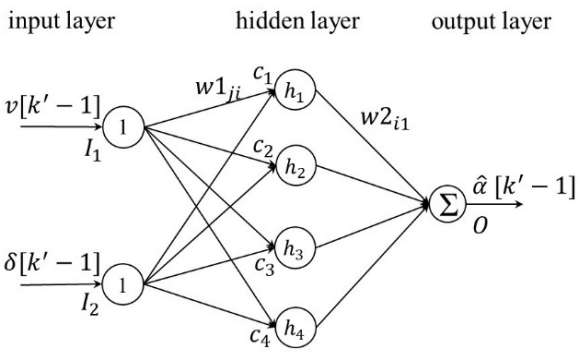
$$s_i[k'-1] = \sum_{j=1}^{2} w1_{ji}[k'-1] \cdot I_j[k'-1] + c_i[k'-1] \quad (8)$$

$$h_i[k'-1] = \frac{1}{1+\exp(-s_i[k'-1])} \quad (9)$$

$$\hat{\alpha}[k'-1] = \sum_{i=1}^{4} w2_{i1}[k'-1] \cdot h_i[k'-1] \quad (10)$$

The observed values of the position coordinates are given to the Neural Network as instruction signal, and online training is performed so as to minimize the cost function (11).

$$I = \sum_{k'=k-W}^{k} \{(\hat{x}[k'] - x[k'])^2 + (\hat{y}[k'] - y[k'])^2\} \quad (11)$$



$(\hat{x}, \hat{y})$ represents the position coordinates of the vehicle model, and (x, y) represents the position coordinates of the actual vehicle. W represents the window width. This neural network is trained so that the position coordinates of the vehicle model $(\hat{x}[k'], \hat{y}[k'])$ from k to W match the position coordinates of the actual vehicle $(x[k'], y[k'])$. Altogether, the parameter $\hat{\alpha}$, which represents the nonlinear properties, is estimated from the position coordinates.

By training and estimating the modeling error due to the nonlinear characteristics online, the behavior of the vehicle can be accurately represented in situations such as acceleration and deceleration, large steering and road surface changes. Because the vehicle is identified in real time, it may be able to respond to changes in vehicle weight, such as changes in the number of passengers. Furthermore, our identification method only uses the wheelbase L as the setting parameter of vehicle model. This means that different types of vehicles can be identified by only changing the wheelbase L. Conventional vehicle models have set parameters (e.g., cornering stiffness, vehicle mass, etc.), which vary for each vehicle. The key feature of this method is that there is only one configuration parameter.

3. Simulation of Fixed Road Surface

In this section, a simulation comparing the Kinematic Bicycle Model (1-4) with the proposed model (1-3,5,6) is described. As in Section 2, the simulation was performed using CarSim installed in the Driving Simulator. It verified whether the center of gravity coordinates of the proposed vehicle model $(\hat{x}, \hat{y})$ matches the center of gravity coordinates of the actual vehicle (x, y). The Kinematics Bicycle Model (1-4) and the proposed vehicle model (1-3,5,6) were given the velocity v and front wheel steering angle δ as inputs, and the trajectory was calculated. The accuracy is checked by comparing the calculated trajectory with the actual vehicle trajectory. The input data and the actual vehicle trajectory are obtained by driving the simulation course shown in Figure 3, which was created in the Driving Simulator (Figure 2) as a driving course with acceleration, deceleration, and steering. This study assumes that the 27 degree of freedom vehicle model in CarSim is the actual vehicle. This simulation assumes a dry asphalt surface (surface friction coefficient $\mu = 0.85$) and drive a B-Class hatchback vehicle (Figure 12).

The driver repeatedly accelerated and decelerated between the two pylons in the course [Slalom] shown in Figure 3. Figure 13 and Figure 14 show the v and δ.The solid line in Figure 15

shows the running trajectory. In this paper, the trajectory is used as the actual vehicle trajectory. The dashed lines in Figure 15 show the trajectory when the velocity v (Figure 13) and the front wheel steering angle δ (Figure14) were given as inputs to the proposed model. The front wheel tire steering angle $\hat{\delta}$ is shown in Figure 16. The calculated position error between the calculated trajectory and the actual vehicle trajectory is shown in Figure 17 and the estimated modeling error $\hat{\alpha}$ is shown in Figure 18. As shown in Figure 17, the maximum positional error is 0.05m, which is considered to be within the practical range.



Figure 12: B-Class hatchback vehicle in CarSim

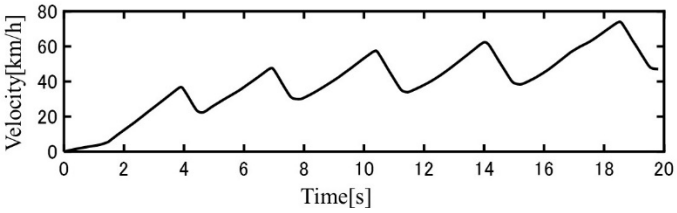


Figure 13: Velocity v □ input□

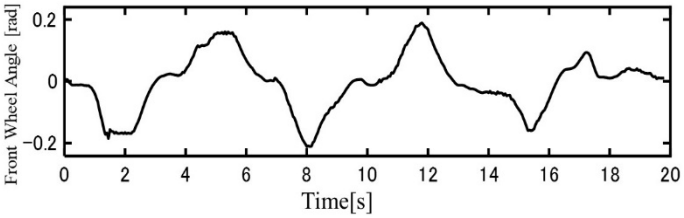


Figure 14: Front wheel steering angle δ □ input□

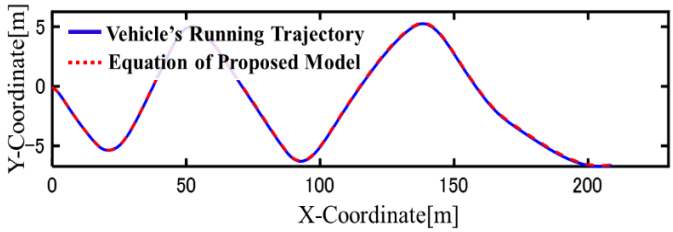


Figure 15: Driving trajectory［Slalom］

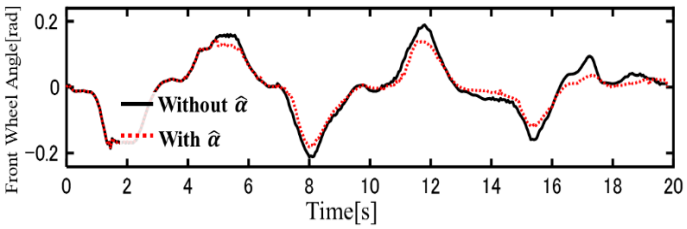


Figure 16: Front wheel steering angle δ and front wheel tire steering angle $\hat{\delta}$

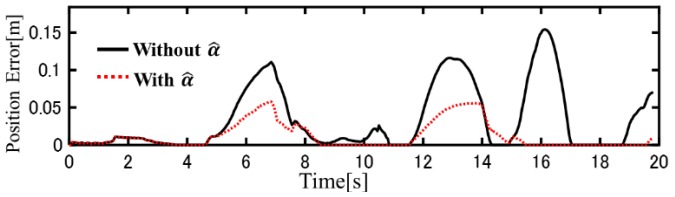


Figure 17: Position error

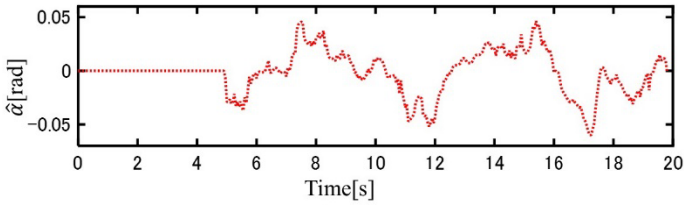


Figure 18: Modeling error $\hat{\alpha}$

The results show that the behavior of the vehicle can be identified with high accuracy. This means that the proposed model (1-3,5,6) can contribute to the controller design of autonomous driving systems using model-based control. However, at this stage, this study has only validated a single driver driving a B-Class hatchback in CarSim several times around the track in a simulation experiment and have obtained good results. In order to prove the effectiveness of the proposed method, it is necessary to conduct similar tests on various courses and vehicle models, and this is a subject for future study.

4. Simulation of Road Surface Change

This section presents additional examples of situations that more closely resemble actual driving situations in order to verify the usefulness of the proposed model. As in Section 3, the experiments were conducted using the Driving Simulator shown in Figure 2. The course used is shown in Figure 19. This course was designed to simulate a mirror burn. Mirror burn is a phenomenon in which the surface of the road is polished by the traffic and becomes very slippery at a part of the intersection. This course was driven by the vehicle (B-class hatchback) in CarSim. This course is designed to have surface friction coefficient μ=0.2 at the center of the intersection and μ=0.5 outside the center of the intersection. This simulation assumes driving on the left side of the road because it is based on Japanese roads. The trajectory of the vehicle on this course is treated as the center of gravity coordinates of the actual vehicle (x, y). In addition to vehicle dynamics, this simulation allows us to verify whether the vehicle can adapt to changing road conditions. The trajectory of the vehicle $(x[k],\ y[k])$ while driving on the course is shown by the solid line in Figure 22. The dotted lines in Figure 20 to Figure 25 indicate the boundary of the surface friction coefficient. Figure 20 and Figure 21 show the v and δ.The solid line in Figure 22 shows the running trajectory. In this paper, the trajectory is used as the actual vehicle trajectory. The dashed lines in Figure 22 show the trajectory when the v (Figure 20) and δ (Figure 21) were given as inputs to the proposed model. The front wheel tire steering angle $\hat{\delta}$ is shown in Figure 23. The calculated position error between the calculated trajectory and the actual vehicle trajectory is shown in Figure 24 and the estimated modeling error $\hat{\alpha}$ is shown in Figure 25. As shown in Figure 24, the maximum positional error is 0.01m, which indicates that the proposed model is able to adapt to the changes in the road surface.

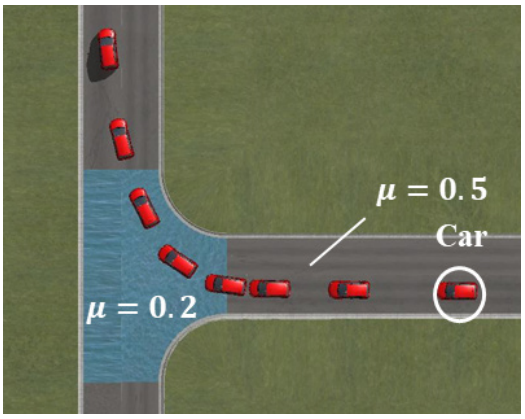


Figure 19: Simulation course［Mirror Burn］

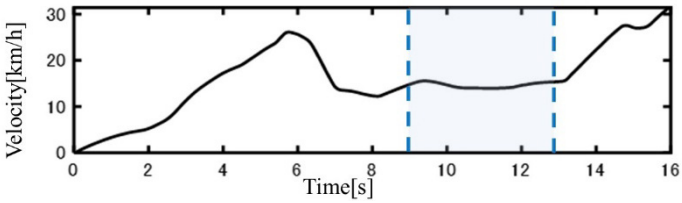


Figure 20: Velocity v（Input）

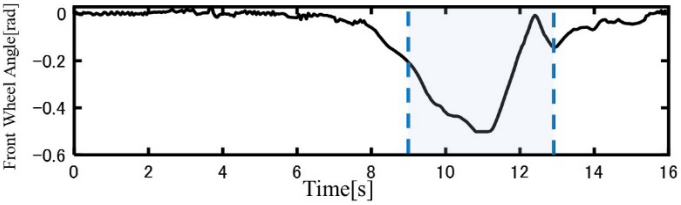


Figure 21: Front wheel steering angle δ □ Input□

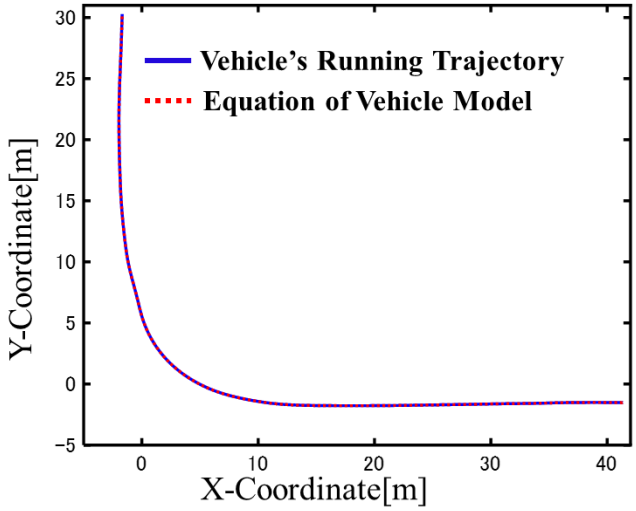


Figure 22: Vehicle's driving trajectory［Mirror Burn］

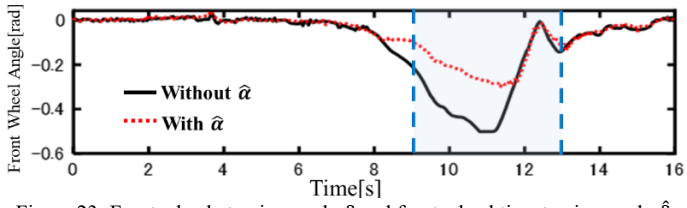


Figure 23: Front wheel steering angle δ and front wheel tire steering angle $\hat{\delta}$

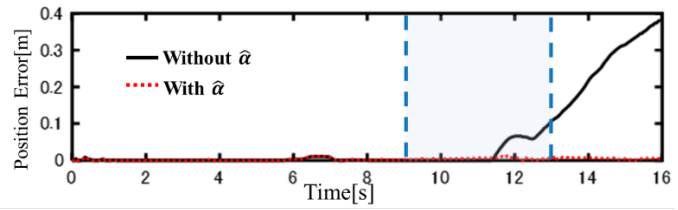


Figure 24: Position Error

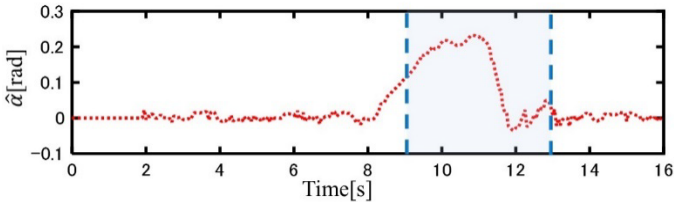


Figure 25: Modeling Error $\hat{\alpha}$

5. Conclusion

The purpose of this paper is to identify an autonomous vehicle with high accuracy in real time. This paper proposed a simple vehicle model that represents the error in the center of gravity between the actual vehicle and the vehicle model as the deviation $\hat{\alpha}$ of the front wheel steering angle. This paper also proposed the method to estimate the $\hat{\alpha}$ in real time using neural network, and simulation experiments using CarSim showed the usefulness of the method in situations that require acceleration and deceleration, large steering, and road surface changes. The authors emphasize that the method can represent the nonlinear characteristics of the vehicle as it is learning online and that the only parameter to be identified in advance is the wheelbase. In other words, this study can eliminate the process of identifying multiple parameters beforehand and contribute to the design of control controllers that is robust to ever-changing environmental disturbances.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Bojarski, D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. Jackel, M. Monfort, U. Muller and J. Zhang, et al. "End to end learning for self-driving cars." arXiv preprint arXiv:1604.07316, 2016. https://arxiv.org/abs/1604.07316

[2] F. Codevilla, M. Müller, A. López, V. Koltun and A. Dosovitskiy, "End-to-end driving via conditional imitation learning." in 2018 IEEE International Conference on Robotics and Automation (ICRA), BNE, AU, 2018. https://doi.org/10.1109/ICRA.2018.8460487

[3] B. Paden, M. Čáp, S. Yong, D Yershov and E. Frazzoli. "A survey of motion planning and control techniques for self-driving urban vehicles." IEEE Transactions on intelligent vehicles, **1**(1), 33-55, 2016. https://doi.org/10.1109/TIV.2016.2578706.

[4] T Ariyoshi. "パーソナルカー 自動運転システムとそのめざす価値." SICE 57(7), 488-492, 2018, in Japanese. https://doi.org/10.11499/sicejl.57.488

[5] S.Adachi. "物理と情報と制御." Calsonic Kansei technical review, 11, 6-14. 2014, in Japanese.

[6] Y. Yamauchi, M. Saito and T. Ono "Adaptive identification method of vehicle modeling according to the fluctuation of road and running situation in autonomous driving", 2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), HIJ, JP, 2019. https://doi.org/10.23919/SICE.2019.8859845

[7] R. Rajamani, Vehicle Dynamics and Control, ser. Mechanical Engineering Series. Springer, 2011.

[8] R Bosch "Automotive Handbook" 8 ed. John Wiley & Sons Ltd, 2011.

[9] J. Kong, M. Pfeiffer, G. Schildbach and F. Borrelli. "Kinematic and dynamic vehicle models for autonomous driving control design." in IEEE Intelligent Vehicles Symposium (IV), SEL, KR, 2015. https://doi.org/10.1109/IVS.2015.7225830

[10] P. Polack, F. Altché, B. d'Andréa-Novel and A. Fortelle. "The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles?." in IEEE Intelligent Vehicles Symposium (IV), SEA , USA, 2017. https://doi.org/10.1109/IVS.2017.7995816

[11] K.Kabe and N. Miyashita" A study of the cornering power by use of the analytical tyre model." Vehicle System Dynamics, **43**(sup1), 113-122, 2005. https://doi.org/10.1080/00423110500139999

[12] Y. Mitsuhashi and S. Fujii. "Validation of Fiala model for motorcycle with FEM tire model.(article in Japanese with an abstract in English)" Transactions of the Society of Automotive Engineers of Japan, **44**(1), 75-80,2013. https://doi.org/10.11351/jsaeronbun.44.75

[13] Y. Uno, R. Quirynen, S. Di Cairano, B. Karl. "Vehicle Integrated Control by Nonlinear Model Predictive Control in GNSS Based Autonomous Driving System." IJAE, **50**(4), 1176-1181, 2019, in Japanese. https://doi.org/10.11351/jsaeronbun.50.1176

[14]] K. Berntorp and S. Di Cairano, "Tire-stiffness and vehicle-state estimation based on noise-adaptive particle filtering," IEEE Transactions on Control Systems Technology, **27**(3), 1100-1114, 2018. https://doi.org/10.1109/TCST.2018.2790397

[15] B. Karl, R. Quirynen, and S. Di Cairano. "Steering of autonomous vehicles based on friction-adaptive nonlinear model-predictive control." 2019 American Control Conference (ACC), PA, USA, 2019. https://doi.org/10.23919/ACC.2019.8814448

[16] F. Borrelli, P. Falcone, T. Keviczky, J. Asgari and D. Hrovat. "MPC-based approach to active steering for autonomous vehicle systems." International journal of vehicle autonomous systems **3**(2-4), 265-291 2005. https://doi.org/10.1504/IJVAS.2005.008237

[17] M. Bujarbaruah, X. Zhang, H. Tseng and F. Borrelli. " Adaptive mpc for autonomous lane keeping." 14th International Symposium on Advanced Vehicle Control (AVEC), BJS, CN, 2018. https://arxiv.org/abs/1806.04335

[18] X. Qian, I. Navarro, A Fortelle and F. Moutarde. "Motion planning for urban autonomous driving using Bézier curves and MPC." in 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), RIO, BR, 2016. https://doi.org/10.1109/ITSC.2016.7795651

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 710-718 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Eliminating Target *Anopheles* Proteins to Non-Target Organisms based on Posterior Probability Algorithm

Marion Olubunmi Adebiyi*, Oludayo Olufolorunsho Olugbara

ICT and Society Research Group, Luban Workshop, Durban University of Technology, P.O Box 1334, Durban 4000, South Africa

ARTICLE INFO

Article history:
Received: 03 September, 2020
Accepted: 07 January, 2021
Online: 05 February, 2021

Keywords:
Alignment
Anopheles
Homology
Organism
Probability

ABSTRACT

Capturing similarity in gene sequences of a target organism to detect significant regions of comparison will most likely occur because genes share a related descendant. Local sequence alignment for the targeted organisms can help preserve associations among sequences of related organisms. Such homologous genes possess identical sequences with common ancestral genes. The genes may be similar to common traits, and varying purposes, but they descend from a common ancestor. Basic local alignment search tool (BLAST) from the National Center for Biotechnology Information. (NCBI) has been used by different researchers to resolve the various forms of alignment problems. However, much literature to bare the efficacy of standard protein-protein BLAST (BLASTp) on the MATLAB platform has not been seen. In this study, a position-specific iteration BLASTp of 20 anopheles insecticide target protein sequence was performed on NCBI Ensembl against genomes of Anopheles (target organism), then against humans, fruit-fly, zebrafish, and chicken genomes (non-target organisms) to eliminate the targets with homology to non-target organisms. Furthermore, the same iteration was repeated for the genomes of Anopheles and non-target organisms using a posterior probability algorithm built into MATLAB as a tool for protein to protein search BLAST. Outputs from NCBI and MATLAB were put forward to determine the optimality of an optimized search algorithm on MATLAB. The MATLAB-Blastp method based on the application of posterior probability has helped to avoid errors occurring in the early stages of alignment. Moreover, the same results were obtained for the sought features on NCBI Blastp with a refined understanding of how feature values are generated from MATLAB posterior probability built-in algorithm for position-specific BLAST.

1. Introduction

Local alignment algorithms are suitable for unrelated sequences that are assumed to comprehend areas of comparable sequence motifs within a larger sequence framework [1]. Alignment is a mutual procedure of two sequences exhibiting positions where sequences are similar or dissimilar. A sequence alignment establishes residue-to-residue correspondences among sequences such that the order of residue in each sequence is preserved. Sequence analysis is a subject of deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or peptide sequence for wide variations of analytical techniques to comprehend its purpose, structure, and development [2].

Many dynamic programming algorithms such as Smith-Waterman, FASTA, and BLAST algorithms were developed for accomplishing local alignment. The Smith–Waterman algorithm accomplishes the task of local alignment of sequences for defining comparable neighborhoods of regions between two nucleotides or protein sequences. The algorithm captures the segments of all likely lengths and optimizes a comparative measure, instead of stretching over the entire sequence being process [3]. FASTA program was written for comparing protein sequences but it was later modified to conduct searches on DNA [4]. FASTA software uses the principle of finding similarity between two sequences statistically. This software matches one sequence of DNA or protein with the other by local sequence alignment method. It searches for local region for similarity but not the best match between two sequences. Since this software compares localized similarities at a time, it can come up with a mismatch. FASTA takes a small part of a sequence known as k-tuples where a tuple can be from 1 to 6, matches it with k-tuples of other sequence and once a threshold value of matching is reached it generates result. It is a program that is used to shortlist prospects of matching large sequences because it is very fast.

*Corresponding Author: Marion Olubunmi Adebiyi, mariona@dut.ac.za

BLAST is frequently used for relating data sequences, recovering, and extracting sequences from databases in bioinformatics [5]. It has shown useful contributions in molecular biology, computational biology, and molecular genetic [6]. BLAST presents reliable and fast statistical reports, flexible search algorithm, and heuristic search methods [7]. BLAST based algorithms seek to extract a snippet of a query sequence that has a perfect alignment with a fragment of a targeted sequence found in a database. In the original BLAST algorithm, the chopped fragment is becoming the input to extend alignment in both query and subject database. BLAST searches for short sequences in an input query that matches short sequences in a database [8]. The nucleotide-nucleotide search, megaBLAST, BLASTN, BLASTP, BLASTX, TBLASTN are other architypes of BLAST algorithms [9].

Dynamic programing through the BLAST algorithm with various variants has made homologous search on genome databases of organisms possible. The aim of such analysis was to predict genes that are homologous in similar or dissimilar species. Gene prediction depends mostly on comparing a genomic sequence with a complementary DNA (cDNA), or protein database. However, most results are inaccurate for several reasons, including incomplete reference databases and lack of contribution to analysis of species. Position specific iterative BLAST (PSI-BLAST) is a program that finds distance relative to a protein. It creates a list of all closely related proteins that were combined into a general "profile" sequence, and summarizes significant features found in protein sequences. A query against a protein database is performed with in-built profile to extract larger group of proteins. The posterior probability is implemented using Bayesian theorem that involves revising a prior profile sequence that is extracted by psi-blast. This algorithm takes into consideration a new sequence profile information, which is a larger profile sequence group used to construct a profile and the process was iterated for four identical non-target organisms. It is believed that PSI-BLAST is much more sensitive in picking up distance based evolutionary relationships than a standard normal protein-protein BLASTP [10].

In this study, a local alignment algorithm has been developed based on BLASTP (protein query sequence against protein database search) and PSI BLAST (for more sensitive protein–protein similarity searches) to determine the effectiveness of the two search algorithms. A literature review to extract one of the best local search algorithms (Smith Waterman algorithm) was conducted. The NCBI BLAST was used to implement a protein query against protein database with a known essential protein sequences with PSI-BLASTP combined with posterior probability to generate an updated list of corresponding homologous protein from four other closely identical organisms. NCBI BLAST is widely used to implement sequence alignment, but very few works have implemented the local BLAST on MATLAB environment. Consequently, we are proposing a distinct way of elucidating homologous genes of a target organism when compared to selected non-target organisms through sequence alignment. This work is relevant for predicting genes that must be targeted in anopheles when formulating new compounds for drug target. If such gene is tampered with in the target anopheles during insecticidal spray, what happens to non-target organisms such as human, fruit flies, chickens, and fishes. There is a high tendency of harming non-targeted organisms during insecticidal spray if homologous genes are not eliminated in the target gene list before insecticidal compounds are formulated and recommended for use. Local sequence alignment for a targeted specie can help preserve associations among sequences of related specie. Such homologous genes possess identical sequences with common ancestral gene and are useful for function prediction and characterization.

2. Related Works

Orthologs are homologous genes that diverge after a speciation event, and still have their main functions conserved. A homologous gene is inherited from two species by a common ancestor and homologous genes can be similar in sequence. But homology is the existence of the same body morphological structures in different organisms. It is an important concept of evolution and comparative biology [11]. The availability of genome sequence of several species has provided an opportunity to elucidate the effects of evolution on every nucleotide and protein in a genome [12-14]. It is easy to identify nucleotide sets that descended from a common ancestral nucleotide with sequence alignment because the problem of identifying an evolutionary related nucleotide and protein is the sequence alignment [15]. Strategies for aligning multiple and entire genomes of organisms include ‘local’ alignment and ‘global’ alignment [13, 15]. Their work has demonstrated a typical example of evolutionary scenario that involved the replication of double-stranded DNA in a parent cell (target organism) and division into two child cells. Their result showed two positions of an undirected duplication because of slippage replication, and occurrence of a directed duplication involving an RNA intermediate.

Studies on efficiency of alignment algorithm for homologous sequence similarity search in a genomic database was performed on a single [3]. The dynamic programming was deployed to isolate similar regions in gene sequences as a form of comparative analysis [16]. The authors discovered a suitable technique to calculate similarity in protein gene sequences within and across related organisms. They proposed a technique that computes sequence similarity, and their algorithm was evaluated using data sets from various species. The ‘best-in-genome’ method has been introduced [17], where a pairwise local alignment [18] between rat (target genome) and human genome (non-target genome) was initiated. The filter kept the best alignment for each position in rat genome and generated many to one relationship between rat genome and human genome but did not capture all orthologous relationships. The result was a reference-based multiple alignment with a property that gave every column, at most one position from each genome. A vertebrate local aligner with a faster nucleotide and more sensitive cross-species protein alignments has been constructed [19]. A web-based BLAST server for human genome makes homologous search possible and serves the purpose multiple genome alignments for yeast, insects, and vertebrates [20]. Aligning human, mouse, and rat genome [21] with progressive extension pairwise alignment orchestrated for human to mouse alignment has been reported [22]. Researches have combined strategies and tools for whole genome alignments [23-27], but none of the previous works have been found to specifically provide the Psi-Blastp by Bayes theorem

posterior probability for generating homologous alignment of anopheles to human, fruit-fly, zebra-fish and chicken genomes as provided in this work.

3. Experimental

This work has been implemented using minimum hardware and software requirements of a computer system with at least 16GB of RAM, 1TB hard disk capacity, Intel Core i7 Microprocessor, with VGA monitor compatible of at least 640/480 resolution and enhanced keyboard with a mouse. The software requires windows operating 7, or higher version, a Blosum62, BLAST standalone database, MATLAB R2016B, and online Ensembl NCBI Database [28] and [29]. The 20 essential genes identified to be potential insecticidal targets for malaria vector were used as input data for Mosquito *Anopheles Gambiae* [30]. The original data for analysis were extracted from Kyoto Encyclopedia of Genes and Genomes (KEGG) database and AnoCyc database on BioCyc, https://biocyc.org/organism-summary?object=ANO [31]. The detailed summary of Anopheles gambiae, version 24.1 KEGG described a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. BioCyc is a swarm of about 5700 pathway/genome databases (PGDBs) specific to various organisms. Each PGDB houses a predicted metabolic network and full genome of a specific organism, including reactions, metabolites, metabolic pathways, enzymes, proteins, genes, and lots of other components [32]. The protein sequence of these genes was extracted from protein database at NCBI as flat file or FASTA format from Genebank. Table 1 shows the dataset features of twenty genes as enzyme name, protein name, gene name, enzyme commission (EC) number and gene identity (ID).

The implementation of a basic local alignment method using a query against protein database was performed in this study using posterior probability function in MATLAB. The goal was to ascertain the creation of position specific score profile matrix from an alignment. The BLASTp program in MATLAB Bioinformatic tool was designed to map sequences of 20 previously identified *Anopheles* gambiae insecticidal target genes unto all available protein sequence databases of a specified organism, disease, population, or proteome. Bayesian posterior probability was calculated from psi-blast output to capture the related proteins from distance species during its search. This has resulted into a larger profile sequence group used to construct the final profile information. The process was repeated for the four non-target organisms investigated in this study.

Table 1: Dataset Features

S/N	AnoCyc (BioCyc) Enzyme Name	Uniprot Protein Name	Gene Name (Uniprot)	EC Number	Gene ID
1	Alkyl hydroperoxide reductase subunit C Thiol specific antioxidant	Thioredoxin-dependent peroxidase	TPX1	1.11.1.15	AGAP000396
2	Cytochrome P450 B-class	AGAP012295-PA	CYP9L1	1.14.14.1	AGAP012295
3	Cytochrome P450	AGAP002429-PA	CYP314A1	1.14.99.22	AGAP002429
4	Betaine aldehyde dehydrogenase	AGAP003578-PA	1274242	1.2.1.3	AGAP003578
5	Ribosomal RNA adenine dimethylase	rRNA adenine N (6)-methyltransferase	1274612	2.1.1.183	AGAP004465
6	Methyltransferase type 11	2-methoxy-6-polyprenyl-1,4-benzoquinol methylase, mitochondrial	coq5	2.1.1.201	AGAP010488
7	tRNA (guanine9-N1)-methyltransferase	tRNA methyltransferase 10 homolog A	1271937	2.1.1.221	AGAP000324
8	MT-A70-like	AGAP002895-PA	1273072	2.1.1.62	AGAP002895
9	Cholineethanolamine kinase	AGAP000010-PA	1272266	2.7.1.82	AGAP000010
10	Diacylglycerol kinase catalytic domain	Sphingosine kinase	1270104	2.7.1.91	AGAP006995
11	Zn (II)-responsive transcriptional regulator	Phenylalanyl-tRNA synthetase beta subunit	1274174	6.1.1.20	AGAP003517
12	Galactose-binding domain-like	Beta-galactosidase	1281056	3.2.1.23	AGAP002055
13	Carbon-nitrogen hydrolase	AGAP012662-PA	1269132	3.5.1.3	AGAP012662
14	Formylmethionine deformylase	Peptide deformylase	1271597	3.5.1.88	AGAP003861
15	Threonyl-tRNA synthetase class IIa	Prolyl-tRNA synthetase	1274253	6.1.1.15	AGAP003589
16	Leucyl-tRNA synthetase	Leucyl-tRNA synthetase	1277687	6.1.1.4	AGAP008297
17	Ribosomal RNA methyltransferase Spb1 C-terminal	23S rRNA (uridine2552-2'-O)-methyltransferase	1274000	2.1.1.166	AGAP004177
18	Farnesyl diphosphate synthase	Polyprenyl synthetase	1269998	2.5.1.1	AGAP007104
19	Glucosaminegalactosamine-6-phosphate isomerase	6-phosphogluconolactonase	1271093	3.1.1.31	AGAP010866
20	FAD-binding type 2	Alkylglycerone-phosphate synthase	1274507	2.5.1.26	AGAP004358

Block substitution matrix (BLOSUM), which is the default matrix for a protein BLAST algorithm is a substitution matrix used for sequence alignment of protein. The matrix is based on local alignment. Based on block comparisons of sequence from database blocks, it contains multiple aligned un-gapped segments that correspond to the top-most conserved protein regions. The goals are to identify "biologically significant" patterns in protein families by emphasizing regions that are thought to be important to protein function. To look for good "discriminators" that emphasize and identify known family members, while excluding known non-members and to pro-site patterns of "motifs". The standalone BLAST is a suit of programs that were designed to mimic the NCBI BLAST server, and include "blastall"," megablast", and "blastp" that exist in NCBI BLAST suit. Its ease of use and user friendliness features are strong inspiration for its application in this study.

4. Results and Discussion

The protein sequences for each of the 20 insecticidal target genes was blasted against protein databases of four non-target organisms. The organisms are *Homo Sapiens* (Taxonomy ID: 9609) - human genome, *Drosophila* (Taxonomy ID: 7227) – fruit fly genome, *Danio Rerio* (Taxonomy ID:7955) – Zebra fish genome and G*allus Gallus* (Taxonomy ID: 9031) – chicken genome. The selection procedure certifies that homologous gene of target organism (anopheles) exists in these four non-target organisms. There is possible homology match, which was identified and should be catered for in case of gene inhibition during insecticidal development. In fact, about 30 sequences were blasted against the whole genome per single iteration. The BLAST was matched with parameters before commencing on further BLASTP commands. The BLAST of proteins was run against the protein database of four targeted organisms. The BLASTP algorithm was implemented on Blosum62 Ensembl database section. During this process we were looking for homologs, which is a measure of relationship between two genes that descended from a common ancestral protein. Their various e-values and percentage identities were identified as the unique selection criteria. It confirmed the literature [33] that homologous genes can be predicted and validated by sequence alignment method.

The e-value measures level of a likelihood that any match in sequences is purely by chance. Consequently, a lower e-value determines the less significant matches made but gives an idea of potential relations among query organisms and database organisms. This is a result of random chance and therefore the most significant gene matches in those non-target organisms were selected as the homologous. Selecting a homolog like the target sequence analyzed, the match with the lowest e-value and highest percentage was classified a significant match or hit. The e-value threshold was set at 0.00001 as a standard with NCBI Ensembl database that was deployed, while the position specific iteration BLAST threshold was set at 0.005. Figures 1 and 2 represent two sample result pages of 20-protein sequences blasted against the four non-target genome databases investigated with e-values and IDs of its homologous gene.

Different values were collected for various organisms blasted on ensembl platform. These values are the homolog, e-value, and percentage identity. Tables 2 and 3 show results of human and zebra fish homologous genes of the four organisms blasted in this study. The extracted database files from the Ftp site for the Anopheles was extracted from ftp://ftp.ncbi.nlm.nih.gov/genomes/ and available on Figure 3 shows this result for the case of human genomic information. The figure constitutes the precise data files and human genome information from NCBI genome databases and similar results were obtained for Zebrafish, Fruit-fly, and Chicken. The databases from the Ftp site were downloaded as GenBank files and converted to Microsoft (MS) Access database. The MS access file was linked as input to the MATLAB for implementation. BLASTP search was performed using MATLAB. A position specific iteration blast for protein query against protein (PSI–BLASTP) was completed for each sequence of anopheles against non-target organism (Human, Zebrafish, Fruit-fly, and Chicken) genome database. This was done to eliminate targets with homolog to non-target organism.



Figure 1: BLAST result page on ensembl.org

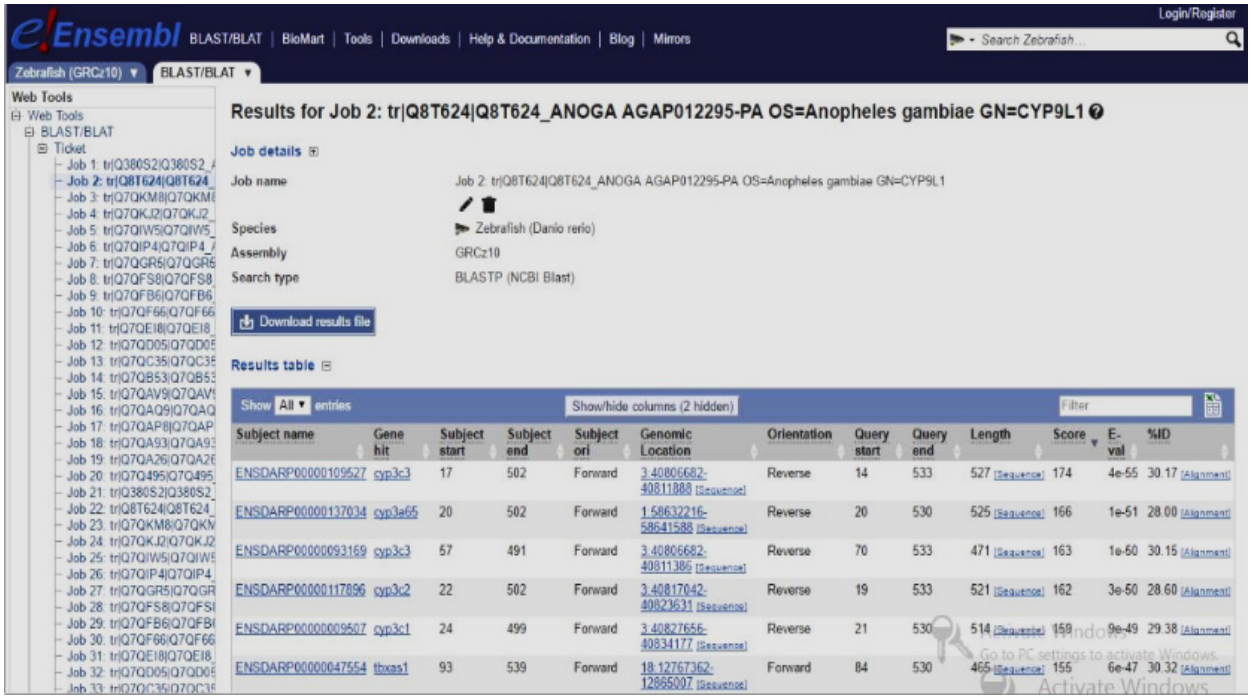


Figure 2: Extracted result with e-value and ID of its homologous gene

Finally, 20 anopheles protein targets were screened with two target genomes of Homo sapiens and Drosophila melanogaster. This was to identify isoforms from two organisms when screened with genomes of anopheles. The result of anopheles protein isoforms when screened with homo sapient and drosophila genomes is presented in Table 4. These isoforms can formulate a potential class of protein and can elucidate more molecular and functional variations ciphered in the genome by further functional analysis.

The search for protein isoform in this study may be out of scope because we do not have plans to conduct clinical or computational proteomics analysis and transcriptomic analysis. However, protein isoform is seen as the same protein existing in many different forms, rather it is a new class of protein that may be useful as biomarkers for early diagnostic of clinical proteomics [34]. Studies have shown that traditional methods of protein isoform determination have proved that isoforms can only be determined quantitatively at the transcript level, not in the protein level. Moreover, that condition came with several other disadvantages, high throughput analysis has made it possible, but the data used in this study are not confirmed transcriptomic dataset.

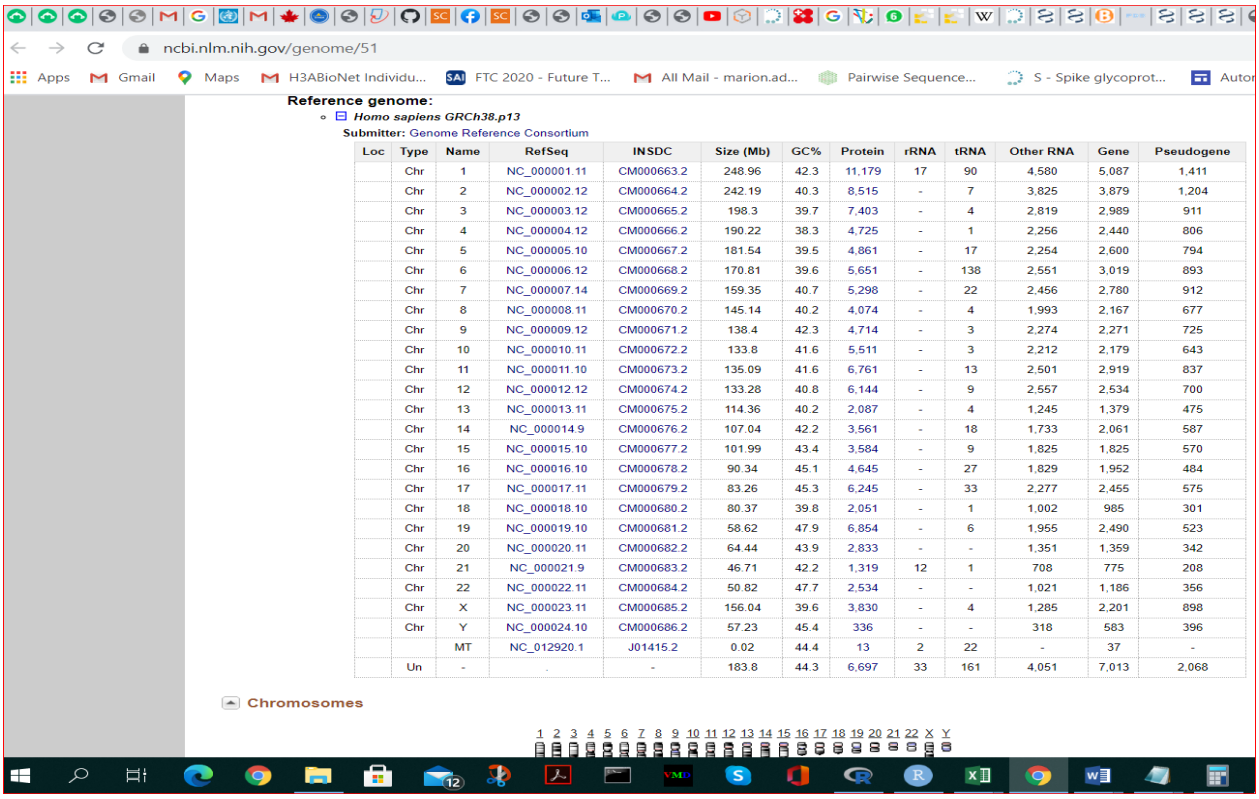
Table 2: Human (homo sapient) homologous gene

S/N	AnoCyc (BioCyc) Enzyme Name	EC Number	KEGG Gene ID	Homolog Gene ID	E-value	% ID
1	Alkyl hydroperoxide reductase subunit C Thiol specific antioxidant	1.11.1.15	AGAP000396	ENSP00000389047	1.00E-64	66.27
2	Cytochrome P450 B-class	1.14.14.1	AGAP012295	ENSP00000228606	1.00E-12	31.63
3	Cytochrome P450	1.14.99.22	AGAP002429	ENSP00000368079	1.00E-18	23.11
4	Betaine aldehyde dehydrogenase	1.2.1.3	AGAP003578	ENSP00000438296	1.00E-157	57.73
5	Ribosomal RNA adenine dimethylase	2.1.1.183	AGAP004465	ENSP00000421754	1.00E-107	77.91
6	Methyltransferase type 11	2.1.1.201	AGAP010488	ENSP00000449933	3.00E-11	58.82
7	tRNA (guanine9-N1)-methyltransferase	2.1.1.221	AGAP000324	ENSP00000423628	1.00E-14	35.77
8	MT-A70-like	2.1.1.348	AGAP002895	ENSP00000440598	1.00E-43	34.93
9	Cholineethanolamine kinase	2.7.1.82	AGAP000010	ENSP00000398091	1.00E-37	47.57
10	Diacylglycerol kinase catalytic domain	2.7.1.91	AGAP006995	ENSP00000471180	1.00E-14	38.25

11	Zn(II)-responsive transcriptional regulator	6.1.1.20	AGAP003517	ENSP00000367498	1.00E-20	30.49
12	Galactose-binding domain-like	3.2.1.23	AGAP002055	ENSP00000407365	1.00E-06	46.94
13	Carbon-nitrogen hydrolase	3.5.1.3	AGAP012662	ENSP00000356986	1.00E-37	33.21
14	Formylmethionine deformylase	3.5.1.88	AGAP003861	ENSP00000288022	1.00E-40	39.35
15	Threonyl-tRNA synthetase class IIa	6.1.1.15	AGAP003589	ENSP00000358060	1.00E-05	22.77
16	Leucyl-tRNA synthetase	6.1.1.4	AGAP008297	ENSP00000447763	1.00E-07	21.05
17	Ribosomal RNA methyltransferase Spb1 C-terminal	2.1.1.166	AGAP004177	ENSP00000384423	1.00E-22	37.41
18	Farnesyl diphosphate synthase	2.5.1.1	AGAP007104	ENSP00000417865	1.00E-05	27.04
19	Glucosaminegalactosamine-6-phosphate isomerase	3.1.1.31	AGAP010866	ENSP00000471446	2.00E-11	37.63
20		2.5.1.26	AGAP004358	ENSP00000417011	1.00E-11	25.58

Table 3: Zebra fish (danio rerio) homologous gene

S/N	AnoCyc (BioCyc) Enzyme Name	EC Number	Gene ID	Homolog Gene	E-value	% ID
1	Alkyl hydroperoxide reductase subunit C Thiol specific antioxidant	1.11.1.15	AGAP000396	ENSDARP00000120934	1.00E-49	72.5
2	Cytochrome P450 B-class	1.14.14.1	AGAP012295	ENSDARP00000122647	1.00E-06	34.07
3	Cytochrome P450	1.14.99.22	AGAP002429	ENSDARP00000091260	1.00E-19	24.52
4	Betaine aldehyde dehydrogenase	1.2.1.3	AGAP003578	ENSDARP00000012767	1.00E-154	57.91
5	Ribosomal RNA adenine dimethylase	2.1.1.183	AGAP004465	ENSDARP00000124704	1.00E-04	30.69
6	Methyltransferase type 11	2.1.1.201	AGAP010488	ENSDARP00000131342	1.00E-04	25.22
7	tRNA (guanine9-N1)-methyltransferase	2.1.1.221	AGAP000324	ENSDARP00000109893	1.00E-36	40.41
8	MT-A70-like	2.1.1.348	AGAP002895	ENSDARP00000022188	4.00E-154	48.34
9	Cholineethanolamine kinase	2.7.1.82	AGAP000010	ENSDARP00000019763	2.00E-93	46.37
10	Diacylglycerol kinase catalytic domain	2.7.1.91	AGAP006995	ENSDARP00000117613	1.00E-07	33.33
11	Zn(II)-responsive transcriptional regulator	6.1.1.20	AGAP003517	ENSDARP00000069614	4.00E-17	31.07
12	Galactose-binding domain-like	3.2.1.23	AGAP002055	ENSDARP00000047190	1.00E-120	38.75
13	Carbon-nitrogen hydrolase	3.5.1.3	AGAP012662	ENSDARP00000121828	1.00E-69	44.09
14	Formylmethionine deformylase	3.5.1.88	AGAP003861	ENSDARP00000013949	3.00E-43	40
15	Threonyl-tRNA synthetase class IIa	6.1.1.15	AGAP003589	ENSDARP00000103726	1.00E-04	22.57
16	Leucyl-tRNA synthetase	6.1.1.4	AGAP008297	ENSDARP00000106738	5.00E-127	63.76
17	Ribosomal RNA methyltransferase Spb1 C-terminal	2.1.1.166	AGAP004177	ENSDARP00000138990	1.00E-15	27.8
18	Farnesyl diphosphate synthase	2.5.1.1	AGAP007104	ENSDARP00000059927	1.00E-74	44.48
19	Glucosaminegalactosamine-6-phosphate isomerase	3.1.1.31	AGAP010866	ENSDARP00000124115	2.00E-13	29.92
20		2.5.1.26	AGAP004358	ENSDARP00000136279	1.00E-177	52.42



Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	1	NC_000001.11	CM000663.2	248.96	42.3	11,179	17	90	4,580	5,087	1,411
	Chr	2	NC_000002.12	CM000664.2	242.19	40.3	8,515	-	7	3,825	3,879	1,204
	Chr	3	NC_000003.12	CM000665.2	198.3	39.7	7,403	-	4	2,819	2,989	911
	Chr	4	NC_000004.12	CM000666.2	190.22	38.3	4,725	-	1	2,256	2,440	806
	Chr	5	NC_000005.10	CM000667.2	181.54	39.5	4,861	-	17	2,254	2,600	794
	Chr	6	NC_000006.12	CM000668.2	170.81	39.6	5,651	-	138	2,551	3,019	893
	Chr	7	NC_000007.14	CM000669.2	159.35	40.7	5,298	-	22	2,456	2,780	912
	Chr	8	NC_000008.11	CM000670.2	145.14	40.2	4,074	-	4	1,993	2,167	677
	Chr	9	NC_000009.12	CM000671.2	138.4	42.3	4,714	-	3	2,274	2,271	725
	Chr	10	NC_000010.11	CM000672.2	133.8	41.6	5,511	-	3	2,212	2,179	643
	Chr	11	NC_000011.10	CM000673.2	135.09	41.6	6,761	-	13	2,501	2,919	837
	Chr	12	NC_000012.12	CM000674.2	133.28	40.8	6,144	-	9	2,557	2,534	700
	Chr	13	NC_000013.11	CM000675.2	114.36	40.2	2,087	-	4	1,245	1,379	475
	Chr	14	NC_000014.9	CM000676.2	107.04	42.2	3,561	-	18	1,733	2,061	587
	Chr	15	NC_000015.10	CM000677.2	101.99	43.4	3,584	-	9	1,825	1,825	570
	Chr	16	NC_000016.10	CM000678.2	90.34	45.1	4,645	-	27	1,829	1,952	484
	Chr	17	NC_000017.11	CM000679.2	83.26	45.3	6,245	-	33	2,277	2,455	575
	Chr	18	NC_000018.10	CM000680.2	80.37	39.8	2,051	-	1	1,002	985	301
	Chr	19	NC_000019.10	CM000681.2	58.62	47.9	6,854	-	6	1,955	2,490	523
	Chr	20	NC_000020.11	CM000682.2	64.44	43.9	2,833	-	-	1,351	1,359	342
	Chr	21	NC_000021.9	CM000683.2	46.71	42.2	1,319	12	1	708	775	208
	Chr	22	NC_000022.11	CM000684.2	50.82	47.7	2,534	-	-	1,021	1,186	356
	Chr	X	NC_000023.11	CM000685.2	156.04	39.6	3,830	-	4	1,285	2,201	898
	Chr	Y	NC_000024.10	CM000686.2	57.23	45.4	336	-	-	318	583	396
		MT	NC_012920.1	J01415.2	0.02	44.4	13	2	22	-	37	-
	Un	-	-	-	183.8	44.3	6,697	33	161	4,051	7,013	2,068

Figure 3: Database file genome information for Humans (Homo Sapiens)

Table 4: Anopheles protein isoforms when screened against *homo sapient* and *drosophila* genomes

Anopheles Gene ID	Number of Isoforms for Anopheles	Human Homolog/ Uniprot Gene ID	Number of Isoforms for Human homolog gene	Drosophila Homolog/ Uniprot Gene ID	Number of Isoforms for Drosophila homolog gene
AGAP000396	0	ENSP00000389047/ PRDX1	3	FBpp0082927/ Prx3	0
AGAP012295	0	ENSP00000228606/ CYP27B1	3	FBpp0088127/ Cyp9b1	0
AGAP002429	0	ENSP00000368079/ CYP4V2	2	FBpp0088437/ Cyp12d1-p	2
AGAP003578	0	ENSP00000438296/ ALDH1A2	10	FBpp0079406/ Aldh	2
AGAP004465	0	ENSP00000421754/ DIMT1	4	FBpp0291478/ mtTFB1	0
AGAP010488	0	ENSP00000449933/ COQ5	6	FBpp0073568/ Coq5	2
AGAP000325	2	ENSP00000423628/ TRMT10A	3	FBpp0077043/ trmt10a	0
AGAP002895	0	ENSP00000440598/ METTL3	6	FBpp0079219/ Mettl14	0
AGAP000010	0	ENSP00000398091/ ETNK2	7	FBpp0310123/ eas	5
AGAP006995	0	ENSP00000471180/ SPHK2	8	FBpp0073346/ Sk1	2
AGAP003517	2	ENSP00000367498/ LRRC47	2	FBpp0084021/ beta-PheRS	0
AGAP002055	0	ENSP00000407365/ GLB1	8	FBpp0078861/ Gal	2
AGAP012662	0	ENSP00000356986/ NIT1	5	FBpp0081507/ Dmel\CG8132	0
AGAP003861	0	ENSP00000288022/ PDF	0	FBpp0081794/ Dmel\CG31373	0
AGAP003589	0	ENSP00000358060/ TARS2	6	FBpp0072825/ ProRS-m	2
AGAP008297	0	ENSP00000447763/ VARS2	13	FBpp0291534/ IleRS-m	0

AGAP004177	0	ENSP00000364423/ PTCH1	12	FBpp0082832/ CG5220	0
AGAP007104	0	ENSP00000417865/ GGPS1	3	FBpp0087266/ Fpps	0
AGAP010866	0	ENSP00000471446/ PGLS	4	FBpp0297905/ Oscillin	4
AGAP004358	0	ENSP00000417011/ LDHD	3	FBpp0070365/ D2hgdh	3

5. Conclusion

In this study, BLAST which is a frequently used to determine sequence similarity by querying various sequences type against databases of various datatypes was experimented. BLAST with 20 *Anopheles* protein sequence was queried against four non-target organism databases. The results of this study have revealed the same e-values, which indicates that there is no significant homology (all e–values > 0.001) between the data on *Anopheles* when compared to humans, fruit-fly, zebrafish, and chicken databases. Posterior probability in MATLAB was used for experimentation in this study. The study results have shown obvious insecticidal targets in *Anopheles gambiae* with no significant homology to humans, fruit-fly, zebrafish, and chicken. Further analysis like synthesizing these targets in an experimental scenario can help close dangling ends in search for new insecticide compounds. This may be a useful endeavor for future research.

6. Data Availability Statement

The original source of data for this study is from a standard and structured public repository KEGG and AnoCyc on BioCyc, (https://biocyc.org/organism-summary?object=ANO). Protein sequence of these genes was extracted from the protein database at NCBI as flat file or FASTA format from Genebank. In addition, contents were extracted from UniProt Knowledgebase @ UniProtKB (https://www.uniprot.org/database/DB-0023).
This represents some details from Cross-referenced databases on UniProtKB, Protein knowledgebase and UniParc Sequence archive. This is because our analysis involved fetching data from various public databases to generate outputs as the algorithm required in pursuit of research objective. Links to all data supporting the conclusions of this study is publicly available as indicated by web links.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to appreciate the postgraduate research directorate for postdoctoral sponsorship to the first author.

References

[1] V. G. Tumanyan, V. O. Polyanovsky, and M. A. Roytberg, “Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences,” Algorithms for Molecular Biology, **6**(25), 1748-7188, 2011, https://doi.org/10.1186/1748-7188-6-25.

[2] E. S. Donkor, N. T. K. D. Dayie, T. K. Adiku. “Bioinformatics with Basic Local Alignment Search Tool (BLAST) and Fast Alignment (FASTA)”, Journal of Bioinformatics and Sequence Analysis, **6**(1), 1-6, 2014, https://doi.org/10.5897/IJBC2013.0086.

[3] T. O. Oladele, O. M. Bamigbola, and C. O. Bewaji, “On Efficiency of Sequence Alignment Algorithms,” African Scientist, **10**(1), 9-14, 2009.

[4] D. J. Lipman, and W. R. Pearson, "Rapid and sensitive protein similarity searches", Science, **227**(4693), 1435-1441, 1985, https://doi.org/10.1126/science.2983426.

[5] C. A. Kerfeld, K. M. Scott. “Using BLAST to teach “E-value-tionary Concepts”, PLoS Biology, **9**(2), 1-11, 2011.

[6] G. G. Syngai, P. Barman, R. Bharali, and S. Dey, “BLAST: An Introductory Tool for Students to Bioinformatics Applications”, Keanean Journal of Science, **2**, 67-76, 2013.

[7] R. S. Neumann, S. Kumar, and K. Shalchian-Tabrizi, “BLAST output visualization in the new sequencing era”, Briefings in Bioinformatics, **15**(4), 484-503, 2014, doi: 10.1093/bib/bbt009.

[8] D. W. Kim, N. R. Kim, D. S. Kim, S. H. Choi, S. H. Chae, H. S. Park, “easySEARCH: A user-friendly bioinformatics program that enables BLAST searching with massive number of query sequences”, Bioinformation, **8**(16), 792–794, 2012, doi: 10.6026/97320630008792.

[9] M. Thomas. “The BLAST Sequence Analysis Tool”, The NCBI Handbook, 2013.

[10] G. L, Rosen, R. Polikar, R., D. A. Caseiro, S. D. Essinger, B. A. Sokhansanj, “Discovering the unknown: Improving detection of novel species and genera from short reads”, Journal of Biomedicine and Biotechnology, 2011, https://doi.org/10.1155/2011/495849.

[11] B. K. Hall, "Homology and embryonic development", In M. K. Hecht, R. J. Macintyre, M. T. Clegg, eds, Evolutionary Biology, Springer, Boston MA, **28**, 1–37, 1995, https://doi.org/10.1007/978-1-4615-1847-1_1.

[12] D. S. Moore "Importing the homology concept from biology into developmental psychology", Developmental Psychobiology, **55**(1), 13–21, 2013, https://doi.org/10.1002/dev.21015.

[13] I. Brigandt, "Essay: Homology", In: The Embryo Project Encyclopedia, ISSN: 1940-5030, 2011, https://embryo.asu.edu/view/embryo:124921.

[14] R. W. Scotland, "Deep homology: a view from systematics", BioEssays, **32**(5), 438–449, 2010, doi: 10.1002/bies.200900175.

[15] N. D. Colin, P. Lior, “Evolution at the nucleotide level: the problem of multiple whole-genome alignment”, Human Molecular Genetics, **15**(1), R51–R56, 2006, https://doi.org/10.1093/hmg/ddl056.

[16] B. Shankar, D. Vinod, M. Basavaraj, and P. Manjunath. “Comparative analysis of dynamic programing algorithms to find similarity in gene sequences”, International Journal of Research in Engineering and Technology, **2**(8), 312-316, 2013.

[17] W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, “Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes”, Proc. Natl Acad. Sci. USA, **100**(20), 11484 – 11489, 2003, doi: 10.1073/pnas.1932072100.

[18] S. Batzoglou, “The many faces of sequence alignment”. Briefings in Bioinformatics, **6**(1), 6–22, 2005, doi: 10.1093/bib/6.1.6.

[19] W. J. Kent, “BLAT—the BLAST-like alignment tool”, Genome Res., **12**(4), 656–664, 2002, doi: 10.1101/gr.229202.

[20] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D. J. Thomas, et al., “The UCSC

Genome Browser Database", Nucleic Acids Research, **31**(1), 51–54, 2003, https://doi.org/10.1093/nar/gkg129.

[21] M. Brudno, A. Poliakov, A. Salamov, G. Cooper, A. Sidow, E. Rubin, V. Solovyev, S. Batzoglou, and I. Dubchak, "Automated whole-genome multiple alignment of rat, mouse, and human", Genome Res., **14**(4), 685–692, 2004, doi: 10.1101/gr.2067704.

[22] O. Couronne, A. Poliakov, N. Bray, T. Ishkhanov, D. Ryaboy, E. Rubin, L. Pachter, and I. Dubchak, "Strategies and tools for whole-genome alignments", Genome Research, **13**(1), 73–80, 2003, doi: 10.1101/gr.762503.

[23] M. Brudno, C. Do, G. Cooper, M. Kim, E. Davydov, E. Green, A. Sidow, S. Batzoglou, "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA", Genome Res., **13**(4), 721–731, 2003, doi: 10.1101/gr.926603.

[24] S. Zhao, J. Shetty, L. Hou, A. Delcher, B. Zhu, K. Osoegawa, P. de Jong, W. Nierman, R. Strausberg, and C. Fraser, "Human, mouse, and rat genome large-scale rearrangements: stability versus speciation", Genome Res., **14**(10a), 1851–1860, 2004, doi: 10.1101/gr.2663304.

[25] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, G.M Weinstock, R. K Wilson, R. A Gibbs, W. J Kent, W. Miller, D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes", Genome Res., **15**(8), 1034–1050, 2005, doi: 10.1101/gr.3715005.

[26] E. H. Margulies, M. Blanchette, D. Haussler, E. D. Green, "Identification and characterization of multi-species conserved sequences", Genome Res., **13**(12), 2507–2518, 2003, doi: 10.1101/gr.1602203.

[27] W. M. Fitch, "Distinguishing homologous from analogous proteins", Syst. Zool, **19**(2), 99-113, 1970, doi.org/10.2307/2412448.

[28] M. Mathur and Geetika, "Multiple Sequence Alignment Using MATLAB," International Research Publications House, **3**(6), 497-504, 2013.

[29] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research, **25**(17), 3389–3402, 1997, doi: 10.1093/nar/25.17.3389.

[30] M. Adebiyi, J. Oghuan, S. Fatumo, E. Adebiyi, R. Jason, "A Functional Workbench for Anopheles gambiae Micro Array Analysis", Proceedings - UKSim-AMSS 7th European Modelling Symposium on Computer Modelling and Simulation, 138-143, 2013, doi:10.1109/EMS.2013.24.

[31] R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti, P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes, a 2019 update", Nucleic Acids Res, **48**(D1), 445–453, 2020, doi: 10.1093/nar/gkz862.

[32] R. A. Holt, G. M. Subramanian, A. Halpern, G. G. Sutton, R. Charlab, D. R. Nusskern, P. Wincker, A. G. Clark, J. M. Ribeiro, S. L. Hoffman et al., "The genome sequence of the malaria mosquito Anopheles gambiae", Science, **298**(5591), 149-159, 2002, doi: 10.1126/science.1076181.

[33] T. Wiehe, S. Gebauer-Jung, T. Mitchell-Olds, R. Guigo. "SGP-1: prediction and validation of homologous genes based on sequence alignments," Genome Research, **11**(9), 1574-1583, 2001, doi.org/10.1101/gr.177401.

[34] F. Zhang, and Y. C. Chen, "A method for identifying discriminative isoform-specific peptides for clinical proteomics application", BMC genomics, **17**(S7), 522, 2016, doi: 10.1186/s12864-016-2907-8.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 750-756 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Johnson Noise and Optical Characteristics of Polymer Nanocomposites based on Colloidal Quantum Dots and *in-situ* Nanoparticles Formation

Fatin Hana Naning[*,1], S. Malik[2], Lee Feng Koo[1], Tze Jin Wong[1], Pang Hung Yiu[1]

[1]*Department of Basic Science and Engineering, Faculty of Agriculture and Food Sciences, Universiti Putra Malaysia Bintulu Campus, 97008 Bintulu, Sarawak, 93010, Malaysia*

[2]*Physics Department, Faculty of Science, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, 31400, Malaysia*

ARTICLE INFO

Article history:
Received:10 August, 2020
Accepted: 11 September, 2020
Online: 12 October, 2020

Keywords:
Johnson Noise
Polymer Nanocomposite
Cadmium sulfide

ABSTRACT

Electrical and optical properties of polymer nanocomposite thin films have been analyzed to study their reliability and competency as a component for optoelectronic devices such as LED and solar cells. Polymer nanocomposite encounters various challenges, such as the dispersion of nanoparticles in the matrix that hinders their efficiency for potential devices. In this paper, two types of polymer nanocomposites have been fabricated, and their Johnson noise, current density-voltage, and optical have been measured. The first type of nanocomposite produced through an in-situ method, that is by impregnating CdS or CdSe nanoparticles in conjugated polymer, P3HT (NP-CdX:P3HT). The nucleation of the nanoparticles was done using gas exposure. The second type is by directly adding CdS or CdSe quantum dots into P3HT (QD-CdX:P3HT). Both kinds of polymer nanocomposite thin films were fabricated using modified Langmuir-Blodgett technique. Results showed that for frequency above 10 Hz, the Johnson noise was less than 1×10^{-27} A2/Hz, regardless of the quantity of quantum dots or nanoparticles. The J-V results show (NP-CdX:P3HT) electrical performance compared with QD-CdX:P3HT. High polymer crystallization of NP-CdX:P3HT thin films is revealed by UV-Vis absorbance spectra. The quantum confinement effect is evidence through peak shifting and depreciation of absorption. The photoluminescence intensity of thin films decreased when they were exposed to the gas. It can be concluded that the NP-CdX:P3HT nanocomposites can be further studied as they have greater potential to be exploited in optoelectronic devices.

1. Introduction

Polymer nanocomposites can be prepared through various methods, including physical adsorption, ligand-metal interaction, or by blending nanoparticles with the desired polymers. The inability of nanoparticles to spread uniformly in matrices is one of the key obstacles in the development of polymer nanocomposites for optoelectronic devices. Besides the size of the nanomaterials, the degree of mixing of the materials influences the polymer nanocomposites characteristics. If the nanomaterials are poorly dispersed, the mechanical properties of the thin films may degrade [1]. The nanomaterials blend affects the movement of electrons and the transport of holes. Current and holes motion cause excess noise in all conducting materials. Various types of electrical noises are present in an electronic system such as Johnson noise, 1/*f* noise, and generation-combination noise.

Johnson noise, also referred to as thermal noise, is caused by the movement of electrons in a material that produces a fluctuating electromotive force (emf) [2]. The electron scattering creates noise in terms of small net current flow. As the temperature and resistance of the material increases, the noise increases as well.

Johnson noise theory derived by Johnson and Nyquist by stating

$$S_V(f) = 4kRhfN(f,T) \quad (1)$$

*Corresponding Author: Fatin Hana Naning, UPM Bintulu Sarawak Campus, fatinhanaz@upm.edu.my

where $S_V(f)$ is the power spectral density (PSD) of the voltage noise, k is Boltzmann constant, R is the resistance, h is Planck constant, and f is the frequency. $N(f,T)$ is Planck number representing the energy quanta hf, at frequency f and temperature, T [3,4]. It can also be expressed as,

$$N(f,T) = kT/hf \quad (2)$$

The noise-voltage power spectral density under open-circuit condition is given by

$$S_V(f) = 4k^2TR \quad (3)$$

Alternatively, the noise-current power spectral density $S_I(f)$ can be calculated using

$$S_I(f) = 4k^2T/R \quad (4)$$

Equations (3) and (4) valid for frequency below tens of gigahertz over a wide range of temperatures [5].

Johnson noise arises from the random motion of free electrons that leads to temporary electrons agglomeration and voltage fluctuation [6]. The noise can be estimated by analysing the power spectral density (PSD) obtained from low-frequency electrical noise measurement. It is a conventional technique used to study the reliability and quality of electronic devices as the procedure is able to differentiate failed devices. Usually, the impaired devices hold excess noise [7] because PSD of noise increases with stress and damage [8]. Deterioration process of Organic Light Emitting Diode (OLED) utilizing noise measurement has been studied by Rocha et al. and Carbone [9, 10]. Both concluded that noise increases with damage level. Another research group reports that the intrinsic transport mechanism greatly influences low-frequency noise [11].

Polymer nanocomposites are commonly prepared by mixing the polymer with the nanomaterials through various tedious steps. The small introduction of nanomaterials into conjugated polymer matrices causes the modification in the polymers' properties due to the substantial surface-area-to-volume ratio of nanoparticles [12]. Researches showed that mechanical and optical properties of the polymer were altered as well [13, 14].

However, this method has some drawbacks that hamper the efficiency of the thin film due to the presence of ligand that inhibits charge transfer. Furthermore, the usage of cosolvent to mix the nanomaterials and polymer negatively affect the orientation of the polymer [15]. One way to impede these issues is by growing the nanomaterials *in-situ* in the matrices.

In the present study, we used poly(3-hyexylthiophene) (P3HT) as the polymer matrix, and the nanomaterials were cadmium sulfide (CdS) and cadmium selenide (CdSe). The nanocomposites were prepared by directly adding CdS or CdSe quantum dots (QD-CdX:P3HT) into P3HT, or by impregnating CdS/CdSe nanoparticles in the polymer matrix *in-situ* (NP-CdX:P3HT).

Poly(3-hyexylthiophene) is one of the well-studied conjugated polymers. It possesses a relatively strong absorption coefficient, low energy band gap, consequently high absorption even in very thin films [16]. P3HT also has greater hole mobility as compared to poly(phenylene vinylene) (PPV), another material that widely used for solar cells and LED [17]. P3HT is based on thiophene rings (four carbon atoms and one sulfur) and highly soluble in common organic solvents. Studies showed that regio-regular P3HT (rr-P3HT) has a higher degree of polymer chain ordering and electrical conductivity as compared with regio-random P3HT (rra-P3HT) [18, 19]. The CdS and CdSe were chosen because they possess lower HOMO (Highest Occupied Molecular Orbital) and LUMO (Lowest Unoccupied Molecular Orbital) than that of poly(3-hexylthiophene) [20].

The computation of Johnson noise engaged low-frequency electrical noise measurement (LFeNM), and the analysis was done at 25 ºC. LFeNM is a method generally employed to verify the reliability of any electronic devices. The relationship of Johnson noise to *J-V* and optical characteristics will be discussed as well.

2. Experimental Setup

This research involved four types of polymer nanocomposite thin film. The films were composed of II-VI semiconducting materials, CdX (CdS or CdSe), and conjugated polymer Poly(3-hexylthiophene) (P3HT).

2.1. Materials and Sample Preparation

Modified Langmuir-Blodgett method has been adopted to fabricate two types of thin films. The first type is the *in-situ* method, made by nucleating cadmium sulfide (CdS) and cadmium selenide (CdSe) nanoparticles (NP) directly in the P3HT matrice (NP-CdS:P3HT and NP-CdSe:P3HT). The second type of thin film was made by adding CdS and CdSe quantum dots (QDs) into P3HT (QD-CdS:P3HT and QD-CdSe:P3HT). Before these QDs were incorporated into P3HT, they were treated by refluxing them for 15 minutes in hexanoic acid at 115 ºC to remove trioctylphosphine oxide (TOPO) that capped the QDs. All materials were purchased from Sigma Aldrich. The solvent used to dissolve P3HT and stearic acid is chloroform (Mallinckrodt). The substrates used were ITO coated glass (5 Ω/sq) from Nanocs. ITO coated glass also acts as an electrode for electrical characterization. All materials used were as-received unless stated otherwise.

The first type of nanocomposite, NP-CdX:P3HT thin films were prepared by dispensing 0.4 ml P3HT solution (0.2 g/l) and 0.1 stearic acid solution (0.14 g/l) on water subphase that contained 0.5 mM cadmium (II) chloride ($CdCl_2$). Stearic acid possesses amphiphilic properties; hence it acts as a stabilizer to P3HT because P3HT is a non-amphiphilic material, thus not suitable for LB deposition. The stearic acid mixture will lessen the domination of the hydrophobic part of P3HT, and reduce the likelihood of supermonolayer structures formation. It also functioned as a capping agent of a divalent cation, cadmium ion, into its headgroup structure (COOH).

The films were deposited when the Langmuir layer surface pressure is continuously at 27 mN/m at an angle of 45º. The modified LB technique allows us to construct large-area polymeric thin films in a controlled order. After every deposition, each transferred layer was put in the oven for 15 s to dry the film and induce cross-linking in the polymer. The temperature of the oven was 50 ºC. Thirteen layers of film (~20 nm) were deposited on the ITO coated glass substrate. After the last film was transferred and

dried, the film was put in the chamber of hydrogen sulfide (H_2S), or hydrogen selenide (H_2Se) for 6 hours. This process was done to nucleate nanoparticles - CdS or CdSe, within polymer layers. Finally, the exposed films were annealed for 10 minutes in an oven at 120 ºC.

The second type of nanocomposite, QD-CdX-P3HT thin films, the QDs were added at 25 wt% or 50 wt% of P3HT solution (0.2 mg/ml). The solution was then carefully injected onto the surface of water subphase. Solid-phase optimized for the film 17 mN/m. Similar to the *in-situ* method, thirteen layers of film were deposited and dried. However, immediately after the last transfer, the film was straight put into an oven for annealing process. All samples were prepared in the cleanroom ISO 6 environment. Samples that will be tested for electrical characteristics were deposited with 150 nm of aluminum layer that serves as an electrode.

2.2. Measurement

Johnson noise measurement utilized current-to-voltage low-noise/low input impedance preamplifier (Signal Discovery), 35670A Dynamic Signal Analyser (Agilent). An interference shield box was used to encase the specimen in order to protect them from external wave or oscillation. The gain of the measurement has been set at 10^{-8} low noise (LN). The low-level noise signal measured from the sample was then fed to an amplifier, which converts the current power (A^2) to voltage power (V^2). The amplifier output would later become the DSA input and processed using LabView. The noise current power spectral density, *SI(f),* was evaluated within the 10 Hz to 1000 Hz frequency range. The noise reading was repeated 200 times.

Keithley 2636A source measure unit (SMU) was used to measure current density-voltage (*J-V*) characteristics. The SMU was attached to a four-point probe and a solar simulator with a light density of 100 mW/cm^2 that complies with the AM 1.5G spectrum. The measurement was carried out by continuously sweeping the feed voltage from -1.5 to +1.5 V.

The spectral absorption measurements were performed using JASCO V570 UV-Vis (Ultraviolet-visible) spectrophotometer operated at a resolution of 1 nm. Perkin Elmer LS 55 Fluorescence Spectroscopy was used to study the PL characteristic of the samples. The samples were excited at 430 nm, and the excitation slit was set at 10 nm, while the laser emission slit was 15 nm. Scan speed was 500 mm/min. The parameters were determined by optimizing the value of a reference sample.

3. Results and Discussion

Figure 1 (a) and (b) show the Johnson noise of NP-CdS:P3HT and NP-CdSe:P3HT, respectively. The unexposed film has the greatest resistance expressed as the highest noise level compared to the films exposed to H_2S or H_2Se gas environment.

The unexposed films are highly resistive due to the presence of stearic acid in the P3HT matrix. Stearic acid hinders efficient electrons movement. However, exposure of the thin film to H_2S or H_2Se gas, causing nanoparticles to nucleate in between P3HT layers. These nanoparticles facilitate the flow of electron, thus lowering the spectral density of noise. The high noise fluctuation for NP-CdS:P3HT thin film indicates that there could be an aggregation of the charges resulting in more substantial resistance at high frequency [21].

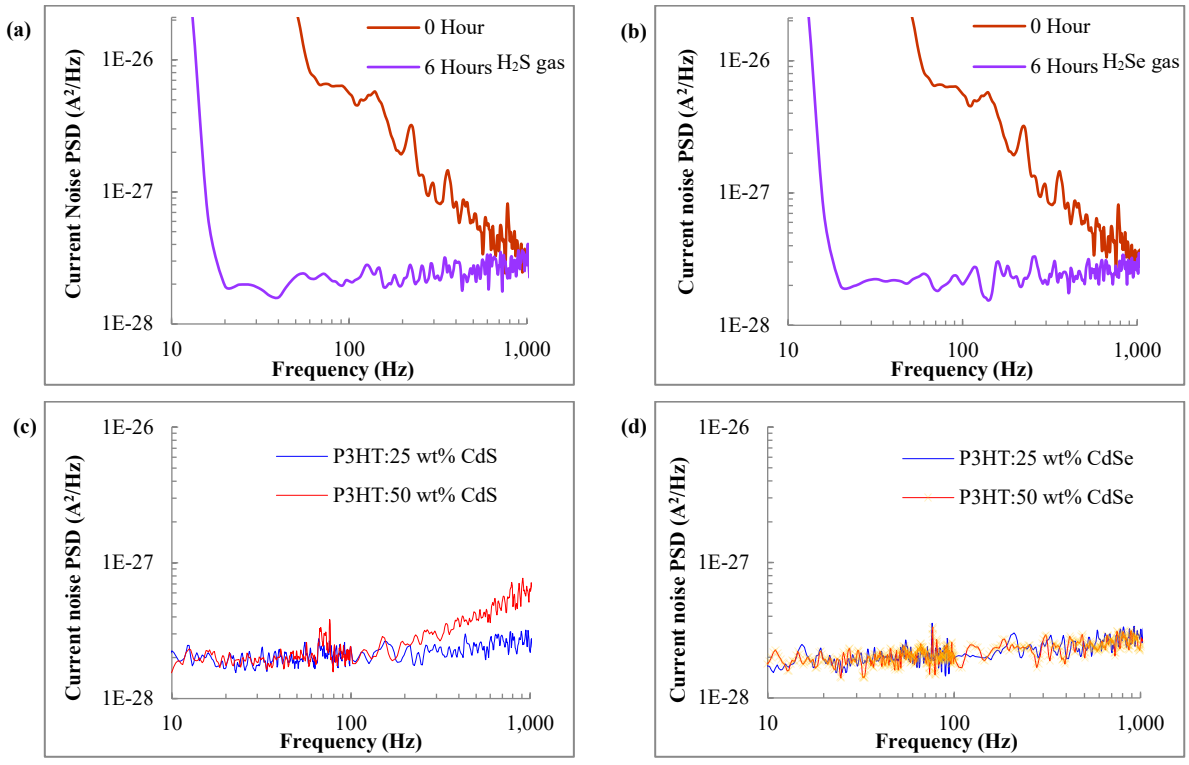


Figure 1: Johnson noise current spectral density of (a) NP-CdS:P3HT (b) NP-CdSe:P3HT (c) QD-CdS:P3HT (d) QD-CdSe:P3HT

On the other hand, as shown in Figure 1 (c) and (d), QD-CdX:P3HT has lower Johnson noise as compared with NP-CdX:P3HT thin films. Among the benefits of thin films formed by directly combining quantum dots with P3HT is that both elements are an exceptional conductor. Therefore the noise for this device is substantially low and stable irrespective of the amount or types of quantum dots. Nonetheless, QD-CdX:P3HT thin films display a growing increase in noise spectra at a frequency greater than 100 Hz, suggests that the thin film is prone to high-frequency degradation.

The noise results were supported by current density-voltage (*J-V*) measurement. The electrical properties studied were open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), fill factor (FF), and maximum power output (P_{max}). The fourth quadrant of *J-V* graphs for NP-CdX:P3HT thin films under illumination were plotted in Figure 2. Excitons are created when incident photons are absorbed in thin films. Then, the excitons dispersed inside the polymer matrix until they reached the interface between polymer/nanoparticles and disassociated into electrons and holes [22]. Consequently, the electric field formed at the boundary layer, pushing electrons and holes to the opposite sides. The nanoparticle's HOMO and LUMO energy is relatively lower than P3HT. The electrons will then be transported to the cathode while the holes are transferred to the anode, producing photocurrent [23].

The thin films exposed to the gas environment demonstrated greater J_{sc} as compared to film not treated with gas. When no current is supplied or $J = 0$ A, V_{oc} can be determined. Figure 2 demonstrated that the value of V_{oc} for the exposed films was twice as large as compared with unexposed thin films. NP-CdSe:P3HT gives the highest V_{oc} , 0.72 V, while the other films are in the range of 0.2 – 0.6 V. Even the principle of V_{oc} is not yell well known, Brabec et al. suggest that it depends heavily on the ionization potential of the donor and the electron affinity of the acceptor; hence the amount energy offset [24]. Meanwhile, Qi et al. show V_{oc} is determined by the composition of the elementals' mixture even it is not a straightforward relationship [25]. Thus it justifies the low V_{oc} value for thin films before gas treatment due to underdeveloped nanoparticles nucleation in the film matrices.

Short-circuit current density, J_{sc}, occurs when the current flow through the external circuit without connecting to a voltage source, $V = 0$ V. Low J_{sc} value for entire samples imply limited photon absorption by the films. Transportation of the charge is associated with Fill Factor or FF. High charge mobility is desired to prevent significant losses due to recombination [26, 27]. However, calculation shows FF values are small for all samples indicating the considerable recombination level occurs in thin films. The results are as corroborated by low J_{sc} value. The "squareness" of the FF is contributed by maximum power output, P_{max}, that is when the current and the voltage of the device are at maxima, or:

$$P_{max} = J_{MPP} \times V_{MPP} \quad (5)$$

where V_{MPP} and J_{MPP} are the values of voltage and current density at the maximum power point (MPP). Electronic devices must be operated at the maximum power point to generate maximum energy efficiency and minimize losses. In order to achieve maximum energy performance and, at the same time, reduces losses, the electronic appliances are desired to be operated at the highest power point [28].

The exposure of H_2S or H_2Se gas nucleated approximately 9% of nanoparticles. Hence the improvements in P_{max} for the exposed film were insignificant. Thin films mixed with 50 wt% quantum dots, however, posses even smaller P_{max} value. This is contrary to the study by Lek et al. that reports a rise in the weight ratio of nanoparticles would improve electrical performance [29]. Nonetheless, the improvement can only be observed until a certain weight percentage is reached. Whilst the study shows a low amount of quantum dots or nanoparticles causes inefficient pathways for electrons transportation [30], too high nanomaterials loading causing limited mobility of the holes, leading to lower device's performance. The summary of the *J-V* analysis is as tabulated in Table 1. The overall results demonstrated that NP-CdSe:P3HT nanocomposite the most reliable performance in both Johnson noise and J-V evaluation of all thin films.

Table 1: The electrical properties of thin films from *J-V* measurement

	V_{oc} (V)	J_{sc} (mA/cm^2)	FF	P_{max} (mW)
NP-CdS:P3HT	0.2	1.2×10^{-3}	0.18	1.2×10^{-4}
NP-CdSe:P3HT	0.72	1.2×10^{-3}	0.15	1.3×10^{-4}
QD-CdS:P3HT	0.32	1.1×10^{-4}	0.14	4.8×10^{-6}
QD-CdSe:P3HT	0.2	7.5×10^{-5}	0.21	3.2×10^{-6}

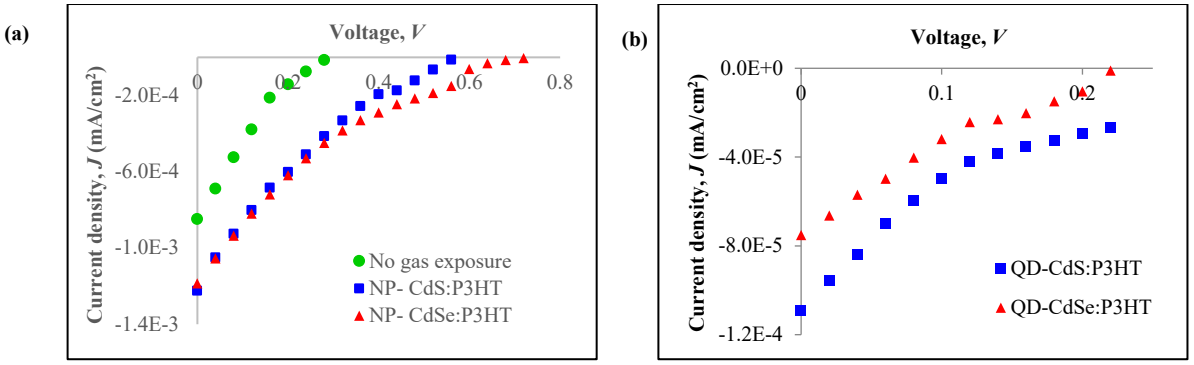


Figure 2: *J-V* graphs for thin films (a) unexposed sample, sample exposed to H_2S gas (NP- CdS:P3HT) and H_2Se gas (NP-CdSe:P3HT), and (b) QD-CdS:P3HT and QD-CdSe:P3HT

Figure 3 (a) displays the absorption spectra of pristine P3HT thin film, thin films before gas exposure, and after exposure to H_2S gas and H_2Se gas. Conjugation system in P3HT causes two peaks at 522 and 556 nm and a shoulder at 605 nm to appear in the UV-Vis graph. A similar result has been obtained by the Liao research group [31]. The peaks attribute to π-π* transition or the order of intermolecular chains of the polymer, while the shoulder is due to crystallization of polymer [32, 33]. When P3HT were mixed with stearic acid, the absorption spectra are lower than that of pristine P3HT, and its peak shifted towards the red spectrum at 529 and 560 nm, and the shoulder becoming less distinguishable. The shift indicates the presence of Cd^{2+} ion capped in the stearic acid matrices, as well as reduced P3HT π-conjugation length.

Before the thin films were exposed to the gas, the observable peak is at 560 nm. Nonetheless, exposure to H_2S gas environment for 6 hours caused the peak to shift to 556.5 nm. The blue shift of the peaks inkling the quantum confinement effect due to CdS nanoparticles formation in the thin film [34, 35]. The shoulder at 605 nm is more perceptible compared with an unexposed thin film. This shoulder is usually due to an increased crystallization of intra-chain interactions. The exposure to H_2S gas causing the absorption intensity to become lesser. This is due to the reaction of cadmium stearate with H_2S gas:

$$(CH_3(CH_{12})_{16}COO)_2Cd + H_2S \rightarrow 2H(CH_3(CH_{12})_{16}COO) + CdS \quad (6)$$

Similar to H_2S gas exposure, the absorption intensity of thin films exposed to H_2Se gas ambiance decreased compared with the unexposed thin film. One possible reason that contributes to such a profile is the shape and amount of CdSe nanoparticles created. The aspect ratio of CdSe nanoparticles is high, in which it resembles rod-like, and some of the nanoparticles agglomerate, creating a larger cluster.

Photoluminescence (PL) is a process of spontaneous light emission from the material under optical excitation. Ideally, the PL spectrum mirroring the image of the absorption spectrum, hence it clearly can be observed from Figure 3(b) the PL spectra resemble the UV-Vis absorption spectra. PL measurement is an indicator of excitons dissociation and its efficiency to transfer charge in the composite materials. The intensity of PL decreased when exposed to H_2S and H_2Se gas. The gas exposure caused the peak shifted towards the blue region, from 720 nm to 718 nm and 714 nm, respectively. The deviation suggested the reduction in P3HT aggregation, causing the PL to quench.

The occurrence of PL depends on the distance from the excitons formed to the donor/acceptor interface [36]. If the distance is lesser than the excitons' diffusion length, charge transfer to the nanoparticle is anticipated, hence the PL is quenched [37]. The more significant range caused recombination to take place. From literature, exciton diffusion length for P3HT is around 8 nm [38]. However, FESEM images in Figure 3 (c) – (e) showed that the distance between nanoparticles was considerably large, ranging from 100 to 150 nm. The number of nanoparticles impregnated in the thin film was small, causing the exciton that is further from the donor/acceptor interface to recombine before reaching the interface.

The UV-Vis spectra when the CdS / CdSe quantum dots were physically mixed into P3HT polymer as shown in Figure 4 (a) and (b). The absorption of pristine P3HT is comparable to the thin film with 25 wt% CdS and CdSe QDs. However, for QD-CdS:P3HT, the absorption is lower when the amount of QDs is doubled. This may be due to a smaller amount of P3HT in the composite [39]. The peaks and shoulders remain at the same wavelength as pristine P3HT. The absence of the CdS QDs absorption band is likely due to the overlapping with the P3HT absorption. While at 50 wt% of CdSe, the absorption is higher, and the shoulder is less distinguishable. The peaks are poorly shifted to red, 524 and 557.5 nm indicating the polymer chains are less crystallized due to QDs added. Both QD-CdS:P3HT and QD-CdSe:P3HT show no significant in amplitude variation with the increment of QDs wt%, comparable to the results reported for P3HT:CdSe TBPO-capped at different weight percentage [40].

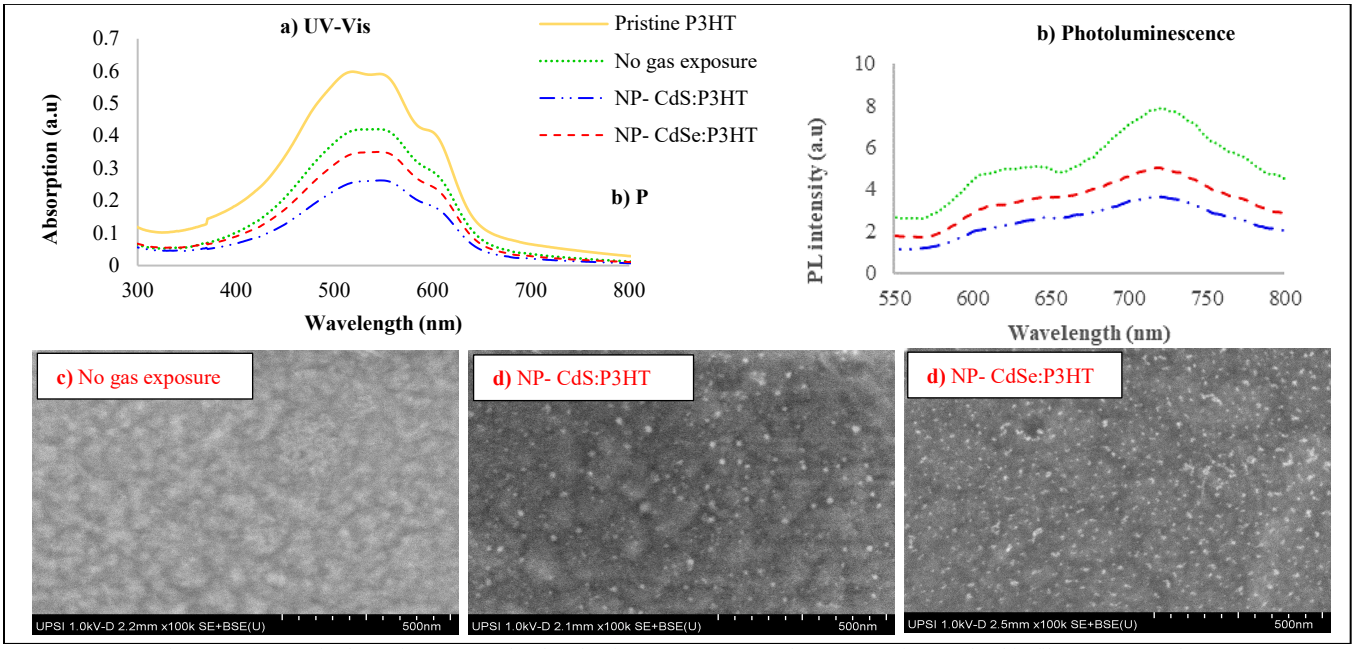


Figure 3: (a) UV-Vis absorption spectra, (b) photoluminescence spectra and (c) FESEM images for thin films not exposed to any gas, (d) exposed to H_2S gas (NP- CdS:P3HT), and (e) H_2Se gas (NP-CdSe:P3HT)

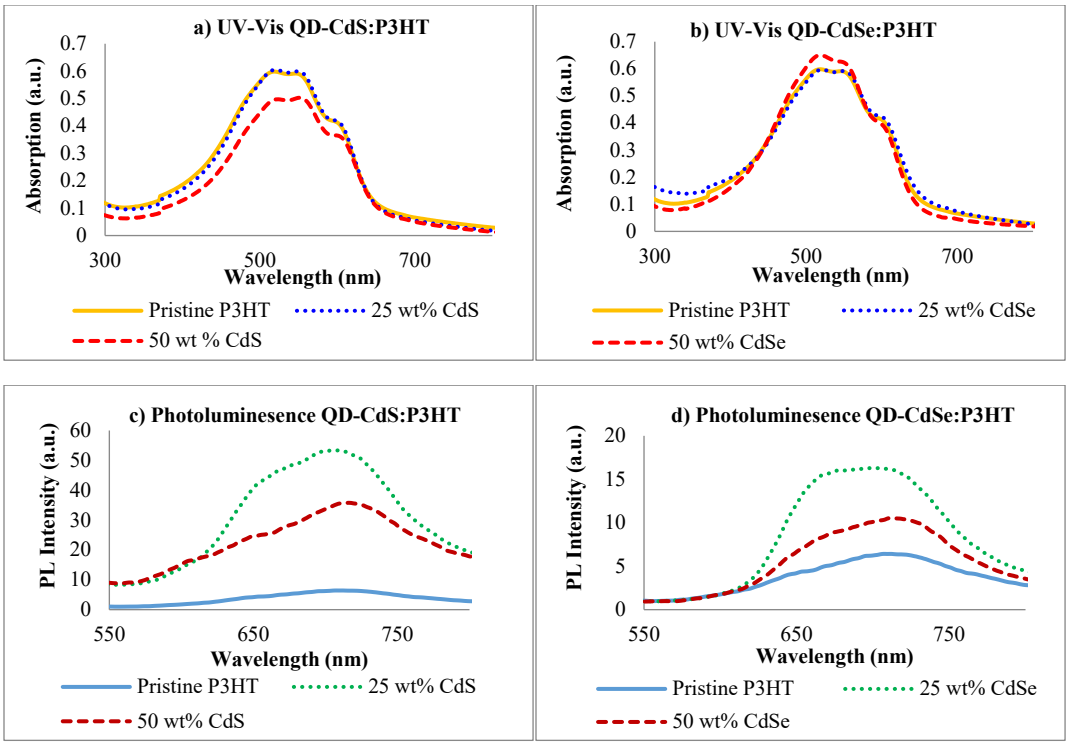


Figure 4: UV-Vis absorption spectra and photoluminescence of QD-CdS:P3HT and QD-CdSe:P3HT

On the other hand, photoluminescence spectra for QD-CdS:P3HT and QD-CdSe:P3HT are shown in Figure 4 (c) and (d), respectively. Both types of thin-film displayed similar characteristics, in which a higher amount of QDs (50 wt%) mixed into P3HT cause the film to quench stronger as compared to films with 25 wt% QDs. The quenching indicates a reduction in the domain size of P3HT blended with nanoparticles [41]. The decrease is owing to a shorter distance for the excitons to travel to the donor/acceptor interface when the weight percentage of nanoparticles in the polymer is doubled. The PL quenching also possibly due to the π-π interaction of P3HT with the nanoparticles [42], creating a new decaying pathway for excitons [43]. Higher excitons dissociation can be justified for better device performance.

Nonetheless, the PL intensity of P3HT:CdS QDs is higher and has a narrower peak than that of P3HT:CdSe QDs. The P3HT:25 wt% CdS has a maximum at 704 nm, while for P3HT:50 wt% CdS, the peak shifted to longer wavelength, 713.5 nm. The shift is also known as Stokes-shift. The redshift implies the increment in polymer aggregation, which can be achieved by the incorporation of QDs into the polymer matrix, as have also been observed by others [44].

4. Conclusion

The analysis has been done on the nanocomposite thin films prepared via two approaches – direct impregnation of nanoparticles in P3HT matrix (*in-situ*), and by physically mixing quantum dots with P3HT (*ex-situ*). Results show that QD-CdX:P3HT thin films possess low Johnson noise due to intrinsic properties of quantum dots and P3HT as an excellent conductor. Consequently, the noise is considerably small and stable for this system regardless of the weight percentage or types of quantum dots. However, the increment of noise can be observed at a frequency above 100 Hz for QD:CdX-P3HT thin films. This signifies the thin film is vulnerable to high-frequency degradation. For nanocomposite produced through the *in-situ* technique, the presence of stearic acid the P3HT matrix curb the free flow of electrons movement.

It can be concluded NP-CdSe:P3HT nanocomposite displayed more reliable performance in both Johnson noise and *J-V* measurement. Absorbance spectra show higher polymer crystallization NP-CdX:P3HT. The quantum confinement effect due to nanoparticles formation is translated by the shift of peaks and depreciation of absorption intensity. The PL intensity decreased when exposed to the gas. Both the thin films fabricated ex-situ displays similar spectra profile even when the amount of quantum dots was doubled. While the amount of quantum dots did not affect the absorbance significantly, the PL intensity decreased with increment of quantum dots' weight percentage. Overall, NP-CdX:P3HT showed better performance compared to QD:CdX-P3HT nanocomposite. Such findings have provided a bright prospective to study the nucleation of nanoparticles through gas exposure technique for optoelectronics applications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors are grateful to Universiti Putra Malaysia for the financial support through IPM Putra research grant 9674000.

References

[1] R. E. Gorga, R. E. Cohen, "Toughness Enhancements in Poly(Methyl Methacrylate) by Addition of Oriented Multiwall Carbon Nanotube"

J.Polym.Sci. Polym. Phys., **42**(14), 2690–2702, 2004. https://doi.org/10.1002/polb.20126
[2] H. Darabi, Radio Frequency Integrated Circuits and Systems, Cambridge University Press, 2015.
[3] J. B. Johnson, "Thermal Agitation Of Electricity in Conductors" Phy. Rev., 32, 97-109, 1928. https://doi.org/10.1103/PhysRev.32.97
[4] H. Nyquist, "Thermal Agitation of Eelectric Charge in Conductors" Phy. Rev., 32, 110-113, 1928. https://doi.org/10.1103/PhysRev.32.110
[5] X. Zhang, D. Chen, "An Integrated Circuit Solution to Johnson Noise Thermometry Using Low-Cost and Fast CMOS Technology" IEEE Sens. J., **19**(9), 3240-3251, 2019. https://doi.org/10.1109/JSEN.2019.2893241
[6] G. Vasilescu, Electronic Noise and Interfering Signals: Principles and Application, Springer-Verlag Berlin Heidelberg, 2005.
[7] L. K. J. Vandamme, "Noise as a Diagnostic Tool for Quality and Reliability of Electronic Devices," IEEE Trans. on Electron Dev., **41**(11), 2176 – 2187, 1997. https://doi.org/10.1109/16.333839
[8] Z. Chobola, L. Hasse, Z. Růžička, "Measurement of Low Frequency Noise of Monocrystalline Silicon Solar Cells," in XVII IMEKO World Congress, Croatia, 2003.
[9] P. R. F. Rocha, H. L. Gomes, L. K. J. Vandamme, D. M. de Leeuw, S. C. J. Meskers, P. van de Weijer, "Low-frequency Noise as a Diagnostic Tool for OLED reliability," in 22nd International Conf. on Noise and Fluctuations, 2013.
[10] A. Carbone, B. K. Kotowska, D. D. Kotowski, "Space-Charge- Limited Current Fluctuations in Organic Semiconductors" Phy. Rev. Lett., **95**, 236601-236605, 2005. https://doi.org/10.1103/PhysRevLett.95.236601
[11] L. Ke, X. Y. Zhao, R. S. Kumar, S. J. Chua, "Low Frequency Optical Noise from Organic Light Emitting Diode" Solid-State Electron., **52**(1), 7-10, 2008. https://doi.org/10.1016/j.sse.2007.07.004
[12] J. J. Luo, I. M. Daniel, "Characterization and Modelling of Mechanical Behaviour of Polymer/Clay Nanocomposites" Composites Science and Technology, **63**(11) 1607–1616, 2003. https://doi.org/10.1016/S0266-3538(03)00060-5
[13] C. L. Wu, Q. Zhang, M. Z. Rong, K. Friedrich, "Tensile Performance Improvement of Low Nanoparticles Filled-Polypropylene Nanocomposites" Compos. Sci. Technol., **62**(10-11), 1327-1340, 2003. https://doi.org/10.1016/S0266-3538(02)00079-9
[14] M. Haridas, S. Srivastava, J. K. Basu "Tunable Variation of Optical Properties of Polymer Capped Gold Nanoparticles" Eur. Phys. J. D, **49**, 93–100, 2008. https://doi.org/10.1140/epjd/e2008-00135-x
[15] Hung-Chou Liao, San-Yuan Chen, and Dean-Mo Liu. "In-Situ Growing CdS Single-Crystal Nanorods via P3HT Polymer as a Soft Template for Enhancing Photovoltaic Performance" Macromol. **42**(17), 6558–6563, 2009. https://doi.org/10.1021/ma900924y
[16] H. Hoppe, "Nanomorphology – efficiency Relationship in Organic Bulk Heterojunction Plastic Solar Cells," Ph.D Thesis, Johannes Kepler University Linz, 2004.
[17] J. Hou, X. Guo, Active Layer Materials for Organic Solar Cells. In W. C. H. Choy (Ed.), Organic solar cells: Materials and Device Physics, Springer-Verlag, 2012.
[18] H. Tang, F. Zhao, F. Zhang, "Investigation on an Optically-Controlled Phase-Shifter Based Organic Semiconductor Poly-(3-hexylthiophene) (P3HT)" J. Phys. Conf. Ser., **276**(1), 012092, 2011. https://doi.org/10.1088/1742-6596/276/1/012092
[19] E. O. Sako, H. Kondoh, I. Nakai, A. Nambu, T. Nakamura, T. Ohta, "Reactive Adsorption of Thiophene on Au(111) from Solution" Chem. Phys. Lett., **413**(4-6), 267-71, 2005. https://doi.org/10.1016/j.cplett.2005.07.086
[20] Y. Zhou, M. Eck, M. Krüger, M. Organic-Inorganic Hybrid Solar Cells: State of The Art, Challenges and Perspectives. In L. A. Kosyachenko (Ed.), Solar Cells - New Aspects and Solutions. IntechOpen, 2011.
[21] F.H. Naning, "Surface Pressure-Area Isotherm, Optical And Electrical Properties Of P3HT/Nanoparticles Thin Films By Modified Langmuir-Blodgett Technique" Ph.D Thesis, Universiti Pendidikan Sultan Idris, 2017.
[22] B. Kippelen, J. L. Bredas, "Organic Photovoltaics" Energy Environ. Sci., **2**, 251–261, 2009. https://doi.org/10.1039/B812502N
[23] N. S. Sariciftci, L. Smilowitz, A. J. Heeger, F. Wudl, "Photoinduced Electron Transfer From a Conducting Polymer to Buckminsterfullerene," Science, **258**(5087), 1474-1476, 1992. https://doi.org.10.1126/science.258.5087.1474
[24] C. J. Brabec, G. Zerza, G. Cerullo, S. De Silvestri, S. Luzzati, J. C. Hummelen, S. Sariciftci, "Tracing Photoinduced Electron Transfer Process in Conjugated Polymer/Fullerene Bulk Heterojunctions in Real Time" Chem. Phy. Lett., **340**(3-4), 232-236, 2001. https://doi.org/10.1016/S0009-2614(01)00431-6
[25] B. Qi, Q. Zhou, J. Wang, "Exploring the Open-Circuit Voltage of Organic Solar Cells Under Low Temperature" Sci. Rep., **5**, 11363, 2015. https://doi.org/10.1038/srep11363
[26] J. A. Bartelt, D. Lam, T. M. Burke, S. M. Sweetnam M. D. McGehee, "Charge-Carrier Mobility Requirements for Bulk Heterojunction Solar Cells with High Fill Factor and External Quantum Efficiency >90%" Adv. Energy Mater., **5**, 1500577, 2015. https://doi.org/10.1002/aenm.201500577
[27] I. Riedel, V. Dyakonov, "Influence Of Electronic Transport Properties of Polymer-Fullerene Blends on the Performance of Bulk Heterojunction Photovoltaic Devices" Phys. Status Solidi A, **201**, 1332–1341, 2004. https://doi.org/10.1002/pssa.200404333
[28] D. Bartesaghi, I. D. Pérez, J. Kniepert, S. Roland, M. Turbiez, D.N. L. J. A. Koster, "Competition Between Recombination and Extraction of Free Charges Determines the Fill Factor of Organic Solar Cells" Nat. Commun., **6**, 7083, 2015. https://doi.org/10.1038/ncomms8083
[29] J. Y. Lek, Y. M. Lam, J. Niziol, M. Marzec, "Understanding Polycarbazole-Based Polymer:CdSe Hybrid Solar Cells," Nanotechnology, **23**(31), 315401, 2012. https://doi.org/10.1088/0957-4484/23/31/315401
[30] Y. Zhou, M. Eck, C. Men, F. Rauscher, P. Niyamakom, S. Yilmaz, …M. Krüger, "Efficient Polymer Nanocrystal Hybrid Solar Cells by Improved Nanocrystal Composition," Sol. Mat. Sol. Cells, **95**(12), 3227-3232, 2011. https://doi.org/10.1016/j.solmat.2011.07.015
[31] H. C. Liao, N. Chantarat, S. Y. Chen, C. H. Peng, "Annealing Effect on Photovoltaic Performance of Hybrid P3HT/In-Situ Grown CdS Nanocrystal Solar Cells" J. Electrochem. Soc., **158**(7), 5251-5455, 2011. https://doi.org/10.1149/1.3585668
[32] U. Farva, M. A. Khan, C. Park, Effect of Post Annealing on the Performance of CdSe/P3HT Bulk Hetero-Junction Solar Cells. In Photovoltaic Specialist Conference, San Diego, USA, 2008.
[33] P. J. Brown, D. S. Thomas, A. Köhler, J. S. Kim, C. M. Ramsdale, H. Sirringhaus, R. H. Friend, "Effect of Interchain Interactions on the Absorption and Emission of Poly(3-hexylthiophene)" Phys. Rev. B, **67**(6), 064203, 2003. https://doi.org/10.1103/PhysRevB. 67.064203
[34] R. R. Prabhu, M. A. Khadar, "Characterization of Chemically Synthesized CdS Nanoparticles" Pramana J. Phys., **65**, 801-807, 2005. https://doi.org/10.1007/BF02704078
[35] P. Sonar, K. P. Sreenivasan, T. Madddanimath, K. Vijayamohanan, "Comparative Behavior of CdS and CdSe Quantum Dots in Poly(3-hexylthiophene) Based Nanocomposites" Mater. Res. Bull., **41**(1), 198-208, 2006. https://doi.org/10.1016/j.materresbull.2005.07.032
[36] B. Freisinger, "Investigation of P3HT/PCBM Particle-Based Solar Cells," Ph.D Thesis, Johannes Gutenberg-Universität Mainz, 2013.
[37] A. J. Ferguson, N. Kopidakis, S. E. Shaheen, G. Rumbles, "Quenching of Excitons by Holes in Poly(3-hexylthiophene) Films" J. Phys. Chem., **112**(26), 9865- 9871, 2008. https://doi.org/10.1021/jp7113412
[38] B. Annalisa, D. L. Tiziana, B. Carmela, V. Fulvia, H. Saif, M. Carla, "Exciton Dynamics in Hybrid Polymer/QD Blends", Energy Procedia, **44**, 167-175, 2014. https://doi.org/10.1016/j.egypro.2013.12.024.
[39] P. E. Shaw, A. Rusecka, I. D. W. Samuel, "Exciton Diffusion Measurements in Poly(3-hexylthiophene)" Adv. Mater. **20**(18), 3516-3520, 2008. https://doi.org/10.1002/adma.200800982
[40] D. Loubiri, Z. Ben Hamed, S. Ilahi, M. A. Sanhoury, F. Kouki, N. Yacoubi, "Effect of TBPO-Capped CdSe Nanoparticles Concentration on Sub-Bandgap Absorption in Poly(3-hexylthiophene) Thin Films Studied by Photothermal Deflection Spectroscopy" Synth. Met,, **206**, 1-7, 2015. https://doi.org/10.1016/j.synthmet.2015.04.018.
[41] A. J. MacLachlan, T. Rath, U. B. Cappel, S. A. Dowland, H. Amenitsch, A. C. Knall, … S. A. Haque, "Polymer/Nanocrystal Hybrid Solar Cells: Influence of Molecular Precursor Design on Film Nanomorphology, Charge Generation and Device Performance" Adv. Funct. Mater., **25**, 409–420, 2015. https://doi.org/10.1002/adfm.201403108
[42] H. C. Liao, S. Y. Chen, D. M. Liu, "In-Situ Growing CdS Single-Crystal Nanorods via P3HT Polymer as a Soft Template for Enhancing Photovoltaic Performance" Macromolecules, **42**(17), 6558-6563, 2009. https://doi.org/10.1021/ma900924y
[43] M. T. Khan, R. Bhargav, A. Kaur, S. K. Dhawan, S. Chand, "Effect of Cadmium Sulphide Quantum Dot Processing and Post Thermal Annealing on P3HT/PCBM Photovoltaic Device" Thin Solid Films, **519**(3), 1007-1011, 2010. https://doi.org/10.1016/j.tsf. 2010.08.032.
[44] S. N. Sharma, T. Vats, N. Dhenadhayalan, P. Ramamurthy, A. K. Narula, "Ligand-Dependent Transient Absorption Studies of Hybrid Polymer:CdSe Quantum Dot Composites" Sol. Energy Mater. Sol., **100**, 6–15, 2012. https://doi.org/10.1016/j.solmat.2011.10.020

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 757-764 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Transient Response & Electromagnetic Behaviour of Flexible Bow-Tie Shaped Chip-less RFID Tag for General IoT Applications

Muhammad Usman Ali Khan*, Raad Raad, Javad Foroughi

School of Electrical Computer and Telecommunication Engineering, University of Wollongong, Wollongong NSW, 2522, Australia

ARTICLE INFO

Article history:
Received: 11 June, 2020
Accepted: 23 July, 2020
Online: 12 October, 2020

Keywords:
Chipless RFID
Bow-tie
Polymer Substrate
FSS
RCS
PET
SEM
Coupling coefficients
IoT Applications

ABSTRACT

This paper is an extension of Novel Flexible Chip-less Bow-Tie RFID tag in which we presented the design, testing and fabrication of the tag and compared the results with Octagonal Chipless RFID tag. The chipless RFID tag was designed by using simulation software CST microwave studio and fabricated by using laser etching technique on a flexible polymer substrate Polyethylene Terephthalate (PET). The tag operates at frequency ranging from 8 to 18 GHz uses the Frequency Selective Surface (FSS) approach. A series of experiments are performed to measure the Radar Cross Section (RCS) in an anechoic chamber. The tag design is composed of six concentric Bow-Tie shaped loop resonators with one unitary element. In this paper, we demonstrated the Singularity Expansion Method (SEM) based circuit modelling and the transient behaviour of the RFID tag is performed. The coupling coefficients and the induced currents over the surface of Bow tie shaped rings are evaluated. The maximum read range is evaluated and the Bow-Tie RFID tag is proved to be more accurate and efficient with the variation of distance up to 1.8m at 0dBm which is extendable to 2.14m for higher input power. This range is maximum to our knowledge for such a high-frequency range of 8-18GHz. The 4-bits Bow-Tie Chipless RFID tag design is compact and can be deployed commercially for general IoT applications.

1. Introduction

This paper is an extension of work originally presented in the 19th International Symposium on Communications and Information Technologies (ISCIT) 2019 [1].

The Internet of Things (IoT) relies on smart environments comprises of wireless sensor networks [2] and personal area networks [3], [4] in which one interacts with IoT infrastructure must be uniquely identified. However, one big challenge is a need to commence seamless integration of IoT solutions with surroundings and with people [5]. Flexible materials for various IoT applications require a high level of integrity of components and mechanical robustness with repeated rolling and bending capabilities. In particular, flexible smart fitness watches, RFID tags and wearable sensors are good examples of applications that use flexible material. Besides, Elasticity and stretch-ability of materials are the key properties required by electronic devices that require large and reversible deformation. These bendable devices will also need to be versatile and may require the ability to store energy, operate with low power, and integrated with other devices and Internet of Things (IoT) applications [6]-[8].

The growth of wearable technology and the success of wearable devices has been hampered by limits such as the lack of materials and manufacturing techniques for seamless integration with electronic parts [9], the robustness in washing, drying and moulding [10], [11]. To address these challenges flexible polymer-based chipless RFID tag can be a good candidate.

In recent years, the research on wearable wireless communication and flexible devices has increased rapidly due to its usability and applications in personal communication systems [12]. For examples; Flexible antennas and Radio Frequency Identification (RFID) tags are used to keep track of several daily-routine activities, such as heartbeat and blood pressure measurement, the distance walked, the steps are taken and the calories burnt [13].

RFID Technology is growing rapidly and has significant importance in various medical, industrial and commercial applications [13], [14] such as UHF RFID tags, Chipless RFID

* Corresponding Author: Muhammad Usman Ali Khan, Email: muak803@uowmail.edu.au

tags and Near Field Magnetic Coupled tags are being in used in various commercial applications and medical applications [15]. In the traditional RFID technology, the backscattered field from the tags are used to transmit information by using time-variant loading or scatterer modulation [16]-[19] however, this technique requires the physical movement of the object which needs to be coded by using inductive coupling or the impedance modulation. Besides, the conventional RFID tags are active circuit, carrying a small near field chip including a power source and a memory at the centre of the body. Whereas the chipless RFID technology doesn't require memory or a chip to save codes, instead, it uses the physical aspects of the body such as geometry and the shape to send the data [13]. The primary principle of the chipless technology is to get the resonance from the body and to encode the backscattered signal at the receiver, without additional circuitry or communication protocols [20]. Various techniques have been proposed but generally, the chipless RFID technology can be categorised into two groups. The first one is the time domain-based also called time-domain reflectometry-based designs, Surface Acoustic Wave (SAW) is an example of this. The second group is frequency domain-based RFID tags sometimes known as spectral-based tags, Frequency Selective Surface (FSS) is an example of this technique. The SAW-based chipless RFID tags are primarily worked by piezoelectricity, in which Interdigital Transducer (IDT) on the surface of piezoelectric materials are used to convert the electromagnetic wave into the acoustic wave. The acoustic waves generate the identification code while propagating through the metallic surfaces of the tag separated by a certain distance and stored on the tag [21], [22]. The SAW RFID tags are commercially available but some important issues such as the reduction in the size, increase in the read range and the data capacity needs to address 15. Besides, the FSS based frequency domain chipless RFID technique is quite promising, in which an incident electromagnetic wave is completely or partially reflected from the surface of the resonators and matches with the resonance frequency, based on the nature and the shape of the elements. The chipless RFID tags are known as passive tags as no additional power is required to operate, which make these tags promising for general IoT applications.

Passive chipless RFID tags have gained much importance as a good candidate for IoT applications because of no additional circuitry, small size, increased lifetime and the cost-effectiveness [23]. As compared to conventional RFID tags, passive tags have many significant advantages such as they are extremely cheap and can be printed over a piece of paper simple as barcodes. Moreover, the wearable applications in which washing and heating will affect the energy source, the absence of the battery and the electrical circuits make these tags promising [24].

The general structure of a chipless RFID system is an adaptive version of the Ultra-Wide Band (UWB) frequency radar [25], in which an incident electromagnetic wave from a transmitter strikes at the surface of the tag and reflect toward the reader and decoder, see Figure 1. When an incident wave E^i on the surface of the conductive element, a current is induced at the resonant frequency, 9.2GHz in the case of chipless Bow-Tie RFID tag, see Figure 1.

In this paper, a 4-bit Bow-tie shaped chipless RFID tag is designed and the transient behaviour by the application of ElectroMagnetic (EM) wave is explained. In this design, the polarization feature of non-periodic bow-tie shaped cells is taken into account to get destructive and constructive interferences. The projection of the incidents wave and reflected field and decoded in terms of bits assigned to Radar Cross Sections (RCS) [26]. The RCS is a parameter to describe the level of scattering of a target object, in this case, a tag is the target object.

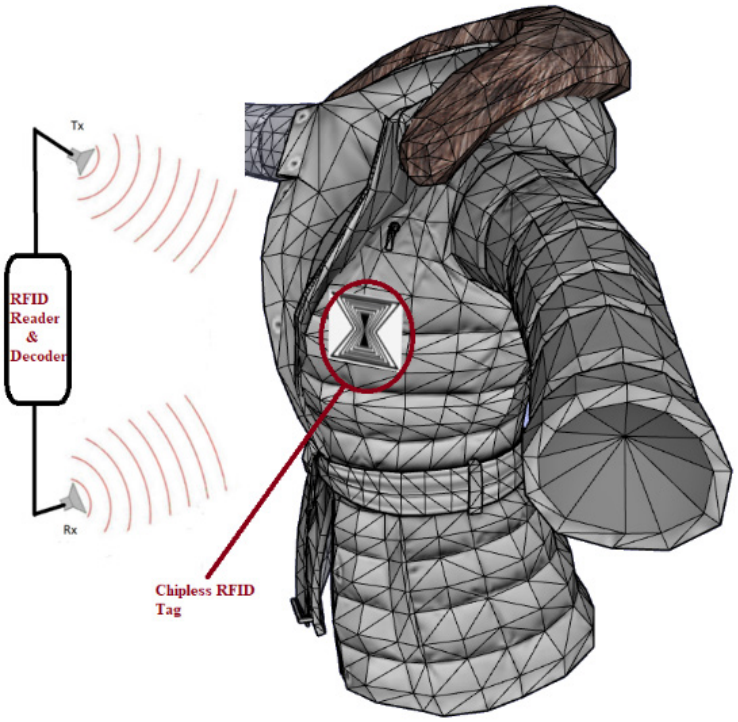


Figure 1: Structural diagram of Bow-Tie chip-less RFID system

This paper employs a simple equivalent model of the tag by using SEM-based circuit modelling. Furthermore, the coupling coefficients and the induced currents over the rings are evaluated and transient response of the Bow-Tie RFID tag is analysed. Finally, the maximum read range is calculated which is 1.8m for 0dBm and extendable to 2.14m for higher incident power, which is maximum to our knowledge for high frequencies of the range 8 to 18GHz.

The rest of the paper is organised as follows: Section II and Section III presents the tag design and its working principle, the material characterisation and optimised designs. Section IV presents the RCS measurements, which further describes the experimental verification of codes, on body RCS measurements, bending analysis and discussion. Finally, the conclusion is presented in Section V.

2. Design and Operation of Bow-Tie Chipless RFID Tag

The Bow-Tie shaped chipless RFID tag proposed in [1] is based on conventional Bow-tie shape which is opted as it has the most intense current flow on the edges of the resonating elements [27], hence cause a more intense backscattered signal. Another reason to opt this design is the fact that it provides more flexibility in designing at higher frequencies unlike octagonal [26], square [28], circular [29] and triangular [30] designs. Moreover, in Bow-tie chipless RFID tag the asymmetrical shape of the tag in which

the distance between the resonators is not constant, it diminished the mutual coupling as compared to the symmetrical surfaces, placed very close to each other [31]. In this paper, a 4-bit Bow-tie chipless RFID tag is proposed and the transient behaviour of the tag is presented. The robust tag is comprised of a periodic pattern of six Bow-tie shaped cells applicable for general IoT and wearable applications.

2.1. Working Principle

The resonant frequency of the tag depends on the inductance and the capacitance between the concentric metallic bow-tie shaped elements [32]. Therefore, when an EM wave strikes at the surface of the tag it induces currents and the corresponding capacitance and the inductances are changed. The extent of the reflected or transmitted wave after falling on the surface, depends upon the placement and the geometry of the rings. Aa results, surface currents (J_{in}) and (J_{out}) are induced on the inner and outer bow-tie shaped rings, respectively, as shown in Figure 2.

In this case, the whole structure including all the surrounding bow-tie rings, interact with each other and cause the creation of the standing waves. The standing waves produced the same polarization with an odd multiple of the difference of π phase by these surface currents which are approximately equal but opposite in phase. Hence, destructive interference results in backscatter wave (Es) and create a specific frequency response. Figure 2. represents the surface current distribution over the elementary cell and all periodic bow-tie shaped rings which tends to be minimum at the inner side of the tag, at the base and ceiling while maximum and almost constant in all vicinities of the Bow-Tie tag. The area of the tag showing minimum concentration correspond to capacitive behaviour and the area where the surface current is intensive corresponds to the inductive behaviour of the tag, see Figure 2.

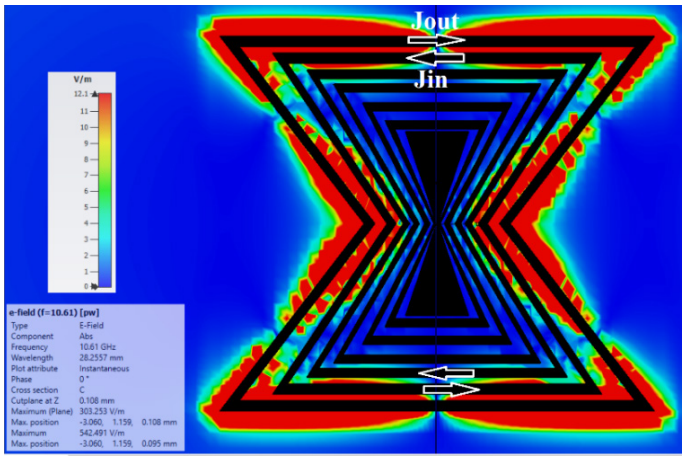


Figure 2: E-Field distribution and the directions of induced surface current (J)

3. Electromagnetic Behaviour of the Bow-Tie RFID Tag

The circuit model of the Bow-Tie chipless RFID tag provides insight into the electromagnetic response of the circuit when an incident EM wave strike at the tag and the induced current re-radiate the scattered fields. The current induced on the tag can be evaluated in different ways. One method is based on Singularity Expansion Method (SEM) in which the current induced on the tag is expanded in terms of the singularity poles of the tag. The concept of SEM has been widely used in circuit theory for a long time. Later on, this technique was used in evaluating EM response and designing of chipless RFID tags [33],[34],[35].

3.1. SEM-Based Equivalent Circuit of Chipless RFID Tag

In this paper, the SEM-based equivalent circuit of the octagonal chipless RFID tag is presented where the solution is evaluated as a collection of poles (Sn), coupling coefficient (Rn) and the entire function (Fe) in the complex frequency domain. The scattering analysis is performed by representing the chipless RFID tag with an equivalent circuit representation, see Fig. For electromagnetic response, the RFID scatterer assumed to be Perfectly Electric Conductor (PEC) in free space. Let's N be a total number of scatterers and an incident electric field E^i strike at the surface of the tag and E^s are scattered electric field with polarization vectors $\hat{a}_i$ and $\hat{a}_s$, respectively. The E^i causes an induced current (J_n) on the surface and the global resonances are demonstrated in the time domain as damped sinusoidal analogous to the Complex Natural Response (CNR) of the scatterer with some weighting residue as coupling coefficients (R_N). The distance between the two scatterers is optimised in such a way that the coupling between one and the alternative scatterer is very small to be negligible. Hence, the coupling between the two scatterers only at a time and ignored the effects of the others on them.

The SEM-based circuit equivalent model of the octagonal chipless RFID tag is presented in Figure 3. where, R_1, R_2…..R_N, L_1, L_2,…..L_N define the resistances and the inductances and C_1, C_2,… C_N is the inductance between two consecutive scatterers for N number of resonators.

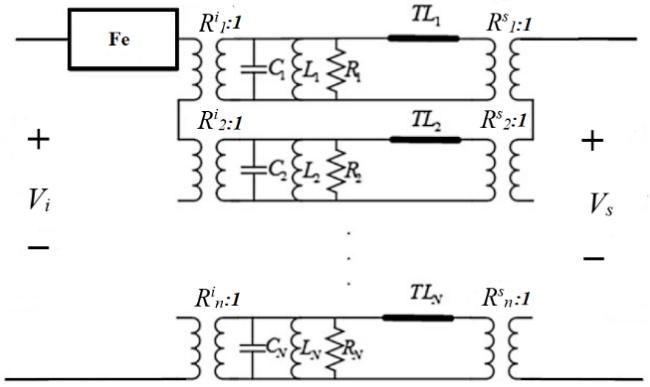


Figure 3: SEM-based equivalent circuit of Bow-Tie RFID Tag with N number of resonators

3.2. Transient Response of Bow-Tie Chipless RFID Tag

The E^i coupled to CNR's by coupling coefficients R^i_1, $R^i_{2,\ldots}$ R^i_n which depends on the direction and the polarization of incident electric filed, whereas R^i_1, $R^i_{2,\ldots}$ R^i_n denotes the coupling coefficients corresponding to E^s. The CNR of the model is presented by a parallel RLC circuit in series with delay line (TL) which designate the turn-on time of CNR, and Fe shows the early time response of the tag. Whereas, the input Vi and output voltages Vs are defined at the transmitting and receiving antennas ports, respectively.

Figure 4. shows the geometry of the Octagonal tag, when E^i strikes at the surface of the tag and E^s electric field is scattered with polarization vectors $\hat{a}_i$ and $\hat{a}_s$ assuming the incident electric field as

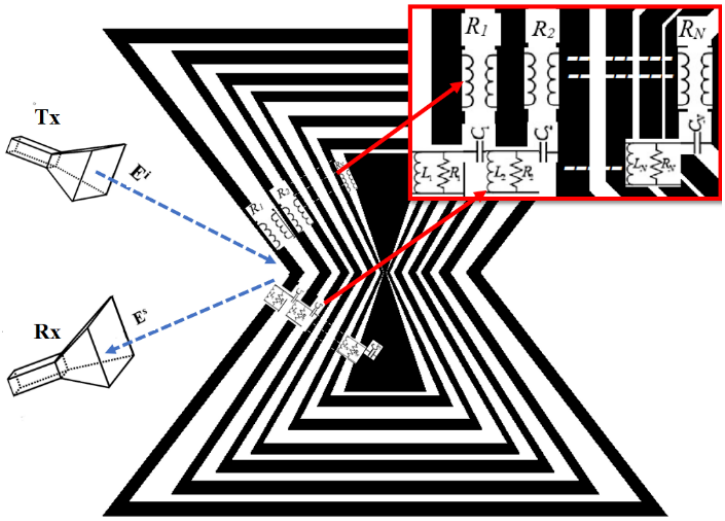


Figure 4: The geometry and the internal circuit modelling of the Bow-Tie RFID Tag Illuminated by an Incident Field Ei

a Dirac-delta function then the transfer function of the tag in the complex s-plane is defined as [15]

$$H_t(\hat{a}_i,\hat{a}_s;s)=\frac{E^s(r;s)}{E^i(r;s)} \tag{1}$$

After some mathematical manipulation, the transfer function in terms of early-time response $H_e(s)$ and late time response $H_1(s)$ with CNR's can be written as [15]

$$H(s)=H_e(s)+H_1(s)$$

$$=H_e(s)+\sum_{n=1}^{N}\left[\frac{A_n}{s-s_n}+\frac{A_n^*}{s-s_n^*}\right] \tag{2}$$

where

$$A_n=\frac{C}{2}\left(1+j\frac{1+\alpha_n}{\omega_n}\right)$$ and CNR's of the circuit is described by

$$s_n=\alpha_n+j\omega_n \tag{3}$$

$$\alpha_n=-\frac{\omega_0}{2Q} \quad \text{and} \quad \omega_n=\frac{\omega_0}{2}\sqrt{4-\frac{1}{Q^2}} \tag{4}$$

where Q represents quality factor which is kept below 0.5 and the RFID first resonator resonates at $f_1=\omega/2\pi$.

3.3. Evaluation of Surface Current Distribution (Jn) and the Coupling Coefficients (Rn).

The current distribution over the first resonator, the outermost octagonal ring is I_1 and V_1 is the induced voltage due to incident fields corresponding to the impedance Z_1 of the resonator then the current distribution over the octagonal scatterers can be evaluated by discretizing the length of the ring into N number of segments [15] as

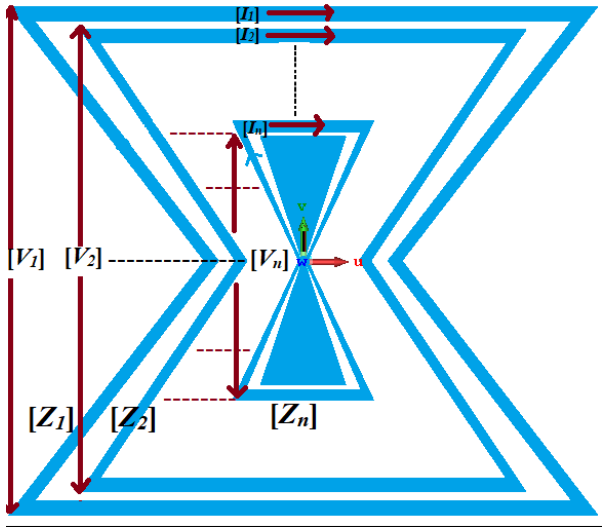


Figure 5: The induced voltages and currents distribution over the surface of the tag

$$[I_1]=[Z_1]^{-1}[V_1] \tag{5}$$

where $[Z_1]$ corresponds to N×N matrix referred to as system impedance matrix, $[I_1]$ and $[V_1]$ are the response vector of N×1corressponding to the incident field, see Figure 5.

The CNR's for the first scatterer is calculated from

$$[Z_1(S_n)][I_1(S_n)]=0$$

$$\Delta_1(S_n)=\det[Z_1(S_n)]=0 \tag{6}$$

The CNR's for the first resonator is obtained by expending $\Delta_1(s_n)$ in complex Taylor series.

$$\Delta_1(s_n)=\Delta_1(s_0)+\Delta_1'(s_0)\,(s-s_0)+\ldots \quad =0 \tag{7}$$

By keeping the first two terms of the equation and ignoring the rest of values the CNR, s_n, is obtained from

$$s_n=s_0-\frac{\Delta 1(s0)}{\Delta 1'(s0)} \tag{8}$$

where s_0 is the initial guess of resonant frequency, by repeating this procedure more accurate values can be evaluated for the first scatterer. For Bow-Tie design, the first resonance was calculated at approximately 9.2 GHz.

The equation (5) is modified to get current distribution for the first resonator of octagonal chipless RFID tag as

$$[I]=\sum_n\frac{R_n}{s-s_n}[J_n] \tag{9}$$

where R_n the residue of nth pole and J_n is the surface current distribution over the surface of the scatterer, which is evaluated in the next section. Also, The time-domain response can be obtained by applying inverse Laplace transform to (9).

To study the mechanism of scattering by a 4-bit octagonal chipless RFID tag, the surrounding medium is supposed as a free space with permittivity ε_0 and permittivity μ_0, the scatterer is PEC and the thickness of the tag is very small hence considered to be zero. Consider an incident field E^i strikes and the induced current over the surface of the tag is (J_n) then the scattered field E^s from the tag is either obtained from Electric Field Integral Equation (EFIE) or Magnetic Field Integral Equation (MFIE) [36]. The EFIE is used which is represented in Laplace form [37].

$$E^s(r;s) = \mu s \iint_A \left[\left(\vec{I} - \frac{1}{k^2} \nabla\nabla \right) G_0(r,r';s) \right] J(r,r';s) dS' \quad (10)$$

where A represents the surface area of the octagonal chipless RFID tag, $\vec{I} = \hat{x}\hat{x} + \hat{y}\hat{y} + \hat{z}\hat{z}$, k represents the propagation constant k=s/c in the complex frequency domain and G_0 defines the scalar Green's function in the free space.

$$G_0(r,r';s) = \frac{e^{-jk|r-r'|}}{4\pi |r-r'|} \quad (11)$$

The coupling coefficients are then evaluated by using the equations

$$R_n(s_n) = -\frac{\left\langle J_n(r), E^i{}_t(r;s_n) \right\rangle_r}{\left\langle \left\langle J_n(r), \left\langle \frac{\partial \vec{G}(r,r';s)}{\partial s}\bigg|_{s=s_n}, J_n(r') \right\rangle_{r'} \right\rangle_r} \quad (12)$$

This equation shows that the coupling coefficient depends upon E^i and natural modes $J_n(r)$ at the resonant frequency.

$$\left\langle J_n(r), E^i{}_t(r;s_n) \right\rangle_r = 0 \quad (13)$$

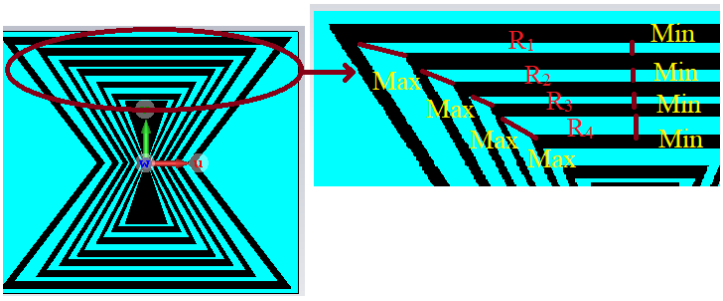


Figure 6: The variation of the distances between consecutive Bow-Tie shaped rings

The evaluation of coupling coefficients (R_1 to R_4) was not easy in the case of chipless Bow-tie RFID. This is because the distance between the two consecutive bow-tie shaped rings is not constant and hence mutual coupling varies, see Figure 6. Therefore, coupling coefficients are evaluated for the values of the maximum and the minimum distances between two consecutive rings at corresponding resonance, see Table 1. The values of coupling coefficients between the first two consecutive rings are -0.010 and -0.021 calculated for the maximum and the minimum spacing of 1.2mm and 0.23mm, respectively for 9.2GHz frequency, see Table 1. Similarly, all values are calculated for corresponding resonant frequencies of 11.3, 14 and 16.9 GHz. The maximum value of R_n from the table is -0.021 which is a sufficiently low value that assures reliable reading of the coded information. To reduce the mutual coupling the average spacing between the resonators can be increased, therefore the size of the tag would be significantly increased.

Table 1: Coupling Coefficients for Corresponding Resonances

n	Resonant Frequency F_{0n} (GHz)	Spacing Between Resonators Min to Max (mm)		Coupling Coefficients Minimum Rn	Coupling Coefficients Maximum Rn
		D_{min}	D_{max}		
1	9.2	0.23	1.2	-0.010	-0.021
2	11.3	0.35	0.9	-0.012	-0.017
3	14	0.17	0.6	-0.009	-0.014
4	16.9	0.52	1.1	-0.014	-0.020

4. Flexibility and Bending Analysis

The Bow-Tie tag is designed on 13.44×11.56mm2 PET substrate of 70µm thickness with conductive gold rings varied in thickness [1].

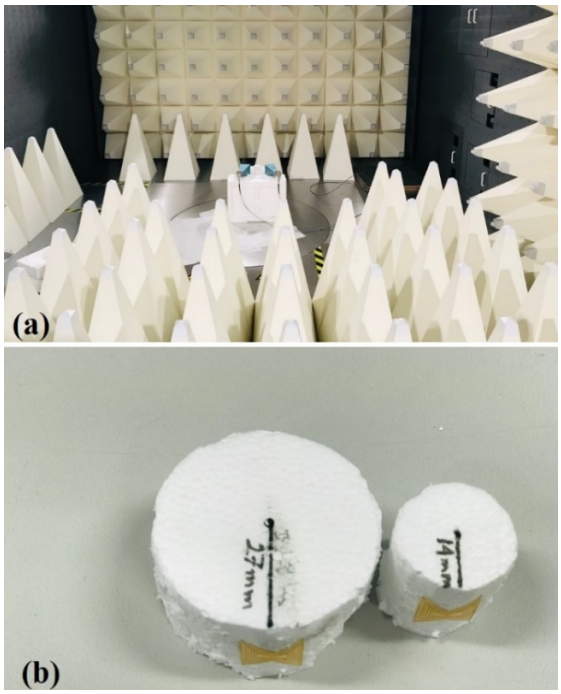
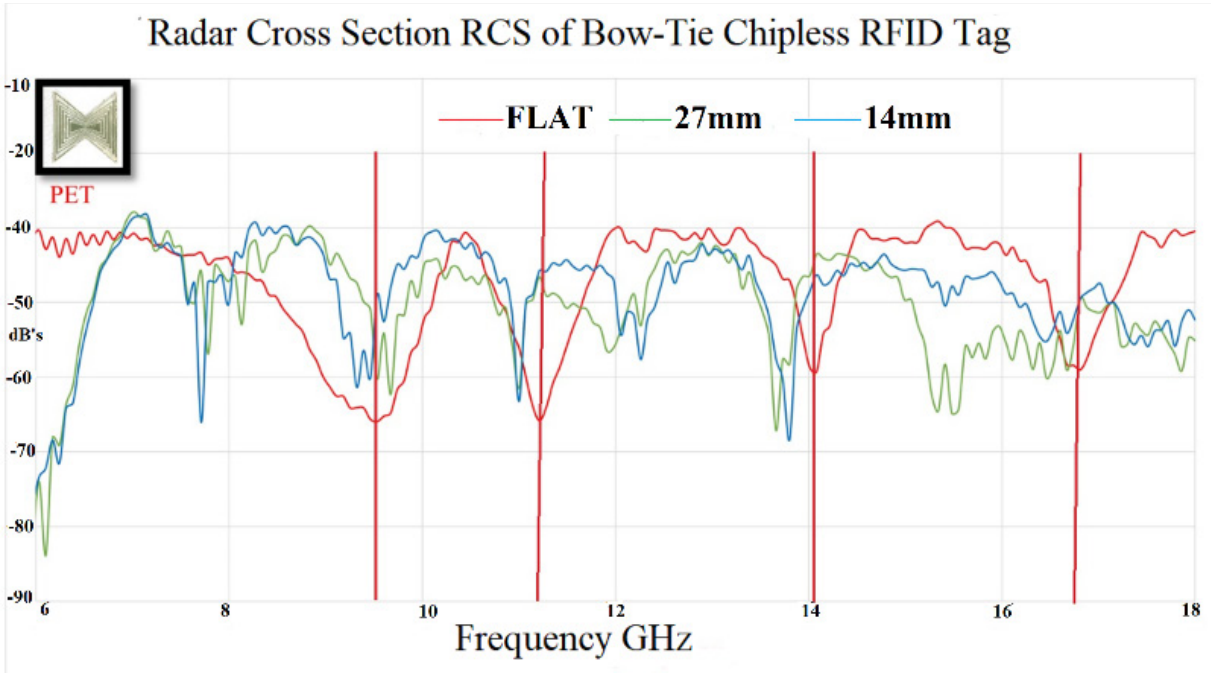


Figure 7: Chipless Bow-Tie RFID tags a) Experimental set-up in chamber b) Bending aspects of fabricated Chip-less RFID Tags at 27mm and 14mm

4.1. Experimental Setup and RCS Results

The chipless Bow-tie RFID tag is printed on a PET substrate by using Universal Laser Systems (ULS). The measurements of



the RCS obtained by the reflections from the tag are analysed by Vector Network Analyser (VNA). For this purpose, Keysigth VNA series E5063A is used with two similar horn antennas operating at 4-18GHz frequencies. To conduct the measurement, the testing is performed in the FRANKONIA anechoic chamber, having a frequency range of 30 MHz to 40 GHz. The compact Bow-Tie RFID tag design is highly flexible and the bit codes are successfully obtained up to the bending radius of 14mm [38]. The Bow-Tie RFID tag undergoes bending for the following cases of curvatures, flat, 14mm and 27mm, see Figure 7 (b).

The Passive Bow-tie RFID tag is placed in the horizontal (H) position to perform the bending analysis for the following cases flat, 14mm and 27mm. All the testing was performed in the Frankonia anechoic chamber. One of the impacts of bending is the signal strength degradation of the backscattered waves and the significant shift in the resonant frequency. In other words |RCS| is lowered while tag undergoes bending. In this case, since the distance of the tag from the reader, is kept very small about 12cm from the transmitter, so a very small reduction in the signal strength is obtained at 0dBm power. However, the percentage shift in the resonant frequency is quite prominent especially at a higher level of bending at 14mm, see Figure 8.

Figure 8. depicts the impact of bending on the RCS obtained in the chamber. A shift of 4% from the resonant frequency is obtained, when a PET substrate-based chipless Bow-tie RFID tag undergoes bending curvature 27mm for the lower range of the frequencies 8-14 GHz, whereas, more than 10% shift is observed for higher range of frequency up to 18GHz. Hereafter, there is no noticeable reduction of the signal strength is observed because of the small distance of tag from the reader. Hence, the experimental results depict that the RCS obtained and encode in frequency signature, from the bent chip-less RFID tag fabricated on the flexible PET substrate.

5. Maximum Read Range Estimation

This section provides the maximum read range of Bow-Tie Chipless RFID tag. The complex RCS feature (σ^{Tag}) of octagonal chipless RFID designs can be obtained by the following equations which are described in [39].

$$\sigma^{Tag} = \left[\frac{S_{21}^{Tag} - S_{21}^{No-Tag}}{S_{21}^{Ref} - S_{21}^{No-Tag}}\right]^2 \sigma^{Ref} \quad (14)$$

Where S_{21}^{No-Tag} is associated with the measurement taken in the empty chamber, S_{21}^{Ref} corresponds to reference value taken from define setup which could be any conductive surface, In this case, a copper square of 50mm x 50mm is taken. σ^{Ref} is associated with the simulation results for the same reference. The results achieved for all the tags show a good agreement between the measured and simulated designs.

The radar equation is used [40] to evaluate the maximum read range of the RFID tag. For this, numerically calculated RCS are considered by using the equation is follows:

$$= \sqrt[4]{\frac{G_T\, G_R \lambda^2 \sigma^{Tag} P_T}{(4\pi)^3 P_{min}}} \quad (15)$$

Where G_T and G_R are the gains of the transmitting and receiving horn antennas, P_T is the power transmitted by VNA to measure RCS, λ is the wavelength. P_{min} is the minimum power receive by receiving antenna and σ^{Tag} is the experimental values of RCS obtained from (14). For maximum read range calculations σ^{Tag} is taken for 1111bit codes, see Figure 8. flat case. The average RCS is obtained from a series of experiments [1] is around 0.02 m^2. The maximum read range is evaluated for -10, -3 and 0dBm transmitted power from VNA and the maximum range observed is 1.80m and extendable to 2.15m for 3dBm power. Hence, a robust Bow-tie chipless RFID tag with transient analysis

is successfully designed fabricated and tested, which is applicable for the IoT applications and the read range of the tag and bit capacity can be increased in future by using the asymmetrical property of the tag.

Table 2: Maximum Read Range Estimation

P_{min}(dB)	G_T (dBi) G_R (dBi)	λ^2 (m)	σ^{Tag} (m^2)	P_T (dB)	R (m)
-70	10	1.06 x 10^{-3}	0.02	-10	1.01
				-3	1.52
				0	1.80
				3	2.15

6. Conclusion

This paper is an extension of Novel Bow-Tie Chipless RFID tag which studies SEM-based circuit modelling of the bow-tie-shaped tag. The electromagnetic response of the tag is presented and coupling coefficients are obtained by transient analysis. The behaviour of the RFID tag is completely studied and experimentally verified for different patterns of bitstreams measured in terms of RCS. The coupling coefficients measured in this analysis evaluated for the values of the maximum and the minimum distances between two consecutive rings at corresponding resonance of 9.2, 11.3, 14 and 16.9 GHz. The maximum value of R_n from the table is -0.021 which is a adequately low value that would not impact the reliable coding of the information. Furthermore, the bending on Flexible Chip-less Bow-Tie RFID tag operating within the range of 8-18GHz are described up to the maximum bend of 14mm and the information stored on a bent chip-less RFID tag made of PET is successfully recovered. Finally, the maximum read range for the RFID tag is evaluated which is 1.8m for 0dBm power but extendable to 2.15m for higher powers. This range is maximum to our knowledge for such a high-frequency range of 8-18GHz.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

I would like to acknowledge the great contribution of my research colleague Panagiotis Ioannis Theoharis, for helping me in the measurements and the fabrications of the tags.

References

[1] M.U.A. Khan, R. Raad, J. Foroughi, F.E. Tubbal, J. Xi, "Novel Bow-Tie Chip-less RFID Tag for Wearable Applications," in 2019 19th International Symposium on Communications and Information Technologies (ISCIT), IEEE: 10–13, 2019, https://doi.org/10.1109/iscit.2019.8905208.

[2] M. Younan, E.H. Houssein, M. Elhoseny, A.A. Ali, "Challenges and recommended technologies for the industrial internet of things: A comprehensive review," Measurement, **151**, 107198, 2020, https://doi.org/10.1016/j.measurement.2019.107198.

[3] A. Čolaković, M. Hadžialić, "Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues," Computer Networks, **144**, 17-39, 2018, https://doi.org/10.1016/j.comnet.2018.07.017

[4] E. Jovanov, "Wearables Meet IoT: Synergistic Personal Area Networks (SPANs)," Sensors, **19**(19), 4295, 2019, https://doi.org/10.3390/s19194295

[5] T.M. Fernández-Caramés, P. Fraga-Lamas, "Towards the Internet of smart clothing: A review on IoT wearables and garments for creating intelligent connected e-textiles," Electronics, **7**(12), 405, 2018, 2018. https://doi.org/10.3390/electronics7120405

[6] Y. Leterrier, "Mechanics of curvature and strain in flexible organic electronic devices," Handbook of Flexible Organic Electronics: Materials, Manufacturing and Applications, 1, 2014, https://doi.org/10.1016/b978-1-78242-035-4.00001-4

[7] N.L. Pira, "Smart integrated systems and circuits using flexible organic electronics: Automotive applications," Handbook of Flexible Organic Electronics: Materials, Manufacturing and Applications, 345, 2014, https://doi.org/10.1016/b978-1-78242-035-4.00014-2

[8] R. Anwar, L. Mao, H. Ning, R. Atakan, H.A. Tufan, S. uz Zaman, C. Cochrane, S.K. Bahadir, V. Koncar, F. Kalaoglu, B.A. Auld, C.E. Baum, D. Betancourt, K. Haase, A. Hübler, F. Ellinger, A.T. Blischak, M. Manteghi, A. Čolaković, L. Corchia, G. Monti, E. De Benedetto, A. Cataldo, L. Angrisani, P. Arpaia, L. Tarricone, F. Costa, S. Genovesi, A. Monorchio, et al., "The internet of things: How the next evolution of the internet is changing everything," IEEE Transactions on Antennas and Propagation, **59**(12), 1–4, 2014, https://doi.org/10.7551/mitpress/10277.003.0004

[9] L. Corchia, G. Monti, E. De Benedetto, A. Cataldo, L. Angrisani, P. Arpaia, L. Tarricone, "Fully-Textile, Wearable Chipless Tags for Identification and Tracking Applications," Sensors, **20**(2), 429, 2020, https://doi.org/10.3390/s20020429

[10] L. Corchia, G. Monti, L. Tarricone, "Durability of wearable antennas based on nonwoven conductive fabrics: Experimental study on resistance to washing and ironing," International Journal of Antennas and Propagation, **2018**, 2018, https://doi.org/10.1155/2018/2340293

[11] R. Atakan, H.A. Tufan, S. uz Zaman, C. Cochrane, S.K. Bahadir, V. Koncar, F. Kalaoglu, "Protocol to assess the quality of transmission lines within smart textile structures," Measurement, **152**, 107194, 2020, https://doi.org/10.1016/j.measurement.2019.107194

[12] Y. Wang, L. Li, B. Wang, L. Wang, "A body sensor network platform for in-home health monitoring application," in Proceedings of the 4th International Conference on Ubiquitous Information Technologies & Applications, IEEE: 1–5, 2009, https://doi.org/10.1109/icut.2009.5405731

[13] K. Finkenzeller, RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication, John Wiley & Sons, 2010, https://doi.org/10.1002/9780470665121

[14] D.K. Klair, K.-W. Chin, R. Raad, "A survey and tutorial of RFID anti-collision protocols," IEEE Communications Surveys & Tutorials, **12**(3), 400–421, 2010, https://doi.org/10.1109/surv.2010.031810.00037

[15] R. Rezaiesarlak, M. Manteghi, CHIPLESS RFID, Springer, 2016, https://doi.org/10.1007/978-3-319-10169-9_3

[16] J.P. Vinding, Interrogator-responder identification system, 1969.

[17] O.E. Rittenbach, Communication by radar beams, 1969, https://doi.org/10.1002/sce.3730530262

[18] R.M. Richardson, Remotely actuated radio frequency powered devices, 1963.

[19] R.F. Harrington, "Field measurements using active scatterers (correspondence)," IEEE Transactions on Microwave Theory and Techniques, **11**(5), 454–455, 1963, .https://doi.org/10.1109/tmtt.1963.1125707

[20] H. Khaleel, Innovation in wearable and flexible antennas, Wit Press, 2014, https://doi.org/10.2495/978-1-84564-986-9/008

[21] D.P. Morgan, "History of SAW devices," in Proceedings of the 1998 IEEE International Frequency Control Symposium (Cat. No. 98CH36165), IEEE: 439–460, 1998, https://doi.org/10.1109/freq.1998.717937

[22] B.A. Auld, Acoustic fields and waves in solids, Рипол Классик, 1973.

[23] V.D. Hunt, A. Puglia, M. Puglia, RFID: a guide to radio frequency identification, John Wiley & Sons, 2007.

[24] J. Foroughi, T. Mitew, P. Ogunbona, R. Raad, F. Safaei, "Smart Fabrics and Networked Clothing: Recent developments in CNT-based fibers and their continual refinement," IEEE Consumer Electronics Magazine, **5**(4), 105–111, 2016, https://doi.org/10.1109/mce.2016.2590220

[25] N.C. Karmakar, R. Koswatta, P. Kalansuriya, E. Rubayet, Chipless RFID reader architecture, Artech House, 2013.

[26] D. Betancourt, K. Haase, A. Hübler, F. Ellinger, "Bending and folding effect study of flexible fully printed and late-stage codified octagonal chipless RFID tags," IEEE Transactions on Antennas and Propagation, **64**(7), 2815–2823, 2016, https://doi.org/10.1109/tap.2016.2559522

[27] A.C. Durgun, C.A. Balanis, C.R. Birtcher, D.R. Allee, "Design, simulation, fabrication and testing of flexible bow-tie antennas," IEEE Transactions on Antennas and Propagation, **59**(12), 4425–4435, 2011, https://doi.org/10.1109/tap.2011.2165511

[28] F. Costa, S. Genovesi, A. Monorchio, "A chipless RFID based on

multiresonant high-impedance surfaces," IEEE Transactions on Microwave Theory and Techniques, **61**(1), 146–153, 2013, https://doi.org/10.1109/tmtt.2012.2227777

[29] M. Martinez, D. van der Weide, "Compact slot-based chipless RFID tag," in 2014 IEEE RFID Technology and Applications Conference (RFID-TA), IEEE: 233–236, 2014, https://doi.org/10.1109/rfid-ta.2014.6934234

[30] S. Rauf, M.A. Riaz, H. Shahid, M.S. Iqbal, Y. Amin, H. Tenhunen, "Triangular loop resonator based compact chipless RFID tag," IEICE Electronics Express, 14.20161262, 2017, https://doi.org/10.1587/elex.14.20161262

[31] M.A. Riaz, H. Shahid, S.Z. Aslam, Y. Amin, A. Akram, H. Tenhunen, "Novel T-shaped resonator based chipless RFID tag," IEICE Electronics Express, 14.20170728, 2017, https://doi.org/10.1587/elex.14.20170728

[32] R. Anwar, L. Mao, H. Ning, "Frequency selective surfaces: a review," Applied Sciences, **8**(9), 1689, 2018, https://doi.org/10.3390/app8091689

[33] A. Blischak, M. Manteghi, "Pole residue techniques for chipless RFID detection," in 2009 IEEE Antennas and Propagation Society International Symposium, IEEE: 1–4, 2009, https://doi.org/10.1109/aps.2009.5172097

[34] A.T. Blischak, M. Manteghi, "Embedded singularity chipless RFID tags," IEEE Transactions on Antennas and Propagation, **59**(11), 3961–3968, 2011, https://doi.org/10.1109/tap.2011.2164191

[35] R. Rezaiesarlak, M. Manteghi, "Complex-natural-resonance-based design of chipless RFID tag for high-density data," IEEE Transactions on Antennas and Propagation, **62**(2), 898–904, 2013, https://doi.org/10.1109/tap.2013.2290998

[36] C.E. Baum, The singularity expansion method, Springer: 129–179, 1976, https://doi.org/10.1007/3540075534_8

[37] J.-M. Jin, Theory and computation of electromagnetic fields, John Wiley & Sons, 2011, https://doi.org/10.1002/9780470874257

[38] M.U.A. Khan, R. Raad, J. Foroughi, F. Tubbal, P.I. Theoharis, M.S. Raheel, "Effects of Bending Bow-Tie Chipless RFID Tag for Different Polymer Substrates," in 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS), IEEE: 1–4, 2019, https://doi.org/10.1109/icspcs47537.2019.9008594

[39] A. Vena, E. Perret, S. Tedjini, "High-capacity chipless RFID tag insensitive to the polarization," IEEE Transactions on Antennas and Propagation, **60**(10), 4509–4515, 2012, https://doi.org/10.1109/tap.2012.2207347

[40] I.Y. Immoreev, "-Signal Waveform Variations in Ultrawideband Wireless Systems: Causes and Aftereffects," Ultrawideband Radar: Applications and Design, 71–104, 2012, https://doi.org/10.1201/b12356-4

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 762-767 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Optimization of Multi-user Face Identification Systems in Big Data Environments

Majdouline Meddad[*,1], Chouaib Moujahdi[2], Mounia Mikram[1,3], Mohammed Rziza[1]

[1]*LRIT Laboratory, Associated Unit to CNRST (URAC 29) Rabat IT Center, Faculty of Sciences, Mohammed V University in Rabat, 10010, Morocco*

[2]*Scientific Institute of Rabat, Mohammed V University in Rabat, 10010, Morocco*

[3]*School of Information Sciences in Rabat, 10010, Morocco*

ARTICLE INFO

Article history:
Received: 29 August, 2020
Accepted: 14 November, 2020
Online: 24 November, 2020

Keywords:
Big Data
Biometric Distribution
Face Identification
Convolutional neural network

ABSTRACT

Computer vision offers several strategies that permit computers to comprehend the substance of inputted data to extract the relevant highlights features. That gives the possibility to develop several successful recognition systems like face identification. One of the enormous difficulties these days is the way to have a prompt identification face in a multi-client identification system. We propose in this paper, an optimization of some face identification systems in big data environments that manages the cost of an appropriate identification time while holding a good performance of the system. The used systems are based on: CNN with Inception V3, PCA, LDA and LBPH. Our experimental results indicate that the performance can be preserved while reducing considerably the running time. LBPH with a Radius equal to 2 and 8 neighbors gives the best accuracy of that is equal to 95,71% with an identification time of 4.44 seconds.

1 Introduction

Computer vision offers several techniques for machines to understand the content of images like human vision. To understand the content of images, machines need some image processing steps to be able to extract features from the images for several applications such as face identification, object recognition and object tracking.

The substantial idea of a biometric system is to utilize acquisitions of behavioral and/or physiological characteristics of human like face, fingerprint, signature and voice for recognition/verification. Currently, several face identification systems use large-scale databases that contain a big number of subjects such as celabA Dataset that contains 10177 identities and 202599 images. Large-scale databases present a big open challenge to create an operational biometric identification that offers an acceptable identification time. The Scalability [1] of such tremendous systems makes the challenge harder to be solved. This work is concerned with the creation and optimization of such biometric identification systems.

This paper is an extension of a previous work presented during 2019 in the 1st International Conference on Smart Systems and Data Science (ICSSD) [2], where we have tested and built several identification systems (i.e., LBPH [3],[4], PCA [5], LDA [5] and deep learning with inception V3 [6]) that use a big database with some hard conditions.

The remainder of the paper is structured as follows. The related works are presented in Section 2. Section 3 presents the different used face identification systems with all optimization steps. Our experimental results are introduced in Section 4 Our conclusion and perspectives are given in Section 5.

2 Related Work

Identification time, efficiency and security are three crucial issues from which large-scale identification systems suffer. Due to these problems. Several published kinds of research deal with such problems. On one hand, some published approaches tackle security issues whom we can cite [7],[8] that implemented Biometric anti-spoofing technique by using the trait randomization technique.

On the other hand, published approaches can be part of two significant classes: learning distribution approach (first class) and Processing distribution approaches (second class). The principal idea of the second class is sharing tasks to several workers in the cluster to handle the issue of treatment large amounts of data in a few

[*]Corresponding Author: Majdouline Meddad, Email: majdouline_meddad@um5.ac.ma

times by employing the big data technologies in the cloud environment for parallel-distribution treatment. Well known, deep learning models are hefty in computing when it needs more iterations to get the best performant model which needs to use an approach of the first class with big data technologies to distribute the learning process.

For the first class, the crucial idea presented in [9] is merging the utility of the neural network algorithm with the MapReduce technique to face the extensive data problem in handwritten characters. The authors in [10] proposed for the handwriting recognition digit by firstly applying an elastic distortion to the entered data then Implemented a computing method with distribution to gain the time cost by mapping the dataset into several machines of the cluster. In [11], the author proposed some parallelization methods which are MRBPNN 1, 2 and 3) based on a MapReduce technique to face exhaustive scenarios in the conditions of the size of data and neurons.

For the second class, [12] displays a MapReduce technique of the process of converting frames of video to grayscale images as a preprocessing technique to extract features from converted images. In [13], the author apply a distributed medical application in two different techniques of distribution such as MapReduce and Sun Grid Engine regarding running time. In [14], the author introduced an approach for storing and processing satellite images by a remote sensing processing tools OTB with the MapReduce technique of Big Data technologies. In [15], the author proposed to use the MapReduce for parallel model execution to extract the similarity of the tumor image. We trust that the principal drawback of this class is hard to manage the cluster to ensure continuity of processing in a distributed platform and Hadoop technology requires a significant memory.

3 Face Identification In Big Data Environment

The conception of a biometric system carries four principal parts as shown in FIG.1.

- The sensor part (Grey Box) that obtains the biometric information to remove a computerized portrayal during enlistment, identification or verification.
- The characteristic extraction part (Green Box) that empowers decreasing the dimensionality of the representation, emitted by the sensor part, by extricating the significant capabilities. This part can possess a sub-part to test the norm of gained biometric information.
- The database part (Red Box) that comprises the biometric formats of the enlisted subjects inside the framework.
- The decision part (Orange Box) that contrasts the test format and the enlisted formats of the database part requires a decision with regards to the sort of the framework.

The correspondence between these parts becomes tedious inside of having huge database for identification systems. To fix this difficulty, we will utilize some technologies for parallelizing these correspondence tasks inside various slaves. Indeed, current datasets are quickly expanding in terms of size and multifaceted nature, also, there is an importunate need to create solutions to extract pertinent data by utilizing methods based on statistics.

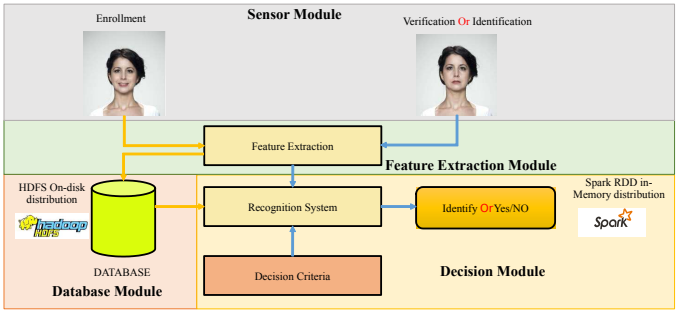


Figure 1: The general conception of a biometric system [16]

An effective biometric framework might be a framework that primarily controls the capacity of the database to be changed in terms of size or scale. they are numerous big data frameworks that are utilized for overseeing a huge data, as Apache Spark framework [17] and Hadoop [18] that run dataflow diagrams over the slaves of a cluster.

In the data-parallel computing process, a client program is arranged into an execution plan graph (EPG), This EPG is the central information structure utilized for distributing tasks, control distributed jobs in the slaves and adapting to non-critical failure. Once a job is running, EPG is unaltered at runtime aside from some restricted adjustments. This makes it hard to utilize dynamic optimization procedures that could significantly improve the disseminated execution dependent on runtime data. Optimus [19] is a framework for progressively rewriting an EPG at runtime.

MillWheel [20] is a system for constructing a data processing with a high size of data in a few time where the user describes the code and the directed computation graph for specific nodes then the framework oversees determined state and recording continuously the flow with non-critical failure in stream processing.

For supporting embedded devices, Sonora [21] is a framework that enhance the stream abstraction for the accompanying reasons. To begin with, streams effectively bind together numerous data activities in cloud and Mobile.

Apache Spark, Naiad [22] and Hadoop run in less calculation time when there is enough memory inside the group of machines to convey the working arrangement of the calculation. Spark keeps up MapReduce's linear adaptability and adaptation to non-critical failure, it is a lot quicker, a lot simpler to program, the distributed data of the scope of computationally heavy tasks, including graphic process, interactive queries, flows and Machine Learning (ML).

In this paper, the jobs of a huge face identification system are parallelized by utilizing Hadoop[23] with Apache Spark introduced by Directed Acyclic Graph (DAG) in figure 3. Every Spark application includes a driver program that completes the user's function and carries out distributes operations on a group of the machine (cluster). The main reflection that Spark furnishes could be resilient distributes dataset (RDD), which could be an assortment of parceled items on a cluster that will run in equal. Users can even request that

Spark remain RDD in memory, permitting it to proficiently reutilize operation in parallel. when a node is failed RDD can automatically be recovered.

3.1 Distributed tasks of LBPH, PCA and LDA in Apache Spark

The utilized identification system is predicated on [3]. However, we have made some changes, to support it inside the Apache Spark, following the stages beneath (see additionally figure 2).

We firstly make first RDD (RDD1) by loading all test images such as Binary Files from HDFS then we split it into sixty parts. Secondly, we apply a transformation of the RDD1 to make RDD2 by transforming the color image to grayscale. Thirdly, we create RDD3 by applying transforms of loaded images in RDD 1 to extract real labels. Fourthly, we made RDD 4 by extracting the content of previous RDDs (2 and 3) by applying a collective action. Finally, the identity of each image is parallelized and compared with the right label extracted in RDD3.

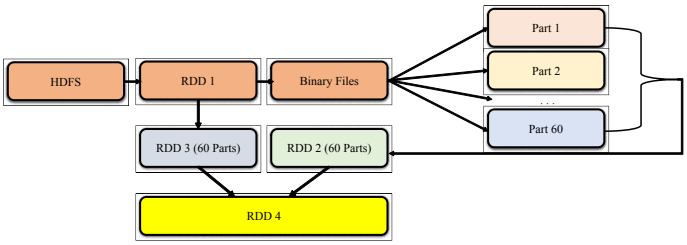


Figure 2: The process of parallelization in Apache Spark

To diminish the running time within the identification stage using the normal PCA, LBPH and LDA identification systems. Initially, the first Apache Spark RDD was created by recovering the test set in 60 parts as binary files from the Hadoop HDFS, and afterward, every thread of our computer changes every picture of every part to grayscale to create the second RDD of grayscale images. Then a third RDD is created by extracting the right label of each test image to collect their identity.

Secondly, we extract the PCA, LDA and LBPH characteristic vectors and compare the vector of the test with those of the learning set.

Thirdly, we predict the identity of the image of the test with the minimal distance overall comparison of the images from the learning set. We used HDFS in the database module to store images and Apache Spark for characteristic extraction and decision parts.

3.2 Pre-trained CNN model, LBPH, PCA and LDA for Multi-user identification system

In this sub-section, we have utilized the Transfer learning procedure where a model produced for tackling an issue is reutilized for a subsequent issue in a timesaving way by starting from patterns that have been learned in the first task, in this manner you raise antecedent learning and keep away from starting from scratch.

We have utilized the Inception V3 as an extractor of pertinent characteristics. It is utilized besides to dispose of the last completely connected layer, at that point, we train the rest of the model as a learning set element extractor for the new dataset. Inception V3 extract 2048 values for every picture. We call these highlights CNN codes. Whenever we have removed the codes for whole pictures. We have prepared a totally connected layer with a Softmax classifier for our system.

A while later, to claim less period, due to the mechanism of spark's build-in broadcast the model is firstly distributed to the cores of our machine in [24]. At that point, this model is used on each core and utilized to identify the identity of an entered image. For this work, the models of a huge identification face system are distributed by utilizing just Apache Spark technology [23].

In our work, every worker takes an image to identify by turning the facial identification system based on PCA, LDA or LBPH to predict the identity of a client by calculating the similarity value with all over the training set as shown in figure 3(a).

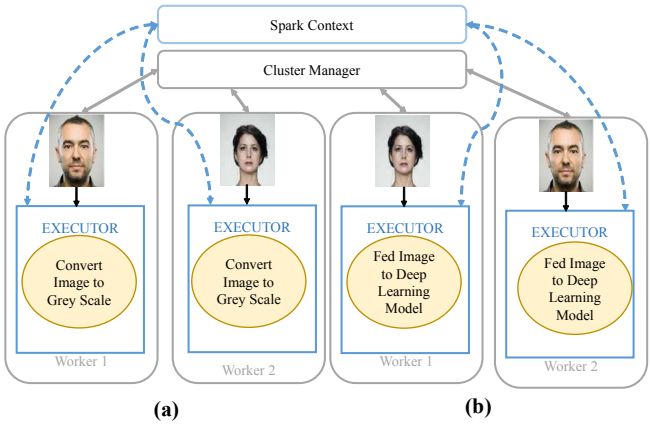


Figure 3: Multi-user identification system: (a) for PCA, LDA and LBPH, (b) for Deep Learning Model

Firstly, the image from RGB (color) to grayscale is converted, afterwards, Every worker compute the eigenvectors, LBP histogram and fisherFace vectors then every worker of the cluster identifies the identity of the tested image from saved PCA, LDA or LBPH models which contain all computed values as shown in figure 3(a) and shared to all cores of the machine.

We employed in this paper a pre-trained CNN Model (Inception V3). Firstly, we have only trained the classification part of the Inception V3 model due to the computational cost of learning to create a model that can identify the person from his/her face image which we need to reutilize the model trained from the Imagenet database to our chosen dataset, we have trained the classification part to induce the correct weights and bias of the fully connected layer (FC) for identification face system, thereafter, we fed the test image into the model to urge the identity, this is often the task of every worker within the cluster as illustrated in figure 3(b) where every worker takes a test image from Hdfs then fed it into the model to extract characteristics that contain 2048 values of every vector then pass it through classification part (Fully connected layer trained to create the facial identification system) then introduce the identity as the last step.

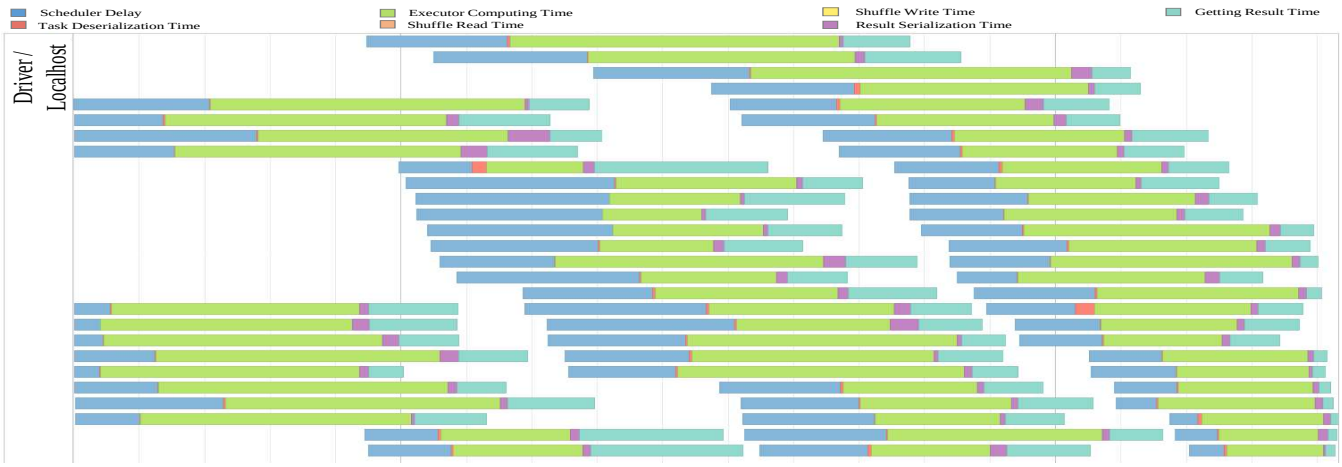


Figure 4: The run time using 6 cores (12 threads) of various tasks

4 Experimental Results

In this section, we evaluate the optimization of various identification systems : the FisherFace [5], LBPH [3][4], , FisherFace [5] and Inception V3 [6] that involve 48 layers can distinguish between 1000 items (e.g., Flower, Fish, Bird, etc) in terms of identification time and the precision. For the evaluation, we have got used the Extended Yale Faces B database [25].

4.1 Database Description and Evaluation

We applied 17 changes (e.g., Rotation, Blur, Noise, etc.) for every subject of the database utilized in this paper [26] to make a bigger database and also to make the identification task more harder.
This database involves 28 subjects with 16128 images with 9 poses and in 64 lighting conditions, In our paper, we are just engaged by 26 subjects. Consequently, the database consists of 107730 images in total. We have suggested splitting the database into three sets as described in figure 5.

4.2 Results and discussion

In the aim to reduce the running time, we run our system using Apache Spark technology by making a collection of RDDs. Every thread of the computer core runs a part of partitioned RDD in 60 parts as illustrated in the figure 4. Each job could be a part of the step, as an example, the primary job contains 60 tasks, in as much as our machine involve 12 threads, 12 tasks are going to be run in parallel (see figure 4). figure 6 illustrates a comparison between various face identification systems that utilize EigenFace, LBPH, FisherFace, and CNN model in Apache Spark [24].

The output of the last convolutional layer of the CNN model is called cnn code extracted from the entered images (Trained images) during the primary part of the learning phase, this stage takes 4 hours and 19 minutes by using Apache Spark broadcasting to calculate all codes. Running the identical stage of the learning step it takes 14 hours without utilizing the Apache Spark broadcasting, is an off-the-cuff term of a feature vector that plays a vital part of identification, which is a representation of a picture in less number of values that will be used for classification.

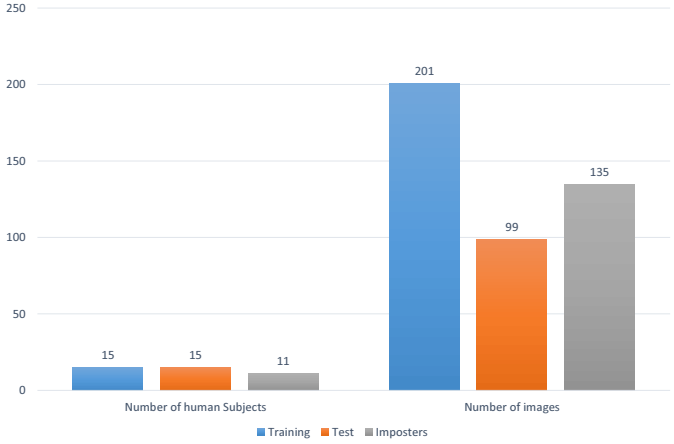


Figure 5: Database Distribution

This final layer was created to give a collection of values that is fitting to the classifier to separate among users. This implies that it needs to be a meaningful and good representation of the images because it should have as many relevant features for the classifier to create an accurate decision by just using a few values.

During the learning phase, every picture is reutilized various times during the learning phase and every code takes an all-inclusive time to be calculated. To quicken the job we conserve these codes on our machine to not recalculate them consistently. On the off chance that we re-execute the framework once more without feed in images to the model.

When the codes are all extracted in the first step, the specific development of the last layer of the neural network model starts. Many series of outputs will be during this step that can show the exactness of the validation and hence the precision of the learning part.

The precision of the learning shows how the model learned from the pre-owned pictures to distinguish between identities. Validation exactness presents the coordinating rate utilizing an arbitrarily

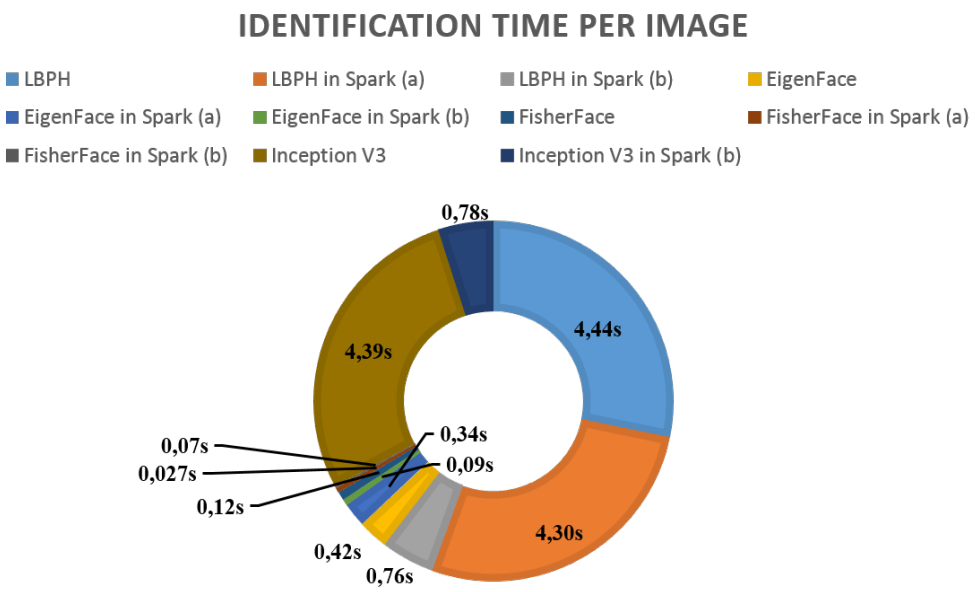


Figure 6: Overall identification time Comparison between The Different Tested Identification Systems with/without distribution

chosen group of pictures from a remarkable learning set. That validation exactness depends on the work of an assortment of data that is not contained inside the learning set if the precision of the learning is high. However, the validation set precision stays low, this suggests that the model is remembering specific highlights inside the learning set pictures. In each validation or learning step, we select 100 arbitrary pictures of the entire validation/learning set, we find its codes inside the cache and brings them into a definitive layer to get identities.

These predictions are contrasted with the specific labels to optimize the classification part and update the bias and weights by utilizing some optimizers like Adam Optimizer and Gradient Descent.

By a learning rate of 0.01 with 250000 iterations using the Gradient Descent Algorithm and Adam Optimizer we respectively get 91, 47% and 91.66%. By a learning rate of 0.1 with 500000 iterations using the Gradient Descent Algorithm and Adam Optimizer we respectively get 91.82% and 92.31%.

By a learning rate of 0.05 with 250000 iterations using the Gradient Descent Algorithm and Adam Optimizer we respectively get 92, 30% and 90.84%. By a learning rate of 0.1 with 500000 iterations using the Gradient Descent Algorithm and Adam Optimizer we respectively get 92.24% and 91.45%.

By a learning rate of 0.001 with 250000 iterations using the Gradient Descent Algorithm and Adam Optimizer we respectively get 87.01% and 91.85%. By a learning rate of 0.1 with 500000 iterations using the Gradient Descent Algorithm and Adam Optimizer we respectively get 88.73% and 92.10%.

We set in the learning phase the proportion of learning set at 80% and we preserved 10% for validation and test. In every iteration we have entered 100 images, after every 10 iterations, we have got evaluated the model using 100 images of the validation set that are chosen arbitrarily. after finishing the learning phase, we fed all the test set through the model to know the precision and how the way the model can classify the inputted faces. After being aware of the good precision of the model in the last part of the learning phase we run our test set for every learning rate and variety of iterations as shown in figure 6. Globally, the precision value is expanded by expanding the number of iterations. If the learning rate is tiny, the precision value is expanded smoothly which makes the learning becomes more reliable, however, the optimization will take longer because steps towards the minimum of the loss function are very small. If the learning rate is high, the precision is expanded rapidly and that we will have the overshoots of the minimum which makes aggravate the loss.

Our main aim is to increase the identification time per image of the test while getting the good performance of the models. This purpose is reached by running our system in Apache Spark (see figure 6). As shown in figure 6 the quick system is FisherFace followed by the EigenFace, then the CNN model, and at last the LBPH.

To advance the exactness of systems, we have made some changes to a few parameters of the LBPH algorithm, such as changing the number of neighbors and the Radius value. For Inception V3, we have made changes to a few parameters simply like the number of iterations, the chosen optimizer algorithm, and subsequently the learning rate value.

For the LBPH algorithm with a Radius equal to 1 with 8, 6 and 4 neighbors, we respectively get 94.05%, 93,74% and 91.91% an Accuracy of identification.

For the LBPH algorithm with a Radius equal to 2 and 8,6 and 4 neighbors, we respectively get 95.71%, 95.43% and 94.02% an Accuracy of identification.

All systems are executed in a server that contains 6 cores with 12 threads and 4.70 GHz Turbo frequency with 12Go of the storage memory.

5 Conclusion and Perspectives

In this paper, we introduced an optimization of multi-user identification face systems in Big Data environments. Hadoop and Apache Spark are two technologies accustomed to parallel tasks and make multi-user systems. This optimization has been examined and evaluated, as far as accuracy and identification time using various classical identification face systems. Our experimental results show that the performance is preserved while offering an affordable identification time for oversized scale systems. In our future work, we have plan to create a multi-view identification face system using a Siamese Convolutional network to tackle the issue of having a few facial acquisitions.

Acknowledgment The financial support of the "Centre National pour la Recherche Scientifique et Technique" CNRST, Morocco, is acknowledged by Majdouline Meddad.

References

[1] K. Janocha, W. M. Czarnecki, "On loss functions for deep neural networks in classification," Schedae Informaticae, 2016, doi:10.4467/20838476SI.16.004.6185.

[2] M. Meddad, C. Moujahdi, M. Mikram, M. Rziza, "A Multi-User Face Identification system with distributed tasks in a Big Data Environment," in ICSSD 2019 - International Conference on Smart Systems and Data Science, 2019, doi:10.1109/ICSSD47982.2019.9003112.

[3] T. Ahonen, A. Hadid, M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, doi:10.1109/TPAMI.2006.244.

[4] D. Huang, C. Shan, M. Ardabilian, Y. Wang, L. Chen, "Local binary patterns and its application to facial image analysis: A survey," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 2011, doi:10.1109/TSMCC.2011.2118750.

[5] A. Khan, H. Farooq, "Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition," **8**(6), 267–270, 2012.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, doi:10.1109/CVPR.2016.308.

[7] K. Okereafor, C. Onime, O. Osuagwu, "Enhancing Biometric Liveness Detection Using Trait Randomization Technique," in 2017 UKSim-AMSS 19th International Conference on Computer Modelling Simulation (UKSim), 28–33, 2017, doi:10.1109/UKSim.2017.44.

[8] K. OKEREAFOR, O. Osuagwu, C. Onime, "Biometric Anti-spoofing Technique Using Randomized 3D Multi-Modal Traits," 2018, doi:10.6084/m9.figshare.12479165.v1.

[9] K. Sun, X. Wei, G. Jia, R. Wang, R. Li, D. C. Oct, "Large-scale Artificial Neural Network : MapReduce-based Deep Learning," .

[10] N. Basit, Y. Zhang, H. Wu, H. Liu, J. Bin, Y. He, A. M. Hendawi, "MapReduce-based deep learning with handwritten digit recognition case study," in Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, 2016, doi:10.1109/BigData.2016.7840783.

[11] Y. Liu, J. Yang, Y. Huang, L. Xu, S. Li, M. Qi, "MapReduce Based Parallel Neural Networks in Enabling Large Scale Machine Learning," Computational Intelligence and Neuroscience, 2015, doi:10.1155/2015/297672.

[12] M. Yamamoto, K. Kaneko, "Parallel Image Database Processing With Mapreduce and Performance," International Journal of Electronic Commerce Studies, 2012, doi:10.7903/ijecs.1092.

[13] T. Ben-Nun, T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," ACM Computing Surveys, 2019, doi:10.1145/3320060.

[14] I. Chebbi, B. W, F. I. R, "Improvement of Satellite Image Classification : Approach Based on Hadoop / Map Reduce," in 2nd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP'2016 March 21-24, 2016, Monastir, Tunisia, 31–34, 2016.

[15] D. D. Manojbhai, K. K. Pradipkumar, R. Rajamenakshi, "Big image analysis for identifying tumor pattern similarities," in Proceedings of 2016 International Conference on Advanced Communication Control and Computing Technologies, ICACCCT 2016, 2017, doi:10.1109/ICACCCT.2016.7831596.

[16] A. K. Jain, P. Flynn, A. A. Ross, Handbook of Biometrics, Springer Publishing Company, Incorporated, 1st edition, 2010.

[17] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, I. Stoica, "Spark: Cluster computing with working sets," in 2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud 2010, 2010.

[18] S. G. Manikandan, S. Ravi, "Big data analysis using apache hadoop," in 2014 International Conference on IT Convergence and Security, ICITCS 2014, 2014, doi:10.1109/ICITCS.2014.7021746.

[19] Q. Ke, M. Isard, Y. Yu, "Optimus: A Dynamic Rewriting Framework for Data-Parallel Execution Plans," in Proceedings of the 8th ACM European Conference on Computer Systems, EuroSys '13, 15–28, Association for Computing Machinery, New York, NY, USA, 2013, doi:10.1145/2465351.2465354.

[20] T. Akidau, A. Balikov, K. Bekiroğlu, S. Chernyak, J. Haberman, R. Lax, S. McVeety, D. Mills, P. Nordstrom, S. Whittle, "MillWheel: Fault-Tolerant Stream Processing at Internet Scale," Proc. VLDB Endow., **6**(11), 1033–1044, 2013, doi:10.14778/2536222.2536229.

[21] X. Chen, I. Beschastnikh, L. Zhuang, F. Yang, Z. Qian, L. Zhou, G. Shen, J. Shen, "Sonora: A Platform for Continuous Mobile-Cloud Computing," Technical report, 2012.

[22] D. G. Murray, F. McSherry, R. Isaacs, M. Isard, P. Barham, M. Abadi, "Naiad: A timely dataflow system," in SOSP 2013 - Proceedings of the 24th ACM Symposium on Operating Systems Principles, 2013, doi:10.1145/2517349.2522738.

[23] A. V. Hazarika, G. Jagadeesh Sai Raghu Ram, E. Jain, "Performance comparision of Hadoop and spark engine," in Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2017, 2017, doi:10.1109/I-SMAC.2017.8058263.

[24] T. Hunter, "Deep Learning with Apache Spark and TensorFlow," 2016.

[25] D. K. K.C. Lee , J. Ho, "Extended Yale Face B," 2016.

[26] A. S. Georghiades, P. N. Belhumeur, D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, doi:10.1109/34.927464.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 768-780 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Surge Protection Device for Ex Application

Teik Hua Kuan*, Kuew Wai Chew, Kein Huat Chua

Department of Electrical and Electronic Engineering, Universiti Tunku Abdul Rahman, Sungai Long 43000, Malaysia

ARTICLE INFO

Article history:
Received: 12 July, 2020
Accepted: 11 November, 2020
Online: 24 November, 2020

Keywords:
Lightning
Surge Protection Device
Hazardous area
Ex area
Explosion Protection Technique
Overvoltage Transient

ABSTRACT

A modern process plant is controlled, operated and safeguarded by the sophisticated distributed control system or a programmable logic controller and an emergency shutdown system. This process plant's operation could be seriously affected by lightning, electrical switching, and power outrage. Lightning propagation can travel kilometers away, which has the risk of disturbing the control system's operation. The process plant's control system could damage when it is subjected to surges induced by lightning, power outrage, or heavy load switching. To protect equipment damage from these surges surge protection device is required. A number of the process plant uses flammable material for its production or during the manufacturing processes, explosive gas or vapor is released created an explosive atmosphere known as hazardous (Ex) area. Surge protection device mounted in the Ex area or used in the Ex control loop needs to be carefully selected to protect the equipment and at the same time, it does not become a source of ignition. This paper provides practical consideration and comprehensive detail in revealing the science and engineering behind Ex application on surge protection devices. The technical presentation is limited to equipment protection; structural protection against lightning will not be discussed here.

1. Introduction

This journal article is an extension of the original conference paper entitled, "Practical approach on surge protection device for Ex application," presented in 2019 1st. International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia [1].

Lightning is a natural phenomenon that can be seen in volcanic eruptions, pyrocumulonimbus clouds, heavy snowstorms, hurricanes and thunderstorms [2]. Lightning-induced voltage surges can enter into the electrical system via the capacitive, inductive and resistive coupling [3]. According to the report from NFPA [4], in the United States alone, from 2007 to 2011, an estimated 22,600 fire cases were caused by lightning, with a $451 million loss in direct property damage per year. The report also cited several explosion incidents caused by the lightning surge. One of them is a coal mine (Ex area) explosion in January 2006 in West Virginia, which claimed 12 lives. A protective circuit like a surge protection device or SPD is used to protect the electrical apparatus and its accessories from the damage created by transient surge activity induced by the lightning and surges generated by switching heavy reactive load or power supply interruption and prevent a fire. SPD has to be carefully selected and with proper installation for its intended function; else, instead of preventing equipment failure, it could become the source of fire or explosion as reported by [5–7].

In the original paper, the following are either not evaluated, missing or too brief as the reader is expected to have the background knowledge to understand the subject. In the present work, the critical area left out by the original paper is being addressed and enhanced. The rectification and enhancements are highlighted below:

- The reasons why SPD is needed for the modern electronics system for the application in a safe, hazardous area and critical control loop application.
- Why is the selection criteria on SPD for safe area application not sufficient for hazardous area application?
- Presentation on the lightning protection zone (LPZ) and the SPD types are too brief on the original paper. Therefore it is decided to separate them into two sub-sections with enhancement and with an application example. Definition and a summary of the key points are also being addressed and tabulated.

*Corresponding Author: Teik Hua Kuan; Adress: Universiti Tunku Abdul Rahman, Sungai Long 43000, Malaysia; Contact No: +60123173650; Email: thkuan@1utar.my or thkuan@klaesb.com

- Application on lightning protection level's (LPL) concept; what needs to be done before implementing LPL and the efficiency level for various LPLs are added.
- The description of the maximum continuous operating voltage has been revised to reflect the practical application.
- The illustration on the concept of the "Temperature Classification" is enhanced with a practical example.
- Elements required to be present for an explosion to occur and the concept of various explosion protection techniques being implemented are added.
- Using "Spark Test Apparatus" to obtain the minimum ignition curve for inductive, capacitive and resistive load at the field. The application of the minimum ignition curves is also added.
- Requirement on the hot permit to carry out maintenance and typical application example is added for flameproof, Exd protection.
- Hot surface investigation and typical applications example for increased safety, Exe protection is added.
- Missing entity safety parameter description on the cable capacitance and inductance for intrinsic safety, Exi protection technique is rectified.
- Surge withstanding capability (SWC) consideration is added as it is missing in the original paper.
- Why are two SPDs needed for a control loop? Why one of the SPD will have the Temperature Classification, the other one is not necessary?

The subject of explosion protection for industrial automation applications (non-military) is a specialist field; hardly any higher education institution conducting the courses. Hence, most of the knowledge on various explosion protection technique is coming from the industry. There are very few industry players who produce multiple Ex-rated products and conduct multiple Ex-product research; most of the manufacturer only specialist in their respective areas of expertise. For example, field device manufacturers (field devices like pressure transmitter, I/P converter, control valve, solenoid valve, etc.) like Honeywell, Emerson or Yokogawa only produce Ex-rated field devices, not other devices like Ex-rated junction box or cable gland. Therefore it is difficult for the industrial user to familiarize themselves with various explosion protection techniques for all the control components. This paper helps the industrial user to implement surge protection in the Ex area systematically. It investigates and evaluates the selection criteria and SPD's application requirement for Ex area application with various explosion protection techniques. Safety calculation and proper installation are also presented. It details down all factors involved in implementing the solutions. Two application examples are quoted in detail.

The paper starts the discussion on the need for the SPD in the modern electronics system in the safe area, hazardous area and the critical control loop application; The working principle of the SPD then moves on to the selection criteria for Ex applications and the common explosion protection technique that could be applied together with SPD. Three explosion protection techniques are selected for evaluation and discussion in detail on its application, namely the application of SPD using Exi (intrinsically safe), Exd (flameproof) and Exe (increase safety) protection method in the common explosion protection technique section. The paper then provides two working examples of how the selection and application of SPD using Exi and Exd methods. Safety parameter calculation, distance requirement, temperature consideration, etc., are presented in the examples. The paper ends with a summary of the key area of consideration and criteria in selecting the SPD for Ex application.

2. Why Is Surge Protection Important In The Modern Electronics System?

Modern equipment demands for more application features, faster speed, smaller physical dimensions and lower power consumption. This makes the circuit designer squeeze more transistors into a smaller IC chip, using the material to operate at a lower voltage and lower power consumption. The progress in developing faster, denser and lower power consumption IC Chip makes the modern electronics circuit more vulnerable to the induce transient surges than the early components that use vacuum tube and discrete transistor. Hence, transient surge protection is becoming more important in the modern electronics system. Figure 1 shows the increase of equipment failure rate from the early vacuum tube technology to the modern IC chip caused by transient surges.

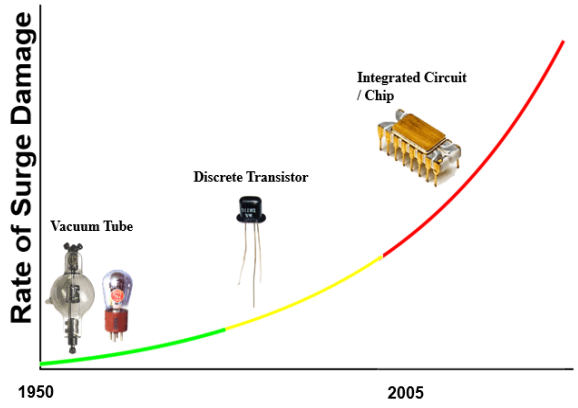


Figure 1: Increased Rate of Equipment Failure [8].

On the other hand, modern electronic systems also rely a lot more high-performance electronic components to execute the critical task that cannot afford to fail. Failure will lead to plant shutdown and production losses, which will cost millions of dollars. For example, emergency shutdown systems (ESD), depending on the process and the safety requirement, can be designed as a standalone, redundant system or triple redundant system to cope with equipment failures. A redundant design plus a logic solver algorithm allows the ESD to detect the real loop failure and decide when to shut down the plant using the voting system. Once the decision is made, the plant is then shut down according to the pre-determined sequence to avoid plant accident or explosion. The redundancy is often achieved by duplicating or triplicating the field sensor, input-output cards, control processors and even the actuators. Almost all the field sensors, except for the wireless system, are connected to ESD by cables, potentially entry

routes for the transient surges. Therefore, SPDs are included in the system design to improve the resilience of such systems from the effect of transients.

When a process plant uses flammable material for its production or during the manufacturing processes, explosive gas or vapor is released created an explosive atmosphere known as hazardous (Ex) area. SPD mounted in this Ex area needs to be carefully selected to protect the equipment from transient surges and at the same time, it does not act as a source of ignition. SPD requirements for the hazardous (Ex) location are often more stringent than in the non-hazardous (non-Ex) locations.

When sizing up the SPD for equipment protection for safe area application, generally, the let-through voltage (should be below the SWC [9–11] of the device to be protected), kA rating, connection method (either series or parallel) of the SPD need to be considered. However, when the application is involved with the Ex area, this is not sufficient as the SPD could become the source to initiate an explosion. The additional consideration for the SPD to use in the Ex area includes the SPD's mounting location, the applicable explosion protection technique, equipment surface temperature, and equipment heat dissipation.

3. Working Principle of SPD

SPD protects the electrical installation, consisting of electrical equipment, consumer unit, wiring and accessories, from electrical power surges known as transient overvoltage. The lightning activity, electricity is interrupted and immediately re-established or the switching of a heavy reactive load could instigate the power surge. Figure 2 illustrates the working principle of a 4-20mA signal line SPD. When there is no transient surge, the SPD does nothing; it is transparent to the circuitry. Whenever a surge appears, the SPD will be triggered instantly. It clamps the SPD's output voltage, depending on the model used, to a pre-determined level and diverts all the excessive current to the ground via a protective earthing terminal.

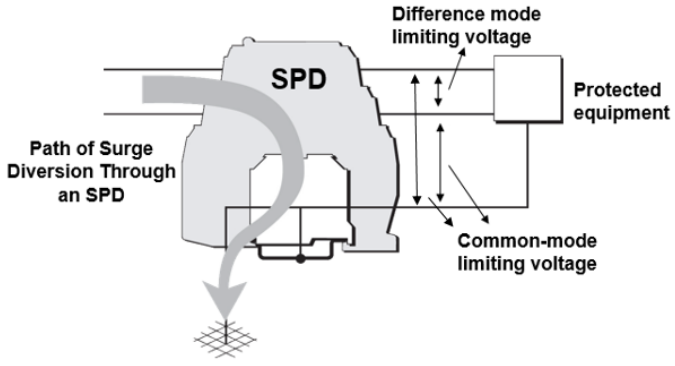


Figure 2: Working Principle of a signal SPD (redrawn from [1] with additional information).

4. Selection Criteria on SPD for Ex Area Applications

When selecting an SPD for use in the Ex area, all the selection considerations and criteria on SPD for the safe area application are applied. Additionally, all other factors that could cause the SPD to trigger an explosion and other ignition sources should be considered as well.

Typically, SPDs are mounted in a panel at the control room to protect the systems and its component against surge. If there is a requirement to protect the field device from surge, additional field mounted SPDs will be required. In other words, the process control loop will require two SPDs. One SPD is mounted as close as possible to the field device (to protect the field device), another one should be mount as close as possible to the controller (to protect the controller) at the control cabinet.

4.1. Types of SPDs

IEC 61643-11:2012 and NFPA 79 standard defined three different types of surge protection devices (SPDs). They are Type 1, Type 2 and Type 3, respectively. The SPD types provide information on where an SPD may be installed, as shown in Figure 3. The energy discharged from the surge is brought down in various stages, starting with the more robust SPD (Type 1) and finer protection Type 2 and Type 3 SPDs. This protection arrangement and co-ordination are represented with the Lightning Protection Zoning (LPZ), which divides the structure based on the lightning strike effects.

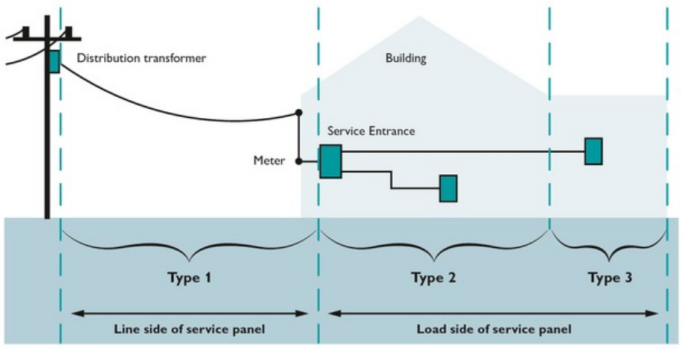


Figure 3: Types of SPD and Its Mounting Location. [12].

Type 1 SPD: Provide lightning protection against transient overvoltage instigates from direct lightning strokes where the lightning discharge voltage could spread from one of the structures earth conductor to the other network conductors. Type 1 SPD is characterized by a 10/350μs surge current waveform per IEC 61643-11:2012

Type 2 SPD: It is the primary protection system for low voltage electrical installations. Protect against transient overvoltage caused by switching or indirect lightning strokes. Typically, it is installed in the electrical distribution board (DB) to prevent transient surges in the electrical installations and protect the loads. Type 2 SPD is characterized by a 8/20μs surge current waveform per IEC 61643-11:2012

Type 3 SPD: Type 3 SPD is aimed to provide local protection on sensitive and critical loads. This SPD has a low discharge capacity. It is only installed as a supplement to Type 2 SPD and it is installed as near as possible to the sensitive loads. They are widely used as hard-wired devices and generally combined with Type 2 SPDs for fixed installations. In many cases, they are built-in and incorporated into a surge-protected socket outlet, surge-protected portable socket-outlets for telecommunication lines and data

protection. Type 3 SPD is characterized by a combination of voltage waveform (1.2/50 µs) and current waveform (8/20 µs).

Table 1 provides a summary of Types of SPD, its test waveform and the current rating, as well as the typical technology used and the Lightning Protection Zone (LPZ). LPZ will be discussed in the following section.

Table 1: Types of SPD, Test Waveform, Applicable Technology and LPZ [13].

	LPZ 0/1	**LPZ 1/2**	**LPZ 2/3**
Typical SPD Installation Point	Service Entrance (e.g. Main distribution board or telecom NTP)	Sub-distribution board or telecom PBX frame	Terminal Equipment (e.g. socket outlet)
Type of SPD	**1**	**2**	**3**
Surge Test Waveform	10/350 µS current	8/20 µS current	Combination 8/20 µS current and 1.2/50 µS voltage
Typical peak test current (per mode)	25KA	40KA	3KA
Typical Technology (Power Line)	Spark Gap	Metal Oxide Varistor (MOV)	Metal Oxide Varistor (MOV)

4.2. Lightning Protection Zone, LPZ

IEC 62305-4:2010 describes the LPZ concept. LPZ identifies and divides a structural area according to the risk level exposed to the Lightning Electromagnetic Impulse (LEMP). It coordinates the type of SPD's kA current rating requirement within the zone. The concept of LPZ taken from IEC 62305-4: 2010 is as shown in Figure 4. ANSI/IEEE C62.41 also has a similar concept, as shown in Figure 5.

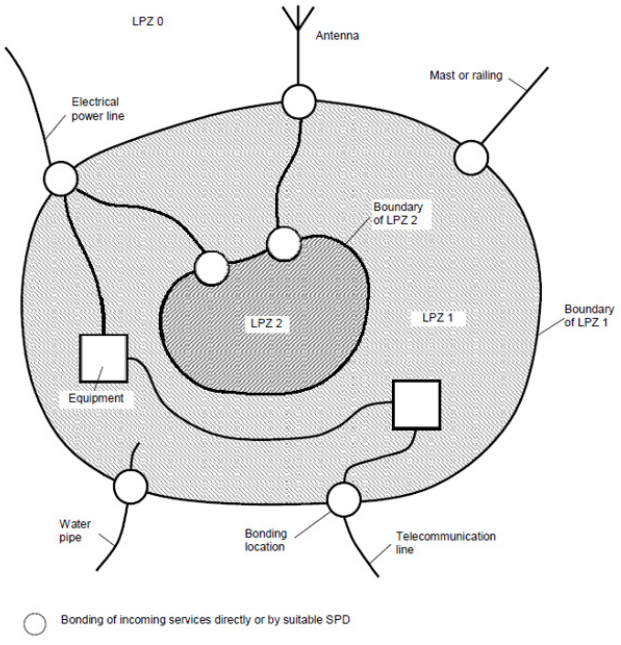


Figure 4: The Principle of LPZ - Division into Different LPZ per IEC 62305 part 4 [14].

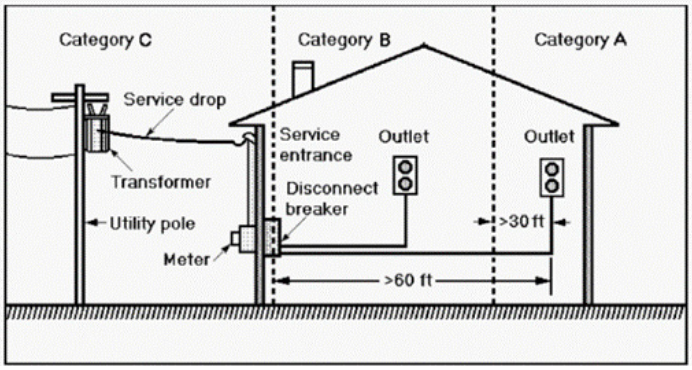


Figure 5: Surge Protective Category as per ANSI/IEEE C62.41 [15].

The LPZs are defined as follows:

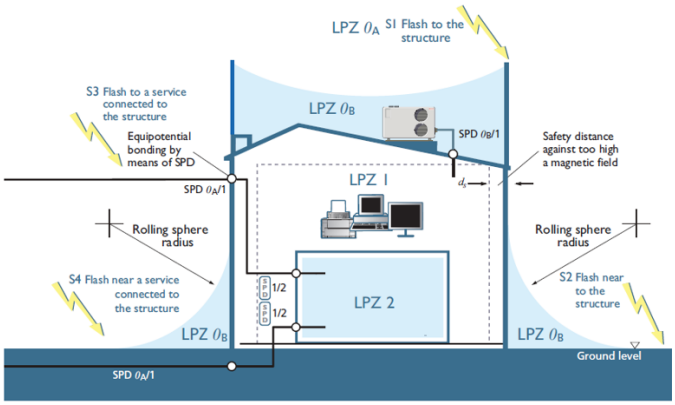
- **LPZ 0_A**: It is an open zone and not protected by the external lightning protection system (LPS). The equipment installed is directly exposed to the direct atmospheric discharges. Hence, the device installed in this zone must be able to handle the entire surge current generated and it is exposed to the whole magnetic induction.
- **LPZ 0_B**: This zone also refers to the open zone but is protected by the external lightning protection system, such as a lightning rod. This method provides direct lightning strikes protection but suffers from total exposure to the magnetic field.
- **LPZ 1**: It is the zone inside a building. At this zone, there is a partial lightning current transmitted from external surges. However, the degree of induced currents is less than zone LPZ 0_A and LPZ 0_B. Between LPZ 0_B and LPZ 1, Type 1 SPD should be installed to protect the downstream equipment.
- **LPZ 2, LPZ n**: This zone refers to an area further inside the building where low surges may occur. The shielding and SPDs, both at the limits of the different zones and protecting the local equipment, allowing further reduction of the induced current.

The LPZs could be view as two categories, as shown in Figure 6, namely two external zones, which are LPZ 0_A and LPZ 0_B; and usually two internal zones, which are LPZ 1 and LPZ 2. If necessary, further zoning can be introduced to further reduce the electromagnetic field and lightning current.

4.3. Lightning Protection Levels, LPL

According to the IEC 62305 standard, LPL can be classified into four levels, namely Level I, Level II, Level III and Level IV. Each level has a set of parameters to predefined the maximum and minimum lightning surge current. Refer to Figure 7, it can be seen that the object or device to be protected will be assigned to one of the four LPL levels before a lightning protection system is implemented. Efficiency in LPL I is the highest at 99% and LPL IV is the lowest at 84%. The maximum lightning surge current set by LPL is used to design and the selection of the SPD. According to the IEC 62305, any lightning surge current parameters exceeding LPL 1 (i.e., 200kA) will be excluded from lightning protection. The chances and probability of occurrence outside the range are less than 2 %. The cost of erecting a lightning protection system, for example, the necessary protection angles, distances from loops, protection profiles and distances from arrestors for

LPL I is much more than that of LPL IV. IEC 62305 Part 1 provides the detail on the LPL probability of occurrence.



LPZ O_A Direct flash, full lightning current, full magnetic field
LPZ O_B No direct flash, but partial lightning or induced current, full magnetic field
LPZ 1 No direct flash, but partial lightning or induced current, damped magnetic field
LPZ 2 No direct flash, or partial lightning, but induced currents, further damped magnetic field
Protected volumes inside LPZ 1 and LPZ 2 must respect safety distances d_s

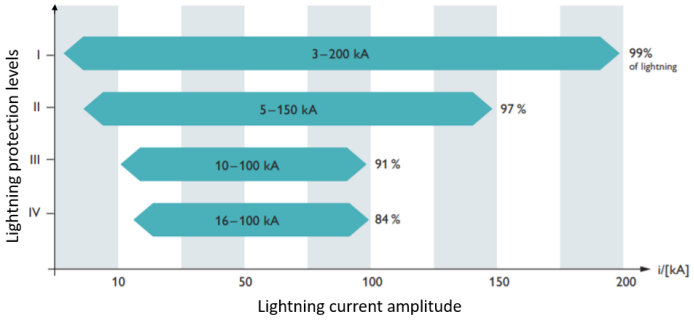


Figure 7: Lightning Protection Levels per IEC 62305-1 with Corresponding Minimum and Maximum Lightning Current Amplitude (redrawn from [16] with additional information).

4.4. Parameter of SPD

Common SPD specifications recognized by IEC 61643 are described below. They are used to select suitable SPD.

- **Nominal Voltage, U_N**
  It is the standard operating supply voltage of the apparatus or system circuit based on the use envisaged for the SPDs.
- **Maximum Continuous Operating Voltage, U_C**
  The maximum voltage that can be permanently applied to the SPD without damaging and triggering it. U_C rating of the SPD is selected such that its triggering voltage is slightly higher than the nominal operating voltage (U_N) of the equipment to be protected. This is to avoid SPD been accidentally triggered during normal operating.
- **Voltage Protection Level, U_P**
  U_p is known as the "Let-Through" voltage or the clamped voltage in the industry. U_P defines the maximum output voltage that could appear at the SPD terminal when tested with a specific waveform per IEC 61643. Generally, U_p will be the maximum output terminal voltage of the SPD when subjected to transient surge. In practice, the SPD's U_p rating should be chosen so that its value is lower than the maximum SWC voltage of the apparatus to be protected.
- **Nominal Load Current, I_L**
  I_L is defined as the maximum continuous RMS load current that is allowed to pass through the SPD without triggering and damaging the SPD, including the capability to withstand the current continuous thermal overload. This term is only applicable when the SPD is connected in series. Since a parallel connection is capable of handling an unlimited load current; hence, this term is not relevant to SPD connected in parallel.
- **Impulse Current, I_{imp}**
  It is the peak current with 10/350μs waveform, direct strike, flowing through the SPD to test mainly for Type 1 SPD.
- **Nominal Discharge Current, I_n**
  Peak current with 8/20μs waveform flowing through the SPD for the test of Type 2 SPD.

4.5. Implementation of SPD in the Ex Area

All guidelines highlighted in sections 4.1 to 4.4 are applicable when applying and installing SPD in hazardous areas. Additional factors described below to ensure the SPD does not become a source of the explosion also need to be considered.

4.5.1 Hazardous Area Classification

Classification of the hazardous area is a risk management tool that handles potentially flammable material to support explosion safety verification. Hazardous areas are defined by the occurrence of the type of hazards present, such as gas, vapor or dust; the probability of that hazard may be present; flammable concentrations on the lower explosion limits (LELs) and upper explosion limits of the explosive substance (UEL) as shown in Table 2. Hence, hazardous area classification is a systematic way to review, analyze and classify plant safety according to the explosive substance. This allows an optimum selection of the electrical apparatus or equipment installed in a hazardous atmosphere. ATEX and IEC standards classify the Ex area into Industry Group, Gas Group, Temperature Classification and Zone. The hazardous area's electrical apparatus and equipment selection is based on the hazardous area classification defined by the ATEX / IEC standard.

Industry Groups:

- **Group I** is the mining industry, where equipment is used in the explosive environment that contains methane gases or vapors of its equivalent hazard, as shown in Table 3.
- **Group II** is the surface industry, as shown in Table 3. It is then further divided into three subgroups, that is Group IIA, Group IIB and Group IIC. Group IIA for the environment containing propane gases or vapors of its equivalent hazard; Group IIB for the environment containing ethylene gases or vapors of its equivalent hazard and Group IIC for the environment containing hydrogen gases or vapors of its equivalent hazard.
- Group III is also for the surface industry but with dust and fibers hazards. They are again subdivided into Group IIIA, IIIB and IIIC according to the nature of the explosive environment, as shown in Table 4.

Table 3 provides the information on Industry Group I and II; and the minimum ignition energy (MIE). MIE is the

minimum energy required to ignite the easiest explosive gas-air mixture.

Table 2: Flammable Properties [17].

Combustible Gas LEL's and UEL's			
		LEL	UEL
Acetylene	C2H2	2.5%	100%
Benzene	C6H6	1.2%	8.0%
Methane	CH4	5.0%	15.0%
Ethane	C2H6	3.0%	15.5%
Ethyl Alcohol	CH2H5OH	3.3%	19.0%
Ethyl Ether	(C2H5)2O	1.7%	36.0%
Ethylene	C2H4	2.7%	36.0%
Hexane	C6H14	1.1%	7.5%
Hydrogen	H2	4.0%	75.6%

Table 3: Industry Group I and II for Mining and Surface Industry [17].

Industry	Gas Group	Substance	MIE	Selection of equipment
Mining Industry Group I	I	Methane	200µJ	Equipment only suitable for mines unless it is additional approved for surface: Marked I/IIC
Surface Industry Group II	IIA	Propane	180 µJ and higher	II; IIA; IIB and IIC certified equipment may be used in an IIA area.
	IIB	Ethylene	60 µJ -180 µJ	II; IIB and IIC certified equipment may be used in an IIB area.
	IIC	Hydrogen	20 µJ – 60 µJ	Only II & IIC certified equipment is allowed to be used in an IIC area.

Dusts and Fibres (Group III) Surface	IIIA	Combustible flying's
	IIIB	Non-conductive dust
	IIIC	Conductive dust

Temperature Classification

Temperature Classification is also known as ignition temperature, auto-ignition temperature (AIT) or T class. It defines the device's maximum surface temperature under a normal or abnormal condition with respect to the ambient temperature of 20°C to 40°C (unless otherwise stated) [18]. Depending on the types of explosion protection techniques used or its combination, for example, Exd, Exe or Exi, the temperature corresponds to the device's external surface's maximum temperature or the maximum temperature of the device's internal surface (for example, Exi device putting into Exd junction box). For a combination of explosion protection techniques used (for example, Exdei – Exd, Exe and Exi are used in the device), the final temperature classification of the device needs to be determined. All flammable materials will have an AIT, the temperature at which ignition will occur without an external source. Apparatus must be selected with its temperature classification below the AIT. This will ensure the Ex environment's apparatus is not exposed to the gas-air mixture with a temperature exceeding its AIT. If several different flammable materials are present within the same area, the material that gives the lowest AIT dictates the device's overall applicable temperature class. The temperature classification only applicable to Industrial Group II (surface industry), as shown in Figure 8. Temperature classes do not apply to Group I (mining) applications. Mining equipment has either a rigid 150°C or 450°C limits [18]. Be noted that the device's temperature class does not represent the operating temperature range of the device, but rather the maximum allowable surface temperature with respect to a default ambient temperature of 20°C to 40°C with the presence of flammable material.

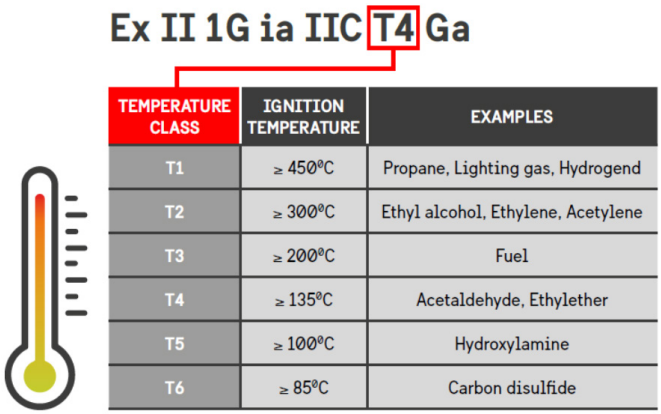


Figure 8: Equipment temperature classification [19].

According to the ATEX directives, all the hazardous area installation apparatus, including SPD, must be marked with its temperature class and assigned an appropriate temperature classification according to the nature of hazard that might be present to maintain its integrity and prevent unexpected explosion due to AIT. For instant, if the hazard present is butane gas, its AIT is 372°C, then all apparatus used in that environment cannot generate a surface temperature exceeding 372°C. In this case, the apparatus rated T2 (300°C) to T6 (85˚C) will meet the requirement. This also means that for this case, all the apparatus used in the butane gas environment cannot have a surface temperature equal to or greater than 372˚C under fault or regular operation. Any equipment that generates 372˚C or more surface temperature could increase the possibility of an explosion.

Table 5: Zone Classification [20].

Zones and Equipment Categories				
Zones		Typical Definitions of Zones (for guidance only)	ATEX Minimum Permissible Equipment Category for Zone	Minimum Equipment Integrity Requirements
Gases and Vapours zones	Dusts			
0	20	Explosive atmosphere is present continuously, for long periods or frequently	1	Equipment must be safe under normal operation, expected and rare malfunction
1	21	Explosive atmosphere is likely to occur under normal operation, occasionally	2	Equipment must be safe under normal operation, expected malfunction
2	22	Explosive atmosphere may occur under abnormal operation and persist for a short period only	3	Equipment must be safe under normal operation.
The higher the probability of an explosive atmosphere occurring and persisting, the lower the allowable potential risk of ignition and the higher the safety integrity requirements of the equipment to be installed. The relationship between zones and equipment categories can be varied following a complete risk assessment				

Zoning Classification

- Zone Classification: Classified concerning the probability of a potentially explosive environment being present in a 3-dimensional region or space and the duration for which it is likely to exist. The available explosion protection methods also depend a lot on the zone classification assessment and its

assignment. The classification of the hazardous area is shown in Table 5.

- This rule of thumb derived from API RP 505 and IP 15 stated if the hazard is continuously present for more than 1000 hours in a year is classified as Zone 0. If the hazard present between 10 to 1 000 hours in a year is classified as Zone 1. If the hazard presents less than 10 hours in a year is classified as Zone 2.

Equipment Marking

- Equipment Marking: All certified apparatus used in the Ex area must carry mandatory marking. The standard uniformly defines the marking requirement. The marking on all equipment intended for use in a potentially explosive environment should provide all necessary and essential information for safe operation and easy identification at the site. Figure 9 and Figure 10 provide a self-explanation and sample of the manufacturer marking according to ATEX Directive 94/9/EC and the IECEx certification scheme, respectively.

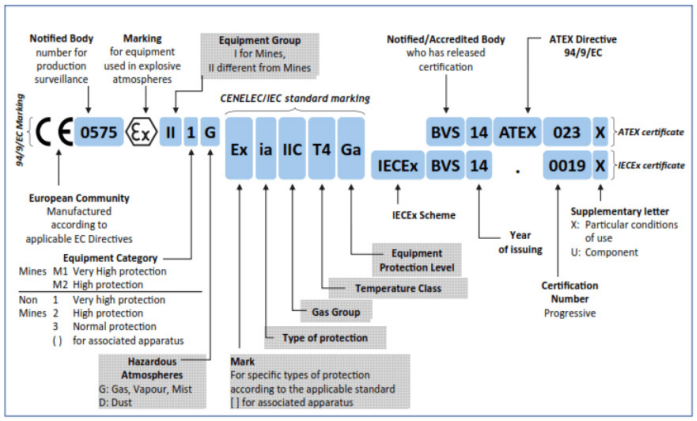
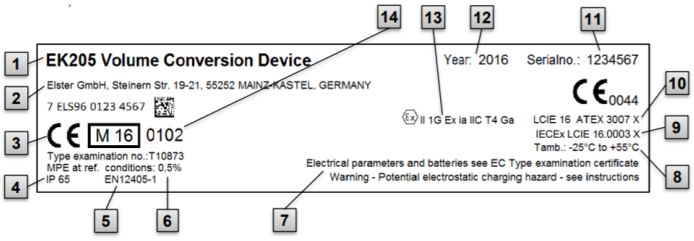


Figure 9: Summary of Apparatus Marking According to ATEX Directive 94/9/EC and IECEx Certification Scheme [21].



1	Type designation	8	Ambient temperature range
2	Manufacturer and address	9	IECEx - EC type examination certificate no.[5]
3	CE marking	10	ATEX - EC type examination certificate no.
4	IP protection class data	11	Serial number
5	Reference to EN 12405-1	12	Year of construction
6	Measurement accuracy data	13	Ex marking
7	Notices	14	Metrology marking

Figure 10: Typical Manufacturer Marking Tag [22].

5. SPD in Common Explosion Protection Technique

For an explosion to occur, three elements must be present at the same time, namely the hazard (flammable or explosive substance), supporter (air or oxygen) and the source of ignition (electrical spark or hot spot). All the explosion protection technique is to remove one or more of this element or contain the explosion so that it will not propagate to a bigger area or outside the enclosure. There are many explosion protection techniques used by the industry. However, only three common techniques and two practical examples based on Exi, intrinsic safety and Exd, flameproof explosion protection technique will be discussed here. Table 6 summarizes the various explosion protection techniques used in the industry, Ex code and its applicable zoning.

Table 6: The IEC Methods of Protection and Permitted Zone of use [23].

Method of Protection	Ex code	Permitted in:		
		Zone 0	Zone 1	Zone 2
Encapsulation	ma	✓	✓	✓
	mb		✓	✓
Oil immersion	o		✓	✓
Powder filling	q		✓	✓
Pressurisation	p		✓	✓
Flameproof	da	✓	✓	✓
	db		✓	✓
	dc			✓
Intrinsic safety	ia	✓	✓	✓
	ib		✓	✓
	ic			✓
Increased safety	e		✓	✓
Type of protection "n"	nA			✓
	nR			✓
	nC			✓

5.1. Intrinsic Safety, Exi

- Permitted to use in Zone 0, 1 and 2.
- Could carry out maintenance repair work without the need to turn off the power supply.
- The use of "Uncertified Simple Apparatus" is allowed when connected through a safety barrier.
- Typical Application with Exi protection is for instrumentation signal, control circuit and other low power apparatus like CCTV.

Exi is a low-energy signaling technique. It brings down the electrical energy level for the Ex area apparatus used below the ignition curve. In this concept, any electrical spark, thermal or hot spot effect in any part of the circuitry intended for use in Ex areas do not have sufficient energy to cause an explosion. Hence, the operation for live maintenance repair work can be carried out and is allowed. The energy levels available for signaling are small but useable and more than adequate for most instrumentation systems.

The minimum ignition energy is obtained experimentally using "SPARK TEST APPARATUS" [24]. The experiments have produced a general agreed "IGNITION CURVES". These are the designer's curves with suitable safety factors for safety parameters calculation (section 6.1 provides the detail on safety parameters calculation). Three curves are resulting from the experiments; they are Resistive, Capacitive and Inductive Ignition Curve, as shown in Figure 11 to 13, respectively.

The resistive ignition curve (Figure 11) provides the entity safety parameter on the maximum open-circuit voltage denotes as Voc and the maximum short circuit current denotes as Isc allowed to go into the hazardous area without causing an explosion for different Gas Group. Any combination of the voltage and current

touching or above the curve could cause an explosion. A value below the curve will not have sufficient energy to cause an explosion. Equations (1) and (2) in section 6.1 provide the detail on the application.

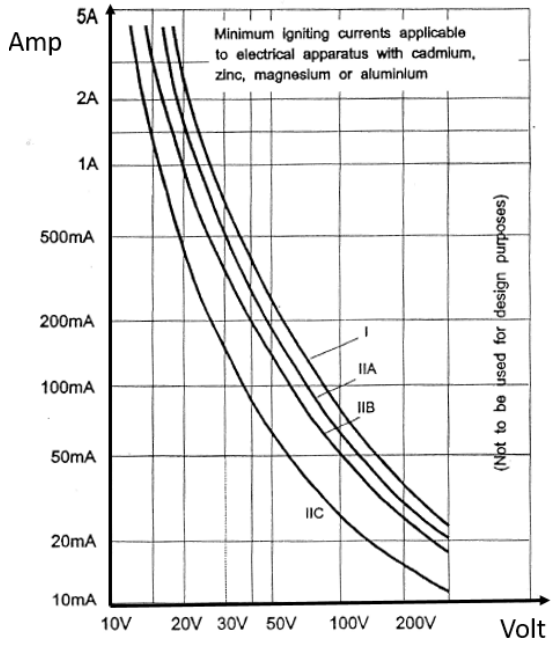


Figure 11: Resistive Ignition Curve [24].

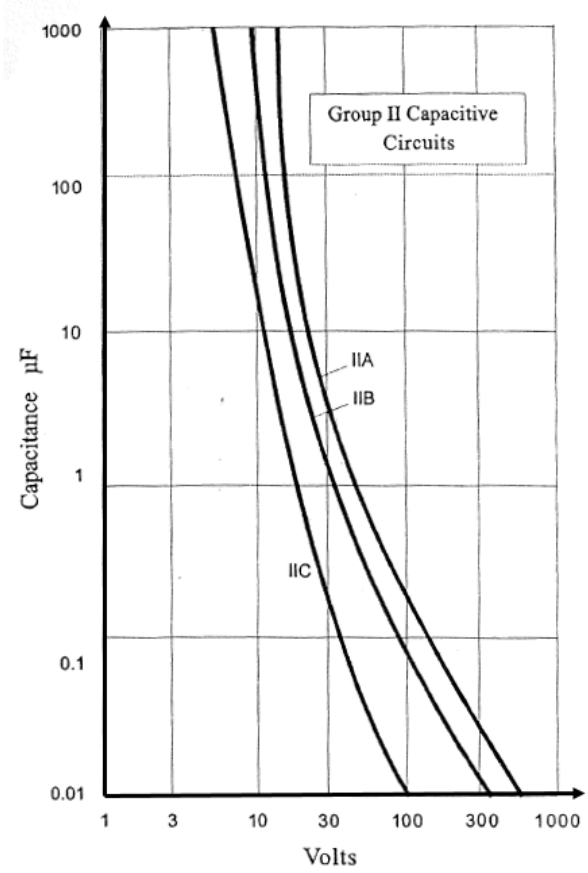


Figure 12: Capacitive Ignition Curve [24].

Capacitive ignition curve (Figure 12) provides the entity safety parameter on the maximum external capacitance (Ca) at the applied voltage allowed to be connected in the hazardous area for different Gas Groups. If the external capacitance (Ca) at the applied voltage is below the curve, it will not have sufficient energy to cause an explosion. Equation (3) in section 6.1 provides the detail on the application.

Inductive ignition curve (Figure 13) provides the entity safety parameter on the maximum external inductance (La) at the applied voltage allowed to be connected in the hazardous area for different Gas Groups. If the external inductance (La) at the applied voltage is below the curve, it will not have sufficient energy to cause an explosion. Equation (4) in section 6.1 provides the detail on the application.

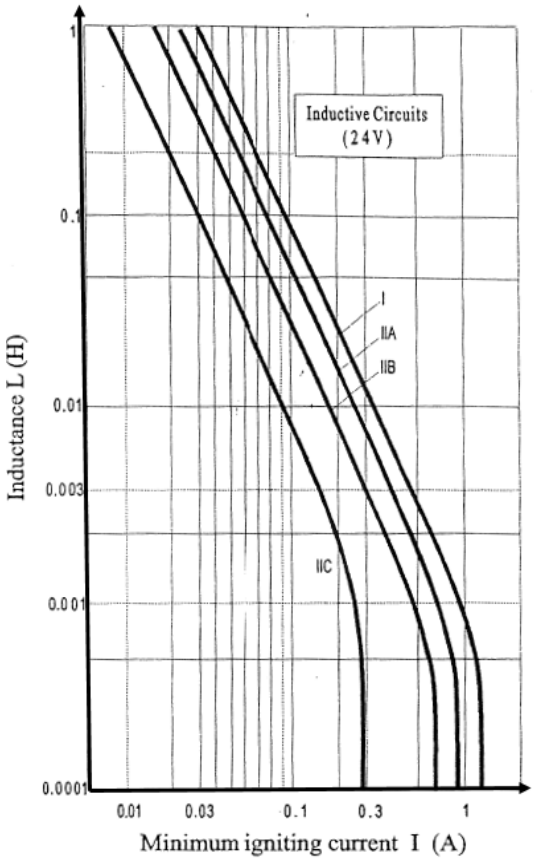


Figure 13: Inductive Ignition Curve [24].

The IEC 60079-11 standard defines three categories of Exi protection. They are "Exia", "Exib" and "Exic" with different fault tolerance levels, as summarized in Table 7. Only Exia category is permitted to apply in Zone 0.

With the introduction of intrinsically safe technique, the uses of “Uncertified Simple Apparatus” are allowed as defined by IEC 60079-25 in a hazardous area without additional certification when the simple apparatus is connected to the field through a certified safety barrier. IEC 60079-25 defines “Simple Apparatus” as an apparatus that does not generate or store energy more than 1,5V, 100mA and 25mW, such as mechanical switches and resistors, thermocouples and PT100. The simple apparatus also required to pass the insulation test at 500 Vrms (one-minute duration) to be used in the Ex area. The application of Simple Apparatus in the Ex area significantly brings down the cost of implementing intrinsically safe systems.

Table 7: Fault Tolerance on Exia, Exib and Exic [23].

Level of protection	Zone of use	Fault Count
Ex ia	Zone 0	Safe with 2 countable faults
Ex ib	Zone 1	Safe with 1 countable fault
Ex ic	Zone 2	Safe with no countable faults (ie safe in normal operation)

5.2. Flameproof, Exd

- Permitted to use in Zone 1 and 2 only.
- The system needs to power down to carry out maintenance work. In other words, no live maintenance is allowed.
- Hot permit is required to carry out maintenance work.
- Wiring enters the Exd enclosure is via a flameproof cable gland or flameproof barrier.
- Typical applications will be rotating machines, lighting, switches (start/stop button) and junction boxes.

The flameproof protection technique relies on housing design to ensure an explosion or ignition inside the enclosure will not propagate outside the enclosure. It is achieved using different flame path designs. There are three possible flame path designs, as shown in Figure 14. The flame path construction plays a significant role. It creates a longer flame traveling distance so that the hazard gas generated during the internal explosion has sufficient time to cool down below the permitted AIT. The enclosure design also must be capable of withstanding the explosion's pressure and, at the same time, preventing the hazard gas from entering the enclosure and contact with the protected equipment inside the enclosure.



Figure 14: Types of Joint

The Exd enclosure usually is supplied in a complete assembly. The manufacturer or its approved agent does all the internal components assembly and wiring work. Both the manufacturer and the approved agent must have a certified facility and trained person to carry out the work. An approved test authority certifies the assembled system as a single entity as system approval (see section 6). Only a single system certificate will be issued. The test procedure for the approval include:

- Ventilation – Consider the available free air volume to ensure sufficient air circulation and ventilation after all the electrical components are fitted and wired.
- Temperature Classification – Consideration of the temperature rises to comply with the temperature classification and define the temperature class;
- Creepage and Clearance: Check all creepage and clearance distance requirements (see Figure 15 on the difference between creepage and clearance distance).
- Pressure Piling – Consideration of the rise in air pressure resulting from the internal enclosure explosion by considering the worst-case scenario on the gas-air mixture.

Future component replacement during maintenance should use back the initially designed component specification. If the replaced component has a larger or smaller dimension than the initially installed part, it will affect the enclosure's internal geometry and ventilation. When a larger than the original component is replaced, pressure piling could happen. If a smaller sized component is replaced, more free air volume will be generated. Meanwhile, the heat conductivity and heat dissipation characteristics of an object are affected by the geometry, ventilation and the material used. This will significantly affect the temperature classification rating. The initial temperature rating of the enclosure could be void.

Drilling and tapping the hole on the enclosure for fitting the cable gland, breather and conduit by a third-party technician is strictly prohibited or else the certification is considered void. The installation of these auxiliary components can only be carried out by the enclosure's manufacturer or certified installer. The type of threads, thread pitch and the clearance tolerance for a particular junction box must be compatible with the cable glands, breather or conduit as flame paths exist at these points.

To improve the ingress protection (IP rating), for example, for weatherproofing applications, non-hardening grease material or non-setting grease or compounds could be applied to the joint. Installation of gaskets to increase the ingress protection is not allowed if they are not part of the original design. It will affect the internal enclosure pressure release efficiency during the explosion. The gasket can only be replaced. Consult the manufacturer and check the installation manual on the type of gasket that could be used.

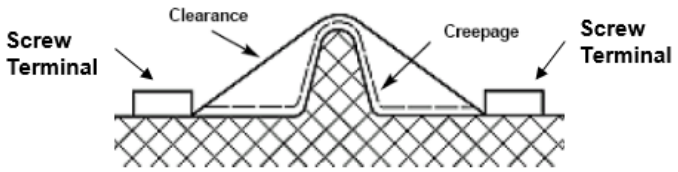


Figure 15: Different between Creepage and Clearance Distance

5.3. Increased Safety, Exe

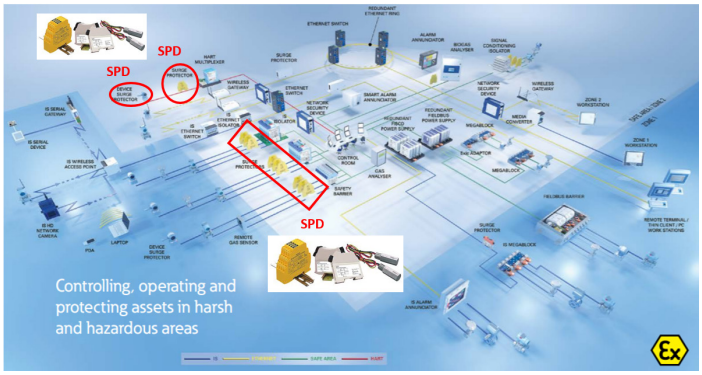
- Permitted to use in Zone 1 and 2 only.
- Required to turn off the power before carrying out maintenance work. In other words, live maintenance is not allowed.
- Minimum IP rating is IP54.
- No hot surface.
- Good mechanical strength (detail refer to EN50014).
- Typical applications will be in lighting, junction boxes and terminal blocks.

Increased Safety, Exe is an explosion protection technique applied to the hazardous area installation that ensures no hot spot or no excessive temperatures and no sparking occurrence from hazardous area electrical equipment and accessories. Equipment and the design that normally cause sparks are excluded from use within this method of explosion protection.

Exe relies on housing construction to withstand mechanical impact with a minimum of IP54 rating. The design of Exe ensures that under any operating mode, the apparatus does not contain normally arcing, sparking devices or hot surfaces that could cause ignition. This protection is realized by using high integrity insulation material. The designer also needs to consider the temperature de-rating factor of the insulation materials and enhanced terminal creepage and clearance distance (made wider than is generally the case in the general industry). The terminal design must be generously dimensioned for the intended cable connection and ensure that conductors are securely fastened to prevent arcing (loose connection can cause arcing). When designing the Exe protection method, a possible spark created by friction should be taken into consideration as well.

6. Method to Protect Assets in Hazardous Area

As shown in Figure 16, the mounting locations of the SPDs in a typical hazardous process plant.



6.1. Application Example on SPD in Exi Control Loop

- Permitted to use in Zone 0, 1 and 2.
- Maintenance work could be carried out with the power on. In other words, live maintenance is allowed
- No further certification is required when using "Uncertified Simple Apparatus" in combination with a certified safety barrier

There are two concepts of Exi applicable to intrinsically safe systems. They are “System Approval” and “Entity Concept”. The meaning of system approval indicates that a single certificate is issued by the approving authority such as BASEEFA or TUV, including the SPD, safety barrier and the field device, as shown in Figure 17 as the components inside the dotted line box. Uncertified control equipment, for example, DCS, PLC or ESD are installing before the safety barrier in the control room; they are not included in the system certificate. In fact, they do not require certification because the energy transmitted to the hazardous area in the field is already clamped down by the safety barrier when there is any fault that happened. Although there is no certification requirement for the uncertified control equipment, they are governed by the installation code, such as the requirement on cable segregation for Ex and Non-Ex cable; creepage and clearance distance for the terminal block; color code and so on. System certificate can only be issued after comprehensive testing and assessment by both the approval authority and the equipment manufacturer. Installation guidelines must be followed strictly according to the certificate/manual's instruction and wiring diagram provided. In the case where the fault happened, all replacement parts will have to be replaced with the exact original model used else; the certification will no longer be valid. This is not so flexible and, in some cases, will post difficulty when any of the parts are obsolete. When the exact replacement part is not available, then replacement on an entirely new set of the system will be required; otherwise, the newly installed component needs to be re-certified.



The most significant advantage of a system certification is to ensure that it is a proven and tested system. The approving authority and equipment manufacturer will be fully responsible for explosion protection in terms of safety. However, this led to the absence of flexibility in changing the component when it is discontinued. The responsibility of the installer is to follow strictly the installation instruction provided by the certificate. As for the entity concept, all intrinsically safe loop components, as shown in Figure 17 as the dotted line box, need to be individually tested and certified to use in Ex area. A separate certificate will be issued individually with all the safety parameters and other special clauses if any. It is the responsibility of the user to match the safety parameter, which will be described later. The utilization of the entity concept provides more flexibility in terms of parts selection, repair and equipment replacement. Regardless of the manufacturer, any product could be used from the safety perspective as long as the safety parameters are satisfied. However, in practice, the device can only be used if they fulfilled both safety and operational parameters. The operational requirement, such as it must have enough voltage and current to drive the field device.

Safety Parameter on Entity Concept

The entity safety parameters are described below:

$$Vmax \geq Voc, \quad (1)$$

Where Vmax is the maximum voltage specified by the field device manufacturer (tested by approval authority) that could be applied safely to the terminal of an intrinsically safe field device, Voc denotes the maximum open-circuit voltage (tested by approval authority) that could appear at the safety barrier’s output terminal when the fault occurs.

$$Imax \geq Isc, \quad (2)$$

Imax is the maximum input current specified by the field device manufacturer (tested by approval authority) that could be applied safely to the terminal of an intrinsically safe field device. Isc denotes the maximum output short-circuit current (tested by approval authority) that could be generated from the safety barrier's output terminal when a fault occurs.

$$Ci + Ccable \leq Ca, \quad (3)$$

Ci is the intrinsically safe field device's internal unprotected capacitance (tested by approval authority). Ccable denotes the cable capacitance.

Ca is the maximum external capacitance (individually tested by approval authority), including all devices at the field that could be connected safely to the hazardous terminal of the safety barrier.

$$Li + Lcable \leq La, \quad (4)$$

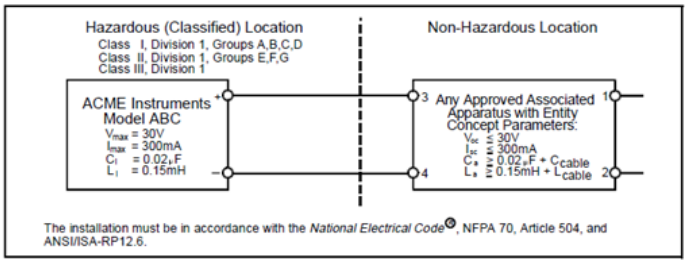
Li is the intrinsically safe field device's internal unprotected inductance (tested by approval authority). Lcable denotes the cable inductance.

La is the maximum external inductance (individually tested by approval authority), including all devices at the field that could be safely connected to the terminal of a safety barrier.

The value for Vmax, Voc, Imax, Isc, Ci, Ca, Li and La could be obtained from the individual intrinsically safe certificate issued by the approval authority. It is usually supplied together with the product, where Ccable and Lcable can be obtained from the cable manufacturer's datasheet.

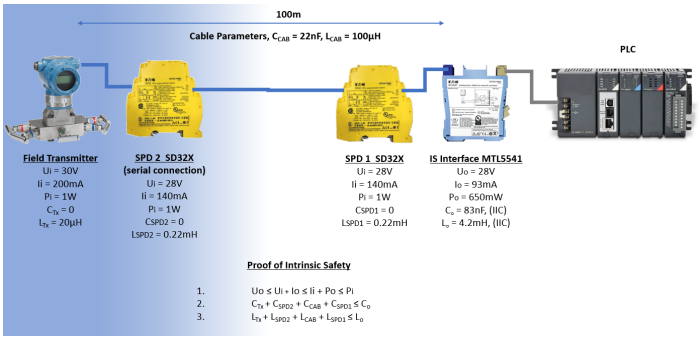
Only the system that fulfills all four mentioned equations, i.e., equations (1), (2), (3) and (4) could be used and safely installed according to the entity concept. However, the user also must take into consideration the operational requirements; for example, it should have sufficient voltage and current to drive the field equipment. Only systems that fulfill both operational and safety requirements will be implemented.

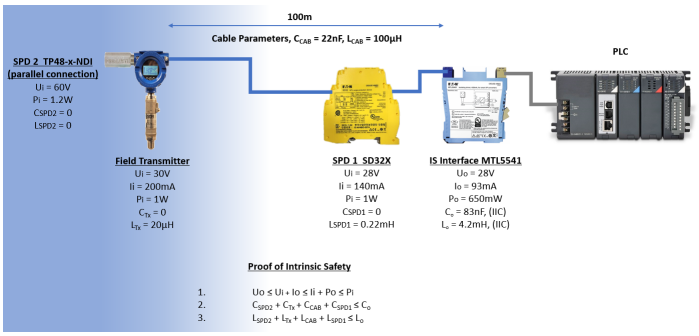
Entity parameters may be found on the control drawing supplied by the equipment manufacturer, along with other pertinent information regarding proper connections, conditions of use, etc. Figure 18 is an example of a control drawing for Exi equipment assessed based on the entity concept.



When an SPD is mounted in the safe area where no hazard substance is present, for example, in the control room, temperature classification and the type of hazard classification such as hazardous gas are not applicable; therefore, it will not be considered here. However, for field mounted SPD where hazard substance is present (Ex area); temperature classification and hazard classification need to be considered.

Figure 19 shows two SPDs connected in series, while Figure 20 shows one of the SPD, field mounted SPD connected in parallel and the other SPD (mounted in a safe area) is connected in series. For the case where all SPDs are mounted in the safe area, then there is no specific requirement for temperature and hazardous area classification. However, temperature and hazardous area classification must be taken into consideration if SPDs are to be mounted in a hazardous area. The hazardous area classification information of the individual Ex rated device can be found in the certifying authority's intrinsically safe certificate. Since all the SPDs in Figure 19 and Figure 20 are part of the Intrinsically Safe loop; therefore, it is a mandatory requirement to perform the entity safety parameter checking (equation (1) to (4)) to ensure its operational safety.





6.2. Application Example on SPD in Exd Control Loop

- Permitted to use in zone 1 and 2 only.
- Need to turn off the power before maintenance work is carried out. In other words, no live maintenance is allowed.
- Hot permit is required to perform any maintenance job.

SPD used in the Ex area is required to mount inside the Exd enclosure. The following information is required to determine the suitable size of Exd enclosure:

- To determine the enclosure's overall size, all the dimension on the individual component to be mounted inside the Exd enclosure is necessary. The system design for heat dissipation and ventilation on top of component size needs to put into the calculation as well.

- Temperature profile or T classification, heat or power dissipation information for the individual component is needed to identify the overall temperature classification of the Exd enclosure. In this case, if the temperature, heat or power dissipation information is not available, the system will need to perform temperature profiling to determine the maximum surface temperature by powered on for an extended period of time using a temperature probe. A bigger system may be required to power on for multiple hours until the temperature rise reaches a stable point where the measured temperature achieves equilibrium, stable and saturated.
- The manufacturer needs to know the type of hazard that could be present, for example, propane, ethylene or others, to determine the kind of joints and its design, either flanged joint, threaded joint or spigot joint. However, in most cases, it is the other way around; the manufacturer states their design is suitable for what kind of hazard. The user's choice is based on the hazard present in their application.
- Hazardous area mounting location or zoning: This is to determine the appropriate type of explosion protection technique. The flameproof (Exd) protection method can only be permitted to apply to Zone 1 or Zone 2 only.

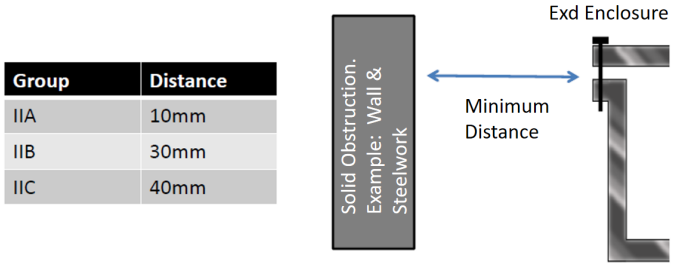


Figure 21: Obstruction of Flame Path (redrawn from [1] with additional information



Figure 22: Sample of the Exd Enclosure

During installation, it is crucial to ensure that there is no blockage such as a wall, brackets, conduit, weather guards or other electrical equipment near the opening of the flame path joint, as shown in Figure 21. The presence of such obstruction could reduce the efficiency of the flame path in the event of an internal explosion. Therefore, the minimum distance requirement between the flame path opening and the obstruction must be satisfied. Figure 21 provides the minimum distance requirement as recommended by IEC 60079-14.

Sample of the Exd enclosures fitted with various Ex and Non-Ex components are shown in Figure 22. It is also a common industrial practice that all the internal assembly and wiring work inside the Exd junction box is carried by the manufacturer or a certified center or an approved agent to ensure compliance.

7. Summary

When selecting SPD for hazardous area application, two criteria involve:

- Equipment protection

The parameters to be considered to protect the equipment from surge:

i. Types of SPD: SPD installation location correspondence to discharge surge energy. This is different from hazardous area classification. It is applied to the installation location for both Ex and non-Ex area.
ii. LPZ and LPL to determine the kA rating of the SPD.
iii. Let-through voltage (Voltage Protection Level, U_P): SPD should be selected where the let-through voltage is below the SWC of the device to be protected.
iv. Bandwidth: only require when it is used for the communication line.
v. Maximum continuous operating voltage, U_C: The U_C rating of the SPD should be selected slightly higher than the U_N (nominal operating voltage) of the equipment to be protected so that during normal operation, the SPD is transparent to the system.
vi. Nominal load current, I_L: For a series connection, the load current needs to be determined so that the SPD's current passing will not damage the SPD. The SPD must be able to withstand the continuous thermal current load. This only applies when the SPD is connected in series whereas SPD connected in parallel is able to handle unlimited load current.
vii. Series voltage drop: This mainly applicable for signal SPD where a series connection is applied. After adding the SPD, the user needs to ensure the voltage drop does not affect the circuit operation.

- Explosion protection

These are the factors that are nothing to do with the equipment protection from surge but rather to ensure the SPD itself does not become the source of ignition.

i. Industry Group: provides information on the type of industry (example: mining or surface industry) and the type hazard substance such as vapor, gas or dust that may be present, and its minimum ignition energy.
ii. Temperature Classification: provides information on the maximum surface temperature of the equipment at fault condition. The temperature classification of the equipment mounted in the Ex area should fall below AIT of the hazardous substances present. For the process loop application, only the SPD mounted at the field will have the temperature classification (Ex are). The other SPD mounted in the control room (non-Ex area) will not have the temperature classification.

iii. Zoning Classification provides information on the probability of a particular hazardous substance being present. The applicable type of explosion protection method depends on the Zoning information and the type of hazard present.

iv. Two approval concepts for safety evaluation could be adopted, entity or system approval. Entity Approval is a technique to combine individually approved apparatus into an intrinsically safe circuit. Two apparatus from different manufacture could be interconnected together as long as it could fulfill the safety calculation spell out by equation (1), (2), (3) and (4), no need for the two apparatus being tested in combination. System Approval is a technique to verify the intrinsic safety of a circuit in which all of the interconnected devices are examined all together as a complete system.

Conflict of Interest

No conflict of interest.

References

[1] T.H. Kuan, K.W. Chew, K.H. Chua, "Practical approach on surge protection device for Ex application," in 2019 IEEE International Conference on Electrical, Control and Instrumentation Engineering, ICECIE 2019 - Proceedings, 2019, doi:10.1109/ICECIE47765.2019.8974781.

[2] J. Kain, Severe Weather 101: Lightning FAQ, NOAA National Severe Storms Laboratory, Nov. 2020.

[3] MTL, Lightning surge protection for electronic equipment-a practical guide, Rev G, Measurement Tecnology Limited, Luton, 2016.

[4] M. Ahrens, LIGHTNING FIRES AND LIGHTNING STRIKES, National Fire Protection Association Fire Analysis and Research Division LIGHTNING FIRES AND LIGHTNING STRIKES, Quincy, 2013.

[5] A. Davis, Surge protector sparks fires instead of preventing them, homeowners say, Click2Houston.Com, 2015.

[6] K.F. Company, Kimberton Fire Company - Chester County, PA, Firehouse Solutions, 2020.

[7] P.˘ Stroch, "ScienceDirect Do not underestimate danger of explosion; Even dust can destroy equipment and kill," Perspectives in Science, **7**, 312–316, 2016, doi:10.1016/j.pisc.2015.11.048.

[8] MTL, Working Principle of Surge Protection Device, Luton, 2019.

[9] F. Martzloff, Surge Withstand Capability of Various Devices, Schenectady NY, 2019.

[10] F. Kloer, G. McDonald, N. Desai, TRICON v10 Nuclear Qualification Project: SURGE WITHSTAND TEST REPORT, Boxborough, Massachusetts, 2008.

[11] I.A. Metwally, A. Gastli, M. Al-Sheikh, "Withstand capability tests of transient voltage surge suppressors," Electric Power Systems Research, **77**, 859–864, 2007, doi:10.1016/j.epsr.2006.07.011.

[12] PHOENIX CONTACT | Information on NFPA 79, 2018 Editi, PHOENIX CONTACT GmbH & Co. KG, Blomberg, Germany, 2018.

[13] Guide to Surge Protection Devices (SPDs): selection , application and theory, BEAMA Ltd, London, 2018.

[14] IEC, INTERNATIONAL STANDARD Protection against lightning-Part 4: Electrical and electronic systems within structures, 2.0, INTERNATIONAL ELECTROTECHNICAL COMMISSION, 2010.

[15] N. Sclater, J.E. Traister, Surge protective categories defined by ANSI/IEEE C62.41, Electrical Engineering Portal, Nov. 2020.

[16] The basics of surge protection, PHOENIX CONTACT GmbH & Co. KG, Blomberg, Germany, 2016.

[17] C.P. Yelland, M.I. Lythe, Selection of explosion protected equipment for hazardous locations, EE Publishers, 2016.

[18] Temperature Classification (T class) - what it means, Control and Instrumentation.Com, Nov. 2020.

[19] IS MY TORCH CORRECTLY CERTIFIED BY ATEX?," LEARN ALL ABOUT ATEX DIRECTIVE 2014/34/EU, 2017, Peli Products Germany, Nov. 2020.

[20] I. Harrison, "HAZARDOUS AREA CLASSIFICATION," When is hazardous area classification needed?, Ex-Consulting, Nov. 2020.

[21] G. T., Understanding Hazardous Area Classification, Instrumentation Tools, Nov. 2020.

[22] Elster, Operating manual Volume conversion device EK205, Mainz-Kastel, Germany, 2016.

[23] P. Saward, Ex ic Intrinsic safety's new protection level, Luton, UK, 2016.

[24] T.H. Kuan, "INTRISIC SAFETY MADE EASY Title: INTRINSIC SAFETY MADE EASY," in Instrumentation & Control Society, ICS, Kuala Lumpur, 2006.

[25] MTL, Controlling, operating and protecting assets in harsh and hazardous areas MTL scope of supply, Measurement Technology Ltd, Luton, UK, 2017.

[26] T. Report, Intrinsically Safe System Assessment Using the Entity Concept, North Carolina, 1995.

[27] MTL, "Entity Concept Calculation," MTL Training Slide, Luton, UK, 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 770-775 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Method for Detecting Human Presence and Movement Using Impulse Radar

Young-Jin Park*, Hui-Sup Cho

Division of Electronics & Information System, DGIST, Daegu, 42988, South Korea

ARTICLE INFO

Article history:
Received:01 July, 2020
Accepted:24 August, 2020
Online: 28 August, 2020

Keywords:
UWB Impulse Radar
Machine Learning
Human Detecting
Noncontact measurement

ABSTRACT

Using non-invasive and non-contact sensors to measure a person's presence or movement helps improve the quality of life for both healthy people and patients. In this paper, a method of measuring the presence and motion of a person is proposed by utilizing UWB Impulse Radar, which is low power consumption and safe to radiate to the human body. The experimental stage of this study is divided into the stage of extracting features by signal processing from radar signals, the stage of generating datasets with 3~6 kinds of labels, and the stage of performing and verifying machine learning by imaging. In this experiment, a small number of images were used because only good quality signals were selected and used by radiating radar signals to the human body. The experiment result show high accuracy when using neural networks such as GoogLeNet and SqueezeNet. Experiments in this study confirmed that radar signals could be used to detect human presence and motion as a result of studies using the proposed method.

1. Introduction

This paper is an extension of the work originally presented in the 2019 International Conference on Information and Communication Technology Convergence (ICTC) [1].

Non-contact measurements are useful for health care, human activity, security, surveillance, etc., and the prevalence of COVID-19 in 2020 is a constant increase in demand for non-invasive, contactless measurements of physiological functions in modern society. Especially, non-contact sensors such as radar technology can be used to identify health conditions and movements without limiting human activities. In commercial applications, radar sensors are widely used for LED control and alarm monitoring in indoor and outdoor environments, as well as in smart homes and cities. A radar sensor can be applied in areas where there are restrictions on the use of sound, infrared, vibration, and camera sensors. For example, acoustic and vibration sensors are very vulnerable to acoustic noise, and infrared sensors frequently generate false alarms in outdoor environments. Moreover, camera sensors are relatively expensive, require high signal processing, have low performance at night, and have lens contamination problems. On the other hand, radar sensors are robust against weather conditions and their performance does not decrease at night. In addition, since signal processing is relatively easy and effective for detecting a target, it is possible to effectively utilize radar sensors in various indoor and outdoor environments [2]. Compared with continuous-wave radar systems, ultra-wideband (UWB) radars have localization capability, consume less power, and can monitor multiple subjects [3]. Radar technology is aimed at detecting targets in aviation and military areas. Recently, studies have been conducted to detect human bodies at close range, and to detect heart rate and respiration. UWB impulse radars are used to obtain biological information due to their low risk of exposure to electromagnetic waves and low power consumption. It has become an emerging technology for indoor localization and tracking. UWB radar has many advantages, including high spatial resolution, ability to mitigate interference, through-wall visibility, simple transceiver, and low cost [4].

In this study, radar signal processing and machine learning are applied in a system to detect human presence and movement. In addition, the machine learning data set utilizing most contactless sensors is generated as a data set by data processing techniques of various statistical calculation methods. However, this study does not use a statistical calculation method, but converts the signal-processed result into images to generate a data set, machine learning, and shows the result. As a method of research, data is received from radar in chronological order, and features are extracted through signal processing. The extracted feature data is divided into six labels, and machine learning is used to determine the existence and movement of the current person through different experiments for each label. As a result of the research, a total of three experiments were conducted: changing the epoch sizes and changing the composition of the label. In addition, even though

*Corresponding Author: Young-Jin Park, yjpark@dgist.ac.kr

this is the result of research on data sets using fewer images, the result is satisfactory in the research of converting radar signals into images and detecting the presence or absence of subject and movement.

In Section 3, the machine learning process is discussed by extracting features with a dataset configured through the preprocessing of radar signals. Section 4 presents the experimental results using several test sets. Finally, the paper is concluded in Section 5.

2. Related Works

The most common method is to use image sensors to detect the presence or movement of a subject. However, there are scenarios when this method cannot be used, such as conditions involving personal privacy infringement. In order to overcome problems, the UWB impulse radar signal was used in this study. Machine learning algorithms is used to determine and classify subject presence and movement. The features for the machine learning model are extracted using signals from the UWB impulse radar module, as shown in Figure 1. The impulse radar module has one transmitter (Tx) and receiver (Rx). The Tx sends very narrow pulses, and the Rx receives the reflected pulses. The received signal passes through several signal processing steps to extract the target signal. However, this target signal is generally perturbed by clutter, noise, and attenuation. Therefore, the removal of unwanted signals and signal compensation are crucial tasks for improving the detectability of a target [5]. The UWB impulse radar emits short pulses through antennas, and a radar transceiver [6] that digitizes pulses returning from the target using sampling methods is used in the experiment. A pulse with a Gaussian envelope and sine wave less than 0.4 ns in width has a 6.8 GHz frequency and a 2.3 GHz bandwidth. The sampler collection read from the radar chip in particular is called the frame, and the delay between individual samplers yields the equivalent sampling rate of the frame [7].

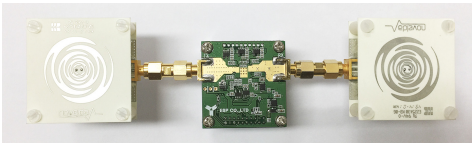


Figure 1: UWB impulse radar hardware module

In this study, the experiments receive frames sequentially from the impulse radar and the process of performing basic signal processing is the same. In the previous work [1], a feature set was created using the signal distribution by calculating standard deviation, root mean square (RMS), etc. However, in this study, the signal is processed in the frequency domain to generate an image, and the dataset used as input for machine learning is converted into an image. Furthermore, support vector machine (SVM) was used in the previous experiment to classify patterns. The major strength of SVM is that training is relatively easy and it has no local optimum, unlike neural networks. It scales relatively well to high-dimensional data, and the tradeoff between classifier complexity and error can be explicitly controlled. This weakness is mitigated by a good kernel function [8-13]. However, in this study, deep learning (GoogLeNet) was used for image classification rather than SVM. In 2014, Google published its network, GoogLeNet, to the imageNet large scale visual recognition challenge. Its performance (6.7%) is slightly better than that of VGGNet (7.3%). The main attractive feature of GoogLeNet is that it runs very fast due to the introduction of a new concept called the inception module, thus reducing the number of parameters to only 5 million, which is 12 times less than that of AlexNet. It also uses lower memory and power [14].

3. Method

3.1. Preprocessing

Figure 2 shows the overall flowchart of the proposed algorithm. First, a raw signal from the radar is collected to the frameset. A feature set is then made by extracting the signal characteristics through signal processing. Next, a dataset is created by extracting the characteristics based on each action label—no one, in front of the radar, moving in front of the radar, and so on—and then converting them into images to be used in machine learning.

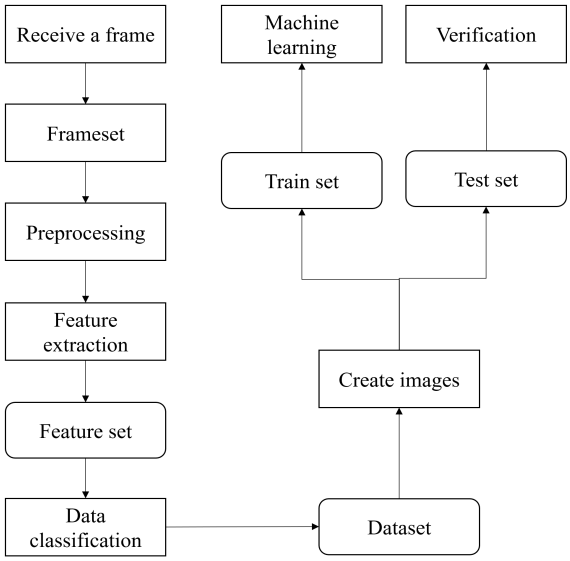
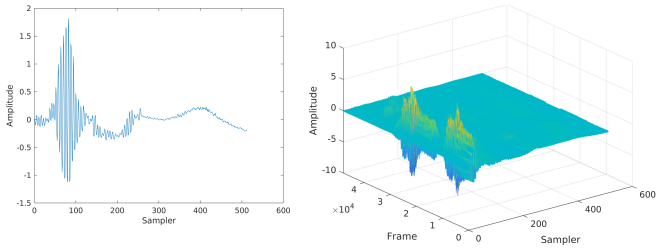


Figure 2: Flowchart

The created images are stored for each label and divided into train and test sets for machine learning and verification, respectively. The train and test sets are required when constructing models in machine learning. The training set is used to construct a model that gives a result that is close to the expected actual value, and the test set is used to check if the model constructed is reasonable. In other words, if a suitable prediction coefficient for the machine learning algorithm is found using the training set, the performance of the model can be verified with the test set. The results of the experiment suggest that the model can classify subject presence and movement using images created with radar signals.

In this study, one frame with 512 samplers is received every 50 ms and its appearance is shown in Figure 3(a). The frame set is a collection of frames accumulated in chronological order, and a single frame with a total of 48,000 frames (40 min.) used in this study is shown in Figure 3(b). For your information, three frames can be used for breathing and heart rate extraction, and can be used in various ways, such as identifying a person's path of travel.



(a) A frame (b) A frameset

Figure 3: Chronologically ordered a feature set

Note that Figure 4 shows the feature extraction process. The raw signal is a frame that is received directly from the radar. Remove the background, etc. of the signal and store it in chronological order in the frameset. In this experiment, frameset is used to extract characteristic information of signals by means of signal processing, such as frequency analysis, digital filtering, etc., and moreover, because the subject's position is fixed, all areas of frameset are not used.

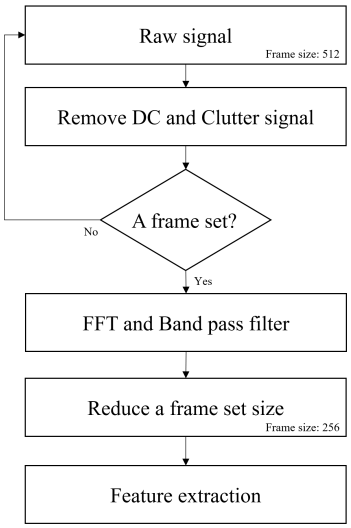
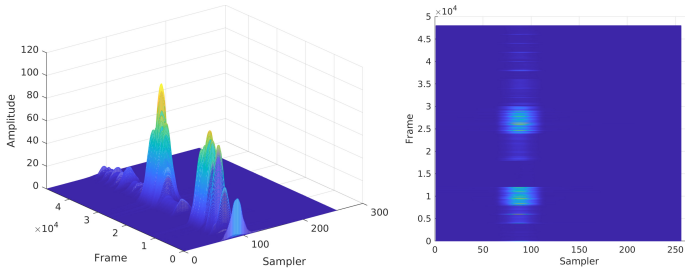


Figure 4: Chronologically ordered a feature set



(a) 3D view (b) Bird's-eye view

Figure 5: Chronologically ordered a feature set

The original radar signal contains noise and unstable signals, so accurate machine learning results cannot be obtained. Therefore, a feature set that can be used can be obtained only after the signal processing process. For reference, as shown in Figure 5, the feature set was used in half the size of a frameset that make up the frame, using only 256 samplers, or half of the total, to remove unnecessary data information. Figure 5(b) presents the strength of the overall signal when the signal is seen from a bird's eye view. The signal strength appears in six forms, which are mapped to the six labels to be used in the experiment. The x-axis represents samplers within the dataset, and the y-axis represents collected each frame.

Each axis in Figures 5 represents the x-axis representing the sampler, the y-axis representing the frame set and the z-axis representing the amplitude of the signal.

3.2. Machine Learning

The total size of the frame set is 24,576,000 double types. In this experiment, the subject is always in a fixed position (within one meter) and is located approximately 70 to 120. A region of interest (ROI) was set to be used to reduce the size and create a data set in the Figure 6, using only the data in a specific area. Therefore, only those areas are set as areas of interest, and the rest of the data is not needed. The data set for creating images using feature sets uses only the area of interest, reducing the amount of data to be used.

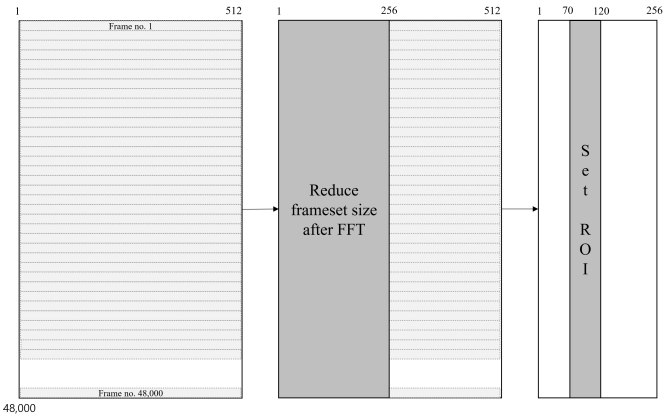


Figure 6: Reduce process of image size

In addition, the labels used to predict the condition of the subjects were divided into six, as shown in Table 1 below.

Table 1: Classification labels

Label	Amount of Data	Location of the subject
0	10,000	No one in front of the radar
1	10,000	Fix in front of the radar
2	10,000	Moving in front of the radar
3	6,000	Back
4	6,000	Left 30°
5	6,000	Right 30°

The dataset was created to generate images for machine learning by classifying the data of the same labels from scattered data within feature set. Due to the characteristics of radar signals, various behaviors (each label) are mixed, such that it was difficult to generate images for machine learning. Hence, a block consisting of data of the same label was created, and the size of the data for each label was compared, as shown in Figure 7. The x-axis represents six labels, and the y-axis represents the number of data within each label. In Figure 7, the 48,000-frame dataset has a complete configuration for each label, and six blocks are created with each label. This is because the data need to be classified under the same label to generate images for machine learning.

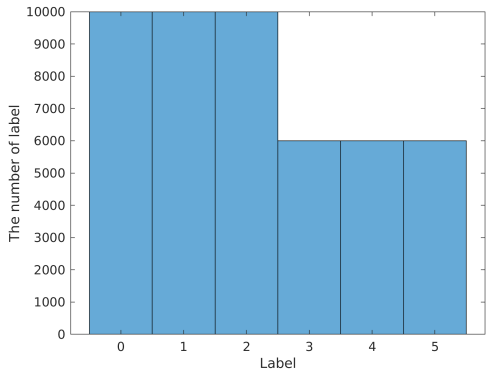


Figure 7: Composition of each label of a data set

Figure 8 shows the composition of the data shape for each label. The x-axis represents samplers of a specific area within the dataset with a corresponding label, and the y-axis represents the amplitude. This section describes the process of creating a dataset to generate an image required for machine learning. Particularly, the quality of the image depends on signal processing. This is the most important aspect because machine learning results vary depending on the quality of the image. The training set was used to implement the machine learning model, and the test set were used to test the model accuracy. If only the front part of the image, which represents the position of the actual subject, was converted into an image, a slightly faster performance is possible, but it is difficult to achieve noticeably better results. However, when experimenting with more images or with embedded equipment, it is necessary to reduce the size of the basic information of the images, even if they are damaged.

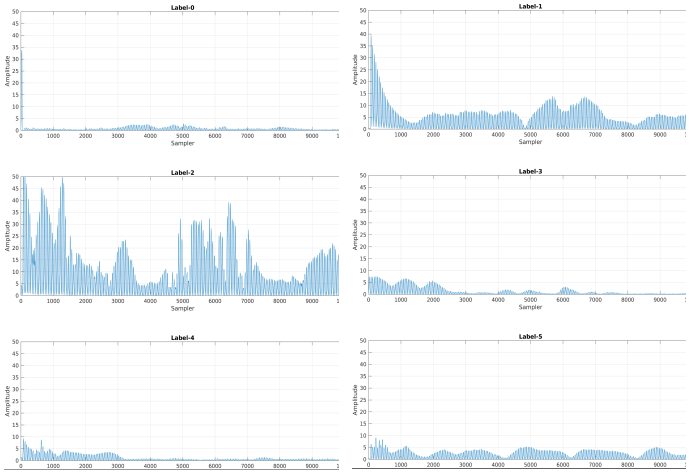


Figure 8: Reconfiguration dataset using labels

The experiments show that machine learning can determine the presence and movement of a person using images generated by radar signals. The current experimental setting and results are better than expected at the beginning of the experiment. However, if the situation changes, such as the distance and location of the radar, person, and different physical conditions of several subjects, the results can be significantly different. Therefore, it is necessary to collect more experimental data and to improve pretreatment algorithms to obtain cleaner images.

4. Experiment

In the preprocessing stage, the radar signals were processed to produce a basic feature set. This is the most important process because machine learning results may vary depending on the feature set produced in the preprocessing stage. Next, we created a dataset from the feature set using classification labels and then created images for machine learning. The images generated were divided into a training set for model construction and test set for verification. The generated images were also used as input data for machine learning. The experiments conducted in this study showed a high level of classification accuracy. However, datasets pretreated only by signal classification seemed to be quite sensitive to the quality of the data when used as features for machine learning. The algorithm was implemented in hardware and software environments, as summarized in Table 2.

Table 2: System environment

Environment	Specification and Version
Processor	Intel i7-8700k
Memory	32 GB
GPU	Single Nvidia Titan Xp
Operating System	Ubuntu 18.04 LTS
Language	Matlab (2019b)
Radar module	NVA6201 (Novelda)

The algorithm proposed was used to generate an image using the dataset created in the previous step. The image was then stored separately by label. During the machine learning process, the images were loaded and learned, and the results were shown. The experiment uses frames received for 40 minutes using a total of 48,000 signals received every 50 milliseconds. The total number of images created using the proposed algorithm is 240 and the train set and test set are used in machine learning at a ratio of 8:2. The images for machine learning were converted into a spectrogram, which combines the waveform and spectrum features to visualize the wave, using the hamming window method. Furthermore, an RGB (Red, Green, Blue) image with an actual size of 224×224×3 was created and stored by label. The created image appeared in the form shown in Figure 12, and features obtained by signing the radar's signal at the front of the image exist using spectrogram that is a visual tool of representing the signal length. Meanwhile, features that appear in front of the radar are required. However, the images were not randomly cut or resized for the experiment. When more than one person is present in future studies or multiple radars are used, new targets appear in the blank area as they are currently being signaled. Therefore, the original image was used because the accuracy could vary depending on whether signal processing was performed or not.

Table 3 summarize experimental environment and accuracy. For experiment- 1, it is a measure of how well six labels measured using radar can be distinguished. In experiment-2, label-4 and label-5 are experiments on subjects in fixed positions, such as Label 1. However, since only the angle of measurement is different, label-4 and label-5 can be integrated with label-1. In other words, it is an experiment that combines data set where similarities exist into a single family of labels. In experiment-3, the simplest experiment was conducted, and although the actual result

is 100%, the accuracy will be decreased as the number of images increases. Depending on the experimental conditions, it can be confirmed that the results of the three experiments vary depending on the composition of the label. As mentioned above, label-1, label-4, and label-5 in experiment-2 are datasets for subject who have different angles but are not moving. Therefore, in experiment-2, we used the image of having three labels on label-1.

Table 3: Experiment condition and accuracy

	Experiment-1	Experiment-2	Experiment-3
Network	GoogLeNet		
Image size	224×224×3		
Train set	192		
Test set	48		
Epochs	32		
Used label Types	0, 1, 2, 3, 4, 5	0, 1(include 4 and 5), 2, 3	0, 1, 2
Accuracy	91.7%	97.6%	100%

The Figure 9, Figure 10 and Figure 11 show a graph that monitors train and validation for machine learning, and the graph at the top represents accuracy, and the graph at the bottom represents loss. The x-axis represents iteration of machine learning, and the y-axis represents the accuracy (%) and loss rate. The black dotted line is the validation result for the actual train set. Figure 7 shows result of the first experiment with GoogLeNet using six labels.

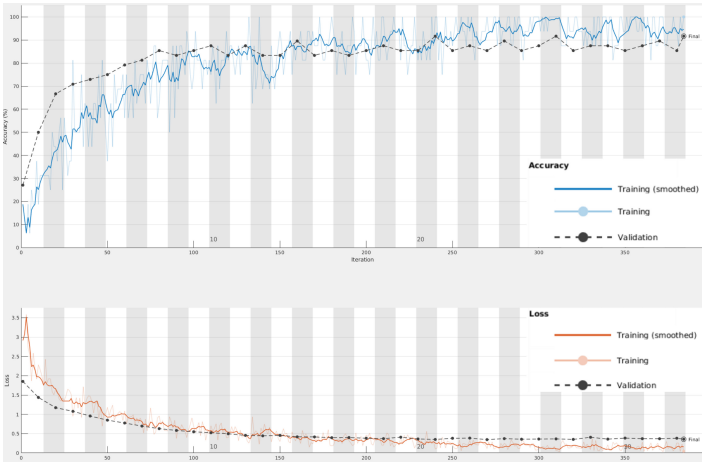


Figure 9: Accuracy and loss function for the experiment–1

Figure 10 shows the result of second experiment performed in a similar manner as the experiment-1. In the experiment, classification by label was slightly simplified. In the case of label-0 and label-2, no change was observed. For label-4 and label-5, radiated radar pulses from the left and right sides of the subject were integrated into the radiated radar pulses from the front of the subject (label-1). Label-3, which radiates radar signals from the back of the subject, including poor results from previous experiments, was excluded from the experiment and machine learning was conducted.

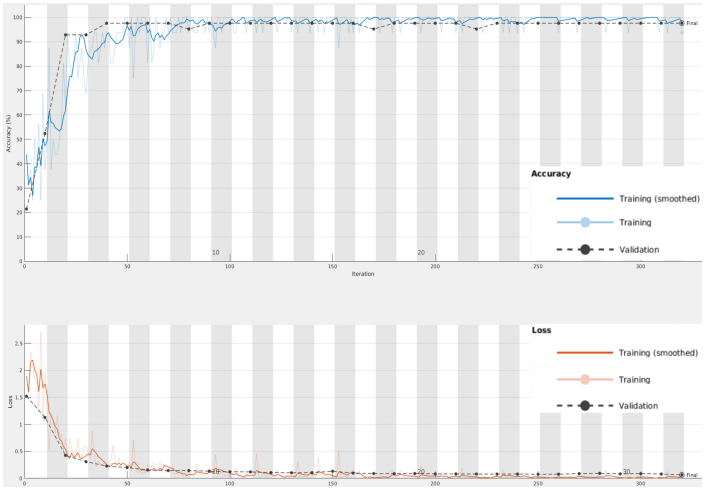


Figure 10: Accuracy and loss function for the experiment–2

Figure 11 shows the machine learning result using only label-0, label-1, and label-2. The results show that label-3 is difficult to recognize. Furthermore, label-4 and label-5 had high predicted rates when integrated into label-1, indicating that these three labels may have similar features.

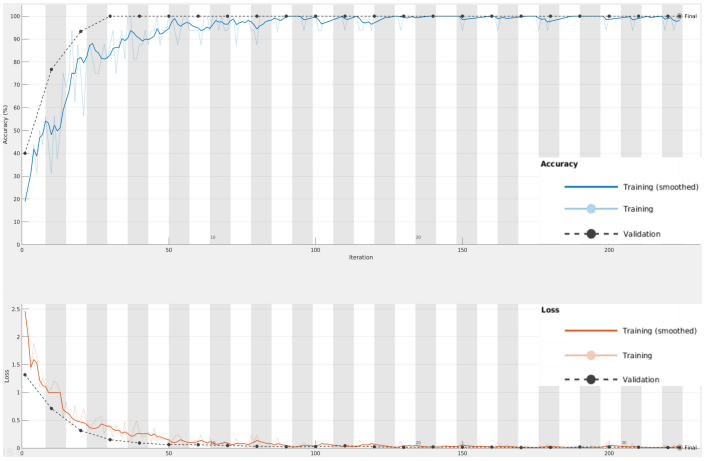


Figure 11: Accuracy and loss function for the experiment–3

Figure 12 shows the result of some validated images for the machine learning model, with high predictive results for label-0 indicating when no one is present and for label-1 and label-2, when a person are located in front of the radar. For label-3, the radar pulse was emitted by the back of the subject, but this was insufficient to judge the presence of a subject. Label 4 and label-5 is the data that is changed only from left to right under the same conditions as label 1, and high predictions can be seen. Information about each label can be found in Table 1. In fact, the use of images representing signal strength within frequencies conducted in this study alone cannot determine whether the subject is human. Therefore, in the future, research will be needed to process biometric signals such as respiration at the same time to determine if they are human or not.

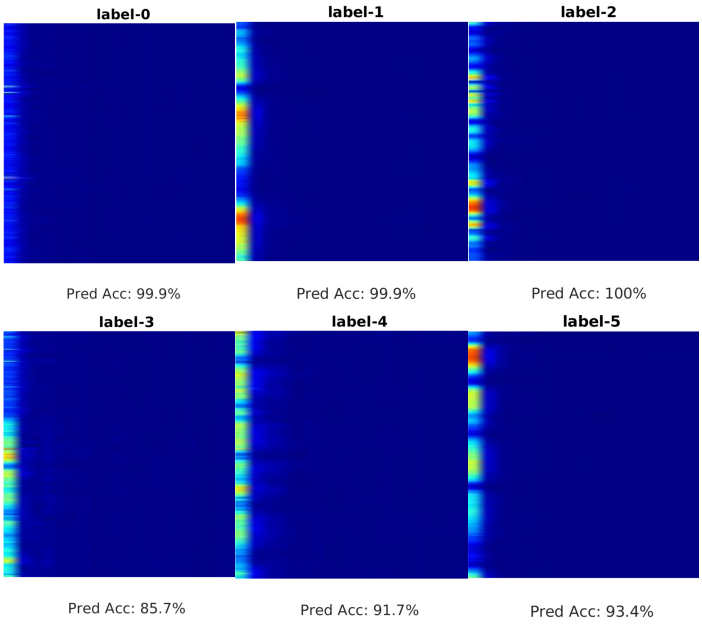


Figure 12: Result images created by the experiment-1

In addition, experiments using other neural networks like Resnet series and SqueezeNet were conducted. In fact, in the case of SqueezeNet, it has the advantage of learning speed, similar to GoogLeNet, and the results were similar. However, experiments using the Resnet series (Resnet18, Resnet50 and Resnet101) of neural networks with more network layers to improve accuracy showed that the results were rather lower at around 80 percent. The reason is thought to be the result of poor learning, passing through many layers because of the small number of images used in the experiment. In this experiment, a total of 240 images (train set: 192 test set: 48) are used. Although much more data was collected than was used in this experiment, many data were discarded due to errors and inaccuracies in the data. Therefore, it is not easy to filter out quality data by analyzing radar signals obtained by emitting them to living things. In order to have higher accuracy, it is necessary to have the integrity of the radar signal and to obtain at least five times the image used in this experiment.

5. Conclusion

Research and development using machine learning is ongoing worldwide, particularly in the prediction of subject activities. In this study, subject presence and movement were determined via machine learning, using an ultra-wideband impulse radar signal. Using a set of data imaged with electromagnetic signals in machine learning, experimentation using several neuronal networks resulted in high classification prediction results. An analysis of human-to-electromagnetic interactions shows that UWB impulse radars have low risk associated with human exposure. Therefore, in this paper, the UWB Impulse Radar signal was released to the human body and processed and imaged the signal. In addition, a new method was proposed to detect subject presence and motion by utilizing the created images as data sets of machine learning. Experiments have shown that subject presence and motion are sufficiently possible to be detected. Based on the results, it is judged that if the proposed method is further developed, it can be used in a variety of applications, including medical care and crime prevention.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the DGIST R&D Program of the Ministry of Science, ICT, and Technology of Korea (20-IT-09).

References

[1] Y. J. Park, H. S. Cho, “An experiment of human presence and movement using impulse radar and machine learning”, 2019 International Conference on Information and Communication Technology Convergence (ICTC), 2019, doi:10.1109/ICTC46691.2019.8939858.
[2] J. Kwon, S. Lee, N. Kwak, “Human detection by deep neural networks recognizing micro-doppler signals of radar”, in 15th European Radar Conference (EuRAD), 26-28, 2018, doi:10.23919/EuRAD.2018.8546605.
[3] L. Ren, Y. S. Koo, Y. Wang, A. E. Fathy, “Noncontact heartbeat detection using UWB impulse Doppler radar”, in Proceedings of the IEEE Topical Conference on Biomedical Wireless Technologies, Networks, and Sensing System, 14–16, 2015, doi:10.1109/BIOWIRELESS.2015.7152128.
[4] R. J. Fontana, “Recent system applications of short-pulse ultra-wideband (UWB) technology”, IEEE Trans. Microw. Theory Tech., 52, 2087-2104, 2004, doi:10.1109/TMTT.2004.834186.
[5] V. H. Nguyen, J. Y Pyun, “Location detection and tracking of moving targets by a 2D IR-UWB radar system”, Sensors **15**(3):6740-62, 2015, doi:10.3390/s150306740.
[6] Novelda, NVA620x Preliminary Datasheet, Novelda AS, Oslo, Norway, 2013.
[7] Y. J. Park, H. S, Cho, “Detecting human movement during vital sign measurement using heterogeneous sensors”, ITC-CSCC2019, 2019, doi:10.1109/ITC-CSCC.2019.8793461.
[8] V. Jakkula, “Tutorial on Support Vector Machine (SVM)”, School of EECS, Washington State University, Pullman 99164.
[9] Tutorial slides by Andrew Moore. http://www.cs.cmu.edu/~awm
[10] C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, **2**, 121-167, Boston, 1998, doi: 10.1023/A:1009715923555.0.
[11] N. Cristianini, J. Shawe-Taylor, “An Introduction to Support Vector Machines and other Kernel-based Learning Methods”, Cambridge University Press, 2000, doi:10.1017/CBO9780511801389.
[12] J. P. Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.
[13] Tutorial slides by Andrew Moore. http://www.cs.cmu.edu/~awm
[14] M. Sewak, Md. R. Karim, P. Pujari, “Practical Convolutional Neural Networks - GoogLeNet architecture”, 2018, https://subscription.packtpub.com/book/data/9781788392303/4/ch04lvl1sec31/googlenet-architecture.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No.5, 770-780 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Using Classic Networks for Classifying Remote Sensing Images: Comparative Study

Khalid A. AlAfandy[*,1], Hicham Omara[2], Mohamed Lazaar[3], Mohammed Al Achhab[1]

[1]*ENSA, Abdelmalek Essaadi University, Tetouan, 93002, Morocco*

[2]*FS, Abdelmalek Essaadi University, Tetouan, 93002, Morocco*

[3]*ENSIAS, Mohammed V University in Rabat, 10000, Morocco*

ARTICLE INFO

Article history:
Received: 12 August, 2020
Accepted: 25 September, 2020
Online: 12 October, 2020

Keywords:
Remote Sensing Images
ResNet
VGG
DenseNet
CNN
NASNet
Deep Learning
Machine Learning

ABSTRACT

This paper presents a comparative study for using the deep classic convolution networks in remote sensing images classification. There are four deep convolution models that used in this comparative study; the DenseNet 196, the NASNet Mobile, the VGG 16, and the ResNet 50 models. These learning convolution models are based on the use of the ImageNet pre-trained weights, transfer learning, and then adding a full connected layer that compatible with the used dataset classes. There are two datasets are used in this comparison; the UC Merced land use dataset and the SIRI-WHU dataset. This comparison is based on the inspection of the learning curves to determine how well the training model is and calculating the overall accuracy that determines the model performance. This comparison illustrates that the use of the ResNet 50 model has the highest overall accuracy and the use of the NASNet Mobile model has the lowest overall accuracy in this study. The DenseNet 169 model has little higher overall accuracy than the VGG 16 model.

1. Introduction

With rapid development of communication sciences especially satellites and cameras, the remote sensing images appeared with the importance of processing and dealing with this type of images (remote sensing images). One of these important image processing is classification which done by machine learning technology. Machine learning is one of the artificial intelligence branches that based on training computers using real data which result that computers will have good estimations as an expert human for the same type of data [1]. In 1959 the machine learning had a definition that "The machine learning is the field of study that gives computers the ability to learn without being explicitly programmed", this definition was made by Arthur Samuel [2]. Where the declaration of machine learning problem was known as "a computer program is said to learn from experience E with respect to some task T and some performance measure P", this declaration was made by Tom Mitchell in 1998 [3]. Deep learning is a branch of machine learning that depends on the Artificial Neural Networks (ANNs) which can be used in the remote sensing images classification [4]. In the recent years, with the appearance of the latest satellites versions and its updated cameras with high spectral and spatial resolution, the very high resolution (VHR) remote sensing images appeared. The redundancy pixels in the VHR remote sensing images can cause an over-fitting problem through training process where used the ordinary machine learning or ordinary deep learning in classification. So it must optimize the ANNs and have a convenient feature extraction from remote sensing images as a preprocessing before training the dataset [5-6]. The convolution Neural Networks (CNNs) is derived from the ANNs but its layers are not fully connected like the ANNs layers; it has an excited rapid advance in computer vision [7]. It is based on some blocks can applied on an image as filters and then extract convolution object features from this image, these features can be used in solving many of computer vision problems, one of these problems is classification [8]. The need of processing the huge data, which appeared with the advent of the VHR remote sensing images and its rapid development, produced the need of CNNs deep architectures that can produce a high accuracy in classification problems. It caused the appearance of the classic networks. There are many classic networks that are mentioned in the research papers which are established by researchers. However, this paper will inspect four of the well-known classic networks; the DenseNet

[*]Corresponding Author: Khalid A. AlAfandy, +212635549566, khalid_yuosif@yahoo.com

169, the NASNet Mobile, the VGG 16, and the ResNet 50 models. These four network models are used in the classification researches for the remote sensing images in many research papers. In [9], authors used the DenseNet in their research to propose a new model for improved the classification accuracy. In [10], authors used the DenseNet model to build dual channel CNNs for hyper-spectral images feature extraction. In [11], authors proposed a convolutional network based on the DenseNet model for remote sensing images classification. They build a small number of convolutional kernels using dense connections to produce a large number of reusable feature maps, which made the network deeper, but did not increase the number of parameters. In [12], authors proposed a remote sensing image classification method that based on the NASNet model. In [13], authors used the NASNet model as a feature descriptor which improved the performance of their trained network. In [14], authors proposed the RS-VGG classifier for classifying the remote sensing images which used the VGG model. In [15], authors proposed a combination between the CNNs algorithms outputs, one of these algorithms outputs is the outputs of the VGG model, and then constructed a representation of the VHR remote sensing images for resulting VHR remote sensing images understanding. In [16], authors used the pre-trained VGG model to recognize the airplanes using the remote sensing images. In [17], authors performed a fully convolution network that based on the VGG model for classifying the high spatial resolution remote sensing images. They fine-tuned their model parameters by using the ImageNet pre-trained VGG weights. In [18], authors proposed the use of the ResNet model to generate a ground scene semantics feature from the VHR remote sensing images, then concatenated with low level features to generate a more accurate model. In [19], authors proposed a classification method based on collaborate the 3-D separable ResNet model with cross-sensor transfer learning for hyper-spectral remote sensing images. In [20], authors used the ResNet model to propose a novel method for classifying forest tree species using high resolution RGB color images that captured by a simple grade camera mounted on an unmanned aerial vehicle (UAV) platform. In [21], authors proposed an aircraft detection methods based on the use of the deep ResNet model and super vector coding. In [22], authors proposed a remote sensing image usability assessment method based on the ResNet model by combining edge and texture maps.

The aim of this paper is to compare the using of the deep convolution models classic networks in classifying the remote sensing images. The used networks in this comparison are the DenseNet 196, the NASNet Mobile, the VGG 16, and the ResNet 50 models. This comparative study is based on inspecting the learning curves for the training and validation loss and training and validation accuracy through training process for each epoch. This inspection is done to determine the efficient of the model hyper-parameters selection. It is based also on calculating the overall accuracy (OA) of these four models in remote sensing images classification to determine the learning model performance. There are two use datasets in this comparative; the UC Merced Land use dataset and the SIRI-WHU dataset.

The rest of this paper is organized as follow. Section 2 gives the methods. The experimental results and setup are shown in section 3. Section 4 presents the conclusions followed by the most relevant references.

2. The Methods

In this section the used models, in this comparative study, will explained with its structures. The classic networks that used in this study will illustrated in brief as literature review, ending with how to assess the performance of the learning models.

2.1. The Used Models

The feature extraction of remote sensing images is provided an important basis in remote sensing images analysis. So, in this study the deep classic convolution networks outputs considered as the main features that extracted from the remote sensing images. In these four networks, we used the ImageNet pre-trained weights because the train of new CNNs models requires a large amount of data. We transfer learning, add full connected (FC) layers with the output layer containing neurons number that equal the dataset classes number (21 for the UC Merced land use dataset and 12 for the SIRI-WHU dataset), and then train these (FC) layers.

In the DenseNet 169 model we transfer learning to the last hidden layer before the output layer (has 1664 neurons) and get the output of this network with the ImageNet pre-trained weights, then adding FC layer (output layer) with softmax activation.

In the NASNet Mobile model we transfer learning to the last hidden layer before the output layer (has 1056 neurons) and get the output of this network with the ImageNet pre-trained weights, then adding FC layer (output layer) with softmax activation.

In the VGG 16 model we transfer learning to the output of last max pooling layer (has shape 7, 7, 512) in block 5 that before the first FC layer and get the output of this network with the ImageNet pre-trained weights, then adding FC layer (output layer) with softmax activation.

In the ResNet 50 model we transfer learning to the last hidden layer before the output layer (has 2048 neurons) and get the output of this network with the ImageNet pre-trained weights, then adding FC layer (output layer) with softmax activation.

2.2. The Convolution Neural Networks (CNNs)

The CNNs are taken from the ANNs with exception that it is not fully connected layers. The CNNs are the best solution for computer vision which based on some of filters to reduce the image height and width and increase the number of channels together, then processing the output with full connected neural network layers (FCs) which reduce the input layer neurons, reduce training time, and increase the training model performance [23-25]. These filters values are initialized with many random functions which can be optimized.

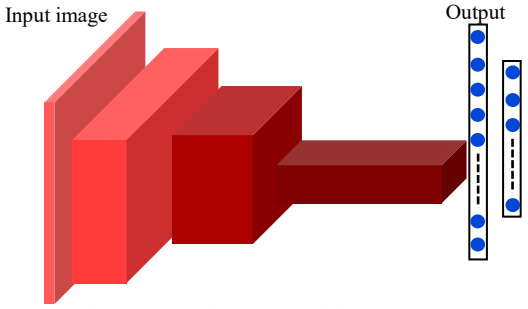


Figure 1: One of a CNNs model structures [25]

The filters design is based on the use of self-intuition as with ANNs structure design and learning hyper-parameters coefficient choice, which increasing the difficulty to reach the best solution for the learning problems [25]. Figure 1 shows an example of a CNNs model structure [24].

2.3. The DenseNet Model

In 2017, the DenseNets was proposed in the CVPR 2017 conference (Best Paper Award) [26]. They started from attempted to build a deeper network based on an idea that if the convolution network contains shorter connections between its layers close to the input and those close to the output, this deep convolution network can be more accurate and efficient to train. Other than the ResNet model which adds a skip-connection that bypass the nonlinear transformation, the DenseNet add a direct connection from any layer to any subsequent layer. So the lth layer receives the feature-maps of all former layers x_0 to x_{l-1} as (1) [26].

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

where $[x_0, x_1, \dots, x_{l-1}]$ refers to the spectrum of the feature-map produced in the layers $0, 1, 2, \dots, l-1$.

Figure 2 shows the 5-layers dense block architecture and Table 1 shows the DenseNet 169 architectures for ImageNet. [26].

Table 1: the DenseNet 169 model architectures for ImageNet [26]

Layers	Output Size	DenseNet 169
Convolution	112×112	7×7 conv, stride 2
Pooling	56×56	3×3 max pool, stride 2
Dense Block (1)	56×56	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times 6$
Transition Layer (1)	56×56	1×1 conv
	28×28	2×2 average pool, stride 2
Dense Block (2)	28×28	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times 12$
Transition Layer (2)	28×28	1×1 conv
	14×14	2×2 average pool, stride 2
Dense Block (3)	14×14	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times 32$
Transition Layer (3)	14×14	1×1 conv
	7×7	2×2 average pool, stride 2
Dense Block (4)	7×7	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times 32$
Classification Layer	1×1	7×7 global average pool
	1000	1000D fully-connected, softmax

2.4. The NASNet Model

In 2017, the NASNet (Neural Architecture Search Net) model was proposed in the ICLR 2017 conference [27]. They used the recurrent neural networks (RNNs) to generate the model description of the neural networks and trained the RNNs with reinforcement learning to improve the accuracy of the generated architectures on a validation set. The NASNet model is based on indicting the previous layers that elected to be connected by adding an anchor point which has N-1 content-based sigmoid using (2). Figure 3 shows one block of a NASNet convolutional cell [27].

$$P(Layer\ j\ is\ an\ input\ of\ layer\ i) = sigmoid(v^T \tanh(W_{prev} * h_j + W_{curr} * h_i)) \quad (2)$$

where h_j refers to the hidden state of the controller at anchor point for the jth layer and $j = [0, 1, 2, \dots, N-1]$.

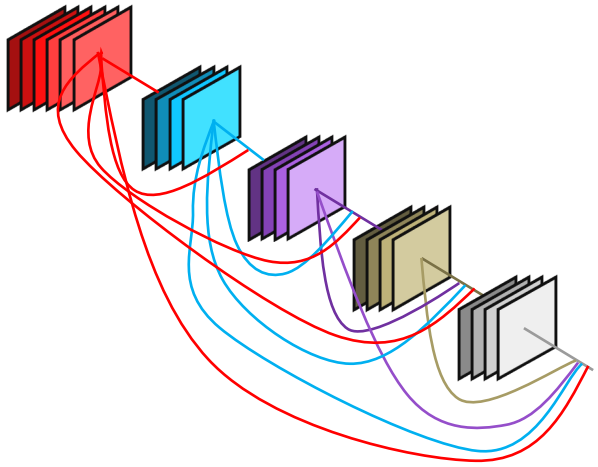


Figure 2: The 5 layers Dense block architecture [26]

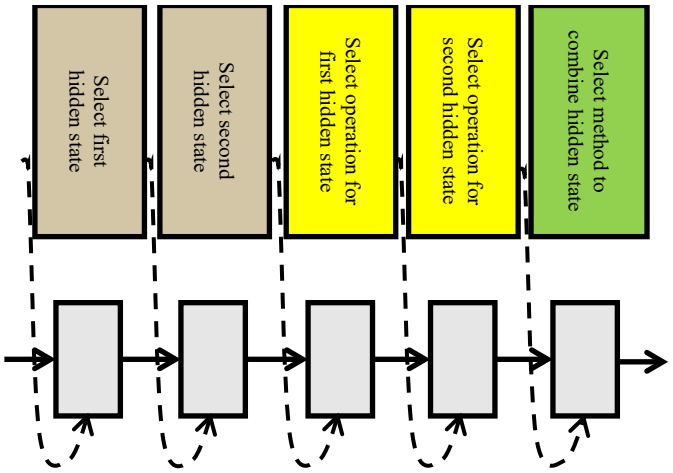


Figure 3: One block of a NASNet convolutional cell [27]

2.5. The VGG Model

In 2015, the VGG network was proposed in the ICLR 2015 conference [28]. They began with investigated the effect of the convolution network depth on its accuracy in the large-scale images. Through they evaluated the deeper networks architecture using (3×3) convolution filters, they showed that an expressive growth on the prior-art configurations can be achieved by pushed the depth to 16-19 weight layers. Table 2 shows the VGG 16 model architectures for the ImageNet [28].

Table 2: The VGG 16 model architectures for ImageNet [28]

Block	Layers	Output Size	VGG 16
Input		224×224×3	
Block 1	Convolution	224×224×64	3×3 conv 64, stride 1

	Convolution	224×224×64	3×3 conv 64, stride 1
	Pooling	112×112×64	2×2 max pool, stride 2
Block 2	Convolution	112×112×128	3×3 conv 128, stride 1
	Convolution	112×112×128	3×3 conv 128, stride 1
	Pooling	56×56×128	2×2 max pool, stride 2
Block 3	Convolution	56×56×256	3×3 conv 256, stride 1
	Convolution	56×56×256	3×3 conv 256, stride 1
	Convolution	56×56×256	3×3 conv 256, stride 1
	Pooling	28×28×256	2×2 max pool, stride 2
Block 4	Convolution	28×28×512	3×3 conv 512, stride 1
	Convolution	28×28×512	3×3 conv 512, stride 1
	Convolution	28×28×512	3×3 conv 512, stride 1
	Pooling	14×14×512	2×2 max pool, stride 2
Block 5	Convolution	14×14×512	3×3 conv 512, stride 1
	Convolution	14×14×512	3×3 conv 512, stride 1
	Convolution	14×14×512	3×3 conv 512, stride 1
	Pooling	7×7×512	2×2 max pool, stride 2
FC		4096	
FC		4096	
Output		1000, softmax	

The VGG network is a deeper convolution network that trained on the ImageNet dataset. The input images of this network is (224×244×3). This network consists of five convolution blocks, each block is containing convolution layers and pooling layer, then ending with two FC layers (each layer has 4096 neurons) then the output layer with softmax activation. The VGG network doesn't contain any layers connections or bypasses such as the ResNet, the NASNet or the DenseNet models, and at the same time gives high classification accuracy with the large scale images [28].

2.6. The ResNet Model

In 2016, the ResNet was proposed in the CVPR 2016 conference [29]. They combined the degradation problem by introducing a deep residual learning framework. Instead of intuiting each few stacked layers directly fit a desired underlying mapping. The ResNet is based on skip connections between deep layers. These skip connections can skipping one or more layers. The outputs of these connections are added to the outputs of the network stacked layers as (3) [29].

$$H(x) = F(x) + x \quad (3)$$

where H(x) is the final block output, x is the output of the connected layer, and F(x) is the output of the stacked networks layer in the same block.

Figure 4 shows the ResNet one building block. Tables 3 shows the ResNet 50 network architectures for the ImageNet [29].

Table 3: the ResNet 50 model architectures for ImageNet [29]

Layers	Output Size	ResNet 50
Conv 1	112×112	7×7 conv 64, stride 2
Conv 2_x	56×56	3×3 max pool, stride 2
		$\begin{bmatrix} 1\times1 \text{ Conv } 64 \\ 3\times3 \text{ Conve } 64 \\ 1\times1 \text{ Conv } 256 \end{bmatrix} \times 3$
Conv 3_x	28×28	$\begin{bmatrix} 1\times1 \text{ Conv } 128 \\ 3\times3 \text{ Conv } 128 \\ 1\times1 \text{ Conv } 512 \end{bmatrix} \times 4$
Conv 4_x	14×14	$\begin{bmatrix} 1\times1 \text{ Conv } 256 \\ 3\times3 \text{ Conve } 256 \\ 1\times1 \text{ Conv } 1024 \end{bmatrix} \times 6$
Conv 5_x	7×7	$\begin{bmatrix} 1\times1 \text{ Conv } 512 \\ 3\times3 \text{ Conve } 512 \\ 1\times1 \text{ Conv } 2048 \end{bmatrix} \times 3$
Classification Layer	1×1	7×7 global average pool
	1000	1000D fully-connected, softmax

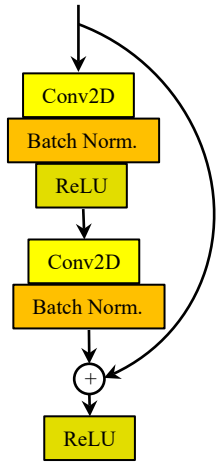


Figure 4: The ResNet building block [29]

2.7. The Performance Assessment

There are many matrices for gauge the performance of the learning models. One of these is the OA. The OA is the main classification accuracy assessment [30]. It is measure the percentage ratio between the corrected estimation test data objects and all the test data objects in the used dataset. The OA is calculated using (4) [30, 31].

$$\mathrm{OA} = \frac{Number\ of\ Correctly\ Estimations}{Total\ number\ of\ Test\ Data\ Objects} \times 100 \qquad (4)$$

The learning curves for the loss and the accuracy are very important indicator for determining the power of learning models through training process by using the corrected hyper-parameters [32]. By using these curves, you can determine if a problem exist on your learning model such as the over-fitting or the under-fitting problems. These curves represent the calculation of the loss and the accuracy values for the learning model at each epoch through the training process using training and validation data [32, 33].

3. Experimental Results and Setup

This comparative study is based on calculating the OA and plotting the learning curves to determine the power of the used hyper-parameter to achieved the better learning model performance. The UC Merced Land use dataset and the SIRI-WHU dataset are the used datasets in this study. The details of these datasets will introduce in this section and then the experiments setup details and the results.



Figure 5: Image examples from the 21 classes in the UC Merced land use dataset [34]

3.1. The UC Merced Land Use Dataset

The UC Merced Land use dataset is a collection of remote sensing images which has been prepared in 2010 by the University of California, Merced [34]. It consists of 2100 remote sensing images divided into 21 classes with 100 images for each class. These images were manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the USA. Each image in this dataset is Geo-tiff RGB image with 256×256 pixels resolution and 1 square foot (0.0929 square meters) spatial resolution [34]. Figure 5 shows image examples from the 21 classes in the UC Merced land use dataset [34].

3.2. The SIRI-WHU Dataset

The SIRI-WHU dataset is a collection of remote sensing images which the authors of [35] used this dataset in their classification problem research in 2016. This dataset consists of two versions that must complete each other. The total images in this dataset are 2400 remote sensing images divided into 12 classes with 200 images for each class. These images were extracted from Google Earth (Google inc.) and mainly cover urban areas in China. Each image in this dataset is Geo-tiff RGB image with 200×200 pixels resolution and 2 square meters (21.528 square foot) spatial resolution [35]. Figure 6 shows image examples with 12 classes in the SIRI-WHU dataset [35].

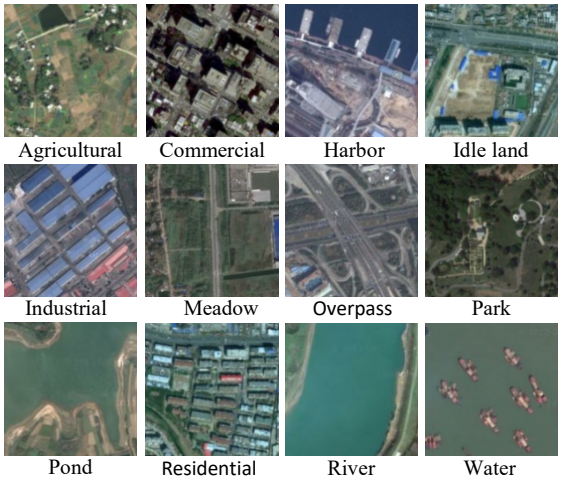


Figure 6: Image examples from the 12 classes in the SIRI-WHU dataset [35].

3.3. The Experimental Setup

The aim of this paper is to present a comparative study of using the classic networks for classifying the remote sensing images. This comparison is based on plotting the learning curves for the loss and the accuracy values, with every epoch, that be calculated through training process for the training and the validation data to determine the efficiency of the hyper-parameters values. The comparison is also based on the OA to assess each model performance. The used datasets in this study are the UC Merced land use dataset and the SIRI-WHU dataset where their details are stated in sections 3.1 and 3.2. All tests were performed using Google-Colab. The Google-Colab is a free cloud service hosted by Google inc. to encourage machine learning and artificial intelligence researches [36]. It is acts as a virtual machine (VM) that using 2-cores Xeon CPU with 2.3 GHz, GPU Tesla K80 with 12 GB GPU memory, 13 GB RAM, and 33GB HDD with Python 3.3.9. The maximum lifetime of this VM is 12 hours and it will be idled after 90 minutes time out [37]. Performing tests has been done by connecting to this VM online through ADSL internet line

with 4Mbps communication speed. This connection was done using an Intel® core™i5 CPU M450 @2.4GHz with 6 GB RAM and running Windows 7 64-bit operating system. This work is limited by used the ImageNet pre-trained weights because the train of new CNNs models needs a huge amount of data and more sophisticated hardware, this is unlike the lot of needed time consumed for this training process. The other limitation is that the input images shape is mustn't less than 200×200×3 and not greater than 300×300×3 because of the limitations of the pre-trained classic networks. The preprocessing step according to each network requirements is necessary to get efficient results; it must be as done on ImageNet dataset before training the models and produce the ImageNet pre-trained weights. The ImageNet pre-trained weights classic networks that used in this paper have input shape (224, 224, 3) and output layers 1000 nodes according to the ImageNet classes (1000 classes) [38, 39]. So, it must perform modifications on the learning algorithms that used these networks to be compatible with the used remote sensing images datasets as stated in section 2.1. The data in the used datasets were divided into 60% training set, 20% validation set, and 20% test set before training the last FC layers in each network model. The training process, using the training set and the validation set, is done for the model with a supposed number of epochs to determine the number of epochs that achieved the minimum validation loss. Thus, we consider that the assembling of the training set and the validation set are the new training data and then retrain the model with this new training set and the predetermined number of epochs that achieved the minimum validation loss. Finally test the model using the test set. It must be notice that, the learning parameters values, such as learning rate and batch size, were determined by intuition with taking in consideration the learning parameters values that used through training these models with the ImageNet dataset through producing the ImageNet pre-trained weights [26-29], where the used optimizer, the number of epochs, the additional activation and regularization layers, and the dropout regularization rates were determined by iterations and intuition. The classifier models are built using python 3.3.9, in addition to the use of the Tensorflow library for the preprocessing step and the Keras library for extracting features, training the last FC layers, and testing the models.

- In the DenseNet 196 model, the image resizing was done on the dataset images to have shape (224, 224, 3) to be compatible with the pre-trained DenseNet 196 model input shape, perform transfer learning to the 7×7 global average pooling layer that above the 1000D FC layer, and then add a FC layer, which has number of neurons equal to the used dataset classes, with softmax activation layer. The last layer weights were retrained with learning rate = 0.001, Adam optimizer and batch size = 256 with 100 epochs. The normalization preprocessing must be done on the dataset images before using on the DenseNet 169 model. Figure 7 shows the flow chart of the experimental algorithm that used the DenseNet 169 model.
- In the NASNet Mobile model, the image resizing was done on the dataset images to have shape (224, 224, 3) to be compatible with the pre-trained NASNet Mobile model input shape, perform transfer learning to the 7×7 global average pooling layer that above the 1000D FC layer, add a ReLU activation layer, dropout regularization layer with rate 0.5, and then adding a FC layer, which has number of neurons equal to the used dataset classes, with softmax activation layer. The last layer weights were retrained with learning rate = 0.001, Adam optimizer, and batch size = 64 with 200 epochs. The normalization preprocessing must be done on the dataset images before using on the NASNet Mobile model. Figure 8 shows the flow chart of the experimental algorithm that used the NASNet Mobile model.

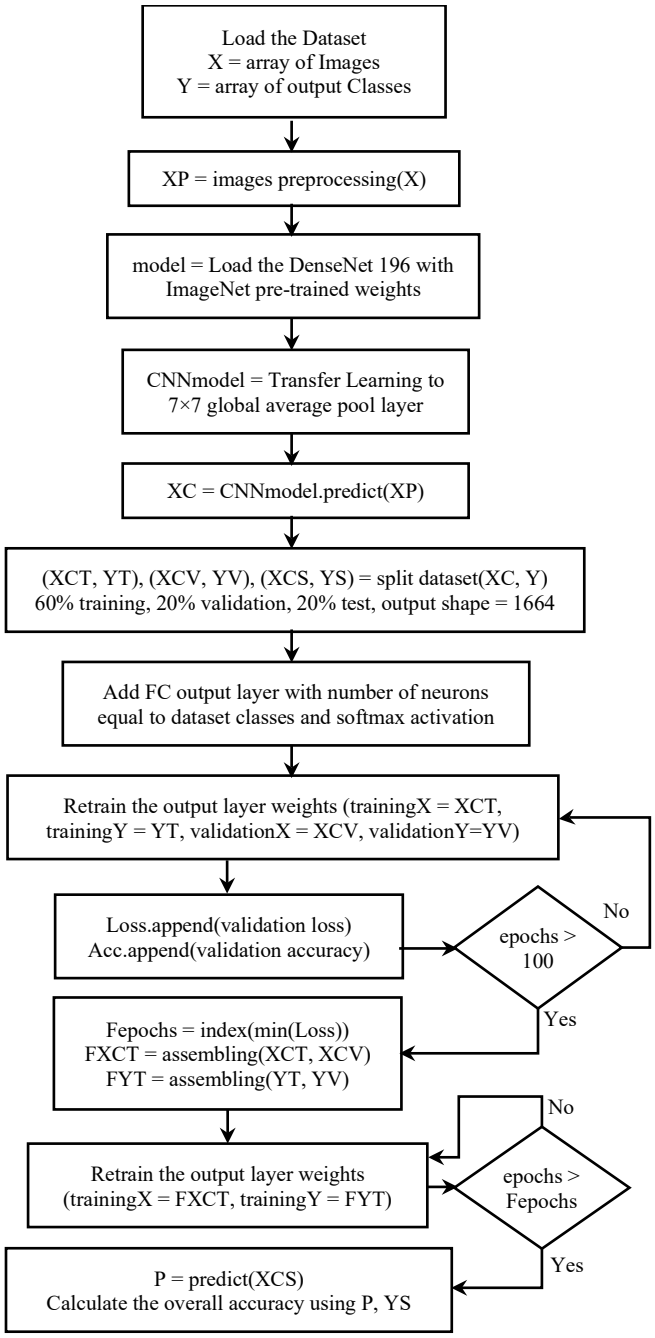


Figure 7: The flow chart of the experimental algorithm that used the DenseNet 169 model

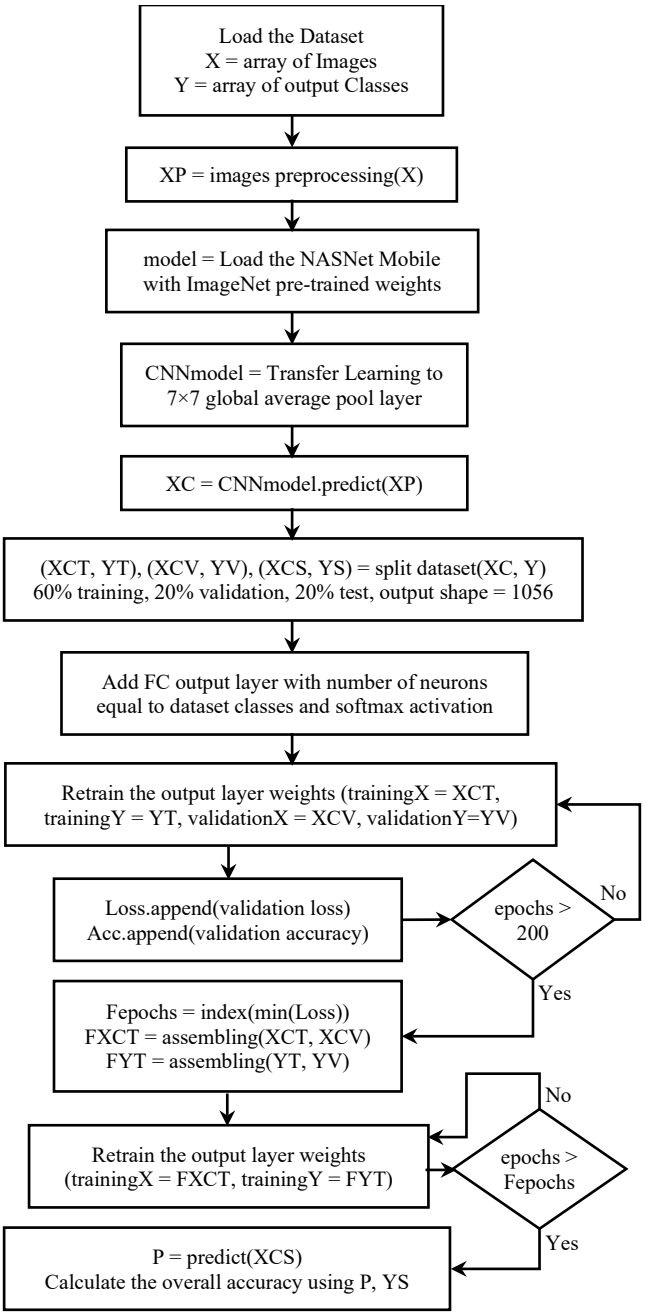


Figure 8: The flow chart of the experimental algorithm that used the NASNet Mobile model

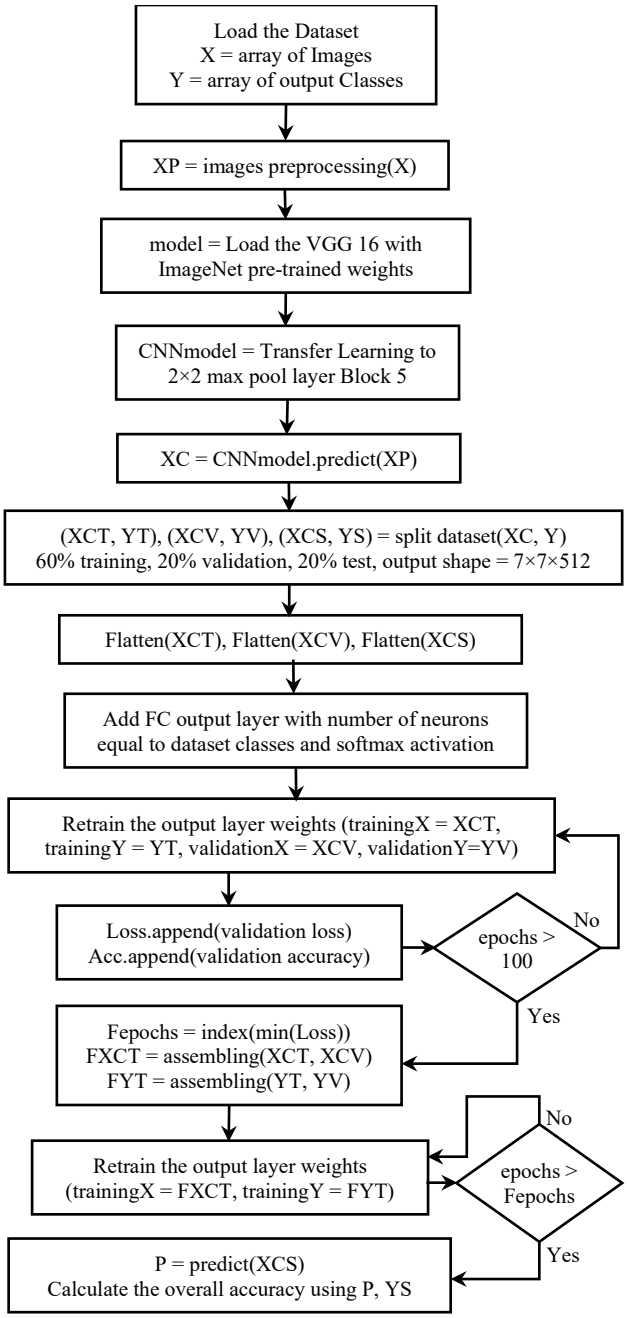


Figure 9: The flow chart of the experimental algorithm that used the VGG 16 model

- In the VGG 16 model, the image resizing was done on the dataset images to have shape (224, 224, 3) to be compatible with the pre-trained VGG 16 model input shape, perform transfer learning to the 2×2 max pooling layer in block 5, flatten the pooling layer output, add a ReLU activation layer, dropout regularization layer with rate 0.77, and then add a FC layer, which has number of neurons equal to the used dataset classes, with softmax activation layer. The last layer weights were retrained with learning rate = 0.001, Adam optimizer, and batch size = 64 with 100 epochs. The image conversion to the BGR mode preprocessing must be done on the dataset images before using on the VGG 16 model. Figure 9 shows the flow chart of the experimental algorithm that used the VGG 16 model.

- In the ResNet 50 model, the image resizing was done on the dataset images to have shape (224, 224, 3) to be compatible with the pre-trained ResNet 50 model input shape, perform a transfer learning to the 7×7 global average pooling layer that above the 1000D FC layer, and then add a FC layer, which has number of neurons equal to the used dataset classes, with softmax activation layer. The last layer weights were retrained with learning rate = 0.1, Adam optimizer, and batch size = 64 with 200 epochs. The image conversion to the BGR mode preprocessing must be done on the dataset images before using on the ResNet 50 model. Figure 10 shows the flow chart of the experimental algorithm that used the ResNet 50 model.

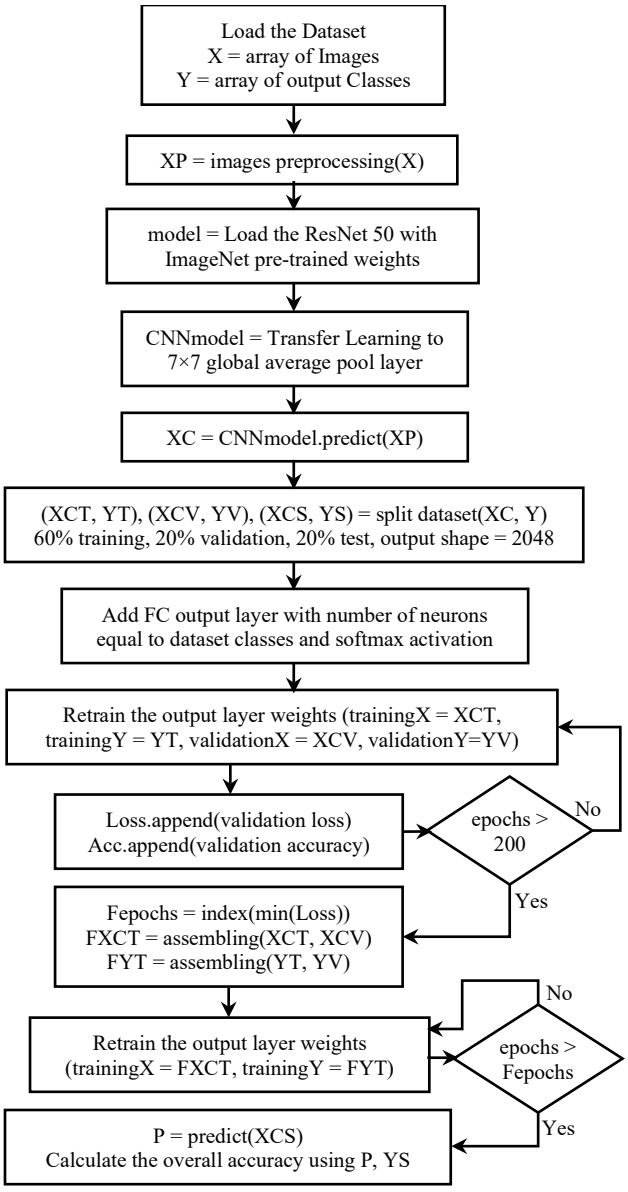
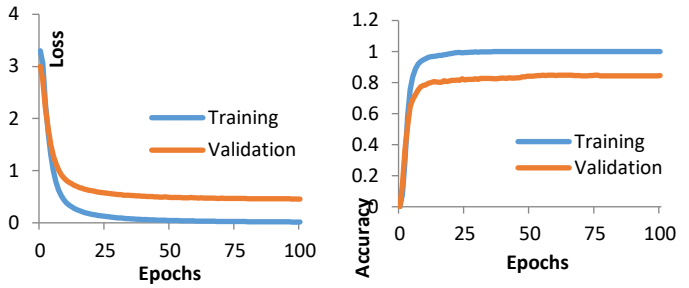
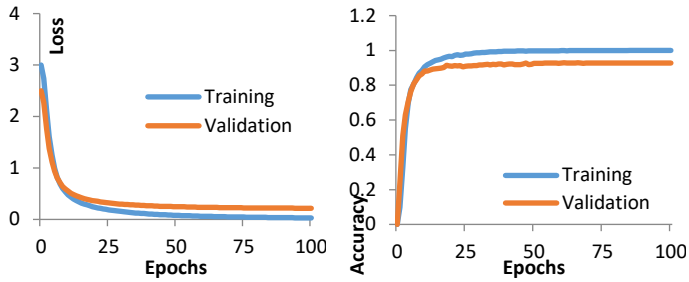


Figure 10: The flow chart of the experimental algorithm that used the ResNet 50 model

With the use of the stated convolution models in this study for classifying remote sensing images, through the training process we used the training data (60% from the used dataset) and the validation data (20% from the used dataset). Figure 11 shows the loss and the accuracy learning curves respectively for training the DenseNet 169 model, (a) for using the UC Merced land use dataset and (b) for using the SIRI-WHU dataset. Figure 12 shows the loss and the accuracy learning curves respectively for training the NASNet Mobile model, (a) for using the UC Merced land use dataset and (b) for using the SIRI-WHU dataset. Figure 13 shows the loss and the accuracy learning curves respectively for training the VGG 16 model, (a) for using the UC Merced land use dataset and (b) for using the SIRI-WHU dataset. Figure 14 shows the loss and the accuracy learning curves respectively for training the ResNet 50 model, (a) for using the UC Merced land use dataset and (b) for using the SIRI-WHU dataset.

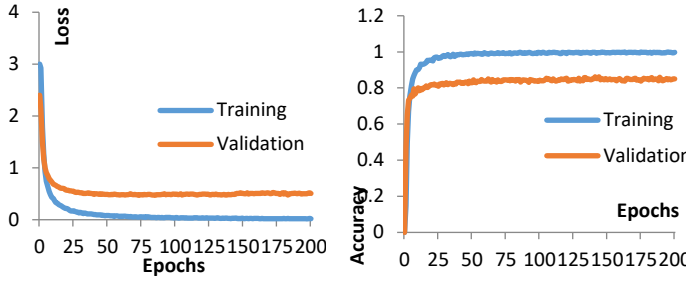


(a) Using the UC Merced land use dataset.

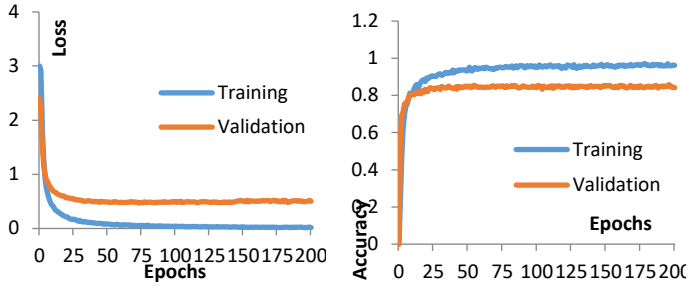


(b) Using the SIRI-WHU dataset.

Figure 11: The loss and the accuracy learning curves for training the DenseNet 169 model.

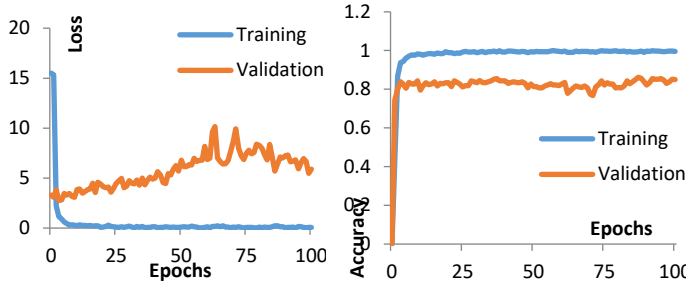


(a) Using the UC Merced land use dataset

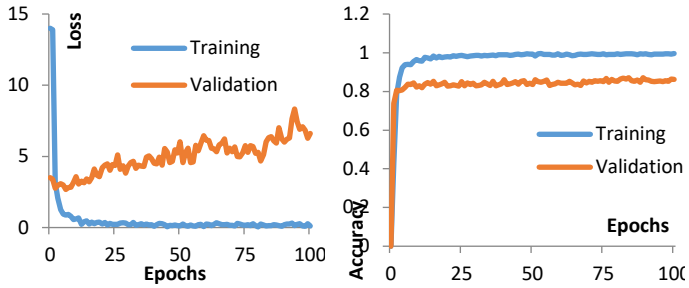


(b) Using the SIRI-WHU dataset

Figure 12: The loss and the accuracy learning curves for training the NASNet Mobile model.

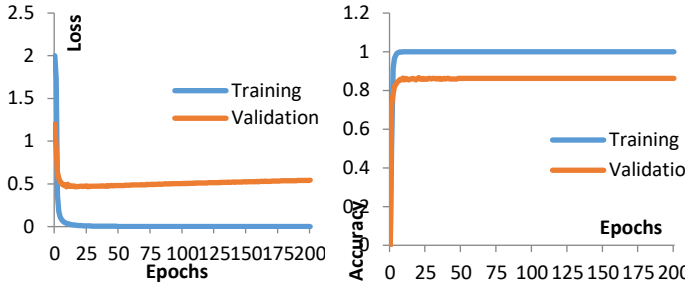


(a) Using the UC Merced land use dataset

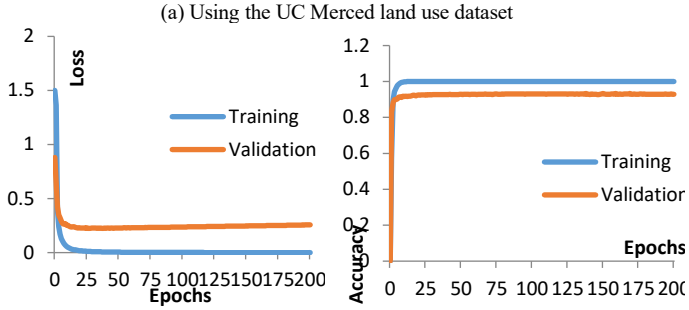


(b) Using the SIRI-WHU dataset

Figure 13: The loss and the accuracy learning curves for training the VGG 16 model.



(a) Using the UC Merced land use dataset



(b) Using the SIRI-WHU dataset

Figure 14: The loss and the accuracy learning curves for training the ResNet 50 model.

From the learning curves it can be determine the number of epochs that achieved the minimum validation loss in each model from the four models that discussed in this study. So, we consider that the training data and the validation data, 80% from dataset, are the new training data. Then, we repeat the same training process for each model with the same hyper-parameters values and the both datasets, and then calculate the OA using the predictions of the test data, 20% from the training set, to assess the performance of each model. Table 4 and figure 15 show the OA for each model using the both datasets.

Table 4: The OA for each model using the both datasets

	The UC Merced land use Dataset	The SIRI-WHU Dataset
The DenseNet 169 model	0.91	0.933
The NASNet Mobile model	0.876	0.896
The VGG 16 model	0.902	0.929
The ResNet 50 model	0.924	0.956

As shown from these results, the ResNet 50 model had the higher OA in this comparative study where the NASNet Mobile model had the lowest OA. In the other hand the OA for the DenseNet 169 model had little higher OA than the VGG 16 model. The use of the SIRI-WHU dataset had higher OA than the use of the UC Merced land use dataset. These results illustrated that the OA had an opposite relation with the dataset image resolution and the dataset number of classes, so the use of the SIRI-WHU dataset which has 12 classes, image resolution 200×200 pixels, and spatial resolution 2 square meters gave higher OA than the use of the UC Merced land use dataset which has 21 classes, image resolution 256×256 pixels, and spatial resolution 0.0929 square meter. The deeper convolution networks give considerable accuracy but the connections between layers may have another influence. The VGG 16 model gave good OA so it had efficient results but its learning curves had some degradation in its validation curves and near to the over-fitting in its training curves. So it may give better results with some additional researches that adjust the optimizations and regularization hyper-parameters. In the other hand the ResNet 50 model, that gave the higher OA in this study, had good learning curves with some little over-fitting. So, with more research and adjusting the regularization hyper-parameters, it is not easy to achieve higher results using this model without any major development. The ResNets are based on the skip layers connections so the layer connections can raise the classification accuracy. The DenseNets may have more connections but still the ResNet 50 model had the higher OA in this comparative study. The NASNets models have layers connections but not more as the DenseNets models, these connections only to determine the previous layer, so its results are low compared with other models in this study. In the other hand the NASNet's model learning curves are good, no degradation and no over-fitting, but with increasing the epochs the validation loss may be raised, so it must have an attention observed for epochs and validation curves through training this model. As a total the deeper convolution networks may give better accuracies but the deeper networks that have layers connections may give the best accuracies.

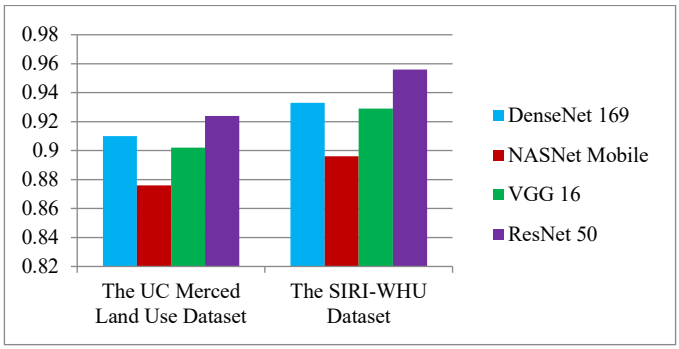


Figure 15: The OA for each model using the both datasets

4. Conclusions

This paper presented a comparative study for the use of the deep convolution models classic networks in classifying the remote sensing images. This comparison illustrated that what the classic network was more accurate for classifying the VHR remote sensing images. The used classic networks in this study were the DenseNet 169, the NASNet Mobile, the VGG 16, and the ResNet 50 models. There were two used datasets in this study; the UC Merced land use dataset and the SIRI-WHU dataset. This comparison was based on the learning curves to check that how much effectiveness of the hyper-parameters values selection and

the overall accuracy to assess the classification model performance. This comparison illustrated that the ResNet 50 model was more accurate than other models that stated in this study, which the overall accuracy of the DenseNet 169 model was little higher than the VGG 16 model. The NASNet Mobile model had the lowest OA in this study. The learning curves elucidated that the adjustment of the hyper-parameters of the VGG 16 model can lead to better overall accuracy where it is not easy to achieve better overall accuracy in the ResNet 50, the DenseNet 169 and the NASNet Mobile models without major developments in these models. The overall accuracy had an opposite relation with the remote sensing images resolution (pixel or spatial) and the number of dataset classes.

In the future, the FC layers can be replaced by other classifiers, and then train these models.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Michael Bowling, Johannes Furnkranz, Thore Graepel, and Ron Musick, "Machine Learning and Games", Machine Learning, Springer, **63**(3), 211-215, 2006. http://doi.org/10.1007/s10994-006-8919-x

[2] Issam ElNaqa and Martin J. Murphy, "What is Machine Learning", Machine Learning in Radiation Oncology, Cham, Springer, 3-11, 2015. https://doi.org/10.1007/978-3-319-18305-3_1

[3] Tom M. Mitchell, "The Discipline of Machine Learning, Pittsburgh", PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department, **9**, 2006.

[4] Kyung Hwan Kim and Sung June Kim, "Neural Spike Sorting Under Nearly 0-dB Signal-to-Noise Ratio Using Nonlinear Energy Operator and Artificial Neural-Network Classifier", IEEE Transactions on Medical Engineering, **47**(10), 1406-1411, 2000. http://doi.org/10.1109/10.871415

[5] T.Blaschke, "Object Based Image Analysis for Remote Sensing", ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier, **65**(1), pp. 2-16, 2010. https://doi.org/10.1016/j.isprsjprs.2009.06.004

[6] Minu Nair S., and Bindhu J.S., "Supervised Techniques and Approaches for Satellite Image Classification", International Journal of Computer Applications (IJCA), **134**(16), 1-6, 2016. . https://doi.org/10.5120/ijca2016908202

[7] Waseem Rawat, and ZenghuiWang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review", Neural computation, **29**(9), 2352-2449, 2017. http://doi.org/10.1162/NECO_a_00990

[8] Khalid. A. Al-Afandy, Hicham Omara, Mohamed LaZaar, and Mohammed AL ACHHAB, "Artificial Neural Networks Optimization and Convolution Neural Networks to Classifying Images in Remote Sensing: A Review", in The 4th International Conference on Big Data and Internet of Things (BDIoT'19), 23-24 Oct, Rabat, Morocco, 2019. http://doi.org/10.1145/3372938.3372945

[9] Yiting Tao, Miaozhong Xu, Zhongyuan Lu, and Yanfei Zhong, "DenseNet-Based Depth-Width Double Reinforced Deep Learning Neural Network for High-Resolution Remote Sensing Images Per-Pixel Classification", Remote Sensing, **10**(5), 779-805, 2018. http://doi.org/10.3390/rs10050779

[10] Gefei Yang, Utsav B. Gewali, Emmett Ientilucci, Micheal Gartley, and Sildomar T. Monteiro, "Dual-Channel DenseNet for Hyperspectral Image Classification", in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 05 Nov, Valencia, Spain, 2595-2598, 2018. http://doi.org/10.1109/IGARSS.2018.8517520

[11] Jianming Zhang, Chaoquan Lu, Xudong Li, Hye-Jin Kim, and Jin Wang, "A Full Convolutional Network based on DenseNet for Remote Sensing Scene Classification", Mathematical Biosience and Engineering,, **6**(5), 3345-3367, 2019. http://doi.org/10.3934/mbe.2019167

[12] Lingling Li, Tian Tian, and Hang Li, "Classification of Remote Sensing Scenes Based on Neural Architecture Search Network", in 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), 19-21 Jul, Wuxi, China, 176-180, 2019. http://doi.org/10.1109/SIPROCESS.2019.8868439

[13] Ali Bahri, Sina Ghofrani Majelan, Sina Mohammadi, Mehrdad Noori, and Karim Mohammadi, "Remote Sensing Image Classification via Improved Cross-Entropy Loss and Transfer Learning Strategy Based on Deep Convolutional Neural Networks", IEEE Geoscience and Remote Sensing Letters, vol. **17**(6), 1087-1091, 2019. http://doi.org/10.1109/LGRS.2019.2937872

[14] Xuan Liu, Mingmin Chi, Yunfeng Zhang, and Yiqing Qin, "Classifying High Resolution Remote Sensing Images by Fine-Tuned VGG Deep Networks", in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 22-27 Jul, Valencia, Spain, 7137-7140, 2018. http://doi.org/10.1109/IGARSS.2018.8518078

[15] Souleyman Chaib, Huan Liu, Yanfeng Gu, and Hongxun Yao, "Deep Feature Fusion for VHR Remote Sensing Scene Classification", IEEE Transactions on Geoscience and Remote Sensing, **55**(8), 4775-4784, 2017. http://doi.org/10.1109/TGRS.2017.2700322

[16] Zhong Chen, Ting Zhang, and Chao Ouyang, "End-to-End Airplane Detection Using Tansfer Learning in Remote Sensing Images", Remote Sensing, **10**(1), 139-153, 2018. http://doi.org/10.3390/rs10010139

[17] Gang Fu, Changjun Liu, Rong Zhou, Tao Sun, and Qijian Zhang, "Classification for High Resolution Remote Sensing Imegery Using a Fully Convolution Network", Remote Sensing, **9**(5), 498-518, 2017. http://doi.org/10.3390/rs9050498

[18] Mingchang Wang, Xinyue Zhang, Xuefeng Niu, Fengyan Wang, and Xuqing Zhang, "Scene Classification of High-Resolution Remotely Sensed Image Based on ResNet", Journal of Geovisualization and Spatial Analysis, Springer, **3**(2), 16-25, 2019. http://doi.org/10.1007/s41651-019-0039-9

[19] Yenan Jiang, Ying Li, and Haokui Zhang, "Hyperspectral image classification based on 3-D separable ResNet and transfer learning", IEEE Geoscience and Remote Sensing Letters, **16**(12), 1949-1953, 2019. http://doi.org/10.1109/LGRS.2019.2913011

[20] S. Natesan, C. Armenakis, and U. Vepakomma, "ResNet-Based Tree Species Classification Using UAV Images", International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, **XLII**, 2019. http://doi.org/10.5194/isprs-archives-XLII-2-W13-475-2019

[21] Jiachen Yang, Yinghao Zhu, Bin Jiang, Lei Gao, Liping Xiao, and Zhihui Zheng, "Aircraft Detection in Remote Sensing Images Based on a Deep Residual Network and Super-Vector Coding", Remote Sensing Letters, Taylor & Francis, **9**(3), 229-237, 2018. http://doi.org/10.1080/2150704X.2017.1415474

[22] Lin Xu and Qiang Chen, "Remote-Sensing Image Usability Assessment Based on ResNet by Combining Edge and Texture Maps", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **12**(6), 1825-1834, 2019. http://doi.org/10.1109/JSTARS.2019.2914715

[23] Ahmed Fawzy Gad, "Practical Computer Vision Applications Using Deep Learning with CNNs", Apress, Springer, Berkeley, CA, 2018. https://doi.org/10.1007/978-1-4842-4167-7

[24] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks", IEEE Transactions on Geoscience and Remote Sensing, **54**(10), 6232-6251, 2016. http://doi.org/10.1109/TGRS.2016.2584107

[25] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez, "Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification", IEEE Transactions on Geoscience and Remote Sensing, **55**(2), 645-657, 2017. http://doi.org/10.1109/TGRS.2016.2612821

[26] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely Connected Convolutional Networks", in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 Jul, Honolulu, HI, USA, 2261-2269, 2017. http://doi.org/10.1109/CVPR.2017.243

[27] Barret Zoph and Quoc V. Le, "Neural architecture search with reinforcement learning", in International Conference on Learning Representations (ICLR 2017), 24-26 April, Toulon, France, 2017. https://arxiv.org/abs/1611.01578

[28] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition", in International Conference on Learning Representations (ICLR 2015), 7-9 May, San Diego, USA, 2015. https://arxiv.org/abs/1409.1556

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition", in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 Jun, Las Vegas, NV, USA, 770-778, 2016. http://doi.org/10.1109/CVPR.2016.90

[30] Gebhard Banko, "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory", in International Institution for Applied Systems Analysis (IIASA), Laxenburg, Austria, IR-98-081, 1998. http://pure.iiasa.ac.at/id/eprint/5570/

[31] Weijia Li, Haohuan Fu, Le Yu, and Arthur Cracknell, "Deep Learning Based Oil Palm Tree Detection and Counting for High-Resolution Remote Sensing Images", Remote Sensing, **9**(1), 22-34, 2017. http://doi.org/10.3390/rs9010022

[32] Jan N. van Rijn, Salisu Mamman Abdulrahman , Pavel Brazdil, and Joaquin Vanschoren, "Fast Algorithm Selection Using Learning Curves", in International symposium on intelligent data analysis, **9385**, Springer, Cham,

22-24 Oct, Saint Etienne, France, 298-309, 2015. http://doi.org/10. 1007/978-3-319-24465-5 26

[33] MartinWistuba and Tejaswini Pedapati, "Learning to Rank Learning Curves", arXiv:2006.03361v1, 2020. https://arxiv.org/abs/2006.03361

[34] Yi Yang and Shawn Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification", in the 18th ACM SIGSPATIAL international conference on advances in geographic information systems, 2-5 Nov, San Jose California, USA, 270-279, 2010. http://doi.org/10.1145/1869790.1869829

[35] Bei Zhao, Yanfei Zhong, Gui-Song Xia, and Liangpei Zhang, "Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery", IEEE Transactions on Geoscience and Remote Sensing, **54**(4), 2108-2123, 2016. http://doi.org/10.1109/TGRS.2015.2496185

[36] Ekaba Bisong, "Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners", Apress, Berkeley, CA, 2019.

[37] Tiago Carneiro, Raul Victor Medeiros Da Nobrega, Thiago Nepomuceno, Gui-Bin Bian, Victor Hugo C. De Albuquerque, and Pedro Pedrosa Reboucas Filho, "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications", IEEE Access, **6**, 61677-61685, 2018. http://doi.org/10.1109/ACCESS.2018.2874767

[38] Olga Russakovsky and Li Fei-Fei, "Attribute Learning in Large-Scale Datasets", in European Conference on Computer Vision, **6553**, Springer, Berlin, Heidelberg, 10-11 Sep, Heraklion Crete, Greece, 1-14, 2010. http://doi.org/10.1007/978-3-642-35749-7_1

[39] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei, "What Does Classifying More Than 10,000 Image Categories Tell Us? ", in The 11th European conference on computer vision, **6315**, Springer, Berlin, Heidelberg, 5-11 Sep, Heraklion Crete, Greece, 71-84, 2010. http://doi.org/10.1007/978-3-642-15555-0_6

ASTES

Advances in Science, Technology and Engineering Systems Journal 5, No. 4, 776-782 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Ethics as a Motivation Indicator in Second Language Vocational Digital Teaching

Natalya Viktorovna Matveeva*, Ludmila Vladimirovna Makar

The Linguodidactics Department, Institute of Humanities, Russian University of Transport (MIIT), 9, bld. 9, Obrazcova Street, Moscow, 127994, Russia

ARTICLE INFO

Article history:
Received:31 March, 2020
Accepted:12 May, 2020
Online: 28 August, 2020

Keywords:
College Students
Computer Games Development
Ethical Teaching Skills
Foreign Language
Out-Of-Classroom Autonomous Activity
Teacher-Student Rapport

ABSTRACT

The non-selective second language course at vocational colleges and universities makes teachers strive at fostering students' motivation to learn by choosing from a variety of enhancing factors. Teacher's personality and skills if they comply with pedagogical ethics are considered to be inspiring for students to learn. The aim of this piloting study was to collect empirical data and ascertain the actual motivational power of ethical teaching skills when organizing and conducting autonomous learning activities for college students on the example of educational computer games.

The ethical teaching skills applied when creating and using computer games as out-of-classroom voluntary activities were conceptualized. Their possible influence on learning motivation of college students was analysed with the use of oral and written feedback procedures, and statistical data retrieved from Metrika.Yandex and students' certificates submitted.

This allowed to determine five specific groups of students valuing the activity, which in their turn were united into two larger formations: students with distinct intrinsic motivation in the activities-related areas (60%) and socially and psychologically responsive students (40%). A possible method to evaluate statistically the efficiency of some ethical skills specified is suggested, whereas the other skills need further research under modified parameters.

Thus, ethical teaching skills are effective in the case the college students possess an intrinsic motivation in an area actualized by an activity suggested. Therefore, they are unlikely to be regarded as an independent motivator, but the indicator of a particular intrinsic motivation characteristic to a student.

1. Introduction

This paper is an extension of work originally presented in *Proceedings 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS)*. It stated that there existed the gap between college and university syllabus in the number of professionally oriented themes (not in favour of college students); individual words learnt (≈800 at college vs 600 at university); and in learning methods. In addition, it has been revealed that college students underestimate the role of searching information online, critical thinking, making quick decisions and some other basic competences and skills as essential for their future career. The language computer games to overcome the differences were developed in Adobe Animate, and their plot and the interface design were described [1].

The piloting research was initiated at Serpukhov college (Moscow Region, Russia). The college is comprehensive with approximately 1400 students from the south of Moscow region and adjacent areas of Tula and Kaluga regions mastering a wide range of specialities. All of them were invited to take part in the games. However, due to the geographically dispersed location of the college faculties, for the purpose of adequate data collection in part of classroom discussions we managed to keep in close touch only with the students of engineering specialities who study in one of the premises (450). This is why we consider this group of students the core of our research and the research itself piloting.

The first invitation to bridge the gap through the games was placed on the college official Internet-site at the beginning of the academic year. It offered a brief thematic review of the games and

*Corresponding Author: Natalya Matveeva, Russian University of Transport (MIIT), +79262460543, matveevanv@mail.ru

description of test conditions. Students were informed that we would be glad if they shared the results with us sending screenshots of their game certificates to the contact e-mail. We also organized a public page in a social network in case the students wanted to discuss gameplay. The English teachers of Serpukhov college duplicated information in classes. Even a month later there was no great activity surge, and we found out that this new type of learning activity was not welcome by students with their arms outstretched.

The question followed: as information technology itself and the intriguing form of presenting educational materials such as educational computer game couldn't be the dominating motives for college students in their independent work in learning English as a second language, a teacher should ethically motivate them to study.

Ethics in teaching profession attracts attention of the professional community. The quality of a teacher is directly connected to setting a good moral example for students and the skills of acting in accord with professional ethical standards [2, 3]. In Russia researchers challenge the task of developing pedagogical ethics in order to make the educational process more effective [4]. A need for continuous teacher training to develop ethical competency is highly emphasized [5 – 7]. The teacher's ethical competency is thought to comprise, firstly, the knowledge of ethics and ethical aspects in teaching, secondly, skills of pedagogical interaction and communication, thirdly, the ability to implement the knowledge of ethics in specific pedagogical situations. Among ethically sound personal qualities in teaching, tolerance, poise, goodwill, caring attitude, and the ability to empathize appear to be the most representative ones [8]. Teaching ethically in a digital environment is of great importance as well. Exactly this aspect of teaching ethics is in the focus of our research.

2. Problem Statement

It has been found out that the teacher's personality can have a major long-term impact on students' motivation to learn a second language, including autonomous, or as Lamb & Wedell said, "in their own time, of their own volition beyond the classroom" [9] learning. Melyokhina & Bobohodzhaeva highlighted that learners are particularly inspired by teachers who are content competent and able to build good teacher-student rapport, demonstrate academic literacy and interest in teaching [10]. Thus, the issue of teaching ethics as a factor motivating students to study is becoming a research agenda.

In college and university educational context, where the second language is a mandatory, non-elective job-oriented course and teachers have to face mixed ability and mixed motivation classes, the issue of motivating students is of the utmost importance. A great number of motivators having been scrutinized by the teaching community worldwide [11], still potentially effective ones must be revealed and investigated.

3. Research Questions

Without minimizing the significance of ethical knowledge and ethically sound personal qualities of a teacher in developing students' learning motivation, we tend to attribute a greater deal of success in this field to the practical skills of applying these two in such fundamental pedagogical situations as 1) designing and implementing language learning courses, classes, materials and activities and 2) building relationships with students. As to the designing tasks, it is generally accepted that informing learners about the aims, providing them with information meaningful to them, using facilitating instructional techniques to perform and achieve goals, etc. is ethical. When building rapport with students, it is ethical to respect their rights, dignity and freedom, their autonomy and privacy, to avoid conflicts, provide choice, timely and objective assessment of works, give informational feedback.

Using this line of thinking, we raised the question as to whether these teaching skills with strongly felt ethical background can play into the college teachers' hands to engage the learners into autonomous training aimed to compensate an existing gap between college and university syllabus in the number of 1) professionally oriented themes (not in favour of college students); 2) individual words learnt (≈800 at college vs 600 at university); 3) in learning methods. In addition, it has been revealed that college students underestimate the role of searching information online, critical thinking, making quick decisions and some other basic competences and skills as essential for their future career [12].

4. Purpose of the Study

The objective of this research is to collect empirical data and ascertain the degree of motivational power of ethical teaching skills in conducting out-of-classroom learning activities at college. For the purpose, we 1) conceptualize the ethical teaching skills of primary importance when creating and using an autonomous learning activity in the form of a computer game; 2) analyse their possible influence on second language learning motivation of college students.

5. Research Methods

Below we outline two sets of ethical teaching skills corresponding to the pedagogical situations mentioned which are of high value in digital environment.

5.1. Ethical Skills Important for Designing and Implementing Learning Materials and Activities

Using extra-curricular interests, skills and competences. When designing materials and activities for autonomous language learning, it is essential to provide for bringing different interests and skills students may have a good command of, these may be drawing, love for sport or IT. While choosing the ones we could rely upon, what we were quite sure of is that the absolute majority of our students are quite competent in computers and digital technologies. The rich virtual environment offers what young people appreciate: simulations, games, digital objects, communication tools. It is natural then to choose students' IT competence for designing autonomous English learning activities.

Providing information meaningful for students. Due to the characteristics of their age group college students are not inclined to work for the long-term perspective. Their life and professional trajectory are far from being clear and well defined. In their free time, they are more fascinated with "having fun", which in fact means communicating with friends and trying new or pleasurable things to do.

Our games are quite in this vein. The topics for the games "The Era of Great Geographical Discoveries: Meeting of Civilizations" and "Arctic Convoys of the Second World War" concern the watershed periods in human history. Being new, they suggest a challenge to be engaged in indirect communication with antecedent generations and civilizations through conceptual thinking and problem solving, acquire and share knowledge of them. Moreover, remembering Macmurray [13], who pleaded for online educational systems based on "learning to be human", we can say the games implement the humanistic approach based on human perception, communication and interaction, which are also the key to the field of learning languages.

Instructional context providing, following Walters' words [14], "choice, optimal challenge, informational feedback, interpersonal involvement, and acknowledgement of feelings" can play a great part in motivating L2 learners. The skill to provide the instruction of this kind is an ethical challenge for a teacher.

We provide a full-range instruction set on the splash screen of each game. It briefly describes the theme, plot and gameplay, states the main aim of each game, informs about helpful Internet resources and the systems of controls for successful playing. The instruction gives a clear rationale that allows participants to fully grasp the specific meaning and importance of the activity.

Utility value of the games designed proceeds from the research task to bridge the gap between college and university foreign language curricula. The themes of the games raise issues on sociology, psychology, culture and religion, communication, economic systems, which are not included in the syllabus at colleges, and also concern civilized values in general – war and peace, independence, non-intervention, friendship, cooperation, poverty and prosperity. Besides language and learning skills, the games promote college students' digital literacy, life, research and professional skills such as navigating, interacting, searching information online, quick reading and making decisions. All kinds of skills gained will benefit college graduates both if they decide to continue education at university and if they go straight to work.

5.2. Ethical Skills Important for Teacher-Student Relationship

Respect for autonomy. Autonomy, according to Noels [15] and some other researcher such as Lozovoy & Zashchitina [16], refers to "a sense that one can freely choose to engage in activities that they find personally meaningful. Teacher's autonomy enhancing behavior includes fostering relevance, allowing criticism and providing choice [17]. We followed the advice of Jang, Reeve & Deci [18] to use informational language that provides options for students and identifies the potential meaningfulness of the activity suggested. We highlighted that the games comprised professionally oriented tasks, and the knowledge and skills students had acquired during their studies at college would help them as players to survive in the game worlds currently and as specialists in the future. The tasks were specifically designed for players to form good generalizations about what will work well when they face more complex professional and life challenges.

Commitment to success means a student's willingness to engage in and persist at a task (Wolters, 2003). In order to sustain engagement and persistence in the games we foresaw the possibility for the participants to play through the games again if the score does not satisfy them, simultaneously drilling the words and memorising the answers thus climbing to game and academic success.

Feedback. In-class short discussions were organized to see how much interest the games provoked, find out the difficulties and drawbacks, and promote learners' sense of competence. Students were motivated to express their opinions about the games. Some reactions just needed explanations or encouragement from the teachers, while others resulted in a forthcoming hot round-table discussion on the topics the games offered. Some of the students demonstrated involvement, immediacy and interest. They provided help to those who experienced difficulties whether IT or English course based. The feedback then was threefold: from students to the teachers, from the teachers to the students, and between students.

Evaluation brings up an issue of rewards. To be effective, rewards need to be valued by their potential recipients and timely administered. Making voluntary participation and confidentiality the corner-stones of the activity, we avoided the necessity of extrinsic rewards, which have been found to interfere with the process and quality of learning (Reeve, 2005). Two kinds of evaluation were ensured, namely, direct individual clear-eyed assessment and indirect, implied self-evaluation. The former was provided by scoring and timing programmed for the whole game. The score and time spent in a game are shown in the last frame, which is also the learner's certificate. The latter was supposed to arouse during feedback in-class sessions. Although these sessions pursued a different goal (see above), it is difficult to argue the possible outbreak of latent self-assessment thoughts as a result of comparing one's own actions with those of groupmates. In both cases, students did not have to share their results with anyone keeping them private if timid or unsatisfied.

Confidentiality. Some students felt comfortable and were not shy to share their scores without being forced. Nevertheless, we strictly followed this principle regarding those students who showed no inclination to discuss anything related to the activity. We did not experience personal data collection, we did not pursue students who refused to take part in the activities, just we were happy for those who had passed the games and shared the score with us.

6. Procedure

We were allowed to deliver information about the computer games developed to fill the gap in learning functional lexis during the buzz session with Sergey Chekmaev, the Russian science-fiction writer and a participant of literary adaptation of numerous computer games, *TimeZero, Destiny, StarQuake* being some of them. The audience of the session, as we saw it, was the students who were "computer game players" and we hoped exactly these ones were the vanguard open to innovations.

Students could launch the games any time they wanted during five-month period starting from the first announcement.

Short classroom feedback took place once a fortnight at the lessons, their duration depending on the number of questions on procedural and rule matters the students had.

The issues raised by the topics of the games were discussed during final round table talks when the students discussed and argued the topics-related issues. During these talks students' L2 and human progress was reviewed.

After this debriefing the site was not closed and had more visitors in the spring and during the summer holidays.

Having conceptualised ethical issues of primary importance when creating and using educational computer games as learning activities we monitored their possible influence on students' L2 learning motivation by means of quantitative and qualitative methods.

To do statistics we processed data from 1) Metrika.Yandex, which is the analytical widget installed on the site to count unique IPs and the number of visits, to fix the non-personalised data regarding the traffic, geography, age of the visitors and time spent on the site available and 2) the students' certificates submitted. Only results proved by certificates sent by emails or VKontakte were subject to the analysis. The certificates shown but not sent were neglected.

The qualitative analysis is based on the students' questions and comments expressed orally during classroom feedback sessions or committed in writing (mostly in English!) via e-mail, VKontakte or directly to the teachers and sorting them out in accordance with the ethical aspects listed above. During these sessions the students could express their main motives why they played the games explicitly, and some of them did. This gave us an opportunity to disclose the motivation these students had. Data retrieved are presented in the tables below.

7. Findings

This section presents data retrieved through the investigation: the total number of unique IPs and the total number of visits; students' scores.

Table 1: Statistical data retrieved from Metrika.Yandex

The number of unique IPs	The number of views
275	692

Table 2: Non-personalized students' scores

Certificate	Score, %	Motivation	Time spent online, min.
1	92	Positive attitude to life	6
2	90	Computer game player	1
3	88	Computer game player	1
4	80	Entering a university	-
5	80	Computer game player	16
6	80	Investigation for cheating	-
7	77	Computer game player	-
8	74	Computer game player	1
9	71	Interest to the topics	9
10	68	Interest to the topics	5
11	67	Inclination to success	9
12	61	Computer game player	17
13	60	High interest to learning English	17
14	51	Social activist	5
15	50	General interest to IT as a speciality	12
16	50	High interest to learning English	19
17	49	Computer game player	9
18	48	Computer game player	12
19	48	General interest to IT as a speciality	20
20	47	Investigation for cheating	20
21	47	General interest to IT as a speciality	9
22	44	Curiosity	4
23	40	General interest to IT as a speciality	16
24	31	Interest to the topics	-
25	28	Interest to the topics	-
26	28	Social activist	15
27	26	Interest to the topics	6
28	26	General interest to IT as a speciality	2
29	26	Computer game player	6
30	25	Curiosity	-
31	23	Computer game player	6
32	23	Interest to the topics	5
33	20	General interest to IT as a speciality	4
34	20	Knock-on effect	-
35	16	Social activist	9

Table 3. Students' Questions and Comments Grouped by Ethical Skills Involved

No	Ethical Skill	Students' question and comments
Ethical Skills Important for Designing and Implementing Learning Materials and Activities		
1	Using extra-curriculum (in this research: IT) interests, skills and competences	How should I activate the game? Can I play the game using a mobile phone? What should I do to get rid of advertising banners on the site? What should I do to start playing? I type in the correct word but the score doesn't improve.
2	Providing information meaningful for students	Do you think a peaceful meeting of civilization possible? Is it possible to help the Arctic convoy survive in a game?
3	Instructional context	How to pass the game through? What is it necessary to do there? Where should I

		get information on the topic? Should I know everything included into the game?
4	Utility value	Do we help you (a teacher) in anything? Do you (a teacher) want us to play? Are you (a teacher) preparing a report?
Ethical Skills Important for Teacher-Student Relationship		
5	Respect for autonomy	The control system is too complicated. I didn't manage to catch the point. Is it a compulsory or voluntary activity? No desire to be measured again.
6	Commitment to success	I've passed through the game several times (12) and didn't get a score I can share.
7	Evaluation	Is it possible to cheat?
8	Confidentiality	Have you passed through the game, my groupmate? Do you, teacher, need help?

8. Discussion

Below we suggest our interpretation of the statistical data and students' reactions to the activities received with a glance to their personalities and conclusions we have come to.

Using extra-curriculum interests, skills and competences. Initially we supposed the students' high IT competence in general and in computer games particularly. However, their questions (table 3) displayed that some IT areas were underdeveloped, among them 1) operating some software programmes, namely, Adobe Flash Player; 2) understanding that the Internet is an unrestricted advertising media; 3) unskillful situational manipulating with a keyboard and a mouse.

Providing information meaningful for students. The games content proved to be topical. The possibility of a peaceful meeting of civilization caused a really hot classroom discussion as well as the alternative for the Arctic convoy PQ-17 to do the 'the Russian run' and survive. The students' willingness to put their opinions on the lessons the humankind learnt made us certain that we had chosen information which was of interest to our students provoking critical thinking and arguing the point of view.

Questions dealing with *instructional context* testify to its particular importance and indicate the necessity of its provision more than once and in different forms. In addition, the technological difficulties students experienced with the software, advertising, etc. made us think of the necessity to cover them in a written form, too, thus providing sufficient guidance and support throughout the entire process of participation.

The in-lead about the *utility value* of the games brought the students to speculations about their future jobs and responsibilities, the difference between college and university educational routes and subsequent careers being clearly understood. In their turn, students showed their sincere interest to the utility value of the games for us, teachers. They tried to discover if we had any personal motivation for the work done other than declared. Here we felt the interdependence of teaching and learning motivations, the so-called knock-on effect described in a number of research [19, 20].

As we have repeatedly stressed, we did not press the students to play, it's them who decided. This *respect for autonomy* was a novelty for them used to compulsory activities. Students returned to this aspect, expressing no desire to be measured and assessed again. This is quite in line with Huang, Han, & Schnapp's conclusions [21] that state standardised testing regimes cause a strong concern. Assured that the activity was voluntary, some of such students participated, not sharing the score though.

Commitment to success ranged from very high (some students passed through the game up to twelve times) to extremely low when the students refused to finish the games alleging technical difficulties, for example, complicated control system. It is interesting to note that it took only one minute to make two of the best scores (90% and 88%), while the winner (92%) spent 6 minutes online (table 2). The winners evidently had passed through the game several times, memorizing the answers and drilling the words. The six of the worst results (20 - 26%) were also received with little time spent in a game (2 - 6 minutes), but this possibly shows they passed through the test formally just to see what it looked like.

Some students were concerned with possible cheating, which was quite unexpected, as we planned these activities to be self-regulated. Such remarks undeliberately reminded us that being placed on the web, all teaching materials should be treated as objects of protection [22]. We fixed the vulnerability, gave praise for valuable help, and thanked for passing through the game with a good (probably winded up) score (certificates 6, 20) and doing additional work of investigation. Now the score a learner gets is always correct and can be relied on in *evaluation* procedures.

Teachers strictly followed the principle of *confidentiality* and did not divulge the scores or any personal game information, but students asked each other directly about passing through the game. They were happy to see the like-minded persons, provided help to those who asked for, withholding comments in respect to those students who did not play or did not want to share the score or discuss anything regarding the game.

9. Conclusion

1. Students' independent activity before we started to apply ethical principles in stimulating them to play our games was exceptionally low. It gave us the understanding that neither information technology be itself or educational computer games as an intriguing format of presenting the education materials motivated students to strive for the best results in learning English as a second language. Teachers should ethically motivate students to learn English autonomously.

2. While ethically motivating students we retrieved statistical data of high participation of college students of Engineering (275 out of 450) – more than 50 % of enrollees, even higher than we expected when starting.

3. The games stimulated active behavior from the specific groups of students, presented in Table 4 below in accordance with their strength.

Table 4. Specific groups of students

№	Groups of students	Number	Percentage
1.	students interested in IT and computer games including those two investigating for vulnerabilities	18	51.4 %
2.	students driven by personality underlying motives	8	22.8 %
3.	interested in the topics	6	17 %
4.	with positive attitude to the language course	2	5.8 %
5.	planning to study at university	1	2.9 %

4. Since the activity was voluntary, it was evidently valued by the students with distinct intrinsic motivation in our games-related areas – IT, language, entering university (groups 1, 4, 5 – 60%) and students energized with their personalities (groups 2, 3 – 40%). Hence, the ethical principles are not likely to possess the motivational power by themselves but provide support in developing motivated learners, on the one hand, and on the other hand, provoke an interest from socially and psychologically responsive students contributing to their human and professional growth. And from this standpoint the student's reaction to them can serve as a kind of indicator whether they possess any sort of intrinsic motivation or do not.

5. The impetus with the least driving power appears to be orientation to higher education – only one four-year student openly declared his intention to continue education at university. The others though keeping this possibility in mind are not so definite about their professional plans for the future. This is quite in line with higher educational establishment enrolment quality monitoring data, which highlight that the amount of college graduates annually entering Russian universities is only 6-10% of the total number of enrolees.

6. The ethical skills of *respect for autonomy* and *confidentiality* provide psychological comfort to learners and cannot be measured. Still, the influence of the others on students' motivation can roughly be evaluated. Thus, the statistics and efficiency of *using extra-curriculum interests, skills and competences* may climb to 100 % covering all the participants, *providing information meaningful for students* – 17 % embracing students fascinated by the topics. *Utility value* can be estimated on the basis of such drivers as entering a university, interest to IT as a speciality, knock-on effect giving 20-25%.

7. The research also proves that students do need teachers' personal assessments and interest in their scores and in their personalities in general. Communication with others and possibility to be a member of a community motivate students much more than information technology itself and even educational games as a means of delivering knowledge.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank the staff of the Chair of Linguodidactics of the Russian University of Transport for their support in their work.

Great thanks to the students of Serpukhov college (Moscow Region, Russia) who gave their feedback, asked questions and kindly shared their scores.

References

[1] N.V. Matveeva, "Computer Games to Fill the Gap in Learning Functional Lexis at Russian Colleges and Universities" Proceedings of 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS), 639-643. 2019. http://doi.org/10.1109/ITQMIS.2019.8928440.

[2] R.R Puhan, L. Malla, S.K. Behera, "Current Ethical Issues in Teacher Education: A Critical Analysis on Pre-Service and In-service Emerging Teachers". American Journal of Educational Research, **2**(12A), 1-7, 2014. http://doi.org/10.12691/education-2-12A-1.

[3] M. Sultana, "Ethics in Teaching Profession" ABC Journal of Advanced Research, **3**(1), 44-50. 2014. URL: http://oaji.net/articles/2015/814-1431431772.pdf

[4] O.S. Tsaregorodtseva, "Razvitiye pedagogicheskoi etiki uchitelya kak factor povysheniya effektivnosti obrazovatel'nogo protsessa v shole". Ph.D Thesis, Murmansk, 2009. (In Russian).

[5] V.M. Grebennikova, V.V. Doroshenko "Ethics of Teaching in the Context of New Educational Paradigm", Modern Problems of Science and Education , **1**(1), 1056, 2015. (In Russian).

[6] I.M. Shadrina, "Model of Building the Moral Culture of a Student – Future Teacher", The Volga district pedagogical bulletin, **6**(3), 117-125, 2018. (In Russian).

[7] N.E. Skripova, "The professional ethics for teachers: perspective directions of the professional development system", Professional development system scientific bases, **2**(23), pp. 28-35, 2015. (In Russian).

[8] E.V. Pyaterikova, "Pedagogical Foundations of Ethical Competence for Future Teachers", Nizhny Novgorod State Linguistic University Bulletin. Issue 21. Intercultural communication and teaching foreign languages. Challenges in Education, 181-188, 2013. (In Russian).

[9] M.V.Lamb & M. Wedell, "Cultural contrasts and commonalities in inspiring language teaching", Language Teaching Research, **19**(2), 207–224, 2015. http://dx.doi.org/10.1177/1362168814541716.

[10] E.A. Melyokhina, Ya.R. Bobohodzhaeva, "Motivation to Learning a Foreign Language: Pedagogical Aspect", Teacher in Higher Education in XXI century, 1, 89-93, 2016. (In Russian)

[11] M. Lamb, "The motivational dimension of language teaching", Language Teaching, **50**(3), 301-346, 2017. http://doi.org/10.1017/S0261444817000088.

[12] N.V. Matveeva, "Professionally Oriented Vocabulary in English Classes is Truly Fascinating", Secondary Vocational Education, **12**, 10-15, 2018. (In Russian)

[13] J. Macmurray, "Learning to be Human", Oxford Review of Education, 38, 2012 - Issue 6, 2012: Special issue: Learning to be human: the educational legacy of John Macmurray https://doi.org/10.1080/03054985.2012.745958

[14] C.A. Wolters, "Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning", Educational Psychologist, **38**(4), 189–205, 2003.

[15] K.A. Noels, "Learning Japanese, learning English: Promoting motivation through autonomy, competence and relatedness", In M. Apple, D. Da Silva & T. Fellner (Eds.), Language Learning Motivation in Japan, 15–34, 2013. http://doi.org/10.21832/9781783090518-004

[16] A.Y. Lozovoy, E.K. Zashchitina, "Online Education: Pros and Cons", Proceedings of 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS), 631–633, 2019. http://doi.org/10.1109/ITQMIS.2019.8928455.

[17] A. Assor, H. Kaplan, and C. Roth, "Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teaching behaviors predict students' engagement in schoolwork", British Journal of Educational Psychology, **72**, 261-278, 2002.

[18] H. Jang, J. Reeve, & E.L. Deci, "Engaging students in learning activities: It is not autonomy-support or structure but autonomy-support and structure", Journal of Educational Psychology, **102**, 588-600, 2010.

[19] S. Aydin, "Factors causing demotivation in EFL teaching process: A case study", Qualitative Report, **17**(51), 1–13, 2012.

[20] M. Bernaus, A. Wilson & R. C., "Gardner Teachers' motivation, classroom strategy use, students' motivation and second language achievement", Porta Lingarum, **12**, 25–36, 2009.

[21] J. Huang, T. Han, & K. Schnapp, "Do High-Stakes Test Really Address English Learners' Learning Needs? – A Discussion of Issues, Concerns, and Implications", International Journal of Learning and Development, **2**(1), 499-508, 2012.

[22] V.V. Zolotarev, M.N. Zhukova, "Role Model Features in Educational Serious Games", Proceedings of 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS), 583–586. 2019. http://doi.org/10.1109/ITQMIS.2019.8928292

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 2, 781-785 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Novel Quantum No-Key Protocol for Many Bits Transfer with Error Correction Codes

Duc Manh Nguyen, Sunghwan Kim*

School of Electrical engineering, University of Ulsan, 44610, Korea.

ARTICLE INFO

Article history:
Received: 04 March, 2020
Accepted: 11 April, 2020
Online: 24 April, 2020

Keywords:
Quantum cryptography
Quantum three stage protocol
Quantum error correction code

ABSTRACT

In this paper, a novel quantum transmission protocol which are based on the quantum no-key protocol is proposed. First, the quantum no-key protocol is discussed to show its important and its non-efficient on data transmission over quantum channel. Then, we improve it to carry many data bits via transmission. In addition, the error correction code is discussed and emerged into the proposed protocol to secure the transmission over quantum channel noise. Finally, the evaluation of transfer cost is discussed to show the effective of our proposed system.

1. Introduction

This paper is an extension of work originally presented in 2019 International Conference on Advanced Technologies for Communications (ATC) in Ha-Noi, Viet-Nam [1].

Quantum computation is problem solving and data computation using a system based on quantum mechanics, which is an effort to generalize classical computation. Quantum information systems could be able to transmit data fundamentally securely and solve complex problems better than classical information systems [2]. In 1994, Peter Shor invented a quantum algorithm for factoring integers into prime factors which runs on polynomial time [3]. In addition, in 1996 Lov Grover devised a quantum search algorithm for searching unstructured databases [4]. Hence, quantum algorithms promised big improvements on performance in comparison to classical computation; consequently, many problems have been considered in quantum computation [5]. However, the effects from imperfectly applied quantum gates, decoherence, and other quantum channel noise would affect the practical design of quantum computation. To overcome such problems, quantum error correction code (QECC) based on the theory of classical error correction code, was developed to protect quantum states from noise environment [6]. Since the first QECC was discovered by Shor [7], the method of containing redundancy to circumvent the no-cloning theorem has been popularly used. Therefore, the importance of QECC on practical building of quantum computer is no longer in doubt.

Cryptography is the research of secure communication techniques which allow only sender and intended receivers to view the content of shared messages. With the advantages of quantum mechanism such as entanglement and no-cloning theory, quantum cryptography is an application of quantum mechanics to cryptography. The first quantum cryptography protocol is BB84 which is invented by Dr. C. H. Bennett [8], it is public-key cryptography. In 1980, a novel private key cryptography protocol named Shamir no-key protocol is invented where the sender and receiver do not exchange the key, however the protocol requires the sender and receiver to have their own private keys for encrypting and decrypting the messages with the hard and complex on computation. Since quantum mechanism with no-cloning theorem which can be used to improve no-key protocol; then, in 2002 a quantum cryptography protocol based on Shamir's no-key protocol is considered [9]. In [10], in combination with Hill-cipher algorithm, the quantum no-key protocol was proven to be a scheme for secure direct quantum communications wherein the information is encrypted into a binary string and each bit is exchange between sender and receiver.

The main result of this research is to propose a new quantum no-key protocol which a single quantum state to establish the transfer of many information bits. In addition, the qubit needs to be protected from noise for the correct quantum state exchanged between Sender and Receiver. Hence, we emerge the proposed protocol with quantum error correction codes to establish the secure communication channel which can against the quantum noise, decoherence, and unwanted environments. The organization of this paper is as follows. In Section 2, the basic of quantum computation and quantum error correction codes are reviewed. In Section 3, we first propose the quantum no-key protocol; then, the

*Corresponding Author: Sunghwan Kim, sungkim@ulsan.ac.kr

emerging quantum no-key protocol with error correction code is discussed. Then, the conclusion is presented.

2. Preliminaries

2.1. Basic of Quantum Computations

Quantum theory uses qubit to represent information, the quantum systems with two levels such as: two polarization states of photons, two energy levels of atoms, etc. A qubit $|\psi\rangle = a|0\rangle + b|1\rangle$ is considered to be found in the both basis states $|0\rangle$ and $|1\rangle$ where the probability value we saw that qubit at state $|0\rangle$ is $|a|^2$ and at state $|1\rangle$ is $|b|^2$. It is superposition concept of a qubit, which is one of a main property of quantum information since the amount of information which presented in qubits are no limitation [11,12]. A qubit can be displayed in matrix form as,

$$|\Psi\rangle = \begin{bmatrix} a \\ b \end{bmatrix} = a\begin{bmatrix} 1 \\ 0 \end{bmatrix} + b\begin{bmatrix} 0 \\ 1 \end{bmatrix} = a|0\rangle + b|1\rangle. \quad (1)$$

According to norm condition for a qubit on the Bloch sphere space, the complex numbers a and b satisfy the equation $|a|^2 + |b|^2 = 1$. A n qubits system is constructed by multiple tensor products of some other qubits, it is given as follows,

$$|\phi\rangle = \sum_{i=0}^{2^n-1} a_i|i\rangle = \sum_{i_k=\{0,1\}} a_{i_1 i_2 \dots i_n}|i_1\rangle|i_2\rangle \dots |i_n\rangle. \quad (2)$$

where $i = \sum_{j=0}^{n-1} 2^j i_j$.

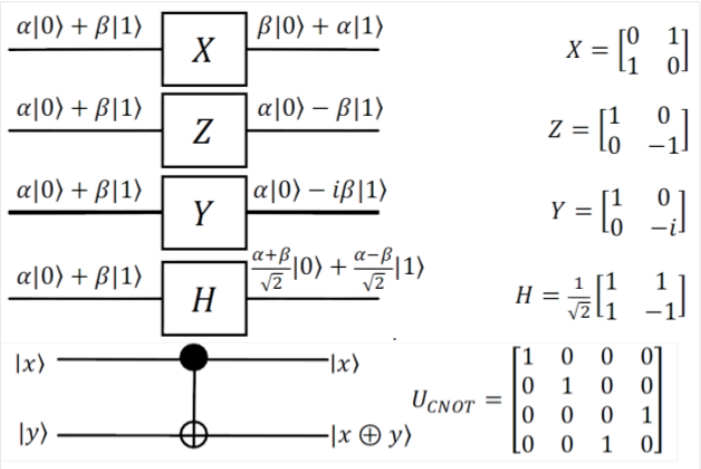


Figure 1: Basic quantum gates with matrices presentation

$E = e_1 \otimes e_2 \otimes \dots \otimes e_n$ where $e_i \in \{\mathbf{I}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$.

2.2. Basic of Quantum Error Correction Codes

The simplest QECC is three qubits repetition code that can correct only one type error, namely bit flip or phase flip [13]. The QECC can be simulated by quantum circuit model; they are showing in the Figure 2, 3 for three qubits repetition code. The difference between them are on the Clifford gate has been used: **X**=**HZH** and **Z**=**HXH**.

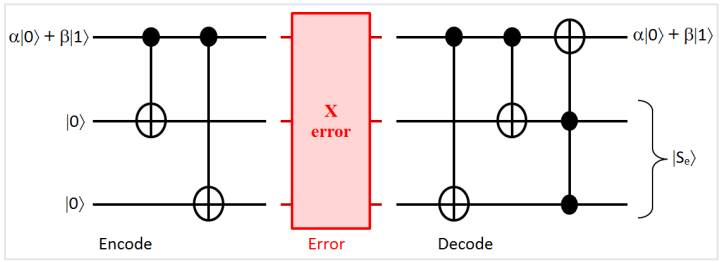


Figure 2: Quantum circuit for **X**-error 3- repetition code

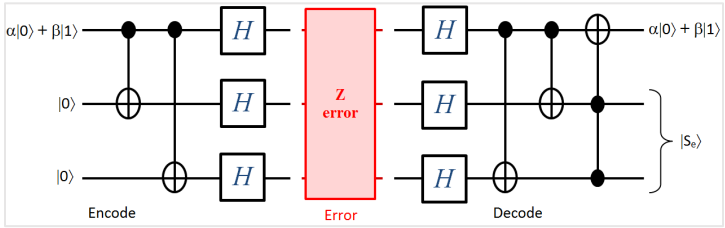


Figure 3: Quantum circuit for **Z**-error 3- repetition code

$$|\mathbf{0}_L\rangle = |000\rangle, |\mathbf{1}_L\rangle = |111\rangle. \quad (3)$$

Then, the quantum gates as previous mentioned must to be declared. Here, the quantum gates **X**, **Z**, and two new types of **CNOT** gates are used. The tensor product was used to extend one qubit to many qubits system.

To extend the first full quantum code, Shor code for 9 qubits is created by Shor, which use both bit-flip correction and phase-flip correction and can correct bit-flip, phase-flip, and their combination [14]. To do so, for one qubit is protected against phase-flip we need extend it to codeword of three qubits. Then, each qubits of that three-qubits need to extend to three-qubits to protect against bit-flip error. Hence, the quantum circuit starts with the initial information, we extend it to the 9-qubits via helps of ancilla 8 qubits of zeros, after transformation by encode step, the logical states or encoded qubits are created. Here, two basis states of codewords 9-qubits repetition as well as

$$|\mathbf{0}_L\rangle = \frac{1}{\sqrt{8}}(|000\rangle + |111\rangle)\otimes(|000\rangle + |111\rangle)\otimes(|000\rangle + |111\rangle), \quad (4)$$

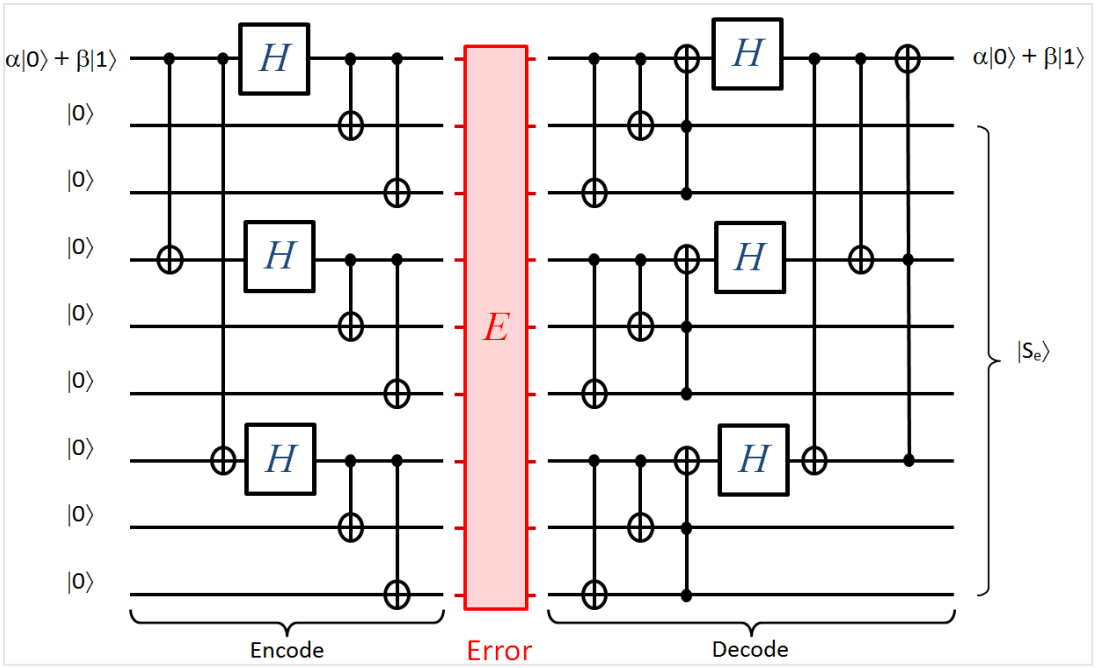


Figure 4: Quantum circuit for Shor code

$$|\mathbf{1}_L\rangle = \frac{1}{\sqrt{8}}(|000\rangle - |111\rangle)\otimes(|000\rangle - |111\rangle)\otimes(|000\rangle - |111\rangle). \quad (5)$$

Using matrices transformation, the states after applying error and decoding can be found. The final states show us the correction state can be recovered the syndrome $|S_e\rangle$ tell us which error has applied to logical states. The full quantum circuit for Shor code is given in Figure 4.

Generally, the quantum error correction code can be denoted as [[*n*,*k*,*d*]] where *n* is length of codeword, *k* is length of information bits, and *d* is minimum distance which corresponds to number of error the code can be detected and corrected. The parameter for Shor code at Figure 4 is [9,1,3], and it is the first quantum error correction code. Since the first proposed of Shor code, many novel quantum codes with diverse parameters have been proposed [15,16], they are called quantum stabilizer codes. The most important of quantum stabilizer codes are that they could be constructed by binary formalism, then the quantum circuit model are easy created which points are discussed detailly in [12].

3.1. Quantum no-key protocol

Quantum no-key protocol was invented by S. Kak [9], which is an emerging of classical no-key protocol with quantum mechanism. In this system, Sender and Receiver do not share the public key, they have their own private keys. The private key on quantum no-key protocol is an unitary transformation which has following form:

$$\mathrm{U}(\varphi) = \begin{bmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{bmatrix}. \quad (6)$$

The private keys in (6) are commutative to each other, it means $\mathrm{U}(\varphi_1)\mathrm{U}(\varphi_2) = \mathrm{U}(\varphi_2)\mathrm{U}(\varphi_1)$. If Sender or Receiver choose $\mathrm{U}(\varphi_3)$ for its private encryption key, the corresponding decryption key will be $\mathrm{U}(-\varphi_3)$ since $\mathrm{U}(\varphi_3)^{-1} = \mathrm{U}(-\varphi_3)$.

We assume that Sender and Receiver want to share the quantum state $|\varphi\rangle$, the protocol is proceeded as Figure 5. The detail of each steps are as follows,

1. Sender randomly choice private encryption key, $\mathrm{U}(\theta_S)$, apply it to quantum state to get $\mathrm{U}(\theta_S)|\varphi\rangle$. Then, transfer that encrypted state to Receiver.
2. Assume that there is no noise on the channel, Receiver gets the correct recipient from Sender, $\mathrm{U}(\theta_S)|\varphi\rangle$. Receiver also randomly choice its private encryption key, $\mathrm{U}(\theta_R)$, apply it to current quantum state to get $\mathrm{U}(\theta_R)\mathrm{U}(\theta_S)|\varphi\rangle$. Then, transfer back encrypted quantum state to Sender.
3. Assume that there is no noise on the channel, Sender gets correct recipient from Receiver, $\mathrm{U}(\theta_R)\mathrm{U}(\theta_S)|\varphi\rangle$. Sender applies the decryption private key according to step 1, $\mathrm{U}(-\theta_S)$, then current quantum state is $\mathrm{U}(\theta_R)|\varphi\rangle$. Sender transfers its state to Receiver.
4. Assume that there is no noise on the channel, Receiver gets the correct recipient, $\mathrm{U}(\theta_R)|\varphi\rangle$. Then, Receiver can get the original message by applying decryption private key according to step 3, $\mathrm{U}(-\theta_R)$. Finally, Receiver gets $|\varphi\rangle$. Sender and Receiver successfully exchange original quantum state.

The original quantum no-key protocol just mentions the exchange one bit, 0 or 1, for each round by attached it to quantum state, $|0\rangle$ or $|1\rangle$. In next subsection, we consider the proposed system for multi-bits transfer for each round. Then, it reduces the time of transferring long message.

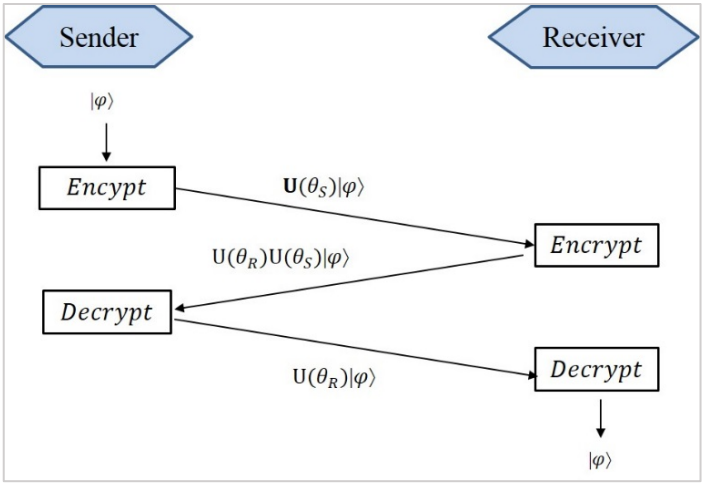


Figure 5: Quantum no-key protocol

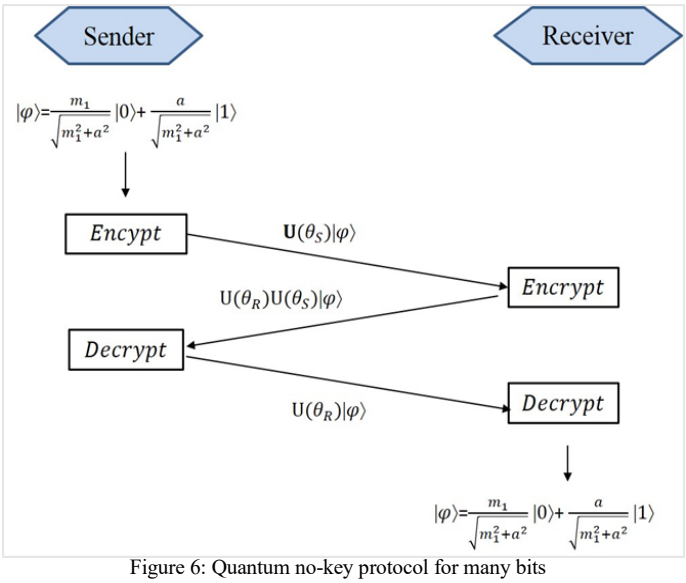


Figure 6: Quantum no-key protocol for many bits

3.2. Proposed quantum no-key protocol for many bits' transfers

We assume that Sender and Receiver wish to exchange the binary string $\mathbf{m}_1\mathbf{m}_2 \dots \mathbf{m}_k$, where each $\mathbf{m}_i$ is binary string of length l, that means the total length of binary string Sender and Receiver want to exchange is $k \times l$. We denote that binary string $\mathbf{m}_i$ has the decimal value m_i. The proposed protocol is as follows.

1. Sender and Receiver first exchange the binary string $\mathbf{m}_0$ by using the quantum no-key protocol. As analysis on Section 3.1, since $\mathbf{m}_0$ has length l, the cost of using quantum no-key protocol for successfully recipient is l. We denote a is the decimal value corresponding to binary string $\mathbf{m}_0$.

2. For binary string $\mathbf{m}_1$, Sender first encodes to the quantum state: $|\varphi\rangle=\frac{m_1}{\sqrt{m_1^2+a^2}}|0\rangle+\frac{a}{\sqrt{m_1^2+a^2}}|1\rangle$, where m_1 is the decimal value of binary string $\mathbf{m}_1$. The detail of using quantum no-key protocol is shown in Figure 6. We assume that there is no error from quantum channel. Then, Receiver gets correct quantum state $|\varphi\rangle=\frac{m_1}{\sqrt{m_1^2+a^2}}|0\rangle+\frac{a}{\sqrt{m_1^2+a^2}}|1\rangle$. Since the value of a is known before, Receiver will know exactly m_1 via measurement. Finally, Receiver will get binary string $\mathbf{m}_1$. The cost for exchange binary string $\mathbf{m}_1$ between Sender and Receiver is one time using quantum no-key protocol.

3. For remaining binary string $\mathbf{m}_2, \mathbf{m}_3, \dots, \mathbf{m}_k$, the same procedure as Step 2 is taken. The cost for successfully recipient is $k-1$.

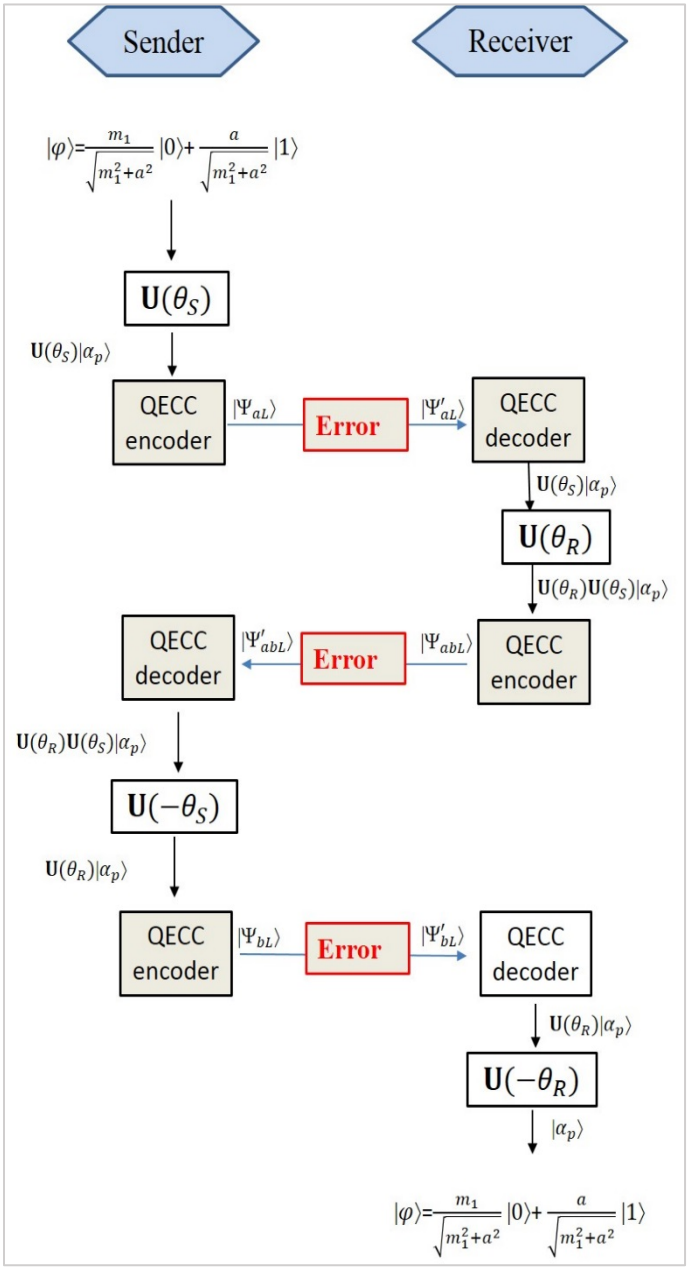


Figure 7: Quantum no-key protocol with error correction code

3.3. Analysis the cost of proposed protocol

For whole binary string $\mathbf{m}_1\mathbf{m}_2 \dots \mathbf{m}_k$, if we use original quantum no-key protocol, it is mentioned in [10], we need to use that protocol $k \times l$ times, since the length of this binary string is $cost = k \times l$. In contrast, if we use proposed protocol, we need the cost is $cost = k - 1 + 1 + l = k + l$. This shows that our propose system is suitable to transfer a long binary string.

In this subsection, we still assume that the quantum channel is perfectly. Hence, Sender and Receiver will transfer quantum state

successfully to each other. In next Section, we will consider the channel have error and we propose the using of quantum error correction code to reduce that problem.

3.4. Proposed quantum no-key protocol with error correction codes

It is proved that in quantum channel quantum state can interact with unwanted environment, which can make bit-flip, phase-flip, or both of them, bit-phase flip. In quantum no-key protocol, we need to transmit quantum state three times, if the error or environment is bad, the result of our system will be go down. Quantum error correction code (QECC) is the solution to solve that problem. The key ideas of QECC is that the quantum state, length k, is extended to have large length, length $n > k$, where $n - k$ redundant qubits we called them parity-check part. The novel QECC is discussed in Section 2.2 where we mention repetition codes for bit-flip and phase flip error, Shor codes for any error, bit-flip, phase-flip, or their combination.

Figure 7 shows the detail step of using QECC on quantum no-key protocol. The difference between Figure 7 and Figure 6 is that the QECC encoder and QECC decoder are inserted three times, since we have three times of transmission between Sender and Receiver.

4. Conclusion

In this research, we have proposed the quantum no-key protocol for many bits transfer which is suitable for long binary string transmission. The cost analysis is given in detail to show the effective of proposed system. In addition, the emerging of QECC on that proposed system is promised the way for secure transmission against the unwanted quantum channel.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the Research Program through the National Research Foundation of Korea (NRF-2019R1A2C1005920).

References

[1] D. M. Nguyen, S. Kim, "A quantum three pass protocol with phase estimation for many bits transfer", in 2019 International Conference on Advanced Technologies for Communications (ATC), Viet Nam. https://doi.org/10.1109/ATC.2019.8924514

[2] D. M. Nguyen, S. Kim, "Multi-Bits Transfer Based on the Quantum Three-Stage Protocol with Quantum Error Correction Codes", International Journal of Theoretical Physics 58 (6), 2043-2053, 2019. https://doi.org/10.1007/s10773-019-04098-4

[3] P.W. Shor, "Algorithms for quantum computation discrete logarithms and factoring", IEEE Computer Society Press. 124-134, 1994. 10.1109/SFCS.1994.365700

[4] L. Grover, "Quantum mechanics helps in searching for a needle in a haystack", Phys. Rev. Lett. 79, 325, 1997. https://doi.org/10.1103/PhysRevLett.79.325

[5] A. M. Zidan et.al, "A Quantum Algorithm Based on Entanglement Measure for Classifying Boolean Multivariate function into Novel Hidden Classes", Results in Physics, 15, 102549, 2019. https://doi.org/10.1016/j.rinp.2019.102549

[6] D. M. Nguyen, S. Kim, "Quantum stabilizer codes construction from Hermitian self-orthogonal codes over GF(4)", Journal of Communications and Networks 20 (3), 309-315, 2018. https://doi.org/10.1109/JCN.2018.000043

[7] P.W. Shor, "Scheme for reducing decoherence in quantum computer memory", Physical Review A, 52, 2493, 1995. 10.1103/physreva.52.r2493

[8] C. H Bennett, G Brassard, "Quantum cryptography: Public key distribution and coin tossing", International Conference on Computers, Systems & Signal Processing, Bangalore, India, Dec 9-12, 1984. http://researcher.watson.ibm.com/researcher/files/us-bennetc/BB84highest.pdf

[9] S. Kak, "A three-stage quantum cryptography protocol". Found Phys Lett 19, 293, 2006. https://doi.org/10.1007/s10702-006-0520-9

[10] A.A. Abdullah et.al, "A realizable quantum Three-Pass protocol authentication based on Hill-Cipher algorithm". Math. Probl. Eng., 2015, 481824, 2015. https://doi.org/10.1155/2015/481824

[11] D. M. Nguyen, S. Kim, "New Constructions of Quantum Stabilizer Codes Based on Difference Sets", Symmetry 10 (11), 655, 2018. https://doi.org/10.3390/sym10110655

[12] D. M. Nguyen, S. Kim, "Construction and complement circuit of a quantum stabilizer code with length 7", in 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN), 2016. https://doi.org/10.1109/ICUFN.2016.7537043

[13] D. M. Nguyen, S. Kim, "The fog on Generalized teleportation by means of discrete-time quantum walks on N-lines and N-cycles", Modern Physics Letters B 33 (23), 1950270, 2019. https://doi.org/10.1142/S0217984919502701

[14] D. M. Nguyen, S. Kim, "A novel construction for quantum stabilizer codes based on binary formalism", International Journal of Modern Physics B, 03/2020. https://doi.org/10.1142/S0217979220500599

[15] D. M. Nguyen, S. Kim, Quantum stabilizer codes based on a new construction of self-orthogonal trace-inner product codes over GF(4), International Journal of Modern Physics B, 34(5), 2050017, 02/2020. https://doi.org/10.1142/S0217979220500174

[16] D. M. Nguyen, S. Kim, New construction of binary and nonbinary quantum stabilizer codes based on symmetric matrices, International Journal of Modern Physics B, 33(24), 1950274, 2019. https://doi.org/10.1142/S0217979219502746

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 781-788 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Development of a Wireless Displacement Estimation System Using IMU-based Device

Tri Nhut Do*, Quang Minh Pham, Hoa Binh Le-Nguyen, Cao Tri Nguyen, Hai Minh Nguyen-Tran

Division of Electronics and Electrical Engineering, Faculty of Engineering, Van Lang University, HCMC 700000, Vietnam

ARTICLE INFO

Article history:
Received: 14 August, 2020
Accepted: 28 September, 2020
Online: 12 October, 2020

Keywords:
Displacement Estimation
Inertial Measurement Unit
Indoor Localization
Wireless Smart Sensor Network
Daily Activities Monitoring System

ABSTRACT

Estimation of displacement is an information required for daily operation monitoring systems to monitor human health or to locate users in buildings, basements, tunnels and similar places which under the same conditions that the global positioning signal (GPS) level is from very weak to completely absent; and is the measurement technique by using multimetric data fusion. Most current displacement estimation methods require a lot of infrastructures and devices such as UWB, wifi access points, cameras. Hence, estimation methods that ultilize inertial measurement unit (IMU) and integrate acceleration to get diplacement are effective alternatives since the three-axis accelerometer embedded in IMU usually low cost, easy to adjust and low noise. The advantage of this approach is that the IMU-based device is compact, easy to install and put on user's body. However, these methods expose some weaknesses when used in large-scale indoor structures such as multi-storey buildings due to the need to compensate azimuth estimation which is drifted overtime and is employed for calculating displacement with refer to earth frame as a base station. This article proposes a low-cost wireless displacement estimation system developed with IMU. The system employs a Kalman-filter type in indirect form for orientation estimates and Median-filter algorithm for classification of motion modes. In order to verify the proposed system in terms of accuracy and feasibility, a device was designed in a wearable form and tetsted on a multi-storey building in university. The wearable device ultilized IMU model MPU9250 and results recored wirelessly via Xbee devices in order to test the system performance in such senarios as climbing/descending staircases only, climbing/descending staircases through one floor combined with walking. Experiments are repeated for root mean square error (RMSE) computation based on the ground-truth. The proposed system performance is evaluated accordingly to RMSE. The experimental results demonstrate RMSE of 3.56%, 1.43%, for climbing/descending staircases only, climbing/descending staircases one floor combined with walking, respectively.

1. Introduction

The wireless displacement estimation system in this article is introduced in order to locate the position of user moving along a path indoor and carrying an wearable device only. Such pre-installed devices as fiducial markers on floor [1], infrared LED landmarks on panel [2], wireless communication devices like radio frequency and wireless local area network [3–5] are not required.

The applications of this system are diverse. High-demand areas of application can be listed including defense applications of tactical planning or strategic coordination for soldiers; social life applications for public safety and elderly care; and firefighter rescue applications when they're stuck and in need of rescue. The common characteristic of the aforementioned applications is the indoor environment with no pre-installed infractructures or devices due to such infractructures as RFID, wireless network, infrared sensor, landmark for vision are unable to installed or have already been destroyed by disaster. Specifically, rescue and security applications often take place in unstable areas that are unsuitable for infrastructure-based systems.

In general, two types of non-infrastructure displacement estimation systems exist and named as the strap-down and the step-and-heading. The strapdown method is integrating the acceleration in order to get displacement. Accumulated errors in the integration process are compensated with reset algorithm in [6]. The step-and-

*Corresponding Author: Tri Nhut Do, Division of Electronics and Electrical Engineering, Van Lang University, 45 Nguyen Khac Nhu Street, District No. 1, Ho Chi Minh City 700000, Vietnam, Cell No. +84938113898, trinhutdo@gmail.com

heading method estimates displacement by combining the step size with the yaw (user heading) which is estimated by quaternion-based filters [7–8].

The infrastructure-free displacement estimation system utilizes sensors such as IMU, force sensor, pressure sensor (barometer), ultra-wide band (UWB), etc. Depending on the specific application needs, the sensors are selected and their positions on user's body can be mentioned as holding in hand [9–10], mounting on foot [11–12], putting at ankle [13], placing at waist [14–16] and attaching at upper torso [17].

All mentioned approaches estimate displacement in two horizontal dimensions. However, Diaz [18] proposed an estimator with IMU put in pocket in order to estimate displacement in the vertical dimension. The inconvenient IMU location for user and the required training are the limitations of Diaz algorithm. Therefore, this article proposes a novel method that doesn't require training but provide a displacement estimation algorithm for indoor three dimensions (3D), which is able to estimate both the horizontal displacement and vertical displacement. The proposed algorithm is based on the double integration of acceleration with reset at step events in order to eliminate accumulate error due to integrating process. The proposed method is tested when a person climbing/descending staircases.

In this article, the IMU is placed at pedestrian's waist as shown in Figure 1 for convenient usage.

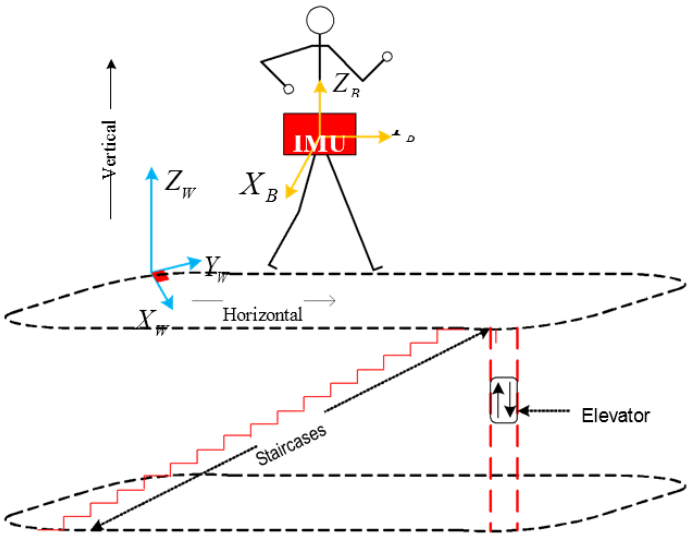


Figure 1: IMU is mounted at pedestrian's waist and used to estimate 3D displacements.

The rest of the article is organized as follows. The proposed system and algorithm to estimate 3D displacements are introduced in section 2. In section 3, the experiment setting is described and the experimental results are shown. Section 4 concludes the article.

2. Proposed Wireless Displacement Estimation System

2.1. System Overview

The proposed wireless displacement system employed the integrating process with novel velocity update algorithm in [6] in order to convert measured acceleration into displacement. Moreover, the system ultilized IMU module MPU9250 for sensing, Xbee devices for wireless communication and Arduino Fio Board for programming the proposed algorithm. The system was implemented within space as in Figure 1 and various experimental scenarios were designed for testing and evaulating as well. People took part in experiments were tasked to wear mobile device at waist, perform multi-activities freely with normal speed. Experiment data was recorded wirelessly in distance of several hundred meters.

2.2. Attitude Estimation System

In order to make the theoretical basis for all math operations in this article, all coordinate systems (or frames) must be assigned and two frames are assigned as illustrated in Figure 1. The subscript notations of W and B indicate for World and Body, respectively. The East-North-Up (ENU) coordinate system is used as the world reference frame ($X_WY_WZ_W$) while the body coordinate system is assumed to rigidly attached to the IMU and the IMU is mounted at the pedestrian's waist.

Let $\hat{\psi}$, $\hat{\theta}$ and $\hat{\varphi}$ be the user heading, IMU pitch and roll angles, respectively. An attitude in three dimensions is estimated based on a quaternion-based Kalman Filter in indirect form by fusing the information from the gyroscope, the accelerometer and the magnetometer.

The employed attitude estimation algorithm has two advantages as follows: 1) reducing computation cost due to dealing with orientation errors instead of dealing direct with orientation which results the state dimension being smaller and its response faster, and 2) two stages update using acceleration and magnetic strength by exploiting two measurements of the difference between accelerometer reading vs the gravity vector and the difference between magnetometer reading vs the magnetic strength vector.

2.3. Classification of Motion Modes

In order to improve displacement estimation accuracy, the motion modes of user carrying IMU must be recognized precisely [19]. Some motion modes including jogging, running, walking with different step sizes have already been done in our previous works [13, 17, 20]. In this article, motion modes of stationary, going up (climbing), going down (descending) staircases and riding a lift are added to update our previous work and enhance our system capabilities. Therefore, the motion modes are organized and indicated by numbers as follows: '0' indicates stationary mode; '1' indicates descending staircases mode; '2' indicates walking on flat surface mode; '3' indicates climbing staircases mode; '4' indicates riding a lift mode; '5' indicates jogging/running mode.

There are many motion modes introduced in [20] and were classified by machine learning based on four decision trees. The device are able to be put in many places including handheld with both portrait and landscape orientations; backpack; purse; belt holder; chest strapped; arm strapped hand still by side; dangling; pocket (consists of pant pocket, shirt pocket, and jacket pocket, in both horizontal and vertical orientations); wrist (ex: smartwatches) and headmount (such as smartglasses).

The algorithm in [20] is a very convenient but it is complicated and time consuming for learning method. Therefore, a simpler method for classification of motion modes is proposed in this article. The entire classification process, as shown Figure 2, consists of: 1) the acceleration in world frame is normalized and compared with a threshold in order to categorize motion modes into stationary, non-stationary and riding a lift modes, and 2) the non-stationary mode is classified into jogging/running, walking, climbing and descending staircases.

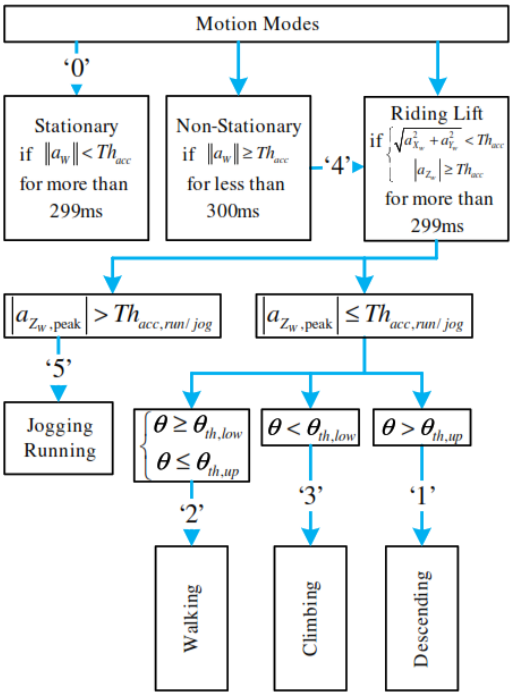


Figure 2: Motion Modes recognized by our work. The notation a_W is the acceleration in world frame and the a_{z_w} is its vertical component. All the Th_{acc}, $Th_{acc,\ run/jog}$, $\theta_{th,\ low}$ and $\theta_{th,\ up}$ are thresholds for motion modes classification.

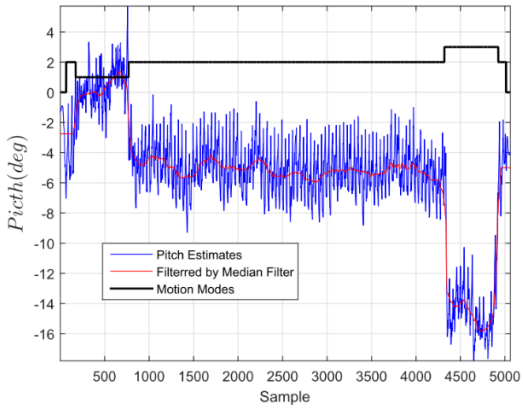


Figure 3: The blue solid line shows pitch angle estimates whereas the red one is the filtered pitch and the black solid line shows the Motion Modes, with values '0', '1', '2', '3' represents modes of stationary, descending staircases, walking and climbing up staircases, respectively.

2.3.1. *Classification for Non-Stationary mode group:*

The mode group of non-stationary consists of walking, jogging/running, climbing/descending staircases. Among them, the climbing/descending staircases motion modes are mentioned in this articles. Basically, classification of climbing/descending staircases is based on the pitch angle thresholds. Firstly, the original pitch estimates (depicted by the blue solid line in Figure 3) are passed through a Median Filter with a 100 samples window before doing a comparison with the pitch angle thresholds. The filtered pitch estimates are depicted by the red solid line in Figure 3. In order to illustrate the feasibility of the classification algorithm, an experiment that a pedestrian performs the following motion modes: stationary, walking, climbing up and descending down staircases with normal step sizes, moving at speed two steps per second, going along with a straight path is conducted and the pitch estimates are shown in Figure 3.

2.3.2. *Classification for Riding Lift mode*

In order to recognize whether a pedestrian riding a lift or not, the surface acceleration and the absolute of vertical acceleration are compared with a threshold. A pedestrian activity is classified into riding a lift mode if the condition shown in Figure 3 is satisfied. A person is tasked to perform a riding lift activity combined with walking then descending staircases in an experiment. Figure 4 proves the feasibility of classification riding a lift mode.

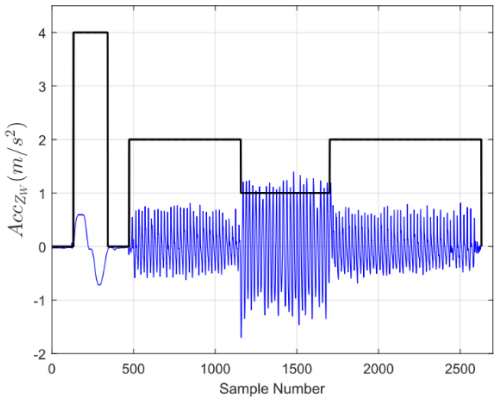


Figure 4: The blue solid line shows vertical acceleration with gravity removal whereas the black solid line shows the Motion Modes, with values '0', '1', '2', and '4' represents modes of stationary, descending staircases, walking and riding a lift, respectively.

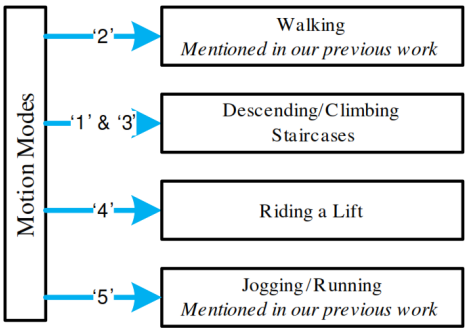


Figure 5: Overview of the proposed in-door localization system. Values '1', '2', '3', '4', '5' represents modes of descending staircases, walking and climbing up staircases, riding a lift and jogging/running, respectively.

2.4. *Displacement Estimation Method*

Authors' previous works [13, 17] described the pedestrian's walking motion and proposed two walking models applied to estimate displacement in two dimensions with the accuracy up to

97%. In order to enhance our displacement estimation system, an estimator applied for displacement estimation in (3D) including vertical and horizontal directions is developed. The details are described in this section.

2.4.1. System Overview

The displacement estimation system in this article consists of different algorithms accordingly to motion modes recognized by above described classification method. Figure 5 shows the block diagram of the proposed system for 3D displacement estimation.

2.4.2. 3D displacement estimation for descending/climbing staircases motion modes

During going up/down, the step events are defined at the extreme peak positions of the vertical acceleration, one component of the global acceleration with gravity removal. The method in [13] is used to detect the peaks.

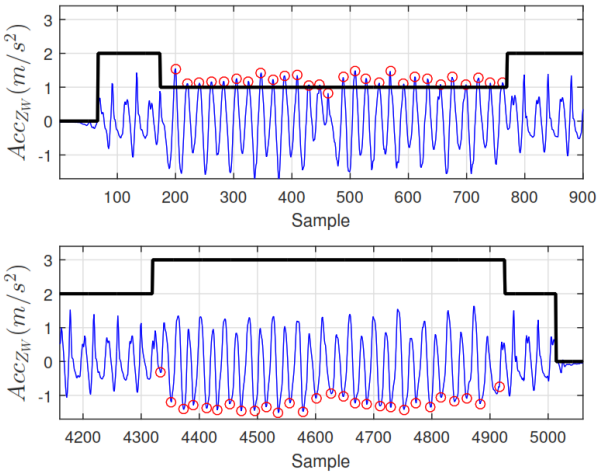


Figure 6: The black solid line shows the Motion Modes, with values '0', '1', '2', '3' represents modes of stationary, descending staircases, walking and climbing up staircases, respectively. The red rings are the extreme peak positions when descending (top) and climbing (bottom).

Figure 6. shows the peaks detected based on the data from the same experiment shown in Figure 3. The blue solid line is the vertical acceleration and the black solid line shows the Motion Modes. The step events are 100% detected.

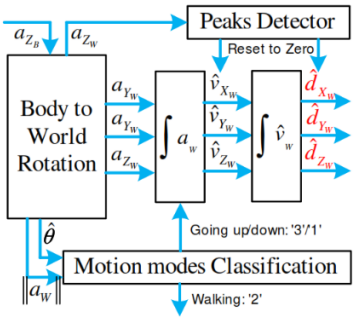


Figure 7: Block Diagram of the estimation algorithm for climbing/descending staircases.

Based on the fact that climbing/descending staircases are the periodic motions and the vertical acceleration is periodic data recored from the IMU mounted on the pedestrian's body, the entire displacement estimation process is able to be considered as the accumulation of segments. Each segment starts from a step event and ended at the next adjacent step event. However, the accumulating error due to integration affects the estimation accuracy as described in [20]. Unlike putting the IMU at the ankle, placing the IMU at pedestrian's waist has no chance of zero velocity interval when the normalized global acceleration is equal to zero. Therefore, the authors propose a method in order to obtain the 3D displacement is that integrating twice from global acceleration with gravity removal and that integrating process is reset to zero for velocity and displacement at the step events. The method is illustrated as in Figure 7. The reseting is called refreshing for segments.

2.4.3. 3D displacement estimation for riding a lift motion mode

The 3D displacement estimation when riding a lift is obtained by double integrating the vertical acceleration with gravity removal in world frame. The method was introduced in our previous work [19]. It is described that the integrated velocity is reset in zero vertical acceleration intervals during riding lift interval. Figure 8 illustrates the integrating process. Firstly, the discrete vertical acceleration with respect to the world frame is in the top plot and is with gravity removal when riding a lift one floor. Secondly, the middle plot is the integrated velocity and the bottom plot is the integrated vertical displacement. The red dot value 'non-zero' in the top plot indicates the zero vertical acceleration intervals where integrated velocity is reset.

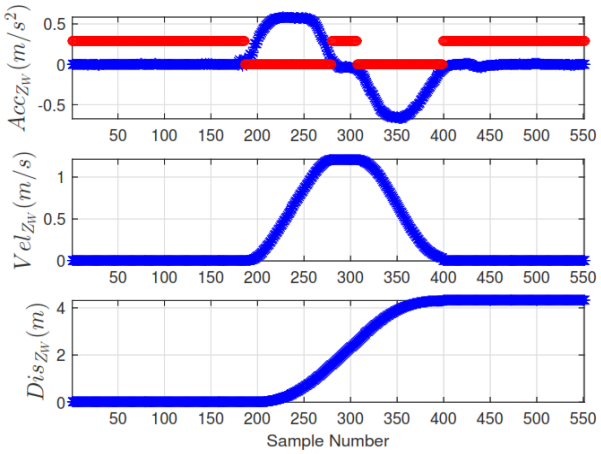


Figure 8: The vertical integration process when riding a lift through one floor.

3. Experiments and Evaluations

The capability of proposed method is tested through experiments conducted by four persons in our research group. In this section, a brief description of the experiment setup is first given, followed by a description of the experimental scenarios and the results obtained.

3.1. Set up

The hardware used to test the proposed method is illustrated in Figure 9, which consists of two parts: 1) the mobile part (attached to the pedestrian) shown in the top portion and 2) the base part shown in the bottom portion. The mobile part consists of an Xbee wireless transmitter, an MPU-9250 IMU sensor board and an Arduino Fio microcontroller Atmega328P board whereas the base part has an Xbee wireless receiver connected to the laptop by

virtual serial port via USB port. The transmitter communicates with Fio board using I^2C communication protocol. The serial baud rate is set at 115200bps.

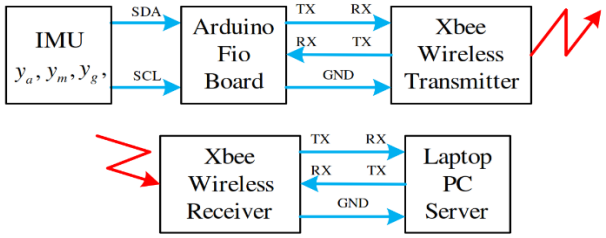


Figure 9: Hardware for experiments, with the mobile part (attached to pedestrian's waist) shown in the top and the base part shown in the bottom.

The proposed algorithm is implemented in the mobile part with the data acquired in every 25 ms and passed through a complimentary filter with 225 ms time constant.

Four experimental scenarios are set and conducted, namely: 1) which is described in section 3.2 is the experiment of going staircases along a straight path with two action modes climbing and descending for 3D displacement accuracy evaluation; 2) which is described in section 3.3 is the experiment of walking along with a triangular path put together with climbing and descending staircases for combination of heading and displacement accuracy evaluation; 3) which is described in section 3.4 is the experiment of riding a lift through floors; and 4) which is described in section 3.5 is the experiment of 3D displacement estimation that consists of many motion modes.

Four people in our research group aging from 25 to 39 with height ranging from 1.6 meters to 1.86 meters are deployed for these experiments, and all the experiments are conducted. Each test is also repeated for four times.

3.2. Results for climbing/descending staircases

For the first experiments scenario, the straight path includes 2m flat path and 15 stairs as shown in Figure 10. Each stair is 15cm in height and 30cm in length. The straight path has 2.25m ground truth in height (vertical direction) and 7.03m ground truth in horizontal direction. It produces 14.06m traveling ground truth in three dimensions.

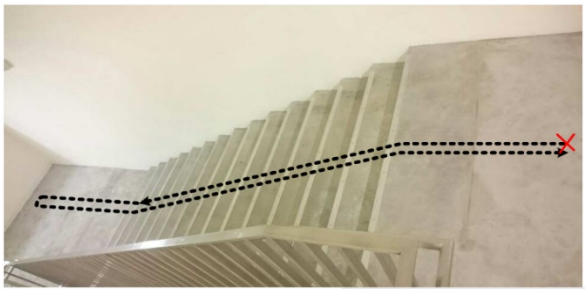


Figure 10: In this figure, the volunteers start from the red mark 'X' and performs motion modes of stationary, descending, climbing along a straight path depicted by the black bold dot line.

On stairs, both horizontal and vertical displacements are obtained by double integration with zero reset at step events. The heading is determined by method in [9]. Figure 11 shows the side view of an experimental result and illustrates the both vertical and horizontal estimates.

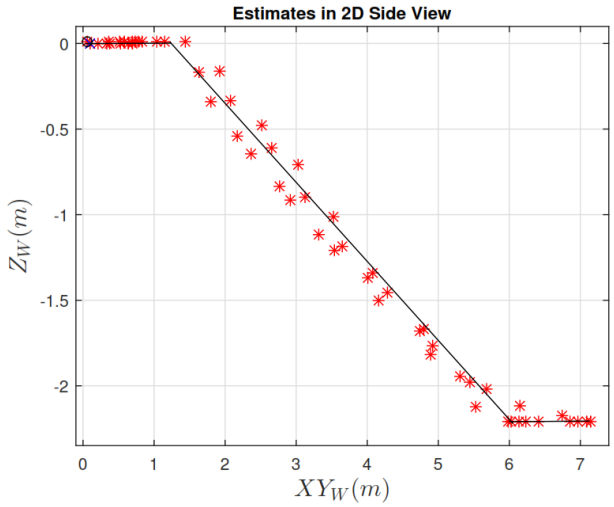


Figure 11: In this figure, both horizontal and vertical estimates are represented by the red asterisk '*' when walking, descending/climbing staircases along with a straight path. (XY_W) is the horizontal direction. The black solid line is the ground truth.

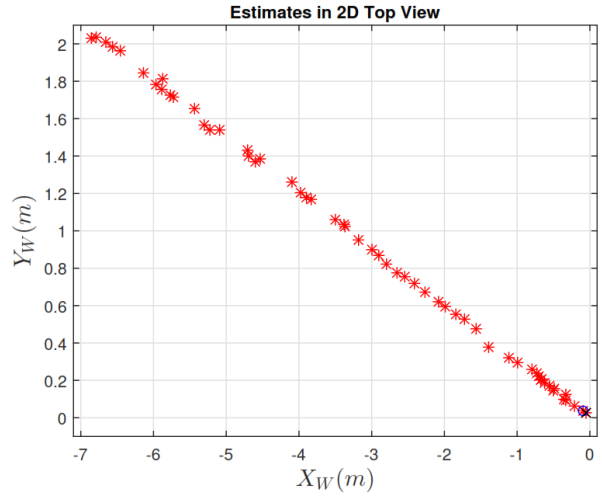


Figure 12: In this figure, the horizontal estimates are represented by the red asterisk '*' when walking, descending/climbing staircases along with a straight path.

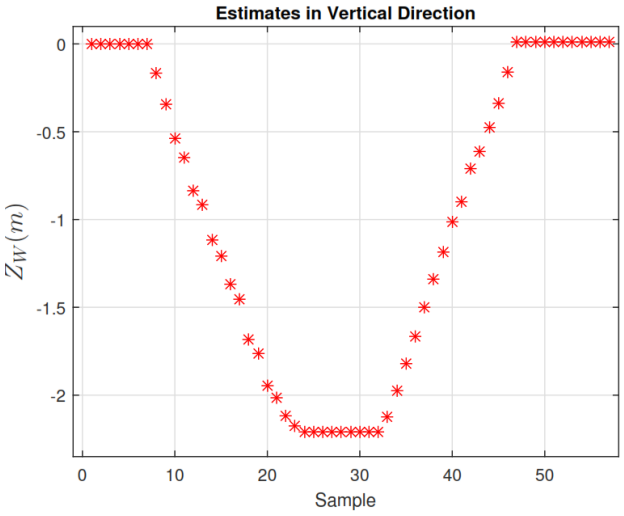


Figure 13: In this figure, the vertical estimates are represented by the red asterisk '*' when walking, descending/climbing staircases along with a straight path.

The vertical estimates are positions of pedestrian in upward direction (Z_W) and the horizontal estimates are computed by taking normalization of pedestrian positions in both north (Y_W) and east (X_W) directions. Let (XY_W) be the representation of horizontal direction.

Next, Figure 12 shows the view from top and illustrates the horizontal estimates only. Finally, Figure 13 illustrates the vertical estimates only.

In the three figures, the red asterisk '*' represents the estimated positions, the blue 'O' represents the starting point and the black 'X' represents the ending point.

3.3. Results for walking and climbing/descending staircases

For the second experimental scenario, a traveling path includes two triangular paths (Figure 14(a)) in two different floors which are connected by 30 stairs of 15cm in height and 30cm in length (Figure 14(c)). The triangle has 3 edges in 11m length and its vertices are circled with 1/3 ring of 1m radius. The pillar in Figure 16(a) is one of triangle vertices. The walking path from a triangle to staircases is 9m for the top level (Figure 14(b)) and 11m for the bottom level (Figure 14(d)). The walking path at middle of staircases is 3m (Figure 14(c)). In order to obtain the highest accuracy in measurement as ground truth for evaluation, the staircase height is measured by Fluke 411D Laser Distance Meter and the number of staircases between two floors are counted. There are 30 staircases with 14.5cm height between two floors at Singapore University of Technology and Design, Building 1. This experimental set produces totally 114.62m traveling ground truth in three dimensions.



(a) Triangular walking path



(b) 9m connecting walking path



(c) 30 Stairs



(d) 11m connecting walking path

Figure 14: The second experimental scenario in the building.

When walking, the walking model and method in [18] is applied for displacement estimation while double intergrating process is employed for displacement estimation when climbing/descending staircases. For illustrating, an experimental result in 3D is shown in Figure 15.

3.4. Evaluation for the first two experimental scenarios

In order to evaluate the method performance, all the Euclidean distances in three dimensions between the starting point and the ending point are computed for both described experimental scenarios. In addition, all the differences between estimated height for each experiment and the staircase height ground truth are also computed for the first experimental scenario. The root means square errors of the Euclidean distances and the height differences are then calculated. The evaluation results are summarized in Table 1.

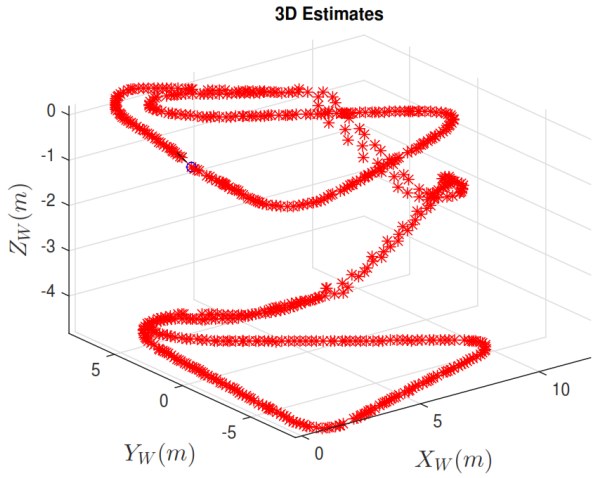


Figure 15: In this figure, the 3D estimates are shown when walking along a triangular path, climbing/descending 30 stairs.

Table 1: Performance of four people conducted experimental scenarios described in Section 3.3

	RMSE [m]	RMSE [%]
3D distance for descending, climbing along with Straight Path	0.5	3.56
Height Differences for descending, climbing along with Straight Path	0.26	11.56
3D distance for travelling on the Triangular Shape Path	1.64	1.43

3.5. Results and evaluations for riding a lift

In this experimental scenario, four members in our research group are tasked to ride a lift upward four times, then downward four times. Three experimental sets are designed such as riding a lift through one floor only as shown in Figure 8, riding lift through multiple floors with one floor stop as shown in Figure 16, riding lift through 40 floors as shown in Figure 17.

Figure 8 shows one of 16 experimental results in total for the first experimental set including 2 upward direction and 2 downward direction rode by 4 persons at Building 1. The experimental results are evaluated by the root mean square error based on the ground-truth of 4.35 meters, which is 0.0334 meters or 0.77%.

Figure 16 shows one of 8 experimental results in total for the second experimental set including 1 upward direction and 1 downward direction rode by 4 persons at Building 1. The experimental results are evaluated by the root mean square error based on the ground-truth of 4.35 meters, which is 0.0384 meters or 0.88%.

Figure 17 shows one of 4 experimental results in total for the third experimental set including 2 upward direction and 2

downward direction rode by 1 person at Block 90. There are 20 staircases with 13.5cm height between two floors of Block 90. The experimental results are evaluated by the root mean square error based on the ground-truth of 2.7 meters, which is 1.7909 meters or 1.66%.

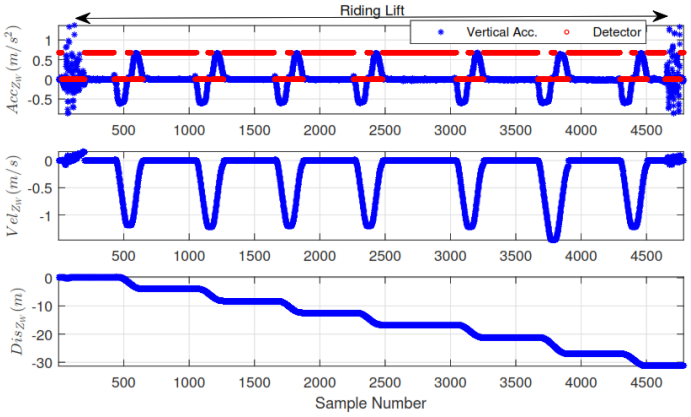


Figure 16: In this figure, the vertical estimates are shown when riding a lift through multiple floors.

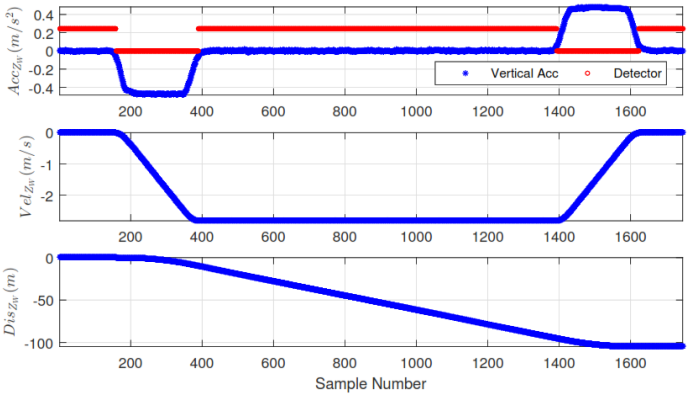


Figure 17: In this figure, the vertical estimates are shown when riding a lift through 40 floors.

3.6. Result and evaluation for 3D displacement estimation

A person is tasked to conduct experiments of walking from starting point, entering a lift, riding a lift upward one floor, followed by a walking out of a lift, descending staircases and walking again to starting point. The experimental result of this scenario is illustrated as in Figure 18.

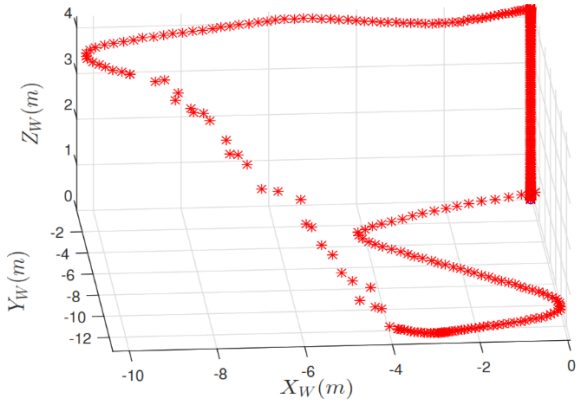


Figure 18: In this figure, the 3D estimates are shown when riding a lift through one floor then walking and descending straircases.

The experiment is repeated 4 times. The Euclidean distances between starting point and ending point are calculated for evaluation, which distance average of four experimental results is 0.71 meters.

4. Conclusion

A simple motion mode classification for climbing/descending staircases comes with an algorithm to estimate both vertical and horizontal displacements by utilizing information from global acceleration with gravity removal have been proposed in this article. The estimation based on integrating process with zero reset at the extreme peaks of the vertical, rather than horizontal, acceleration. Authors have given new method to estimate indoor 3D displacements as a basic for positioning of pedestrian without knowing of stair measurements and provided the most convenient IMU location at waist. In order to verify displacement estimation system's feasibility and evaluate its performance, a range of experiments which are grouped into two sets of experiment including straight path, and triangular shape path are conducted by four subjects. In each set, people who tasked to perform experiments are required to move at normal step size and at speed 2 steps per second. The experimental results have demonstrated the capability of the algorithm via a RMS displacement error of 3.56% for straight path and 1.43% for the triangular shape path.

Finally, the algorithm in this article is only focusing on climbing upstairs and descending downstairs. The displacement estimations when pedestrian acts in other motion modes like standing, walking on escalator and standing in evaluator will be researched in the future work.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This project is supported by Singapore University of Technology and Design (SUTD) Temasek Lab, Pillar of Engineering Product Development (EPD), Changi 487372, Singapore.

In addition, this project is supported by Van Lang University (VLU) Electronics and Electrical Engineering Division, Ho Chi Minh City 700000, Vietnam.

References

[1] T. N. Do, and Y. S. Suh, "Gait Analysis Using Floor Markers and Inertial Sensors" Sensors, **12**(2), 1594-1611, 2012. https://dx.doi.org/10.3390%2Fs120201594

[2] T. N. Do, and Y. S. Suh, "Gait Analysis System Using Infrared LED Landmarks" Journal of Institute of Control, Robotics and Systems, **17**(7), 641-646, 2011. https://doi.org/10.5302/J.ICROS.2011.17.7.641

[3] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems" IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, **37**(6), 1067-1080, 2007. https://doi.org/10.1109/TSMCC.2007.905750

[4] K. El-Kafrawy, M. Youssef, A. El-Keyi, and A. Naguib, "Propagation modeling for accurate indoor wlan rss-based localization" in IEEE 72nd Vehicular Technology Conference Fall, Ottawa, Canada, 2010. https://doi.org/10.1109/VETECF.2010.5594108

[5] L.-H. Chen, E. H.-K. Wu, M.-H. Jin, and G.-H. Chen, "Intelligent fusion of Wi-Fi and inertial sensor-based positioning systems for indoor pedestrian navigation," IEEE Sensors Journal, **40**(11), 4034-4042, 2014. https://doi.org/10.1109/JSEN.2014.2330573

[6] T. N. Do, U-X. Tan, "Novel Velocity Update Applied for IMU-based Wearable Device to Estimate the Vertical Distance" IEEE 1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia, 25-25 Nov. 2019. https://doi.org/10.1109/ICECIE47765.2019.8974671

[7] X. Yun, E. R. Bachman, and R. B. McGhee, "A Simplified Quaternion-Based Algorithm for Orientation Estimation from Earth Gravity and Magnetic Field Measurements" IEEE Transactions on Instrumentation and Measurement, **57**(3), 638-650, 2008. https://doi.org/10.1109/TIM.2007.911646

[8] Y. S. Suh, "Orientation estimation using a quaternion-based indirect Kalman filter with adaptive estimation of external acceleration" IEEE Transactions on Instrumentation and Measurement, **59**(12), 3296-3305, 2010. https://doi.org/10.1109/TIM.2010.2047157

[9] V. Renaudin, M. Susi, G. Lachapelle, "Step Length Estimation using Handheld Inertial Sensors" Sensors, **12**(7), 8507-8525, 2012. https://doi.org/10.3390/s120708507

[10] H. Zhang, W. Yuan, Q. Shen, T. Li, and H. Chang, "A Handheld Inertial Pedestrain Navigation System with Accurate Step Modes and Device Poses Recognition" IEEE Sensors Journal, **15**(3), 1421-1429, 2015. https://doi.org/10.1109/JSEN.2014.2363157

[11] H. M. Schepers, E. H. van Asseldonk, C. T. Baten, and P. H. Veltink, "Ambulatory estimation of foot placement during walking using inertial sensors" Journal of Biomechanics, **43**(16), 3138-3143, 2010. https://doi.org/10.1016/j.jbiomech.2010.07.039

[12] E. Bachmann, J. Calusdian, E. Hodgson and X. Yun, "Insitu heading drift correction for human position tracking using foot-mounted inertial/magnetic sensors" IEEE International Conference on Robotics and Automation (ICRA), 5425-5430, Saint Paul, MN, May. 2012. https://doi.org/10.1109/ICRA.2012.6225007

[13] T. N. Do, R. Liu, C. Yuen, and U-X. Tan, "Design of an infrastructureless indoor localization device using an IMU" IEEE International Conference on Robotics and Biomimetics, 2115-2120, China, Dec. 2015. https://doi.org/10.1109/ROBIO.2015.7419086

[14] A. Kose, A. Cereatti, and U. D. Croce, "Estimation of traversed distance in level walking using a single Inertial Measurement Unit attached to the waist" IEEE International Conference in Medicine and Biology Society, 1125-1128, 2011. https://doi.org/10.1109/IEMBS.2011.6090263

[15] J. C. Alvarez, D. Alvarez, A. Lopez, and R. C. Gonzalez, "Pedestrian Navigation Based on a Waist-Worn Inertial Sensor" Sensors, **12**(8), 10536-10549, 2012. https://doi.org/10.1109/I2MTC.2012.6229714

[16] F. Inderst, F. Pascucci, M. Santoni, "3D pedestrian dead reckoning and activity classification using waist-mounted inertial measurement unit" IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN), 1-9, 2015. https://doi.org/10.1109/IPIN.2015.7346953

[17] T. N. Do, R. Liu, C. Yuen, M. Zhang and U-X. Tan, "Personal Dead Reckoning using IMU mounted on upper Torso and inverted Pendulum Model" IEEE Sensors Journal, 2016. https://doi.org/10.1109/JSEN.2016.2601937

[18] E. M. Diaz, "Inertial Pocket Navigation System: Unaided 3D Positioning" Sensors, **15**(4), 9156-9178, 2015. https://doi.org/10.3390/s150409156

[19] M. Elhoushi, J. Georgy, A. Noureldin, M. J. Korenberg, "Motion Mode Recognition for Indoor Pedestrian Navigation Using Portable Devices" IEEE Transactions on Instrumentation and Measurement, **65**(1), 208-221, 2016. https://doi.org/10.1109/TIM.2015.2477159

[20] T. N. Do, R. Liu, C. Yuen, and U-X. Tan, "Personal Dead Reckoning using IMU device at upper torso for walking to running" IEEE International Conference on Sensors, Orlando, FL, USA, Nov. 2016. https://doi.org/10.1109/ICSENS.2016.7808521

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 781-787 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Optimal Irrigation Strategy using Economic Model Predictive Control

Luisella Balbis*

Department of Computer Engineering, University of Bahrain, Kingdom of Bahrain

ARTICLE INFO

Article history:
Received: 05 August, 2020
Accepted: 17 November, 2020
Online: 24 November, 2020

Keywords:
Soil Moisture
System Modelling
Optimization

ABSTRACT

In many countries, irrigation water is one of the major contributors to water scarcity. In the present study, a novel optimized irrigation system which minimizes water consumption in irrigation is presented. The system is based on a predictive control algorithm, which foresees the water need of the crop, and regulates the time and amount of irrigation to maintain the soil moisture around an optimal level, while taking into account system constraints. The predictive feature of the algorithm requires a model of the soil moisture, which is obtained from the actual meteorological data of the Kingdom of Bahrain. The optimization problem is formulated as an Economic Model Predictive Control (EMPC) problem and implemented using MATLAB. The simulation experiments show that the novel system yields a reduction of water consumption around 8% and 16% compared with the PID and On-off controllers, respectively, while maintaining an optimal soil moisture level.

1. Introduction

Reducing water consumption is a fundamental constituent of water management, particularly in water-scarce areas. In many countries, irrigation water is one of the major contributors to water scarcity, and reducing irrigation water while increasing water efficiency is a way to overcome this issue. This work, which is an extension of the paper originally presented in 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO) [1], looks at different practices used to reduce the use of irrigation water and proposes the use of an advanced optimization technique as possible solution to the problem.

Different mulching practices, irrigation techniques and irrigation strategies can reduce water consumption in irrigation. Mulching refers to the practice of covering the soil with protective material to reduce evapotranspiration. Irrigation techniques (dripping, furrow, sprinkler etc.) refer to the way irrigation water is applied, which also affects evapotranspiration. Irrigation strategies refer to the timing and the amount to irrigation water provided to the crop. Irrigation scheduling, which refers to when and how much to irrigate, is a fundamental concept associated with irrigation strategies [2], [3]. Although mulching practices, irrigation techniques and irrigation strategies are all important tools, in [2] it is found that for optimal results, the irrigation strategy should be improved first, followed by mulching practices and the irrigation technique. Therefore, in the present study we will focus on irrigation strategies and, in particular, on water-efficient irrigation technology.

Technological developments include automated irrigation systems, that irrigate at set times and for a certain amount of time. Although these automated techniques present advantages in terms of increased water efficiency and reduced farmers'intervention, the performances of these open loop control systems in terms of water consumption are not yet optimal. The main reason of their unsatisfactory performance is that they can cause under- or over-watering, since weather and crop conditions are not taken into account. This of course can lead to spoiled crop and water wastage.

Greater expectation is placed on technologies which include sensors and closed-loop watering systems [4]-[6]. The simplest approach uses sensors in an on/off control scheme, where irrigation is turned on or off depending on the reading of a moisture sensor above or below set values.

More complex control strategies, which often are model-based, have been described in different papers [7]-[10]. In general, model-based control strategies require models of the soil moisture dynamics including variables such as climatological data, crop water needs, water saturation etc. [7], [8]. Among such advanced control strategies is Model Predictive Control (MPC), as proposed in the research described in [9] and [10]. The models presented in [9] and [10] are based on the water balance equation, which requires measuring or estimating the initial soil moisture, rainfall,

*Corresponding Author: Luisella Balbis, lbalbis@uob.edu.bh

water capacity of the soil and evapotranspiration (ET). From these measurements, the water balance equation gives an estimation of the soil moisture. In these papers, the MPC algorithm is formulated as an optimization problem, which reduces set point tracking errors.

Similarly to the work presented in [9] and [10], our work describes the development of a soil moisture dynamic model and its application in a predictive control setting. However, the predictive control strategy here proposed is based on Economic Model Predictive Control (EMPC), which aims at maximizing the economic performance of the irrigation system rather than reducing tracking errors. Constraints in the optimization problem are used to guarantee that optimal soil moisture is obtained. Moreover, in the present study the soil moisture model is based on atmospheric conditions of the Kingdom of Bahrain [11].

The remainder of this paper is organized as follows. Section 2 describes the derivation of the soil moisture model based on the water balance equation and the measurement of meteorological data of the Kingdom of Bahrain. Section 3 describe the principles of MPC and presents EMPC as a special case of the MPC policy. In Section 4 the outcomes of employing the EMCP are shown and comparisons are drawn with On-off and PID automated irrigation systems. Finally, conclusion and recommendations are included in Section 5.

2. System Identification

This section presents the grey-box modeling of the soil moisture dynamics using difference equations. Grey-box modeling is a method for identifying the system's dynamics that uses measurement obtained from experimental data combined with physical knowledge of the system. For parameters estimation, state space models are often used, where the state, input and output of the system are related by difference equations.

2.1. Physical Modelling

2.1.1. Water-balance equation

The model developed here is based on the water-balance equation represented graphically in Figure 1 [9], [10]:

$$M(k+1) = M(k) + IR(k) + RAIN(k) + CR(k) - ET_c(k) - DP(k) - RO(k) \quad (1)$$

where k is the sampling time, M is the moisture of the soil, IR is the irrigation water, $RAIN$ is the precipitation, CR is the capillary rise, ET_c is the crop evapotranspiration, DP is the deep percolation and RO is the water runoff.

Equation (1) expresses the variation in soil moisture as the difference between the water influx (capillary rise, precipitation and irrigation) and the water losses (evaporation, transpiration, water runoff and deep percolation) [10].

Geomorphological and meteorological data of the Kingdom of Bahrain show a generally flat land and very scarce precipitations. Therefore, it is reasonable to assume that water runoff, precipitation, and capillary rise are negligible. Under these assumptions, (1) can be reduced to:

$$M(k+1) = M(k) + IR(k) - ET_c(k) - DP(k) \quad (2)$$

Equation (2) reduces the dependency of soil moisture to three variables: irrigation, evapotranspiration and deep percolation.

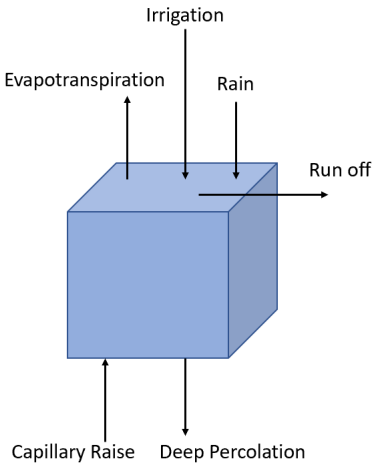


Figure 1: Water Balance Model

2.2. Evapotranspiration Estimation

Evapotranspiration is defined as the movement of water from the soil and the crop into the atmosphere. Many factors influence evapotranspiration rate, such as crop type and growth stage, weather factors (air temperature, wind, humidity etc.), environmental conditions (e.g. soil salinity) and land management (use of fertilizers and pesticides, etc.) [12].

Different approaches exist to estimate evapotranspiration [13]. In the present study, the FAO Blaney-Criddle estimation approach has been used. Despite being a simpler method compared to others, it has been proven to be satisfactory under a variety of weather conditions [14], [15].

Using this method, the actual evapotranspiration ET_c is estimated for a particular type of crop C as:

$$ET_c = K_c \cdot ET_o \quad (3)$$

In (3), the crop coefficient K_c depends on the crop type and the growth stage of the crop. ET_o is the estimation of the evapotranspiration from the reference crop with hypothetical qualities (actively growing green grass of 8–15 cm height). In the present study, ET_o is assumed to be affected by climatic parameters, according to the following empirical FAO Blaney – Criddle equation:

$$ET_o = p \cdot (0.457 \cdot T_{mean} + 8.128) \quad (4)$$

where T_{mean} is the average daily temperature [°C] and p is the average annual percentage of daytime.

2.3. Deep Percolation

Deep percolation (DP) is a hydrologic process, where water penetrates below the roots of the crop. The quantity of percolation during irrigation depends on the characteristics of the soil as well as the soil humidity.

In the present study, DP is assumed to be directly proportional to M with proportionality constant K_p: when M increases, DP

increases due to differential pressure [16]. This is expressed by the following equation:

$$DP(k) = K_p \cdot M(k) \tag{5}$$

In (5), K_p is a constant that depends on the type of soil (sandy, clay, silt, loamy, etc)

Substituting (3) and (5) in (2) we get a difference equation which leads to a model of the water balance dynamic in the form:

$$M(k+1) = K_s \cdot M(k) + IR(k) - K_c ET_o(k) \tag{6}$$

where K_s=1-K_p.

2.4. Model Structure

Equation (6) gives a mathematical model which contains two parameters, K_s and K_c. In our work, these parameters are unknown, meaning that the value of crop coefficient and the type of soil used in our experiment are unknown. Therefore, a gray-box system identification process, which uses a the first principles model of the system and data coming from direct system measurements, was used to estimate the values of the unknown parameters.

The linear difference equation (6) can be re-written in the form of a discrete state space model as:

$$M(k+1) = [K_s]M(k) + [1]IR(k) + [-K_c]ET_0(k) \tag{7}$$

$$M(k) = [1]M(k) \tag{8}$$

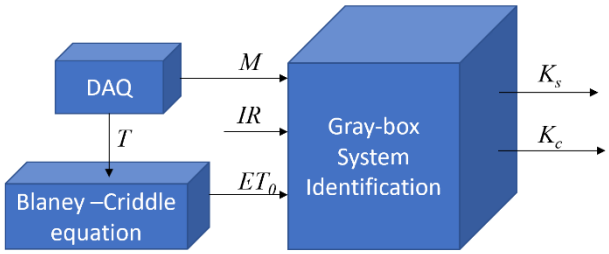
where $M(k)$, $IR(k)$, $ET_0(k)$ are the state vector, the input signal, and the disturbance of the system.

The state-space model (7) and (8) presents two parameters, K_s and K_c, that need to be estimated from measured data.

2.5. Parameters Estimation

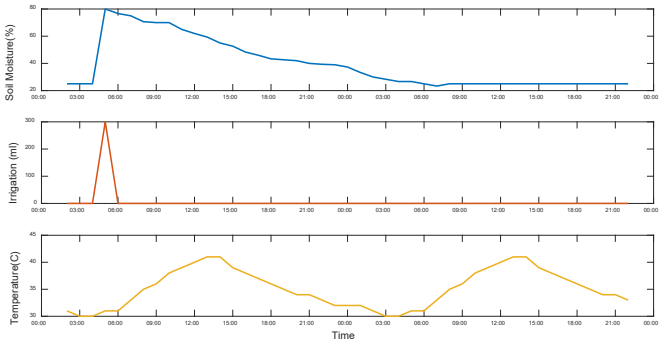
Parameter estimation was carried out using Matlab System Identification Toolbox R2018a, which provides a tool for deriving mathematical models of dynamic systems given measured input-output data. The toolbox was used to estimate the values of K_s and K_c, through a gray-box model estimation process.

Figure 2 shows the process of model identification.



A set of measurements (see Figure 3), called identification data, was collected using a data acquisition system implemented on an Arduino Uno microcontroller board.

The experimental setup consists of a temperature sensor LM35, and a soil moisture sensor FC-28. The temperature sensor outputs a voltage which is directly proportional to the instantaneous temperature. The soil moisture sensor measures the volumetric soil moisture as a function of electric conductivity in the soil. In our experiment, the position of the soil moisture sensor was fixed at 10 cm depth. The sensor was calibrated by taking two measures; one with the soil entirely wet and one with the soil entirely dry. The moisture was expressed in percentage, with the values of the measurements mapped from 0 (dry soil) to 100 (soaked soil).



For our experiments, the soil moisture M and the temperature T near the moisture sensor were synchronously measured every hour over a period of two days. The irrigation IR was modeled as a pulse signal, since it provides a great amount of water but for a limited time [10].

The Blaney-Cradle equation block in Figure 2 estimates ET_o using (4). In our experiment, the value of the variable p in (4) was substituted by the actual percentage of daytime for the time of the year in which the experiment was conducted. In the same way, the variable T_{mean} was substituted by the actual temperature T measured at time k.

2.6. Model Validation

To validate the model obtained in the previous section the output of the simulated model was compared with measured data.

For this purpose, another set of input data was measured and then inputted to the derived model of the system. The input data set used in the validation is shown in Figure 4.

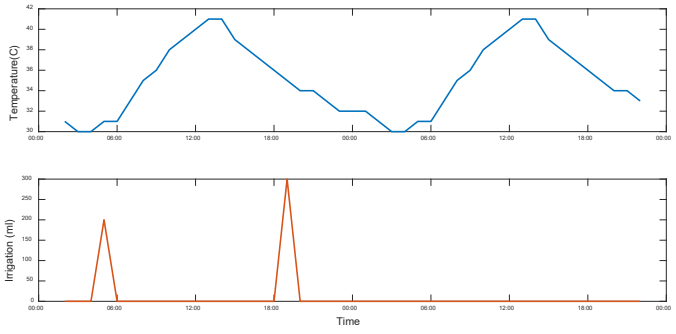


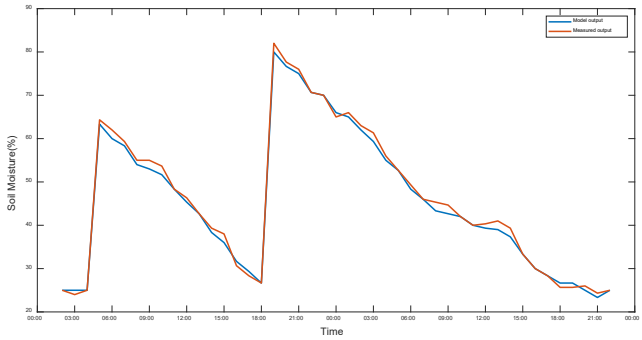
Figure 4: Validation Data

Figure 5 shows the difference between the output of the model and the measured soil moisture. The plots show a good agreement between simulated and measured outputs, with negligible time delay between the irrigation event and the change in soil moisture dynamics.

A further comparison between simulated and measured outputs to verify the reliability of the identified model was done calculating the Mean Absolute Error (MAE) as follows:

$$MAE = \frac{\sum_{k=1}^{T} |\widehat{M}_k - M_k|}{T} \quad (9)$$

where $\widehat{M}_k - M_k$ is the difference between the predicted and the measured output value and T=48 is the total number of samples. The obtained *MAE* value of 2% is considered an indication of a valid model.



Measured and Simulated Soil Moisture

3. Economic MPC

3.1. EMPC Basic Concepts

The term Predictive Control refers to a control design method characterized by [17]:

- use of a model to forecast the system output at future time instants (prediction horizon);
- computation of an input control sequence that optimizes a given objective function;
- receding horizon strategy, where the input control sequence is computed by taking into account the predicted outputs, but only the first control signal of the sequence is applied to the system.

The different algorithms which belongs to the Predictive Control group present differences in terms of the chosen system models (impulse, step, state-space, CARIMA, fuzzy models, etc.), disturbance (constant, decaying, filtered white noise etc.) and in the cost-function to be minimised.

A general MPC mathematical formulation is given by the following optimization problem [18]:

$$J = min \sum_{i=0}^{H_c-1} f\left(x_{k+i|k}, u_{k+i|k}\right) \quad (10)$$

where the variables $x \in R^n$ and $u \in R^m$ denote the state and control action at the instant k and H_c denotes the control horizon. The notation $k+i|k$, with $k \in Z$, indicates the estimation of the future variables based on the measurements at current time instant k.

Traditional MPC controllers minimize a quadratic cost function which tracks output and manipulated variable references.

The general prediction model can be expressed by the discrete equation:

$$x_{k+i+1|k} = g(x_{k+i|k}, u_{k+i|k}) \quad (11)$$

where the predicted states depend on the future states and control actions from instant k to instant $k+H_p$.

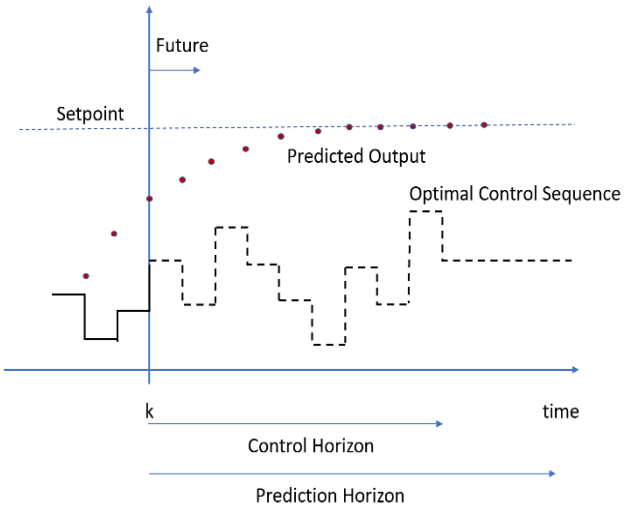
The minimization problem is usually subject to a set of constraints which have to be satisfied over the predictive horizon. These constraints are due to physical or operational limitations of the real system and they take the general form:

$$h(x_{k+i|k}, u_{k+i|k}) \leq 0 \quad (12)$$

At each time instant, MPC measures or estimates current state x_k and predicts the system state evolution over the prediction horizon H_p using the current state and the model of the system.

The optimal control sequence over the control horizon H_c is found solving the minimization problem (10). Only the first control action is applied to the system, and the remaining results are discarded.

In absence of disturbances and plant-model mismatch, the optimization problem could be solved as an open-loop problem, and the input sequence found at k=0 could be applied to the system for all $k \geqslant 0$. In a more realistic scenario, the behavior of the real system differs from the predicted model output. To cater for this mismatch, the optimal control problem is solved again at the next sampling time. Using the state measurement/estimation at time k+1, the cycle of prediction and optimization is repeated, moving the horizon forward (Figure 6).



Receding Horizon

In recent years, the EMPC approach has been proposed with successful applications in various fields [19].

Contrary to traditional MPC, EMPC is employed to maximize profitability, rather than minimize tracking errors. This can be done by defining a cost function that, directly or indirectly, measures the process economics [20]. The formulation of cost function depends on the particular applications, such as the operating profit, production rates, product selectivity, and product yield. In EMPC the cost function, as well as the plant dynamics and the constraints can be linear or nonlinear. Therefore, generic performance cost functions and nonlinear optimization algorithms are used to solve the minimization problem. The structure of the basic EMPC algorithm is shown in Figure 7.

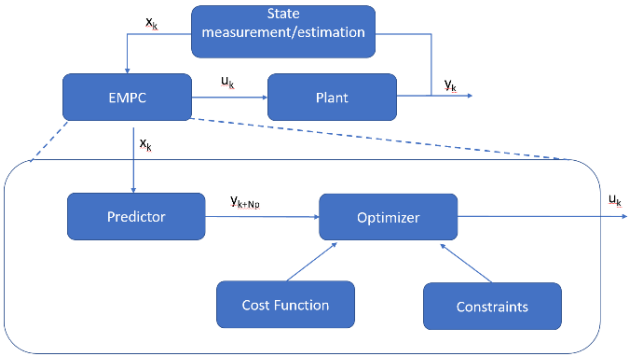


Figure 7: Basic EMPC Structure

3.2. EMPC Problem Formulation

In our work, we utilize an EMPC approach to minimize water consumption while maintaining the soil moisture close to an optimum value. We consider a linear economic stage cost function; additionally, both plant dynamics and constraints are linear. Because of this, our formulation of the EMPC problem is linear. The fundamental steps of the devised EMPC algorithm are the following:

- At the time k, measure $M(k)$ and compute the optimal control sequence $\overline{IR}(k)$ by solving a stochastic optimization problem over the control horizon N_c.
- Apply only the first computed value $IR(k/k)$ as input to the system.
- At the time k+1, measure $M(k+1)$ and repeat the optimization.

At each time instant k, a vector $\overline{M}(k)$ of the moisture predicted values is computed, using the model given by (7) and (8):

$$\overline{M}(k) = [M(k+1/k) \quad \dots \quad M(k+N_p/k)] \tag{13}$$

The stochastic optimization problem becomes deterministic if we assume that ET_o does not change over the prediction horizon N_p., i.e.:

$$\overline{ET_0}(k) = [ET_0(k) \quad \dots \quad ET_0(k)] \tag{14}$$

where $ET_o(k)$ is calculated using (4) and the current measured temperature $T(k)$.

Let's define the future control signals as the vector:

$$\overline{IR}(k) = [IR(k/k) \quad \dots \quad IR(k+N_c-1/k)] \tag{15}$$

Combining (13), (14) and (15) with the model (7) and (8) we obtain:

$$\overline{M}(k) = \Theta M(k) + \Phi \overline{IR}(k) + \Psi \overline{ET_0}(k)) \tag{16}$$

where:

$$\Theta = \begin{bmatrix} K_s \\ \vdots \\ K_s^{Np} \end{bmatrix} \tag{17}$$

$$\Phi = \begin{bmatrix} -K_c & 0 & \dots & 0 \\ -K_c K_s & -K_c & 0 & 0 \\ \vdots & \vdots & -K_c & 0 \\ -K_c^{Nc} K_s & -K_c^{Nc-1} K_s & \dots & -K_c \end{bmatrix} \tag{18}$$

$$\Psi = \begin{bmatrix} -K_c & 0 & \dots & 0 \\ -K_c K_s & -K_c & 0 & 0 \\ \vdots & \vdots & -K_c & 0 \\ -K_c^{Np} K_s & -K_c^{Np-1} K_s & \dots & -K_c \end{bmatrix} \tag{19}$$

The purpose of the designed control system is to obtain optimal economic performance. The following stage cost is assumed to reflect the nominal economic cost of irrigating the system over one sampling period, that is:

$$L(k) = IR(k) \tag{20}$$

The cost function does not consider the cost of operating the valves or the cost of pumping water, since their minimization is not considered an objective in the present study.

The resulting optimization problem adopted in our EMPC controller formulation is:

$$\min_{IR(k),v(k)} \sum_{i=0}^{N_c-1} IR(k+i/k) + \sum_{i=0}^{N_p} \rho v(k+i) \tag{21}$$

In addition to the stage cost, the objective function contains the terms $\rho v(k+i)$ which represent soft economic constraints.

The minimization problem (21) is subject to:

$$\overline{M}(k) = \Theta M(k) + \Phi \overline{IR}(k) + \Psi \overline{ET_0}(k)) \tag{22}$$

$$\overline{M}(k) \le \overline{M}_{opt}(k) + \overline{V}(k) \tag{23}$$

$$\overline{M}(k) \ge \overline{M}_{opt}(k) - \overline{V}(k) \tag{24}$$

where $\overline{M}_{opt}(k), \overline{V}(k)$ are vectors of size N_p defined as follows:

$$\overline{M}_{opt}(k) = [M_{opt} \quad \dots \quad M_{opt}] \tag{25}$$

$$\overline{V}(k) = [v(k/k) \quad \dots \quad v(k+N_p/k)] \tag{26}$$

The value M_{opt} represents the soil moisture that should be kept constant for optimal operating conditions. The variable $v(k)$ is a slack variable that is introduced to allow for small variation of the soil moisture $M(k)$ around the optimal value. That is, we nominally want to satisfy $M(k)$=M_{opt}, and when these constraints are violated, an additional cost $\rho v(k+i)$ is paid, weighted by the penalty parameter ρ. The choice of the parameter ρ is critical; when it is chosen too small, the value of the soil moisture might be far from the optimal one; when it is chosen too big, we might incur into infeasibility issues.

4. Simulation Results

The performances of PID controller, On-off controller and the proposed EMPC are compared. Not surprisingly, the behavior of the three control techniques is very different, as it can be observed in Figure 8 and Figure 9.

The On-off controller uses the measurement from a moisture sensor to switch the irrigation pump on or off based on whether the measurement is below or above maximum and minimum limits. The output from the device is either 100% on or off, with no middle

state. Figure 8 shows the outcome using the On-off controller. In this example, the soil moisture minimum and maximum levels are fixed at 50% and 70%. As expected, the output presents a variation of the moisture level (overshoot) around the optimal value of 65%. Reducing the thresholds could help improving the underwatering issue, but it would imply a higher switching of the actuators and a persistent overwatering problem.

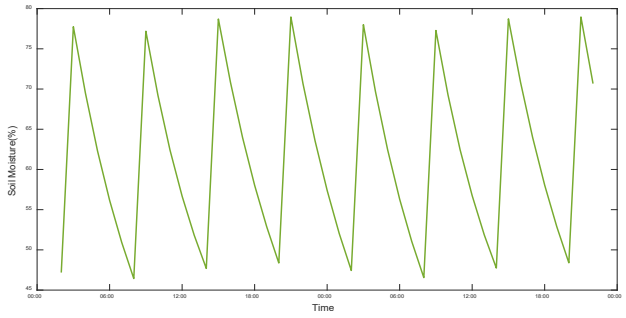


Figure 8: Soil Moisture with On-Off Controller

In Figure 9 the outcomes of the PID and EMPC approaches are shown. It can be noticed that the EMPC maintains the value of the soil moisture around the optimal value better that the PID controller in presence of disturbances, since it considers the effect of $ET_o(k)$ on the moisture level over the prediction horizon.

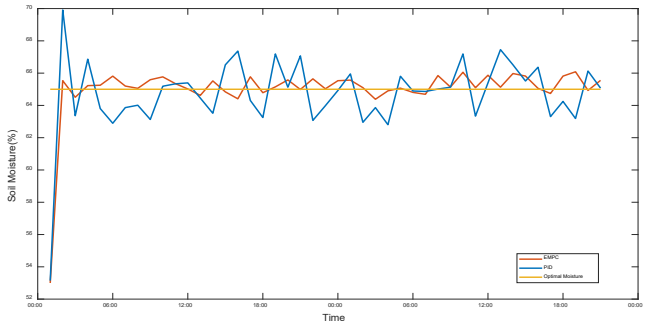


Figure 9: Soil Moisture Dynamics with EMP and PID Controllers

The three irrigation systems are also compared in terms of accumulated error, defined as:

$$CE = \sum_{k=1}^{T} \left| M_{opt_k} - M_k \right| \tag{27}$$

where the error is the difference between the optimal soil moisture and the simulated soil moisture. The cumulative error gives an indication of the ability of the controller to keep the soil moisture close to the optimal value. As it can be seen in Figure 10, the EMPC has smaller cumulative error compared to the other irrigation methods.

The total water consumption over the 48 hours period is around 2400 ml for the On-off controller, 2160 ml for the PID based system and 2000 for the EMPC. Therefore, the EMPC yields a reduction of water consumption around 8% and 16% compared with the PID and On-off controllers, respectively, while keeping the soil moisture within optimal values.

The simulations demonstrate that the EMPC scheme applied to an irrigation system provides benefits compared to conventional irrigation methods, since it maintains the moisture level near the nominal optimum, while providing economic benefit.

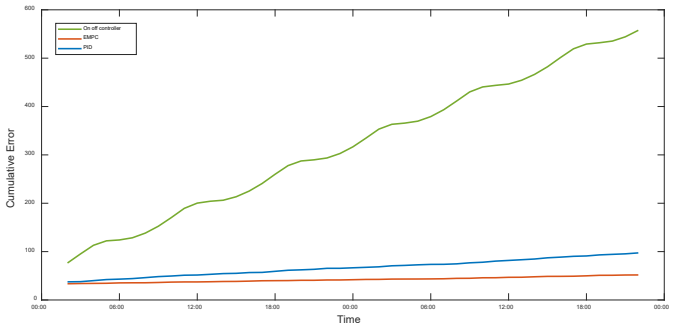


Figure 10: Cumulative Error for On-Off Controller, PID and EMPC

5. Conclusions

This work presents the process of developing an optimal controller regulating the time and amount of irrigation to maintain the soil moisture around an optimal level, while taking into account system constraints. The proposed control strategy, based on EMPC techniques, was implemented in MATLAB and results were compared with traditional irrigation methods based on PID and On-Off controllers. The results suggest that EMPC applied to irrigation systems can lead to increased efficiency and reduced water consumption.

Further improvement of this work would require a more accurate model of the system. A current limitation is given by the model developed, which can be improved by collecting sensors measurements over a sparse area and under different operating conditions, and using more complicated approaches to estimate ETo such as the FAO Penman Montheith technique.

Furthermore, developments on this work could explore the implementation of the proposed EMPC algorithm on hardware and evaluate the feasibility of such strategy in real time.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] L. Balbis, "Economic Model Predictive Control for Irrigation Systems," in 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, 2019. https://doi.org/10.1109/ICMSAO.2019.8880332.

[2] A. D. Chukalla, M. S. Krol, A. Y. Hoekstra, "Marginal cost curves for water footprint reduction in irrigated agriculture: guiding a cost-effective reduction of crop water consumption to a permit or benchmark level" Hydrol. Earth Syst. Sci., **21**(7), 3507–3524, 2017. https://doi.org/10.5194/hess-21-3507-2017, 2017.

[3] A. D. Chukalla, M. S. Krol, A. Y. Hoekstra, "Green and blue water footprint reduction in irrigated agriculture: effect of irrigation techniques, irrigation strategies and mulching" Hydrol. Earth Syst. Sci., **19**(12), 4877–4891, 2015. https://doi.org/10.5194/hess-19-4877- 2015.

[4] R. Aqeel-ur, A.Z. Abbasi, S. N. Islam, Z. A. Shaikh, "A Review of Wireless Sensors and Networks Applications in Agriculture" Computer Standards & Interfaces, **36**, 263-270, 2014. https://doi.org/10.1016/j.csi.2011.03.004.

[5] P. Alagupandi, R. Ramesh and S. Gayathri, "Smart irrigation system for outdoor environment using Tiny OS," in 2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), Chennai, India, 2014. https://doi.org/10.1109/ICCPEIC.2014.6915348.

[6] M.W. van Iersel, M. Chappell, J. D. Lea-Cox, "Sensors for Improved in Efficiency of Irrigation in Greenhouse and Nursery Production",

HortTechnology, **23**(6), 735-746, 2013. https://doi.org/10.21273/HORTTECH.23.6.735
[7] S. Kulkarni, "Innovative Technologies for Water Saving in Irrigated Agriculture" International Journal of Water Resources and Arid Environments, **1**(3), 226-231, 2011. ISSN 2079-7079
[8] J. B. Million, T. H. Yeager, J. P. Albano, "Evapotranspiration-based irrigation scheduling for container-grown Viburnum odoratissimum" HortScience, **45**, 1741-1746, 2010. https://doi.org/10.21273/HORTSCI.45.11.1741
[9] S. K. Saleem, D. K. Delgoda, S. K. Ooi, K. B. Dassanayake, L. Liu, M. N. Halgamuge, H. Malano, " Model Predictive Control for Real-Time Irrigation Scheduling", in 2013 4th IFAC Conference on Modelling and Control in Agriculture, Horticulture and Post Harvest Industry, Espoo, Finland, 2013. https://doi.org/10.3182/20130828-2-SF-3019.00062.
[10] C. Lozoya et al., "Model Predictive Control for Closed-Loop Irrigation", in 2014 19th World Congress, Cape Town, South Africa, 2014. https://doi.org/10.3182/20140824-6-ZA-1003.02067.
[11] L. Balbis and A. Jassim, "Dynamic Model of Soil Moisture for Smart Irrigation Systems," in 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakhier, Bahrain, 2018. https://doi: 10.1109/3ICT.2018.8855748.
[12] M. Todorovic, Crop evapotranspiration, in book: Water Encyclopedia: Surface and Agricultural Water, John Wiley & Sons Publisher, 2005.
[13] R.G. Allen, L.S. Pereira, D. Raes, M. Smith, "Crop evapotranspiration: guidelines for computing crop water requirements" in FAO Irrigation and Drainage, FAO, Rome, Italy ,1998.
[14] H. R. Fooladmand, "Evaluation of Blaney-Criddle equation for estimating evapotranspiration in south of Iran" in African Journal of Agricultural Research. **6**(13), 3103-3109, 2011.
[15] O. E. Mohawesh, "Spatio-temporal calibration of Blaney-Criddle equation in arid and semiarid environment" in Water Resour. Manage., **24**(10), 2187-2201, 2010. https://doi.org/10.1007/s11269-009-9546-7
[16] S.K. Ooi, I. Mareels, N. Cooley, G. Dunn, G. Thoms, "A Systems Engineering Approach to Viticulture On-Farm Irrigation" in 2018 17th World Congress. Seoul, Korea, 2008. https://doi.org/10.3182/20080706-5-KR-1001.01618
[17] E.F. Camacho, C. Bordons Introduction to Model Predictive Control, in book: Model Predictive Control. Advanced Textbooks in Control and Signal Processing, Springer, 2007.
[18] J. M. Maciejowski, Predictive Control with Constraints, Prentice Hall, 2002.
[19] T. Tran, K. Ling, J. M. Maciejowski, "Economic Model Predictive Control - A Review", in 2014 31st International Symposium on Automation and Robotics in Construction and Mining, ISARC, Sydney, Australia, 2014. https://doi.org/10.3182/20080706-5-KR-1001.01618.
[20] J. B. Rawlings, D. Angeli, C. N. Bates, "Fundamentals of economic model predictive control" in 2012 51st IEEE Conference on Decision and Control (CDC), Maui, Hawaii, USA, 2012. https://doi.org/10.1109/CDC.2012.6425822.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 789-798 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design and Implementation of Quad-Site Testing on FPGA Platform

Basavaraj Rabakavi[1,*], Saroja V Siddamal[2]

[1]*Department of Electronics and Communication, Government Engineering College, Haveri, 581110, India*

[2]*School of Electronics and Communication, KLE Technological University, Hubli, 580021, India*

ARTICLE INFO

Article history:
Received: 17 August, 2020
Accepted: 21 September, 2020
Online: 12 October, 2020

Keywords:
Multi-Site Testing
Concurrent Testing
Propagation Delay
Quad IC Tester
VLSI
ATE
FPGA
IC Packages

ABSTRACT

As manufacturing efficiency has become a main focus of today's business, it is very critical to surge the throughput by developing different test strategies. With throughput, testing cost also has been recognized as the major challenge in the future of leading semiconductors. Reducing test time is a significant effort to maximize throughput as the complexity increases in future generation outcomes and devices. So, low-cost Automatic Test Equipment (ATE) with parallel test can be promoted as the obvious solution for challenges said above. In parallel testing, multiple devices-under-test (DUT) can be tested at a time that enhances way of testing by increasing product flow, limiting gross test times, and efficient usage of tester. The proposed Integrated Circuit (IC) tester is used to implement multi-site testing (Quad-Site testing) and concurrent testing. It exhibits multi-site efficiency which substantially enhances the throughput by reducing test time. Modular, re-configurable test system provides cost-saving solution. To confirm these effects, authors have presented experimental results for Quad site testing of different ICs namely Decoder, Buffer, Multiplexer and Logic gates. This portable IC Tester handles variety of IC packages like Dual Inline Package (DIP), Small Outline Integrated Circuit (SOIC), Thin Shrink Small Outline Package (TSSOP). With functional test, the proposed tester also verified the AC Parametric tests (i) Propagation Delay is 20ns (ii) Operating frequency with 50MHz for Decoder IC (74HC138). The proposed IC tester consumes 70% less power and throughput enhanced by 11% compared to existing IC testers.

1. Introduction

This paper is an extension of previous work originally presented in the 3rd IEEE International Conference on Electrical, Electronics, Communication, Computers and Optimization Techniques [1]. Integrated Circuits which are more commonly known as ICs, are the central components of every electronic circuits present in modern times. Though it is one of the most important and vital part of the circuit, these ICs are most prone to be the cause of the malfunctioning of the whole design. With the advent of the complex integrated circuits, the problem arises after production of digital circuits to verify its functionality towards the meeting specifications. While delivering a product to market, testing plays an important role in the overall process. It is impossible to dispatch quality chips to the customer without proper testing. However, testing cost is also part of the overall product cost. Since it is very essential to test the IC before actually using them in any applications, an IC tester has been developed in order to eliminate the complex process of circuit troubleshooting [2, 3].

1.1. Types of testing

There are three types of testing: (i) Functional (ii) Parametric (iii) Structural tests. Functional testing is carried out by assigning specific test vectors to device under test, measuring the circuit response and device's functionality. Parametric tests are performed to determine the basic electrical characteristics of the device. DC parametric tests verify threshold levels of voltages and currents, in addition to open/short circuit tests and DC parameters are independent of time. AC characterization tests verify time dependent properties like propagation delays, transition signals (rise and fall) and operating frequency range. In structural testing, test inputs are designed to find out particular faults that might have occurred in circuit due to manufacturing deficiencies. During a structural verification, it is assumed that the circuit design is proper and the intention of the testing is to detect faults because of failures in processing [4].

*Corresponding Author: Basavaraj Rabakavi & Email: bsrabakavi@gmail.com

1.2. Stages of testing

There are four different stages of Very Large-Scale Integration (VLSI) testing; (i) Characterization (ii) Production (iii) Burn-in (iv) Incoming inspection. A characterization test measures the true operating limit values of the product. Before manufacturing, this test is carried out on a current design. Manufacturing test undergoes for smaller duration but checks all relevant parameters of the product. This test covers maximum number of standard faults but all expected functions may not be covered by test inputs. Burn-in makes sure the safety of devices by verifying the devices for longer duration, either continuously or regularly, and causing the wrong products to really fail. System manufacturers carry out incoming inspections on the purchased devices before incorporating them into the system [5].

1.3. Traditional method of testing

Everywhere digital circuits are growing consistently in case of both quantity and complexity. Meantime, automated circuit testing is going tougher, costly and critical. Manufacturing test plays a key role in ensuring the standard of the ICs supplied to the consumer. While ICs consistently integrating greater Built-In Self-Test (BIST) approaches, external testing instrument is still required in case of manufacture testing such as reading of specific configuration data or IC characteristics for assessment of ICs performance. Manufacturing test is carried out by advanced ATE, designed to work consistently with high quality standards. These High performance ATEs (HATEs) are costly and high-power consuming machines [6].

Now a day, there are two major types of ATE instruments exist in market: Highly sophisticated ATE and Low-price ATE. Instruments belongs to highly sophisticated ATEs are Verigy, Advantest, Teradyne etc. and normally used in testing industries and their functioning is closely related to manufacturing stage. Highly sophisticated ATE manufactured by industries such as Advantest are costlier (millions of dollars) and need specialized technical skills to be utilized with precision [7]. Costs can be decreased much further by using Field Programmable Gate Array (FPGA) based hardware and computer software in combination. The advances in FPGA technology supports the development of FPGA based low cost ATE. Low cost FPGA based ATE case study can perform functional tests of digital circuits [8].

1.4. Advanced method of testing

Present product trends need greater parallelism in testing methods and flows of VLSI. The overall price of both wafer and production testing will be increased by high-end ATE. So, the increased price of System on Chip (SoC) testing and the ATE cost are two major worries for the semiconductor companies. Multi-site testing with available low-price ATEs is the only solution to resolve this issue by reducing the cost of test. In both wafer and package stage testing methods, multi-site testing can be utilized [9]. Increasing the degree of parallelism can reduce testing price by decreasing the recognizable test time for every product. Multi-site testing is a classic parallel test method where many devices are tested at the same time. Concurrent method of testing is another parallel technique, in which two or more operational modules related to same product are verified in parallel. If the product and test design engineers are able to make sure that independence in their designs individually, then the testing can be carried out simultaneously [10].

1.5. Necessity

The VLSI circuit manufacturer cannot guarantee the defect free integrated circuits (ICs). It is not feasible for small-scale industries investing huge amounts of money to buy sophisticated IC testers. Similarly, directly replacing ICs mounted on a board may not be desirable solution to repair a malfunction. The limitations of high end ATE machines are high price instrument, imprecision and need large memory. Therefore, all the limitations indicated above lead to the way for designing a new IC tester with low price, high accurate with maximum speed, otherwise advantages of semiconductor technology are purposeless [11].

1.6. Contributions

In previous work [1], Dual IC testing of digital ICs was implemented with feature of functionality test only. In this paper, work is extended to perform propagation delay test, operating frequency test with Quad IC testing simultaneously and also to analyze the power consumption and throughput of the proposed testing system. This paper describes a FPGA based reconfigurable, inexpensive, high speed, indigenous testing platform for standard digital ICs (74/54- types). This IC tester can be implemented with low budget for small or medium-scale industry customers of these ICs & offers fast and detailed check-out functions with minimal operator action.

Rest of the paper is organized as follows. Section 2 describes motivation for designing the multi-site tester. In section 3, concept of multisite testing is discussed. Proposed methodology is described in section 4. Results are discussed in section 5 and followed by the conclusion.

2. History and Motivation

2.1. Manufacturing Process Defects

Various types of defects can occur during the manufacture of any circuit, this leads to errors. There are four categories of defects as age defects, process defects, material defects and package defects. Process defects and material defects arise due to material imperfection during manufacture of the circuit. Age defects arise due to usage of device for longer duration and also with over time. Physical assembling of the circuit can cause package defects and other defects can be caused due to opening or breakage of pins [12]. Addition to mentioned defects, the person working on testing can also cause defects by testing the devices with incorrect test program or sometimes testing procedure itself can cause faulty components. Test instrument without calibration leads to wrong results. The person working on production lab can put good devices in the incorrect bin and also possible that a good device can be stamped with wrong part number [13].

2.2. Need for FPGA

Earlier digital IC testers were implemented using micro-processor or microcontroller family. Test vectors stored in the memory as lookup table are assigned to every IC to be tested. Reconfigurable logic system has programmable interconnections

to allow the user to perform required operation within it and it is known as Programmable Logic Device (PLD) or FPGA. Now a day, PLD's logic density has become more sufficient to implement lot of data processing operations within it. Even though ASICs achievement is more compared to PLD, when same application is realized in both devices, PLD's re-configurability has many advantages. PLD System with applications having high parallelism can be quicker more than hundred times than the microcontroller-based system. FPGA produces the test patterns by using Test Pattern Generator (TPG), rather than storing the test patterns as a lookup table for specific IC [11, 14].

FPGAs provide a greater number of tested devices in the specified period of duration. FPGA supports high standard of parallelism while generating various test vectors that varies in properties, ranging from simple digital circuits to highly complicated digital products [15]. FPGAs have many benefits which consist of more speed, large number of input and outputs, less power consumption, less size, enhanced time-to-market and increased flexibility and re-configurability. The number of I/O pins differs between one device to another device. Additionally, the interconnection between DUT pins and test equipment also varies from device to device. Hence, to perform the tests, the test device must be wise enough to handle pin configurations and input/output allocation of the DUT. For I/O handling issue, FPGA based ATE with reconfiguration property is the solution [16].

2.3. Tester requirement

All devices are not identical even though they pass in production tests. When actually used in the field, some will function for longer time and some will fail early. Out of four testing stages, incoming inspection is one of them. Incoming inspection can be conducted for arbitrary samples depending on the standard device and the system specifications. As the cost of incoming inspection is much less compared to the cost of testing in assembled system, the incoming inspection of components is much necessary in the industry [5]. To carryout incoming inspection, digital IC testers are required in various electronics laboratories and industries during design, after manufacturing stage of different electronics devices [6]. As small-scale industries or assembly housing units cannot offer large amount of investment for high end ATEs, less price ATE or IC testing device is the solution for incoming inspection. The requirement of low-cost tester is not only in industries but also in the academics. The low-cost tester can be achieved by implementing multisite and concurrent testing.

3. Concept of Multi-site Testing

IC testing is an important method to create insulation between good and poor components to make sure the quality of the outgoing goods. Traditionally single site testing method was used for testing. In single site method one DUT was verified at given instance using ATE [17]. Reduction in the testing cost is the repeated challenge of the semiconductor manufacture, and the test time has become important part of the Cost of Test (CoT). Because of high competency, system design companies have low profit margins in market, a continuous check on the low average price and quick time to market is very important [18]. Multi-site parallel testing provides the greatest enhancement to throughput on test, irrespective of the particular task. The multi-site parallel test concept has throughput advantages over traditional parallel test methods. In traditional parallel testing, multiple devices are tested on the same site. Once all testing on one site is completed, the prober indexes to the next site. By contrast, the multi-site parallel testing method is performed on multiple devices on different (multiple) sites at the same time as shown in the Figure.1. This eliminates all dead time between tests, and means there is no waiting for any one test to complete before the prober indexes to the next site. The test program is written so that at a time all similar DUTs are tested.

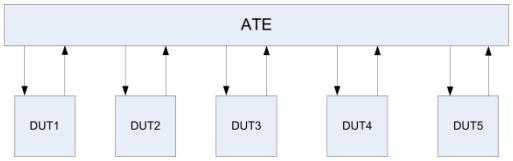


Figure 1: Multi-Site with parallel DUT testing

The total time taken for the group of devices tested at the same time is the addition of the index time t_i and the production test time t_p as shown in the Figure 2. The index time t_i is the time duration to build contact between the bonding pads of the DUTs and the probe interface by making proper positioning. The index time t_i differs on the type of probe station used and it is fixed for each probe station. A normal value of t_i is 0.7sec [19]. The production test t_p is the sum of different tests performed on DUT which may include functionality and characterization test.

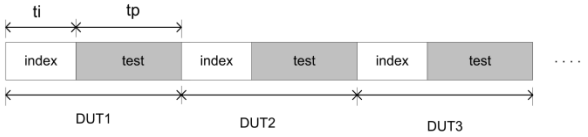


Figure 2: Testing time of DUT consists of index time t_i and test time t_p

If there is no waiting time between replacement of chip being tested is completed with new chip to be tested, then only index time was considered as valid. If immediate replacement of a new chip after the chip being tested is completed, then the indexing time is ignored. If there are no issues like tester downtime and chip contacting problems then test time was considered as valid [20]. The total testing time of a device T_s (Single site test time) can be written as shown in equation (1).

$$T_s = t_i + t_p \quad (1)$$

The Multi-Site-Testing is the test method for simultaneous testing of multiple similar DUTs on single ATE. The test time spent on one device of the Multi-Site-Test is 1/N of the Single-Site-Test. So, Multi-Site-Test testing time is close to time taken by Single-Site-Test. Earlier this test method was utilized for the testing memory devices, but now almost all semiconductor devices are tested with this strategy to decrease the testing cost. The Multi-Site-Test needed N times of the tester resources. Relationship between Multi-Site-Test and Single-Site-Test test timings can be given by Multi Site Efficiency (MSE). MSE is calculated as shown in equation (2).

$$MSE = ((N - T_m/T_s)/(N-1))x100 \quad (2)$$

Where N is the number of devices that can be tested in Multi-Site test, T_m is time consumed for Multi-Site-Test and T_s is time

taken for Single-Site-Test. Generally, Digital circuits MSE is higher than the MSE of analog circuits. i.e. MSE of Digital tests is about 95% and Analog mixed signal MSE is around 60 to 85% [21]. Considering the maximum usage of the ATE, total devices tested per hour i.e. D_{th} for multi-site testing with N number of devices or throughput can be given as in equation (3).

$$D_{th} = (3600 \times N)/T_m \quad (3)$$

4. Proposed Architecture for Quad IC Tester

The proposed architecture of FPGA based Quad IC tester is shown in the Figure 3. It contains FPGA device, devices under test, power supply device and peripheral Interfaces such as switches, LEDs and Displays. In this design, Altera Development board is used for FPGA device (Cyclone II – family) and also for user I/O interfaces. FPGA design consists of various modules as Test Pattern Generation (TPG), Output Response Analyzer (ORA), Input/output compatibility, Frequency Synthesizer and peripheral interface modules.

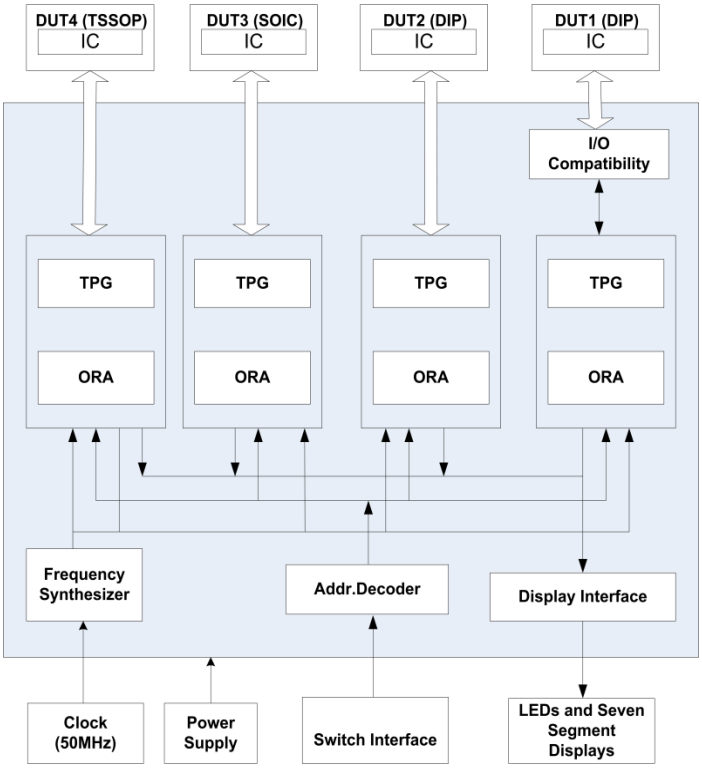


Figure 3: Proposed architecture of Quad IC Tester.

Different test vectors required to test the functionality of the DUT are generated by TPG module. The outputs of the DUT are latched by ORA module to verify with expected ones. If the actual results are similar to the expected results, IC tester displays *P* (PASS), or else *F* (FAIL). Frequency Synthesizer is used to produce various frequencies, which is required for TPG module to generate test vectors and to maintain synchronization among different blocks. As ICs will differ in their Input/Output (I/O) configurations, the I/O compatibility module is required to use single slot for testing different ICs. All the modules are described in the remaining sections.

4.1. Test Pattern Generation

Test Pattern Generation module produces desired test vectors for the device under test. Deterministic, exhaustive, random generations are different ways of producing the test patterns. In present work, exhaustive test patterns are utilized for 100% fault coverage during the testing of IC. In the case of logic gate IC, all quad gates are applied with the same patterns and tested simultaneously. Each different combinational circuit or sequential circuit will have different test patterns generated by Finite State Machine (FSM). Figure 4 shows the FSM for generating test patterns for dual 4:1 multiplexer. This FSM has eight states (S0 to S7) controlled by master clock. FSM is in initial state S0 as long as reset and ebar signals are held high.

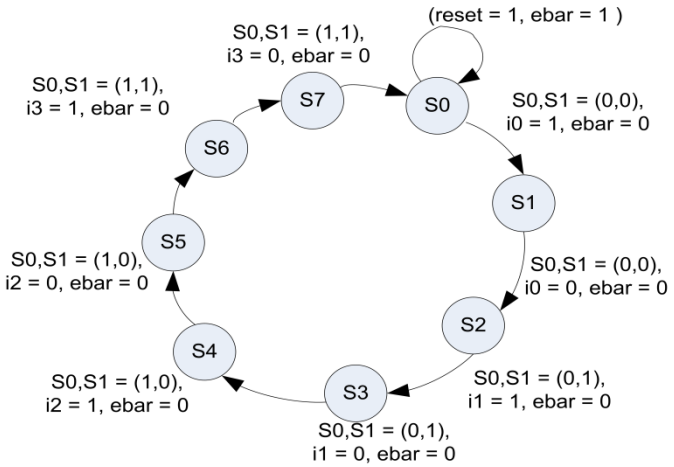


Figure 4: FSM for Generation of Test Patterns for Dual 4:1 MUX

4.2. Output Response Analyzer

The ORA module receives the actual outputs from the DUT and compared with desired values stored in register. Depends on the comparison result, specific bit say R1 will be set as '1' if the result exactly same as desired one, or else it is declared as '0'. In view of complex digital circuits, ORA module is used to verify whole data given in truth table or sequence table of DUT's datasheet with its real outputs, prior to setting of R1 bit. Choice of expected values stored in register is done by setting the switches at the beginning of test, based on the IC inserted in the specific socket. By considering dual 4:1 Multiplexer IC, outputs of multiplexer are captured and stored on every negative edge of input clock into temporarily register i.e. temp [7:0]. This process is continued until input patterns are reached to DUT by FSM in TPG module. Figure 5 shows the FSM for latching the outputs of dual 4:1 MUX.

4.3. Frequency Synthesizer and Input/output Compatibility

The proposed IC tester tests various ICs of different frequencies. Frequency Synthesizer is needed to produce clocks with different frequencies like 2Hz, 0.5MHz, 1.567MHz, 3.125MHz, 6.25MHz, 12.5 MHz, 25MHz. Altera FPGA board consists of 50MHz and 27MHz as master clocks. Binary counters are used to generate various frequencies. Frequency synthesizer module is utilized to calculate propagation delay and operating frequency range of particular IC. Depending upon the modes of

operation like user mode or fast mode, specific frequency is selected for IC testing. IC results are monitored using LEDs with help of 2Hz frequency in user mode. Figure 6 shows frequency generation of 25MHz and 12.5MHz, similarly other frequencies are also generated.

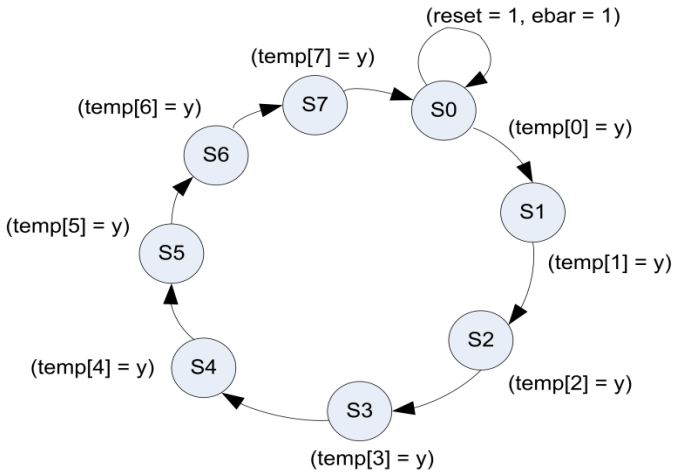


Figure 5: FSM shows output latching of Dual 4:1 MUX

Provision of single slot for testing of various ICs with different I/O connection is the complex problem. If different ICs are to be tested using single socket, then each pin of particular IC need I/O compatibility. This issue is solved by designing I/O compatibility module using multiplexers and tristate buffers. The Figure 7 represents the I/O compatibility logic for single pin. Similarly, all other pins of ICs can be configured as input/output based on the IC selection.

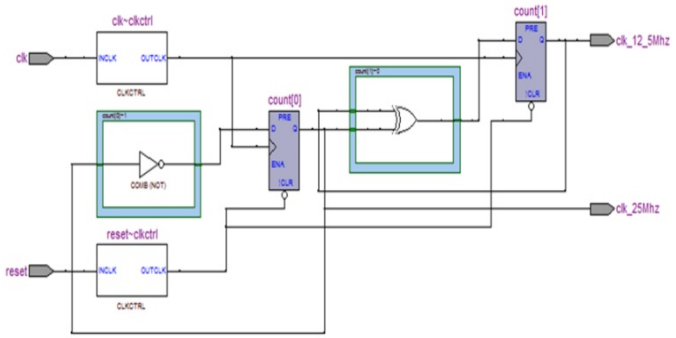


Figure 6: Frequency generation of 12.5MHz and 25MHz

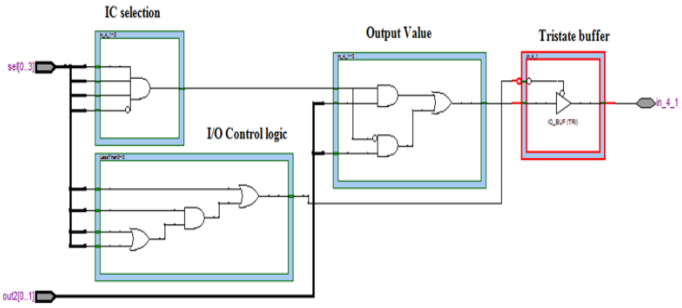


Figure 7: I/O handling logic for single pin of IC

4.4. Propagation delay test and Functional speed test

Propagation delay test is one of the AC Parametric Test of IC. Propagation delay is the addition of switching delay of device and transport delay of interconnects on the track [5]. The process of finding the propagation delay is shown in the Figure 8. Once the test vector of DUT is active i.e. transition from low to high, one pulse is generated. At this time, high frequency counter is started and continues to count till it come across change in output of the device to be tested. The change in output of the device is recognized by pulse generated due to the low to high transition occurs at the device output. Propagation delay is calculated by multiplying the number of counts with time period of master clock used for counter. The equation (4) shows calculation of the propagation delay, where P_d represents propagation delay, C indicates the number of counts from change of input value to the change of output value and T is the time period of the high frequency clock used to run the counter.

$$P_d = C \times T \quad (4)$$

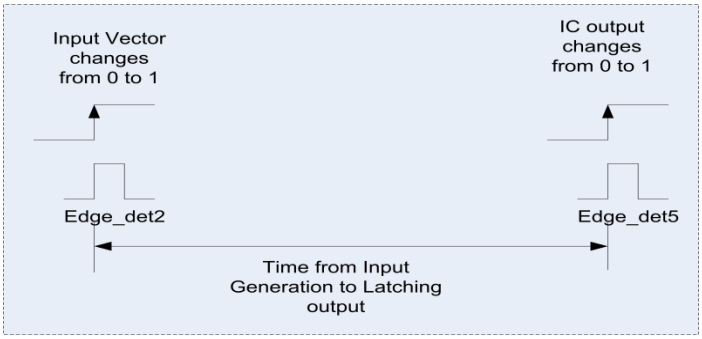


Figure. 8: Process to calculate the propagation delay

Functional Speed Test or Operating Frequency Test of device under test is also one of the AC Parametric tests. Here, the specific test inputs are produced and verified for the performance of the device. The speed at which functionality results of device meet the expected values is called functional speed of the device. The propagation delay test and functionality speed test results are shown in the section 5.

4.5. Quad IC Testing

Most of the existing single IC testers are made up of microprocessor or microcontroller device. The limitations of these testers are sequential execution and a smaller number of input/outputs. In the presented work, the above limitations are overcome by developing test logic with individual FSM for each device. All the FSMs will run in parallel, because of this feature proposed Quad-IC tester can test four different ICs simultaneously.

5. Results and Discussion

The proposed IC tester is designed and developed on target Altera FPGA Development board (Cyclone II – family). The testing logic is developed using Verilog HDL. The design is synthesized using Altera Quartus-II 9.1 tool. Test vectors are produced by TPG modules according to the truth table or sequence table of individual ICs. The functional test of four different ICs, namely Logic gate, Multiplexer, Decoder and

Buffer is performed simultaneously and details of the testing is given in Table 1. Simulation results of TPG and ORA modules are verified for all Digital ICs with respect to its functionality through waveforms. All four ICs are tested in parallel with one-time test and loop test facility by selecting switch as 0/1 respectively. Results are verified through LEDs using IC selection values with fast mode or user mode facility. Status of the result shown in the seven-segment display as IC1P (i.e. First IC is PASS).

Functional test of ICs is carried out with different frequencies by frequency synthesizer to find out the testing speed of IC. Buffer IC is tested with four values namely as 0xff, 0xaa, 0x00, 0x55 in a loop test with 50MHz. Similarly, Decoder IC is tested according to its truth table sequence with 50MHz successfully. Logic gate ICs outputs are verified with 12.5MHz. Multiplexer IC is tested for all combinations of test vectors with 6.25MHz successfully. All the ICs are powered with 5V supply. Table 2 shows the operating frequency and total time required to complete the functionality test of particular IC for all combinations of test vectors.

Propagation Delay test is performed on ten ICs individually and measured values of propagation delays of each IC is given in Table 3. The ICs having 'HC' (High speed CMOS) symbol in their names operate with both voltages '2V' and '5V'. Propagation delay values of ICs differ with respect to supply voltages.

In this work, system clock 50MHz is used to find out the propagation delay. The count value represents the delay between rising edge of test vector to rising edge of gate output. The count value is multiplied with clock period of master clock (i.e. 20ns) which gives the switching delay of the device. In this work, transport delay is considered as negligible (usually in terms of Pico Secs) and it is fixed value. For example, for 74HC86 IC the propagation delay is calculated by using equation (4) i.e. $P_d = C \times T = 2 \times 20ns = 40ns$. In which, C is count value and T is time period of 50MHz clock. Similarly, propagation delay is calculated for all other ICs. Out of ten ICs, eight ICs results are almost similar to the respective datasheet values except two ICs (74HC266 and 74HC153). Reasons may be because of EX-NOR IC has open collector pins, the delay depends on pull up resistor values and MUX IC depends upon its switching activities.

Proposed IC tester also aims to test functionality of ICs with low power consumption. The power consumed by testing logic with respect to each IC is given in the Table 4. Power consumption of developed testing logic is calculated using Powerplay Early Power Estimator Spreadsheet provided by respective FPGA IC vendor (Altera Cyclone II spreadsheet- Now Intel). It will be achieved by filling the utilization of FPGA device i.e. LUTs, FFs, Input/output pins used, clock rate and total fanouts details in the spreadsheet of FPGA family.

The top verilog module or integrated module gives overall power consumption for testing four ICs simultaneously. Power consumption calculated using spreadsheet is static or offline method and for top module it is 169mW. Table 5 shows the power consumed by proposed system in real time during Quad IC testing. The power consumed by FPGA board is calculated by multiplying Icc with Vcc (5V). The current Icc is measured through ammeter placed on power supply path of hardware during Quad IC testing and it is 0.30A. Similarly, board power is calculated before downloading the testing logic onto FPGA device. So, the power consumption of testing logic is calculated by subtracting board power consumption (powered on condition i.e. without code) from power consumed during testing of Quad ICs i.e. 1.5W-1.25W = 0.25W. Total power consumed during testing is 1.5W which is less compared to commercial testers. In the Table 6, the power consumption of the proposed tester is mentioned approximately as 2W by considering enhancement of testing logic to cover variety of ICs.

Table 1: Functional test of Quad ICs

Quad IC Testing (Functionality Test - Simultaneously)								
Sl. No	**Device Name (IC name)**	**Package**	**IC Pins**	**IC Selection**	**Fast mode (1)/User mode (0)**	**Loop test (1)/One time test (0)**	**Expected Result**	**Actual Result**
1	74HC244 (8bit Buffer/Driver)	TSSOP	20	1 (001)	0	0	IC1-PASS	IC1-PASS
2	74HC138D (3:8 Decoder)	SOIC	16	2 (010)	0	1	IC2-PASS	IC2-PASS
3	74HC153N (Dual 4:1 MUX)	DIP	16	3 (011)	1	1	IC3-PASS	IC3-PASS
4	SN74HC86N (Quad XOR Gate)	DIP	14	4 (100)	1	0	IC4-PASS	IC4-PASS

Table 2: Operating Frequency test

Operating Frequency Test			
Sl.No	**Verilog Module name**	**Operating Frequency for Loop Test**	**Time required complete functionality test (Power supply - 5V)**
1	Buffer module	50MHz	80ns (4 values – write and read)

2	Decoder module	50MHz	160ns (8 test vectors)
3	Gate module	12.5MHz	320ns (4 test vectors)
4	MUX module	6.25MHz	1280ns (8 test vectors)

Table 3: Comparison of Real time Propagation Delays with Datasheet values

Propagation Delay Test							
Sl. No	Device (IC)	Values measured using Proposed IC Tester (Actual values)				Values given in Datasheet	
		Power supply (2V)		Power supply (5V)		2V	5V
		Count value	Time taken by devices	Count Value	Time taken by devices	Expected values	
1	74HC86 (EXOR gate)	2	2x20ns = 40ns	1	1x20 = 20ns	40	12
2	74HC138 (Decoder)	2	2x20ns = 40ns	1	1x20 = 20ns	40	15
3	74HC244 (Buffer)	2	2x20ns = 40ns	1	1x20 =20ns	30	11
4	74HC153N (Mux)	4	4x20ns = 80ns	2	2x20 = 40ns	50	17
5	74LS00 (NAND gate)	NA	NA	1	1x20 = 20ns	NA	22
6	74LS02 (NOR)	NA	NA	1	1x20 = 20ns	NA	10
7	74LS08 (AND)	NA	NA	1	1x20 = 20ns	NA	15
8	74LS32 (OR)	NA	NA	1	1x20 = 20ns	NA	15
9	74LS04 (NOT)	NA	NA	1	1x20 = 20ns	NA	10
10	74HC266 (EX-NOR)	18	18x20 = 360ns	18	18x20 = 360ns	150	30

Table 4: Power Consumption of Testing Logic

Power Calculation Through PowerPlay Early Power Estimator Spreadsheet (Offline)						
Individual Modules						
Sl.No	Verilog Modules	Logic		Input/ Output	Clocks	Total (mW)
1	Bufer module	1		36	3	40
2	Decoder module	0		32	3	35
3	MUX module	0		34	3	37
4	Gate module	0		32	3	35
5	Clock module	0		29	3	32
Total = 97 (Pstatic) + 0.179 = 276mW						
Integrated Module						
Sl.No	Module name	P static	Logic	I/O	Clocks	Total (mW)
1	Final module	97	2	63	7	169

5.1 Simulation Results

The Simulation results of TPG and ORA logic modules for each IC under test are verified by using Xilinx ISE 14.3 version through testbench module. Simulation results are described with case study of 74HC138 IC. In TPG module, address and control inputs of decoder are generated at each positive edge of clock. All possible combinations of test vectors are generated as shown in the Figure 9. The outputs from decoder are latched at the falling edge of each clock and stored in register. The 8th output of decoder labeled as y7, latched value shows 0xfe during its active state as shown in the Figure 10.

Similarly, each TPG and ORA module of different ICs are simulated and verified by writing testbench modules. These modules are verified in hardware by implementing on FPGA and tested with physical ICs by sending test vectors to and read the outputs from ICs and compared with expected values.

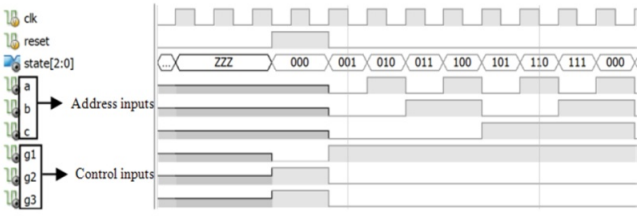


Figure 9: Test Pattern Generation for 3 : 8 Decoder (74HC138 IC)

Table 5: Power consumption in real time testing

Power Values during testing (Real time condition)				
Sl. No	**Module Name**	**Powered on state (No code)**	**Running State (With code)**	**Designed module Power**
1	Final code	0.25Ax5V = 1.25W	0.30Ax5V = 1.50W	(Running code) – (Powered on condition) = 1.50 – 1.25 = 0.25W

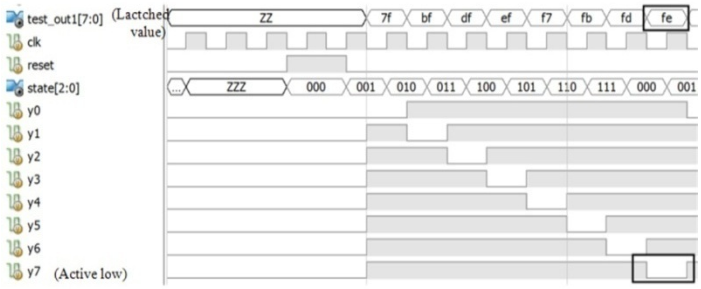


Figure 10: Latching of IC outputs is verified through testbench

Propagation Delay test concept is verified through simulation results as shown in Figure 11. The signal 'edge_det2' shows rising edge of input test vector and signal 'edge_det5' shows rising edge of IC output. The 'count time' register gives the number of counts between the change in the input and output of DUT. The number of clocks marked in the waveform is matched with value stored in the 'count time 'register. This logic is verified through hardware and found the count time value for each IC to be tested. This count time value is multiplied with clock period to get the propagation delay as mentioned in Table 3.

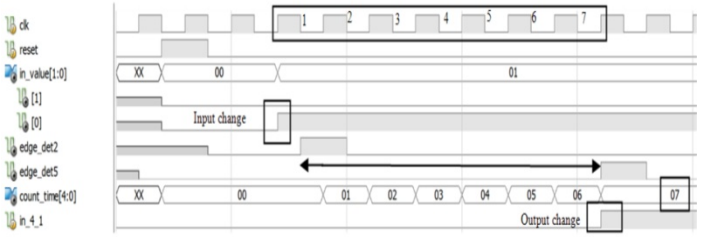


Figure 11. Finding the time between change of input to change of output

5.1. Implementation Results

5.1.1. Dual IC Testing (DIP and SOIC Package)

Dual IC testing presented in paper [1], the Decoder (74138-SOIC) and XOR (7486-DIP) ICs were tested simultaneously. The 7486 – XOR IC was tested by applying all possible test vectors to each gate simultaneously and latched outputs were compared with desired values. 'P' on the seven-segment display represents result status of the individual gate in IC. Figure 12 shows Dual IC testing setup. Results of both ICs were verified through LEDs.

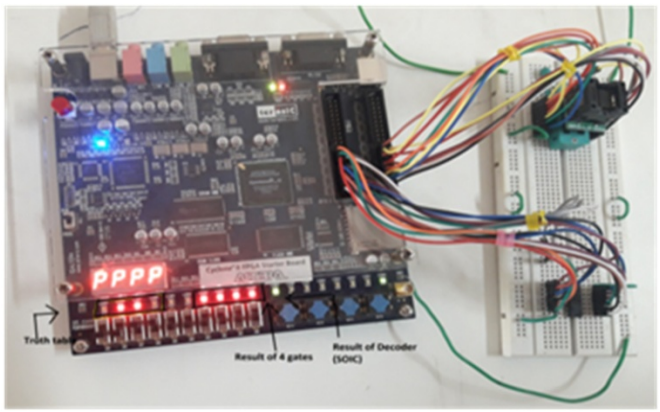


Figure 12: Results of Dual IC tesing (74HC138 and 7486 ICs)

5.1.2. Quad IC testing (TSSOP,SOIC,DIP,DIP- Package)

In Quad IC testing, Buffer - 74HC244 (TSSOP), Decoder - 74HC138 (SOIC), Dual 4:1 MUX - 74HC153 (DIP), XOR Gate - 74HC86 (DIP) ICs with various packages are tested simultaneously. The Figure 13 shows the results of Buffer IC among four ICs. 'IC1P' on the seven segment display represents result status of particular IC as free of error. Buffer IC is tested with fast mode, one time execution method and results shown on LEDs are last combination of test vector i.e. 0x55. IC selection should be set to '001' to monitor Buffer IC result. Similarly, decoder IC also tested in fast and one time execution mode. Both multiplexers in dual 4:1 MUX IC and quad gates in XOR IC are tested at same time. All ICs results are monitored through LEDs by setting switches. Quad IC testing is verified simultaneously without modifying source code.

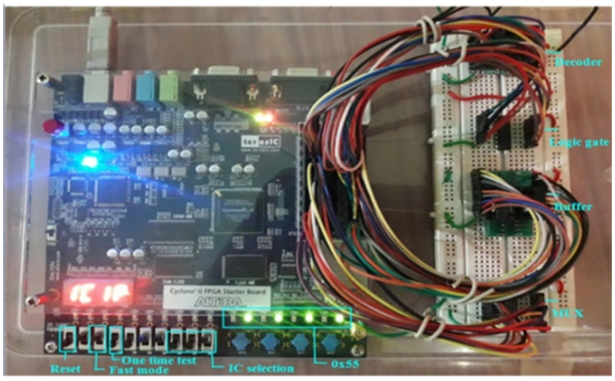


Figure 13: Buffer IC (74HC244) result during Quad IC Testing

The comparison is made between FPGA based Quad-IC tester and available IC testers in market and it is shown in Table 6. The Quad IC tester is having all existing features like multi package support, loop test capability, truth table verification etc. except unknown IC identification. With additional to these features, the Quad IC tester has capability of testing four ICs at a time (parallel execution) with more speed. Throughput calculation is made for single IC and Quad IC testing by using equation (3). The comparison results are demonstrated in the Figure 14 with respect to number of ICs tested per hour. The speed of proposed IC tester is about 0.01sec, which is much lower than speed of existing IC tester i.e. 0.8 sec. Also, Quad IC tester has some more features like

Propagation delay and operating frequency range test. The Altera FPGA Cyclone II EP2C20F484C7 device's (Now its Intel) utilization summary for quad IC testing logic is given in the Figure 15. Implementation of quad IC testing has utilized about 1% of logic and 32% of I/O pins of cyclone II device. So, to provide highly extended support still lot of ICs can be included into tester.

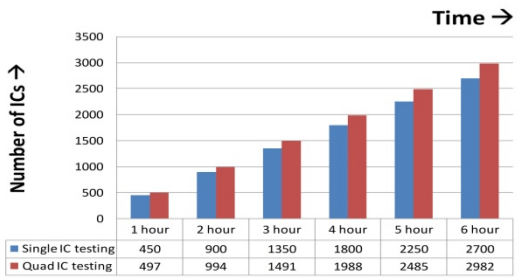


Figure 14: Throughput comparison between Single IC and Quad IC Testing

Quartus II Version	9.1 Build 350 03/24/2010 SP 2 SJ Web Edition
Revision Name	ic_tester_top
Top-level Entity Name	ic_tester_top
Family	Cyclone II
Device	EP2C20F484C7
Timing Models	Final
Met timing requirements	No
Total logic elements	217 / 18,752 (1 %)
Total combinational functions	164 / 18,752 (< 1 %)
Dedicated logic registers	159 / 18,752 (< 1 %)
Total registers	159
Total pins	102 / 315 (32 %)
Total virtual pins	0
Total memory bits	0 / 239,616 (0 %)
Embedded Multiplier 9-bit elements	0 / 52 (0 %)
Total PLLs	0 / 4 (0 %)

Figure 15: Device Utilization summary for Quad IC Testing

6. Conclusion

In this paper, portable Quad-Site Tester for digital ICs is implemented with automatic testing approach on FPGA. Testing has been carried out on Quad ICs simultaneously with features such as Propagation Delay and Operating frequency test along with Functional test. The `concurrent testing with true parallel execution` is the main key of proposed IC tester which greatly reduce the test time and cost compared with traditional testing. The proposed Quad IC tester improves throughput by 11% compared with existing IC testers. In this context, considering index time as 7secs as it is done by manually and testing time is 1sec. If index time becomes less than the execution time, then there will be significant change in throughput. This can be achieved by automatic replacing of ICs using hardware setup.

Also, throughput can be further increased by enhancing the work to Octal-Site testing. The proposed IC tester consumes 70% less power compared to existing IC testers.

In future, proposed work can be enhanced to include analog ICs (ADC, DAC, Opamps, and Analog MUX etc.), peripheral ICs (UART, level translators, drivers etc.) and memories with support of different packages like Thin Small Outline Package (TSOP), Thin Quad Flat Package (TQFP), and Plastic Leaded Chip Carrier (PLCC) etc.

Table 6: Comparison with existing testers

Features	IC Tester (Outside INDIA)	IC Tester (INDIA)	FPGA based Proposed Tester
Quad IC testing (Different type)	No	No	Yes
Package Support	DIP/SOIC	DIP	DIP/SOIC/TSSOP
Parallel Execution	No	No	Yes
Propagation Delay Test	No	No	Yes
Testing Speed	0.8sec	Not shown	0.01sec
Power Consumption	5W	6W	2W
MC/FPGA based	MC based	MC based	FPGA based
Battery operated	Yes	Yes	Can be incorporated in the system
Loop-test capability	Yes	Yes	Yes
Truth table observation	No	Yes	Yes
Counting process observation	No	Yes	Yes
Unknown IC identification	Yes	Yes	Can be enhanced to this feature as proposed IC tester is reconfigurable

Conflict of Interest

The authors declare no conflict of interest.

References

[1] B. Rabakavi, Saroja Siddamal, "Design of High Speed, Reconfigurable Multiple ICs Tester using FPGA Platform", IEEE, ICEECCOT, 909-914, December, 2018. https://doi.org/10.1109/ICEECCOT43722.2018.9001588.

[2] M. S. Zaghloul and M. Saleh, "Implementation of FPGA for decision making in portable automatic testing systems for ICs library & Digital Circuits ," IEEE Applied Imagery Pattern Recongnition Workshop, 2011. https://doi.org/10.1109/AIPR.2011.6176361.

[3] B. Vermeulen, Camelia Hora, Bram Kruseman, Erik Jan Marinissen, Robert van Rijsinge, "Trends in Testing Integrated Circuits", International Test Conference, IEEE, 2004. https://doi.org/ 10.1109/TEST.2004.1387330.
[4] I.A. Grout , Integrated Circuit Test Engineering: Modern Techniques, Springer, 2006.
[5] M.L. Bushnell, Vishwani D. Agrawal, Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits, Kluwer Academic Publishers, 2002.
[6] S. Fransi, G. L. Farré, L. G. Deiros, and S. B. Manich, "Design and Implementation of Automatic Test Equipment IP Module", ETSYM, 2010. https://doi.org/ 10.1109/ETSYM.2010.5512749.
[7] L. Mostardini, L. Bacciarelli, L. Fanucci, L. Bertini, M. Tonarelli, and M. De Marinis, "FPGA Based Low Cost Automatic Test Equipment for Digital Integrated Circuits", IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 21-23 September 2009.https://doi.org/ 10.1109/IDAACS.2009.5343031.
[8] A. A. Bayrakci, "ELATE: Embedded low cost automatic test equipment for FPGA based testing of digital circuits", 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, 2017, pp. 1281-1285.
[9] H. Kim, Yong Lee, Sungho Kang, "A Novel Massively Parallel Testing Method Using Multi-Root for High Reliability", IEEE Transactions on Reliability, **64**(1), March 2015.https://doi.org/10.1109/TR.2014.2336395.
[10] T. Nakajima, Takeshi Yaguchi, Hajime Sugimura, "An ATE Architecture for Implementing Very High Efficiency Concurrent Testing", International Test Conference, IEEE, 2012. https://doi.org/ 10.1109/TEST.2012.6401551.
[11] Gary L. West, H. Troy Nagle, Jr., Victor P. Nelson, "A Microcomputer-Controlled Testing System for Digital Integrated Circuits", IEEE Transactions on Industrial Electronics and Control Instrumentation, **IECI-27**, No. 4, Nov. 1980.https://doi.org/10.1109/TIECI.1980.351644.
[12] J. Dunbar, "FPGA Based Design for Accelerated Fault-testing of Integrated Circuits" , Master's Theses, Bucknell University, 2010.
[13] A. L. Boudreault, "Automatic Test Equipment in the Production Process", IEEE Transactions on Manufacturing Technology, Volume-4, No.2, December 1975.https://doi.org/10.1109/TMFT.1975.1135861.
[14] Takashi Kitagaki, "Flexible ATE Module with Reconfigurable Circuit and Its Application", International Test Conference, IEEE, 1999.https://doi.org/10.1109/TEST.1999.805826.
[15] J. Romoth, M. Porrmann, and U. Ruckert " Survey of FPGA Applications in the period 2000-2015", Center of Excellence Cognitive Interaction Technology, Bielefeld University, Germany. https://doi.org/10.13140/RG.2.2.16364.56960.
[16] S. S. Tripaliya and P. P. Bansod, "FPGA based IC Tester," IJEEDC, ISSN: 2320-2084 , **3**(5), May-2015.https://doi.org/10.1109/ITC-Asia.2019.00018.
[17] N. Velamati, Robert Daasch, "Analytical Model for Multi-site Efficiency with Parallel to Serial Test Times, Yield and Clustering", 27th IEEE VLSI Test Symposium, 2009.https:doi.org/10.1109/VTS.2009.42.
[18] Y. Lee, Inhyuk Choi, Kang-Hoon Oh, James Jinsoo Ko and Sungho Kang, "Test Item Priority Estimation for High Parallel Test Efficiency under ATE Debug Time Constraints", Teradyne inc. 2017. https://doi.org/10.1109/ITC-ASIA.2017.8097131.
[19] S. Kumar Goel, Erik Jan Marinissen, "On-Chip Test Infrastructure Design for Optimal Multi-Site Testing of System Chips" , Proceedings of the Design, Automation and Test in Europe Conference and Exhibition, 2005. https://doi.org/10.1109/DATE.2005.231.
[20] K. Voon Ching, "A Case Study of Return on Investment for Multi-sites Test Handler in The Semiconductor Industry Through Theory of Industry 4.0 ROI Relativity", International Journal of Recent Contributions from Engineering, Science &IT, **7**(3), 23-40, July 2019. https://doi.org/10.3991/ijes.v7i3.11057.
[21] Y. Takahashi, Akinori Maeda, "Multi Domain Test: Novel Test Strategy to reduce the Cost of Test", 2011 29th IEEE VLSI Test Symposium, 2011. https://doi.org/10.1109/VTS.2011.5783738.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 937-943 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design and Implementation of DFT Technique to Verify LBIST at RTL Level

Nagaraj Vannal[1,2,*], Saroja V Siddamal[3]

[1]*Department of ECE, BVBCET Hubballi, VTU Belgaum,590018, India*

[2]*Technical Lead (Level-2), ASIC Division, Einfochips (an ARROW company) Pvt. Ltd Bangalore, 560095, India*

[3]*School of ECE, KLE Tech University, Hubballi,580031, India*

ARTICLE INFO

Article history:
Received: 17 August, 2020
Accepted: 04 October, 2020
Online: 08 December, 2020

Keywords:
DFT (Design for Testability)
LBIST (Logic Built in Self-Test)
SoC (System on Chip)
ECO (Engineering Change Orders)
Netlist
RTL (Register-Transfer Level)
ATPG (Automatic Test Pattern Generator)

ABSTRACT

According to IEC 61805 and ISO 26262 standards requirement inclusion of LBIST (Logic Built in Self-Test) became mandatory to achieve safety critical application such as automotive field. In such systems, once device is switched ON LBIST (Logic Built in Self-Test) is activated and testing of digital logic is performed. After safety subsystem says that the LBIST passed, the SoC (System on Chip) moves into the functional mode otherwise, the SoC moves into a safe state. In this entire start-up sequence the LBIST interacts extensively with the safety sub-system of the SoC. Startup sequence remains un-verified at RTL (Register-Transfer Level) leading to painful ECOs (Engineering Change Orders) and post Silicon issues in some cases. LBIST verification can only run if scan chains are present in design which is not the case at RTL. The paper describes design of a Design-for-Testability (DFT) technique to enable LBIST based system verification with different test approaches at RTL which eliminates the possibility of ECOs by catching most of the issues at RTL level. Simulation results are demonstrating the feasibility of the approach with emphasizing the benefits obtained on significant computational modules.

1. Introduction

LBIST became alternative approach for ATPG (Automatic Test Pattern Generation) due to increased complexity and test cost requirement for functional blocks [1]-[3]. Electronic component general safety-critical applications [4] are referenced in IEC 61805 [5] and Automotive Safety Integrity Level (ASIL) for automotive domain is referenced in ISO 26262 standard [6].

Latent faults (LFs) detection for automotive systems using power on self-test (POST) based on logic built in self-test (LBIST) is executed during engine start -up to test safety – critical devices before staring any functional operations [7]. POST is having limited test application time in automotive systems [8]. Getting required LF coverage (>90%) within less test application time is primary requirement [9].

the pass/fail comparison of the unit [10], [11]. PRPG (Pseudo Random Pattern Generator) is used for generating input stimulus for the design, and MISR (Multiple Input Signature Register) to capture the compact the response. Test patterns which is supplied to internal scan chains are output of PRPG block and response is the input to the MISR compaction. Signature is the compacted response for that cycle of operation [12], [13]. The LBIST implementation is as shown in Figure 1.

LBIST is based on 'scan' methodology to test digital logic [14], this suffers from limitation – The design cycle will be delayed by the time the netlist is ready with Scan insertion, X-bounding [15] and Test-point insertion done [16] to perform LBIST verification. Performing LBIST patterns gate-level simulations early in the design cycle will take weeks to complete since execution is performed in series and are large number [17].

In this paper, we propose an innovative strategy for the in-field testing of automotive devices. Methodology to perform parallel verification of sub-modules of System on Chip (SoC) is described. Section 2 describes proposed design methodology to per-form Logic BIST verification at RTL level with dummy netlist. Section 3 describes implementation details such as scan chain insertion steps, dumpy netlist creation and direct mode entry. Simulation results with debugging analysis details are discussed in section 4 and in section 5 conclusion of the work is described.

*Corresponding Author: Nagaraj Vannal, Email: nagaraj.vannal@gmail.com

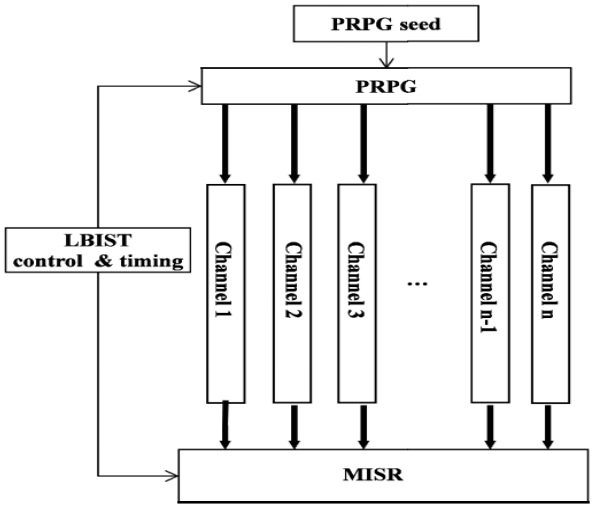


Figure 1: Standard LBIST Implementation

2. Proposed Design Methodology

Logic Built-In Self-Test (LBIST) is implemented by several LBIST controllers which operate independently on different partitions of the device as shown in Figure 2. This is needed to fulfil safety requirements about the independence and diversity of replicated IP, helps in physical flow and helps to avoid exceeding power limits. For applying traditional external scan patterns, LBIST controllers are put into trans-parent mode. Typically scan chains used for LBIST are shorter than scan chains used for compressed ATPG (Automatic Test Pattern Generation). Therefore, in LBIST transparency mode, the short LBIST chains are concatenated during to longer pro-duction scan chains which then are connected to the EDT (Embedded Deterministic Test) scan compression IP.

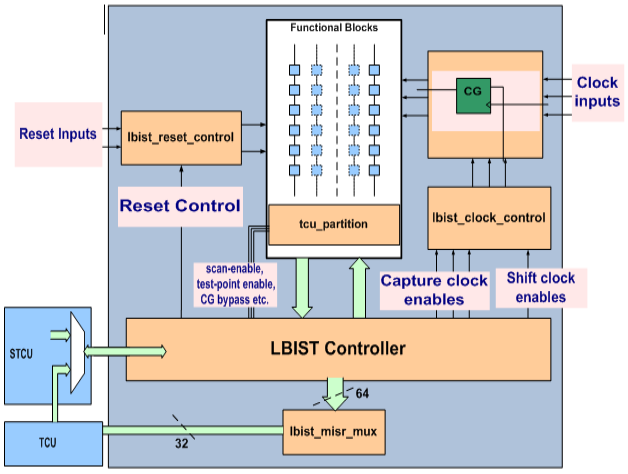


Figure 2: LBIST Architecture

LBIST controller IP itself will become part of the scan chains used for external scan patterns. Certain registers may not change value during LBIST operation but should be part of regular ATPG scan. These registers are put into one or more separate scan chains which are not shifted during LBIST. During LBIST their values are not influenced by shifting. These special scan chains are only used for ATPG. These registers will be identified during implementation. Each of the LBIST controllers is connected independently to the STCU (Self-Test Control Unit). The STCU can either start all LBIST controllers at the same time in parallel or in a sequence defined by the STCU. Only for debug reasons it is possible to access and operate each LBIST controller separately. This is done via JTAG (Joint Test Action Group) and a path through TCU (Test Control Unit) to the selected LBIST controller.

2.1. LBIST Partitioning

According to the safety concept eight LBIST partitions are implemented. They are numbered A0, A1, B0, B1, C0, C1, P0, P1. Each LBIST partition is contained in a separate LBIST island or lake as shown in Table 1. List of blocks where LBIST implementation is exempted:

- JTAG Controller
- TCU
- IO_top level ((Input-Output) / Pad ring
- PMU (Power Management Unit)/ PMU-Controller
- STCU
- PLL (Phase Locked Loop)
- XOSC (Oscillator)

Table 1: Characteristic data of LBIST partitions

	Characteristics				
Partition	**Size**	**LBIST chains**	**PRPG size**	**Clock domains**	**Chain overhead**
A0	77K	1600	46	System, tck	10%
A1	60K	1100	36	System, tck	10%
B0	24K	500	24	System, tck	10%
B1	60K	1100	36	System, tck	10%
C0	45K	1000	34	System, tck	10%
C1	60K	1100	36	System, tck	10%
P0	26K	600	26	System, tck	10%
P1	60K	1100	36	System, tck	10%

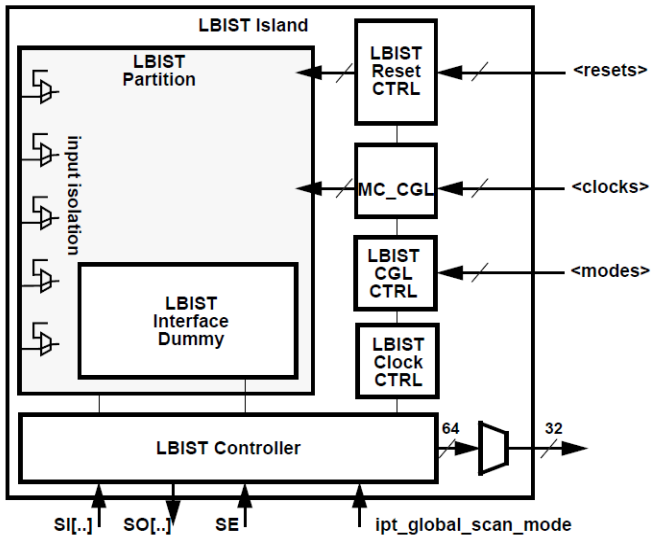
2.2. LBIST Integration

The design is split into several LBIST islands, where each LBIST island contains the actual LBIST partition, i.e. the logic to be covered by LBIST, a dedicated LBIST controller, dedicated clock and reset control blocks as shown in Figure 3. LBIST Controller:

Essential features of the LBIST controller IP used are:

- External clock control.
- Serial and parallel interface (to simplify communication, both functional and test/debug).
- Neg-edge lockup registers after PRPG and before MISR (to simplify LBIST scan chain timing)

With the introduction of lockup registers after PRPG and before MISR all LBIST scan chains must begin and end with a pos-edge flop. With this architecture all LBIST clock domains with less than half a cycle clock skew can be handled safely.



LBIST Controller Hierarchy: The controller needs to be generated with the following parameters:

set rtl hierarchy -Clk_mux Out -external_clk_enables
set lbist Controller -programmable_shift_rate ON
set shift_rate Divisor 8

This generates the LBIST controller IP in a hierarchy. The LBIST clock controller is entirely based on clock enable rather than the clocks generated by the LBIST IP in order not to generate a new clock path. Figure 4 gives an overview of the generated Controller IP and its hierarchy.

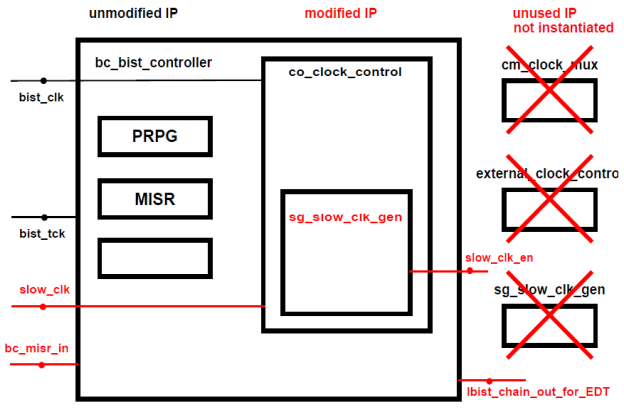


Figure 4: LBIST Controller IP Hierarchy

2.3. LBIST Controller Clocks

The LBIST controller operates on three different clocks, the LBIST engine clocks, bist_clk, and slow_clk, as well as the STCU communication clock bist_tck. The LBIST engine clocks bist_clk and slow_clk must fulfill the following requirements:

- The clock source for 'bist_clk' and 'slow_clk' must be identical.
- The clock source for 'bist_clk' and 'slow_clk' must be the fastest clock used inside the LBIST partition.
- The clock source for 'bist_clk' and 'slow_clk' must be the clock with the highest clock insertion delay used inside the LBIST partition.

2.4. LBIST Interface Dummy

The LBIST interface dummy propagates a standard LBIST interface into the LBIST partition, to ensure, that all relevant signals are present and preserved through synthesis and scan insertion. The LBIST interface dummy combines the following interfaces / features:

- LBIST controller IP interface
- LBIST clock interface
- LBIST clock gating control ('ipt_cg_bypass' and 'ipt_se_gatedclk')
- Test points that can be connected by the physical team to improve coverage if required
- Production Scan: Launch-Off-Shift (LOS) per clock domain and blocking of LBIST control signals for at-speed scan.

The LBIST interface dummy is protected during synthesis because connections to the module are created during scan chain stitching and automatic clock gate insertion only.

2.5. LBIST Interface Dummy

For direct execution of LBIST, the optional parallel interface of Mentor's LBIST controller is used. It provides the following interface signals:

- bist_run_input- triggers LBIST run
- bist_done - flags a finished LBIST run
- misr_value - vector output providing parallel read access to MISR value

Debug and diagnosis require access to LBIST control registers. It is done through LBIST's serial interface. Writing and reading control registers happens in a JTAG-like serial mode. The LBIST serial debug interface is used by the STCU to con-figure LBIST run parameter, such as pattern count. Full debug and diagnosis are performed through the TCU only. The interface is based on a JTAG-like protocol with shift- and update-DR states. Both TCU and STCU share the interface to the LBIST controller IP, where the muxing between TCU and STCU is performed inside the STCU controlled by 'tcu_lbist_direct_ctrl'. The actual TCU access is performed though the TCU external TDR mechanism based on the register selection through the LBIST TDR (LTDR). Expected LBIST signatures are kept in the Non-Volatile Memory (NVM). Through STCU programming the required signature is fetched from the NVM for comparison.

2.6. LBIST Clock and Reset Control

LBIST Clock Control: All LBIST clocks are controlled through the Magic Carpet Clock Gating Logic (MC_CGL) associated with each LBIST partition. The clock control by the MC_CGL is purely based on clock enables rather than the clocks generated by the LBIST IP. Figure 3 shows the modules relevant for LBIST clock control for one LBIST partition 'C0'.

The LBIST controller, as well as the blocks lbist_clock_control, lbist_cgl_control, and cgl_lake exist local to all eight LBIST partitions. The LBIST controller with its modified co_clock_control block is connected to the main lbist_clock_control module, that combines the shift and capture enable signals. The lbist_cgl_control block switches between the clock enables generated by the TCU and by the LBIST controllers depending on the mode setting. The clock gate inside the MC_CGL will generate the slow bist clock based on the enable provided by the LBIST IP. The LBIST clock control will provide punch out clocks, not 50% duty cycle clocks.

Slow Clock Generation: The LBIST slow clock ('slow_clock') is a punch out clock divided down from bist_clk (== fast_clk) using a programmable clock divider inside the 'sg_slow_clock_gen'. It is used to control several key LBIST functions, such as PRPG, MISR, pattern counter etc. The same slow clock divider is used to program the shift clock frequency. The default divider ratio of the 'sg_slow_clk_gen' module is set to 8 but can be reprogrammed via the LBIST control registers.

LBIST Reset Control: To control all resets within the LBIST partition, all resets must be routed through the LBIST reset control block. The block allows reset control for LBIST, while the system resets are not gated. This is to make sure, that a system reset is not blocked during device startup.

LBIST reset requirements are:

- All asynchronous resets of an LBIST partition must be routed through the LBIST reset control block.
- All functional resets must be in off state during LBIST execution.
- No LBIST partition output to MC_RGM/MC_RSL must have an influence on self-test / LBIST execution.

The original reset control logic provided by the LBIST controller IP is an active-high reset (clock) signal multiplexed into the functional reset path. With the non-blocking reset control, where the functional reset is not blocked by LBIST a new reset control is introduced replacing the original one. To pre-vent reset glitches from propagating into the LBIST partition in scan mode, the LBIST reset control must be bypassed by a scan mode reset. The non- blocking reset control dictates, that all functional resets are off during LBIST execution, otherwise the LBIST MISR becomes invalid.

2.7. Making the LBIST Partitions LBIST Ready

The interface of the LBIST partitions need special attention to ensure that there is no X poisoning into the LBIST partition. All critical signals causing problems in the standard X-bounding flow will be X-bound in RTL. Mentor Testkompress fault simulation is used to identify missing X-bounds. The decision on whether an input is to be X-bounded is taken based on the partition input signal categories shown in Table 2.

The X-bounds inserted in RTL will instantiate a MUX controlled by lbist_en from the LBIST controller IP. Whenever possible the partition outputs are used as inputs to the MUX. This increases coverage as it provides toggling and not constant inputs to the MUXes, while at the same time adding observability to the partition outputs.

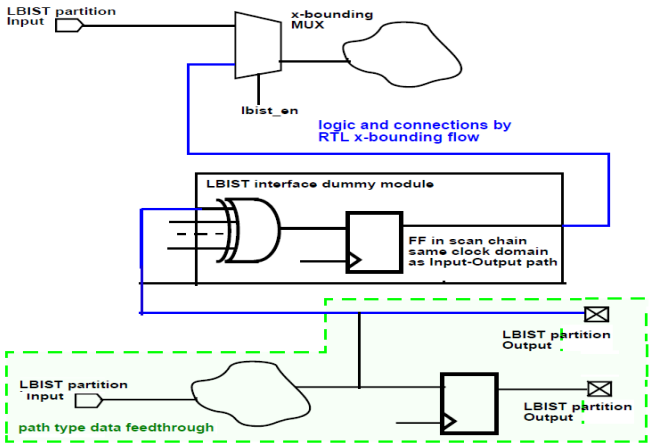


Figure 5: X-bounding to avoid multi-cycle path generation

Figure 5 shows a concept used to avoid multi-cycle path generation on known critical input – output combination. The LBIST interface dummy module will include several XOR gate FF combinations as shown below. The RTL X-bounding flow uses and connects these structures. In a later stage the automatic RTL x-bounding flow might also include instantiation of these structures instead of using predefined ones.

Table 2: LBIST partition input categories

Category	LBIST Control	Bound
Reset input	Via LBIST reset control	No
Clock input	Via LBIST clock control	No
Clock gate control	Via LBIST interface dummy	No
Common Digital Input		Yes
Selected Tied Digital Input		Yes
Unconnected Digital Input	None required	No
Digital Input from TCU – no influence on LBIST		yes
Digital Input from TCU – influence on LBIST	Constant from TCU	No
Bidirectional	NA	No
Analog/Special Nets	All hard IP in LBIST mode	No
LBIST Controls	From LBIST control IP	No

3. Implementation Details

LBIST scan chain insertion and concatenation to top level LBIST chains including lockup-FF insertion followed these basic steps in the DFT and physical flow.

1. DFT flow - Generated EDT controllers everything pos-edge based w/o lockup-FFs
2. DFT flow - Generated LBIST Controller and connected internal LBIST chains at the BIST controller to form ATPG scan chains. Modified LBIST controller to provide ATPG chain outputs with LBIST controller interface. Connecting ATPG chain inputs and outputs of LBIST controller to EDT controllers using IPXACT flow. (Assigned fixed value to EDT controller pins for remaining chains outside LBIST to protect them in synthesis)
3. Physical flow -One step scan insertion: Building LBIST chains, connecting them to the LBIST controller. Building ATPG chains for logic outside LBIST, connect-ed them to the

EDT controller using the pins with fixed value assigned by IPXACT.

4. Physical flow - Post processing step to insert lockup FFs, including required lockup-FFs in concatenated LBIST chains.
5. Checked for missing lockup-FFs on

a. change of clock name or clock polarity in scan cell report - DFT flow

b. clock hierarchy change - Physical flow

Figure 6 depicts concatenation of four LBIST scan chains to form one production chain with input and output available at the LBIST controller boundary for hook up to the EDT controller. Also shows where lockup FFs are placed by the physical flow.

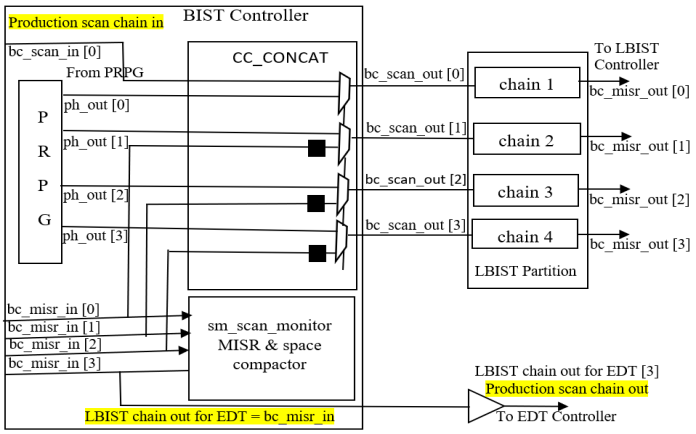


Figure 6: LBIST chain concatenation and lockup latch insertion

3.1. LBIST Dummy Netlist Integration at RTL

Integrating a dummy netlist in the design at RTL level with the same number of scan chains, scan clocks and scan resets is done. This dummy netlist should only be visible to the verification environment, in the LBIST testcases expect the MISR calculated for the dummy netlist.

Flow: Generate a dummy netlist (with scan chains) → Generate the MISR value for the same dummy netlist →Integrate this dummy netlist into the actual design RTL → run the testcases for verifying the complete start-up sequence. Details of each step are described below:

Step1: Generate a dummy netlist using "skeleton design creation".

create_skeleton_design -o lbist_for_verif -i skeleton_design.txt

Step2: Generate MISR for dummy netlist by doing fault simulation using LBIST Architect

*lbistarchitect -fault_simulation *
*lbist_for_verif.v -verilog *
*-lib skeleton.fslib -nogui *
-dofile lbist_for_verif.dofile

Step3: Integrate the dummy netlist into the lbist partition using "ifdef"

Step4: Perform RTL simulations for LBIST based start-up sequence using the com-pilation argument:

+define+LBIST_DUMMY_SIM

3.2. LBIST Direct Control Mode

For LBIST debug purposes there is a 'LBIST direct control' mode, where the different LBIST partitions can be run stand alone or in parallel. Which LBIST partitions are run, as well as LBIST parameters like pattern count, MISR start value etc., are programmed via JTAG. The LBIST Direct Test Mode Entry uses the LBIST Test Data Register (LTDR) to select the LBIST controllers, start the LBIST run. The following sequence shown in Table 3 is used in the Teskompress setup as well is in the Verilog stimuli for LBIST direct mode verification and pattern generation.

Table 3: LBIST direct mode entry

Step	Description
1	Common steps to TCU based mode entry
2	Load and update the Generic TCR
3	Load and update the RGMTDR to disable gating reset phase 3.
4	Load and update the CLK TDR to program the PLL0 clock source for LBIST execution
5	Set LBIST direct control: Load and update LTDR LBIST direct mode: tcu_lbist_testmode=1; tcu_lbist_direct_control=1;
6	Turn TEST mode on to be able to program LBIST Controllers
7	Program LBIST controllers Select the LBIST controllers: Load and update the LTDR: ipt_lbist_sel[4:0]=<controller id>Program controller settings like pattern start/end counter using the steps in Table V.
8	Deassert RESET and apply EXTAL clock pulses to get rgm state machine to stabilize.
9	Start LBIST run on selected controllers: Load and update LTDRipt_lbist_sel[3:0]=<selected controllers> and ipt_lbist_run=1.
10	2 cycles for LBIST finder PRPG and MISR recognition
11	Deassert RESET_B

The following Table 4 describes the sequence of Writing to / Reading from an LBIST controller register in LBIST direct test mode and Table 5 describes the sequence to read out the MISR values after completion of an LBIST run in LBIST direct test mode.

Table 4: Writing to / reading from an LBIST controller register

Step	Description
1	Write register address to LBIST controller dselect register: a. Load and update the LTDR to select dselect register: ipt_lbist_reg_sel=1 b. Send JTAG instruction 6'd5 to access external LBIST controller register. Write value to LBIST controller dselect register. E.g. '0110' to map pattern count end value into data register.
2	Write value to / read value from an LBIST controller data register a. Load and update the LTDR to select data register: ipt_lbist_reg_sel=0. b. Send JTAG instruction 6'd5 to access external LBIST controller register. Write value to / read value from LBIST controller data register

Table 5: Reading out MISR values

Step	Description
1	Read the LTDR to observe LBIST done flags lbist_done [3:0]
2	Turn off EXTAL before switching from bist_clk to TCK
3	Load and update LTDR: ipt_lbist_run=0. This switches from bist_clk to TCK
4	Read MISR registers a. Select the LBIST controller: Load and update the LTDR: ipt_lbist_sel[3:0] b. Read the MISR value using the steps in Table V.

4. LBIST verification results at RTL

Verification of all different partitions are carried out with four different testcases listed below running parallelly with respective MISRs generated during LBIST Fault simulation using LBIST Architect tool [18].

1. Testcase with PLL clock
2. Testcase with EXTAL clock
3. Scan chain mask decoder
4. MISR exposed on PADs

4.1. Testcase with PLL clock frequency

LBIST controller logic/integration reset and clocks of the respective partitions, mode entries are verified at lbist_fast_clk = 320MHz & lbist_slow_clk = 80 MHz frequencies. Figure 7 shows one of the issues identified for a partition.

- MISR mismatching is observed after pattern count value 237, expected MISR value is 415B3579F4B1C0DA, simulation result MISR value is DE2F87E1C111F4D1.
- Analysis: Capture window for pattern count value after 237 is as below but from waveform we can see along with below 4 clk_domain clk_domain_58 is also pulsed which is causing the MISR mismatch.

clk_domain_1 → pulse '10000000'
clk_domain_18 → pulse '00100000'
clk_domain_39 → pulse '00001000'
clk_domain_30 → pulse '00000010'



Figure 7: MISR mismatch issue caught during RTL verification

4.2. Testcase with EXTAL clock Frequency

LBIST controller logic/Integration reset and clocks of the respective partitions, mode entries are verified at EXTAL = 100MHz frequencies. As shown in Figure 8 MISR is observed after pattern count value 260, expected MISR value is B8D5_F506_668B_F517, simulation result MISR value is B8D5_F506_668B_F517.

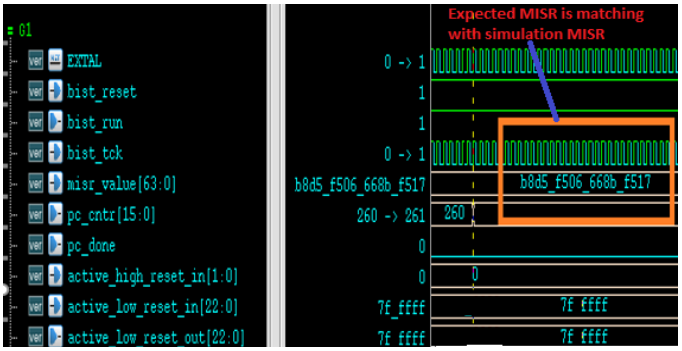


Figure 8: LBIST Verification with EXTAL clock frequency 100MHz

4.3. Scan Chain Mask Decoder Testcase

LBIST controller logic/Integration, reset and clocks of the respective partitions, verified at lbist_fast_clk = 320MHz & lbist_slow_clk = 80 MHz frequencies. As shown in the Figure 8 chain_0 is masked for every LBIST partition and pattern lbist_si[980:0] toggles for mask decoder lbist_si[980:1] and lbist_si[0] will be static 1.

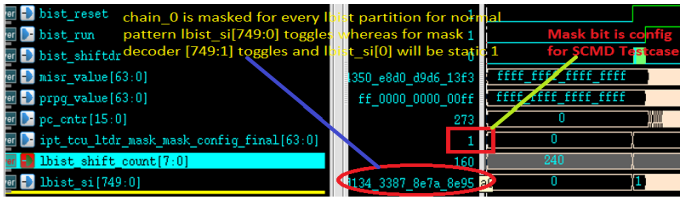


Figure 9: Scan chain mask decoder testcase result

4.4. MISR Expose at PAD Testcase

LBIST controller logic/Integration, reset and clocks of the respective partitions, verified at lbist_fast_clk = 320MHz & lbist_slow_clk = 80 MHz frequencies along with MISR value is exposed at PAD during end of the pattern count using lbist_misr_word_sel bit to expose 1st 32 -bit and later 32 -bit values respectively. As show in Figure 10 MISR value DCCF_7DA6_225C_9279 at program counter value 272 is exposed on PAD pins using lbist_misr_word_sel (0/1).

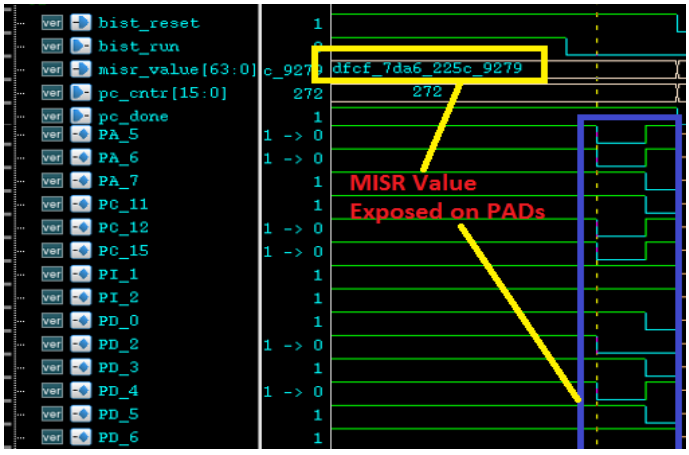


Figure 10: MISR exposed at PAD testcase result

5. Conclusion

The proposed methodology for verifying LBIST implementations at RTL in SoCs is not a replacement for gate

level verification of LBIST implementations, with this approach a significant effort and design cycle time can be reduced by moving the verification process early into the SoC design cycle. The results are proven with measurable improvements on real design where the flow/methodology was implemented and helped in identifying the bugs & fixing them before logic freeze, thereby avoiding ECOs.

References

[1] Andal Jayalakshmi and Tan Ewe Cheong, "A methodology for LBIST logic diagnosis in high volume manufacturing," 2012 4th Asia Symposium on Quality Electronic Design (ASQED), Penang, 249-253, 2012, doi: 10.1109/ACQED.2012.6320510.

[2] K. Butler, "ATPG versus Logic BIST -Now and in the Future," in 2013 IEEE International Test Conference (ITC), Baltimore, Maryland, 1181, 2001. doi: 10.1109/ITC.2001.10048

[3] P. Maxwell, I. Hartanto and L. Bentz, "Comparing functional and structural tests," Proceedings International Test Conference 2000 (IEEE Cat. No.00CH37159), Atlantic City, NJ, USA, 400-407, 2000, doi: 10.1109/TEST.2000.894231.

[4] S. Carbonara, P. Bernardi and M. Restifo, "A Hybrid In-Field Self-Test Technique for SoCs," 2019 14th International Conference on Design & Technology of Integrated Systems In Nanoscale Era (DTIS), Mykonos, Greece, 1-6, 2019, doi: 10.1109/DTIS.2019.8735075..

[5] International Standard - IEC61508 - Functional safety of electrical/electronic/programmable electronic safety-related systems, International Electrotechnical Commission, 2010.

[6] ISO Standard 26262, Part 5, "Road vehicles— functional safety," 24 May 2016.

[7] Y. Sato et al., "A circuit failure prediction mechanism (DART) for high field reliability," 2009 IEEE 8th International Conference on ASIC, Changsha, Hunan, 581-584, 2009, doi: 10.1109/ASICON.2009.5351352.

[8] H. Iwata and J. Matsushima, "Multi-configuration Scan Structure for Various Purposes," 2016 IEEE 25th Asian Test Symposium (ATS), Hiroshima, 131-131, 2016, doi: 10.1109/ATS.2016.32.

[9] S. Wang, Y. Higami, H. Takahashi, H. Iwata and J. Matsushima, "Automotive Functional Safety Assurance by POST with Sequential Observation," in IEEE Design & Test, **35**(3), 39-45, June 2018, doi: 10.1109/MDAT.2018.2799801.

[10] Wu-Tung Cheng, Manish Sharma, T. Rinderknecht, Liyang Lai and C. Hill, "Signature based diagnosis for logic BIST," 2007 IEEE International Test Conference, Santa Clara, CA, 1-9, 2007, doi: 10.1109/TEST.2007.4437703.

[11] M. Abramovici, M. A. Breuer, and A. D. Friedman, "Digital System Testing and Testable Design", AT&T Bell Laboratories and W. H. Freeman and Company, 1990.

[12] H. D. Schnurmann, E. Lindbloom and R. G. Carpenter, "The Weighted Random Test-Pattern Generator," in IEEE Transactions on Computers, C-**24**(7), 695-700, July 1975, doi: 10.1109/T-C.1975.224290.

[13] G. Kiefer and H. -. Wunderlich, "Using BIST control for pattern generation," Proceedings International Test Conference 1997, Washington, DC, USA, 347-355, 1997, doi: 10.1109/TEST.1997.639636.

[14] Y. Maeda, J. Matsushima and R. Press, "Automotive IC On-line Test Techniques and the Application of Deterministic ATPG-Based Runtime Test," 2017 IEEE 26th Asian Test Symposium (ATS), Taipei, 237-241, 2017, doi: 10.1109/ATS.2017.52.

[15] M. Sadi, G. K. Contreras, J. Chen, L. Winemberg and M. Tehranipoor, "Design of Reliable SoCs With BIST Hardware and Machine Learning," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, **25**(11), 3237-3250, Nov. 2017, doi: 10.1109/TVLSI.2017.2734685.

[16] P. Wohl, J. A. Waicukauski, G. A. Maston and J. E. Colburn, "XLBIST: X-Tolerant Logic BIST," 2018 IEEE International Test Conference (ITC), Phoenix, AZ, USA, 1-9, 2018, doi: 10.1109/TEST.2018.8624738.

[17] E. Moghaddam, N. Mukherjee, J. Rajski, J. Solecki, J. Tyszer and J. Zawada, "Logic BIST With Capture-Per-Clock Hybrid Test Points," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, **38**(6), 1028-1041, June 2019, doi: 10.1109/TCAD.2018.2834441.

[18] LBIST Architect process guide Manual Mentor Graphics, Software Version 2018.1.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 795-804 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Non-Ridership Presumption Toward New Bus Rapid Transit Purwokerto-Purbalingga Execution

Fauzan Romadlon*

Industrial Engineering Department, Faculty of Industrial Engineering and Design, Institut Teknologi Telkom Purwokerto, Purwokerto, 53147, Indonesia

ARTICLE INFO

Article history:
Received: 11 June, 2020
Accepted: 17 August, 2020
Online: 28 August, 2020

Keywords:
Non-ridership
Bus Rapid Transit
Presumption
Purwokerto-Purbalingga

ABSTRACT

Bus Rapid Transit in Purwokerto-Purbalingga is a new mass transportation mode. Recently, the execution of the BRT has been going on for three years. In terms of service standards to ridership, the BRT has been fulfilled the requirement. However, during the execution, it shall be supported by the non-ridership (local communities) who get the impact as public engagement. The non-ridership impact is captured by observing their presumptions. This study uses quantitative method and survey technique to collect the data by spreading questionnaires to the non-ridership in Purwokerto and Purbalingga. The collected data is analyzed by Analysis of Variance (ANOVA) and Structural Equation Modelling Partial Least Square (SEM-PLS). The ANOVA results show that gender, age ratio, and residence (living area) are significant presumption factors. According to SEM-PLS model, the R-squared of non-ridership presumption variables toward the BRT execution as excellent public transportation is at 51.8% (moderate level). It is found that the economic variable affects the excellent public transportation variable is at 41.4%, and the social variable have a correlation with the excellent public transportation variable is at 36.2%, but not with the environment variable (5.6% only). Following up these findings, it is recommended that public engagement through the non-ridership presumption will lead the BRT provider to purpose some programs to improve the service and increase the occupancy. So that, the proposed program will attract the sense of awareness and public engagement of the non-ridership toward the BRT execution.

1. Introduction

Developing worldwide Bus Rapid Transit (BRT) system are increased significantly. Especially in Indonesia, BRT has been implemented not only in the capital city but also in many big cities to support urban mobilities. One of the advantages of BRT is to provide public transportation that relies on the ease of use [1], safety, comfort, and convenience service [2]. The level of BRT service is measured based on BRT effectiveness. BRT service satisfaction is achieved by increasing service procedures, service requirements, and service distinction with conventional public transportation [3]. The quality of service is measured by customer experiences through maintaining bus stop average waiting time, evaluating the impact of operation timetable, and considering BRT capacity [4].

Bus stop facilities are compulsory in BRT execution. Bus stops have to integrate with people density which is close to job location to come up with accessibility enhancement [5] and give impacts to potential access informal job opportunities [6]. Besides, the building environment around the bus stops has to consider the density of public facilities and green areas [7]. Affordability of BRT stops shall concern spatial structure [8] and spatial heterogeneity to support BRT planning in service management and infrastructure provision [9]. Therefore, the placement of BRT stops must be considered friendly pedestrian, densely of resident, industrial park connection, and public facilities consideration [10].

In advanced development, BRT has been configured as multimodal transportation networks. It is used to promote the reduction of greenhouse gas emissions and made the diversity of land use [11]. Multimodal and intermodal transportation is enhanced to emphasize of social action while providing reliable and smooth access to the city center [12]. The multimodality index in BRT has a positive relation with BRT ridership by shaping their behavior, land use [13], and transit rate [14]. Ridership perspective

*Corresponding Author: Fauzan Romadlon, , Institut Teknologi Telkom Purwokerto, Email: fauzan@ittelkom-pwt.ac.id

and attitude (behavior) on BRT are the leading indicators to enhance BRT service. Differences between males, elder workers, workers with higher education, and income [15] make the different perspectives of the service. The perspectives are generated motivation challenge of using BRT, such as security, reducing commuting time, staff attitude, and price charges [16]. The impact of land use toward BRT affects high values of properties [17], increased urban development accesses, and optimized time reduction [18].

The success bus implementations are affected overall sustainability such as safety has increase BRT benefit, pro-poor, or lower share of daily earnings [19], and decrease road traffic crashes [20]. On the other hand, BRT implementation faces some issue such as limited government financial support, green transportation issue, and high ridership occupancy [21] and no support form vehicle industries [22]. Sometimes, the execution of BRT has a problem with existing mass transportation, so it requires an understanding of existing transportation to be reorganized [23].

Implementing BRT as public transportation shall include public engagement [24]. The public engagement is to achieve transparent decision with good input from stakeholder [25]. After the public transport is realized, the system of public transportation shall be evaluated whether has economic development impact, sustainability impact, and effect to livable communities [26].

Post-planning or operation step needs more understanding of the connection between the services and public engagement [27]. In some case, public engagement especially from marginal stakeholder who has low power and low interest [28] are less involved. The marginal stakeholder can be identified as three groups: community who has no vehicles and they captive as primary beneficiaries of BRT, community who has vehicle but reluctant, and rich community who has strong voice [29].

Marginal stakeholder engagement can be measured from their presumption. In this study, the marginal stakeholder is non-ridership that lives around BRT stops and BRT routes. Their presumption will describe public engagement through new public transportation execution. The presumption can be set as a tool to capture the existing condition of public engagement toward BRT execution especially in different perspective. The new public transportation of the study is BRT Purwokerto-Purbalingga in Indonesia. It is chosen because it merely has been executed for three years but less public awareness.

The public engagement will lead to a sense of local ownership for the BRT in Purwokerto and Purbalingga. A sense of local ownership is used to identify economic probability and community benefits especially, in the job sector, age groups of worker effect, investment time [30], and less CO2 production [31]. In the other hand, community acceptance increases awareness system to the community, maximizes in implementation, and enhances ridership capacity [32].

2. Methodology

2.1. The Covered Respondent

The covered area of the study is in Purwokerto and Purbalingga. Purwokerto and Purbalingga are placed in Central Java Province, Indonesia. The operation of BRT is conducted from Ministry of Transport in Central Java Province. The route is started from Bulupitu bus station in Purwokerto until Bukateja bus station in Purbalingga. Bulupitu bus station is central bus station in Purwokerto that is the transit hub for ridership to travel intercity and interprovince. Bukateja is bus station that is set final destination and it will be transit hub that close to planned airport in Purbalingga. The total route of the BRT is 26.4 km with 56 bus stops [33].

The study is established by quantitative methods. The questionnaire is set to survey non-ridership who live around the BRT route and bus stops. The respondent is people who live or work around the BRT route and the BRT stop. Data collection technique is conducted by doing direct interviews, face to face, and then the result of these interview is submitted to a Google Form. The proportion of the respondent form both cities are set equal.

2.2. ANOVA Analysis

Demographics spread is set as proportional to both cities. The surveyed criteria are gender, residence, job, age, and routine. The presumption is measured using a Likert scale where the scale is five points; 1) Strongly disagree, 2) Disagree, 3) Moderate, 4) Agree, and 5) Strongly agree. After data collection, the presumption questionnaire result is tested for the validity and reliability. The validity test of the questionnaire uses Pearson Correlation and the reliability uses Cronbach' Alpha. Then, the multicollinearity, outliers, and normality test are performed to check whether the data is followed statistical rules or not.

The significance correlation of the demographics of non-ridership and the presumption data are approached by using ANOVA (Analysis of Variance) with Minitab 19 software (alfa is 5%). The hypothesis is set below.

H1: The demographics of non-ridership has significance correlation with their presumption.

2.3. SEM-PLS Model Evaluation

Structural Equation Modelling Partial Least Square (SEM-PLS) is used to find correlations among the presumptions. The presumption is divided into four categories; social, economic, environment, and excellent public transportation. The modelling is done by Smart-PLS 3 software. After finishing the concept, it is executed by designing a measurement model and set up the path diagram. Then, the last process is evaluated the model and tested the hypotheses.

The SEM-PLS model needs two steps to gain the fit model; model evaluation and structural model analysis. The model evaluation is measured by testing the validity and reliability of the model. The validity test needs three criteria; Convergent Validity, Average Variance Extracted (AVE), and Discriminant Validity. The criterion of reliability is Composite Reliability [34].

Convergent validity is measured based on the correlation between the indicator value and the variable value. The weight value is represented as loading factor or cross loading factor value. The loading factor is high if the value is more than 0.7. But for the initial research, the measurement scale with a loading value of 0.5 to 0.6 is considered as sufficient.

Average Variance Extracted (AVE) is the average of percentage extracted variance value from latent variable that is estimated through the loading standardize. The minimum AVE value is 0.5 (1).

$$\text{AVE} = \frac{\sum_{i=1}^{n} \lambda_i}{n} \quad (1)$$

Where λ is standardize loading factor, *i* means amount of indicator and *n* is amount of data.

Discriminant Validity can be calculated by comparing the square root of Average Variance Extracted (AVE) values. If the square root of AVE value is higher than the correlation value between latent variables, then discriminant validity can be considered as fulfilled. Then, the criterion of reliability model is Composite Reliability. The latent variable can achieve the reliability if minimum the Composite Reliability value is 0.6.

2.4. Structural Model Analysis

The structural model analysis is set up to know the relationship among constructed model, the significance value, and the model itself. The testing criteria is significance probability using R-squared (R^2). R-Squared is a method to measure Goodness of Fit (GoF) and to appraise the influence independent latent variable with dependent latent variable. After finishing the structural model analysis, the bootstrapping process is conducted. The bootstrapping process is functioned to minimize the abnormality of the modelled study.

2.5. Model Hypothesis

The social, economic, and environmental criteria are used to set as an exogen variables and excellent public transportation as an endogen variable. Each category is built up from related presumptions that are set with significant alfa 5%. The hypothesis of the model is,

H2: The economic variable has positive or significance correlation with excellent public transportation system.
H3: The social variable has positive or significance correlation with excellent public transportation system.
H4: The environment variable has positive or significance correlation with excellent public transportation system.

3. Result and Discussion

3.1. Demographics of The Non-ridership BRT

In Purwokerto and Purbalingga, the non-ridership respondent of the survey is 138. The demographics of them can be seen in Table 1.

Table 1: Demographics of Non-ridership BRT

User variable		Distribution (%)
Gender	Men	60.87
	Women	39.13
Residence	Purwokerto	43.48
	Purbalingga	49.28
	Others	7.25
Job	Student	26.09
	Teacher	2.90
	Labor	7.97
	Housewife	5.80
	Others	57.25
Age Ratio	15-25	30.43
	26-35	15.22
	More than 35	54.35
Attempt	Tried	62.32
	Never	37.68

According to Table 1, non-ridership gender is dominated by men at 60.87%, and non-ridership women at 39.13%. The dominated non-ridership is caused by men to work at informal economic sector around the bus stop. Most of them rely on the crowded street to gain the customer. They work as food street sellers, fruit sellers, parking guards, and pedicab drivers. The percentage is high at 57.25% for the job sector. It is described in Table 1 as other job. Meanwhile, most non-ridership women work as students (26.09%), workers (7.97%), housewife (5.8%), and teachers (2.90%). So, their availability around the BRT stops is lower than non-ridership men. Then, most of their age ratio is more than 35. The age ratio is productive worker age who become a family mainstay. The rest is the age ratio between 15 and 25 and between 26 and 35. The age ratio between 15 and 25 is dominated by student and between 26 and 35 is dominated by labor and housewife.

Regarding the residence area, most of them are living in Purbalingga (49.28%), followed by Purwokerto (43.38%) and others (7.25%). Other places mean the non-ridership is living around both cities such as Banyumas, Banjarnegara, and Cilacap Regency. The percentage of Purbalingga is higher than Purwokerto and other places mean that, most of the non-ridership works or stay around the BRT stop. Furthermore, the BRT route in Purbalingga is longer than Purwokerto, so the covered area of the BRT is wider than Purwokerto. People in Purbalingga is more often using the BRT than Purwokerto and others because of the BRT can connect the industrial park from their living residence. The majority of respondents have tried the BRT (62.32%). They tried the BRT for many purposes, such as visiting their family, going to work or school, and just for a fun activity.

Table 2: Non-Ridership Presumptions toward BRT

Code	Presumption
Q1	BRT capacity is suitable to fulfill demands
Q2	BRT is a reliable public transportation
Q3	BRT service is better than other public transportation
Q4	BRT accommodates ridership destination
Q5	BRT routes shall be expanded
Q6	BRT can be alternative public transportation
Q7	BRT provider is followed transportation procedure service
Q8	BRT is environment friendly transportation.
Q9	BRT is on-time schedule transportation.
Q10	BRT can increase economic productivity for local people
Q11	BRT can open economic opportunity for local people
Q12	BRT can open additional job opportunity
Q13	BRT can decrease the traffic congestion
Q14	BRT improve local people mobility
Q15	BRT can reduce traffic accident
Q16	BRT can reduce fossil fuel use.
Q17	BRT facilitates to access intermodal transportation
Q18	BRT stops are able to be added and expanded.
Q19	Ticket payment shall be set e-payment

3.2. Non-Ridership Presumption

The experience of non-ridership lead to their presumption. The presumption depicts society or community knowledge of implementing BRT. The presumption is used to measure non-ridership participation and awareness even though they have never tried the BRT service.

According to Table 2, there are several criteria of non-ridership presumptions that categorized into 19 presumptions. The queries are divided into four categories. The queries are set based on the non-ridership observations toward the BRT services around their living area. First presumption is economic. The presumption is focused on punctuality of the BRT schedule (Q9), economic productivity for local (Q10), economic opportunity for local (Q11), and additional job opportunity for local (Q12).

The second presumption is social. The presumption is set to measure that the BRT can decrease traffic jams in the cities (Q13), enhance local mobility or movement (Q14) and reduce traffic accidents (Q15). Then, the environment is set for seeking the knowledge of local about green transportation (Q8) and reducing fossil fuel usability in transportation (Q16). The last is a non-ridership presumption about common presumption about excellent public transportation. The limitation of excellent public transportation is giving impact to the non-ridership living and their social activities. Excellent public transportation presumption is built up from fulfilling demands based on BRT capacity (Q1), being reliable public transportation (Q2), being excellent service bus than the existing bus (Q3), accommodating ridership destination (Q4). Moreover, BRT route shall be expanded (Q5), the BRT can be alternative transportation (Q6), the BRT provider following the service procedure (Q7), BRT facilitates to access intermodal transportation (Q17), BRT stops are able to be added and expanded (Q18), and implementing e-payment ticket (Q19).

According to Table 3, the result of validity test of the questionnaire (Q1-Q19) are valid. It is shown that the Pearson Correlation values are higher than Pearson Correlation Table (0.455) except Q19. However, Q19 is still included into measurement because the p-value is 0.01 (less than 0.05). The result of reliability test is reliable. It is shown from the table that the Cronbach' Alpha is 0.9113 where the value is higher than 0.7 as standard minimum of reliability.

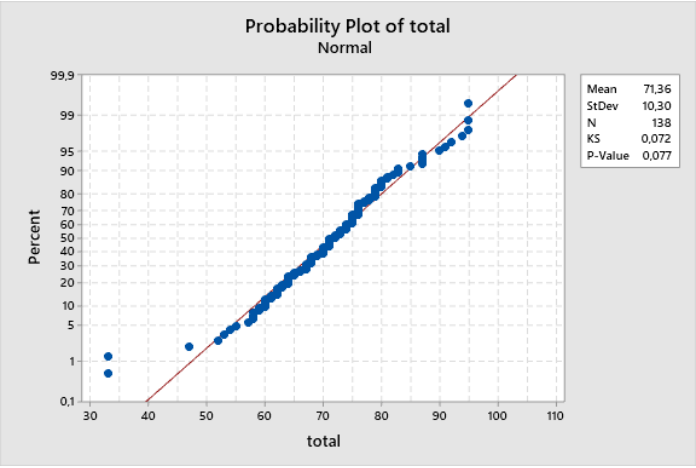


Figure 1: Normality Test Result

Table 3: Validity and Reliability of The Questionnaire Result

Code	Pearson Correlation	P-value	Cronbach' Alpha
Q1	0.631	0.000	0,9113
Q2	0.637	0.000	
Q3	0.619	0.000	
Q4	0.627	0.000	
Q5	0.614	0.000	
Q6	0.747	0.000	
Q7	0.682	0.000	
Q8	0.648	0.000	
Q9	0.553	0.000	
Q10	0.678	0.000	
Q11	0.695	0.000	
Q12	0.640	0.000	
Q13	0.650	0.000	
Q14	0.709	0.000	
Q15	0.559	0.000	
Q16	0.564	0.000	
Q17	0.665	0.000	
Q18	0.562	0.000	
Q19	0.286	0.001	

The multicollinearity and outliers test result can be seen in Table 4. According to the table, all of indicator has Varian Inflation Factor (VIF) value that less than 10. It means that there is multicollinearity among the indicators and it guarantees that there is no multicollinearity problem. The outliers test indicates all of p-value are more than 0.05. It means that there is no outlier data. The normality test of the data is shown in Figure 1. Regarding to the figure, the data is followed normal distribution rules because the p-values is 0.077 or greater than 0.05. So, all collected data is sufficient and qualify referring to the statistical rules.

Table 4: Multicollinearity and Outlier Test Result

Code	Multicollinearity	Outlier	
	VIF	G	P-values
Q1	2,18	2,34	1,000
Q2	2,41	2,72	0,812
Q3	1,69	2,81	0,603
Q4	1,64	4,16	0,266
Q5	1,72	3,54	0,355
Q6	2,42	4,40	0,108
Q7	2,33	4,39	0,110
Q8	2,13	3,44	0,061
Q9	1,57	2,65	1,000
Q10	2,34	2,69	0,901
Q11	2,68	2,56	1,000
Q12	1,53	2,75	0,736
Q13	1,96	2,32	1,000

Q14	1,88	3,94	0,869
Q15	1,83	1,83	1,000
Q16	1,61	3,19	0,159
Q17	1,83	2,23	1,000
Q18	1,75	3,00	0,321
Q19	1,19	1,55	1,000

3.3. ANOVA Result

After getting the presumptions data, ANOVA is used to test the significance between demographic factors and the non-ridership presumptions. The result is shown in Table 5.

Table 5: ANOVA Result

Code	Significance Factor	P-value
Q1	-	-
Q2	-	-
Q3	-	-
Q4	Gender	0.035
Q5	-	-
Q6	-	-
Q7	-	-
Q8	-	-
Q9	Residence	0.035
Q10	Residence	0.001
Q11	Residence	0.008
Q12	-	-
Q13	Age Ratio	0.023
Q14	Residence	0.006
Q15	Residence	0.018
Q16	Residence	0.000
Q17	-	-
Q18	-	-
Q19	Age Ratio	0.000

According Table 5, gender is significant factor for accommodation ridership destination (Q4). In this case, gender correlates with public transportation. Different gender has different travel pattern in term of using public transportation especially, for non-ridership women. Sometimes, women use more frequently use public transportation than men, so they think that the BRT can transport them to their destination.

Age ratio is a significant factor for reducing traffic congestion (Q13) and implementing e-payment for the BRT (Q19). At traffic congestion, the member of age ratio has different experience toward the BRT service. Before implementing the BRT, some of them experienced that the road condition is crowded, then they hope that after implementing the BRT, the congestion will be reduced. Furthermore, they also hope for the person who drive a private car or ride a motorcycle will switch to use the BRT as the main modes of transportation. Moreover, at e-payment system, more than 35 years old age ratio has different choice. They prefer to use cash rather than e-payment or using an e-card that is implemented recently. Regular payment is more interesting because they do not need to install any application in mobile phones or bringing e-card for tapping. The rest is age ratio 15-25 and 26-35 are prefer to use a trended application such as e-money or e-payment because most of them have mobile phones or e-money cards.

In the context of living areas for non-ridership, it has significant factor for many queries. Residence correlates with the punctuality of the BRT schedule (Q9), increasing economic productivity for local (Q10), economic opportunity for local (Q11), improving local mobility (Q14), reducing traffic accident (Q15), and reducing fossil fuel use (Q16).

The different city has different presumption. Related to the punctuality of the BRT schedule, non-ridership who stay at Purbalingga has different presumptions with Purwokerto and the other cities. It means that the schedule is not always precise. According to the BRT service, the bus is mobilized every 15 minutes for regular hours and every 10 minutes for peak hours. The BRT does not have an exact schedule because of some limitations such as the quantity of the bus, road condition, traveling time, and break time. In some cases, the BRT will not pick up the ridership because of full capacity, so the ridership shall wait next BRT. The next presumption is enhancing local economic productivity. Purbalingga and the other cities have different presumptions with Purwokerto. The non-ridership in Purbalingga and the other cities think that BRT will increase their productivity because the places have many factories such as artificial eye flashes, wig, and motorcycle exhaust. Many of them are labor at the industries that access to industrial park. On the other hand, Purwokerto has fewer factories than Purbalingga and the other cities, yet it has more offices and banks. So, most of the people at Purwokerto work as private or public employees rather than labor. Moreover, the non-ridership hopes that there will be opened new economic access that connects government office, industrial park, bus terminal, railway station, and shopping center.

Next presumption is improving job opportunity for local people. The BRT is seen as a new job vacation even though there is a limited job opportunity in the BRT. The non-ridership hope by implementing the BRT, it will involve local participation and support. The non-ridership expects there will be additional linkage for local to access to gain a new job. Furthermore, the idea is similar to improving non-ridership mobility. By increasing local mobility, the non-ridership will have any public transportation a choice. They can take advantage of the BRT existence or not. At least, they have choice for public transportation that regular and low fare, but if their destination cannot be facilitated, they will choose other public transportation or private vehicle. In terms of opportunity for mobilizing people, among the cities has different presumptions. The non-ridership in Purbalingga and outside both cities will rely on the BRT for fulfilling their mobility Purwokerto is not too reliant on it.

The next presumption that has significant factor is reducing traffic accidents. The BRT is set for being alternative mass transportation. The logic is if many people use the BRT, the traffic is reduced because the private vehicle is few. The non-ridership in Purbalingga think that the BRT will reduce the traffic accident because many labours rely on the BRT. The BRT is more safety than using private vehicles but, the rest living area said that the BRT are not too rely on because the traffic is still lower than Purbalingga. So, in this case, private car is more suitable for going

to work or back home. The last presumption is the BRT is reduced fossil fuel consumption. This is one of a promotion for being sustainable transportation. The case is similar to reducing traffic accidents. Non-ridership thinks by reducing private vehicle the fossil fuel consumption will be low. Practically, the BRT is still not covered all destinations especially in some public area such as the central traditional market and railway station. So, it needs to develop more routes to reach the uncovered area, so that additional ridership can be facilitated. The non-ridership presumptions in Purbalingga is more agree than non-ridership who are living in the outside both cities and Purwokerto.

3.4. Convergent Validity

After setting up the hypothesis, we develop a model for analysing the correlation among environment, social, and economic criteria with excellent public transportation criteria using SEM- PLS. The model is constructed using Smart PLS 3 software (Figure 2).

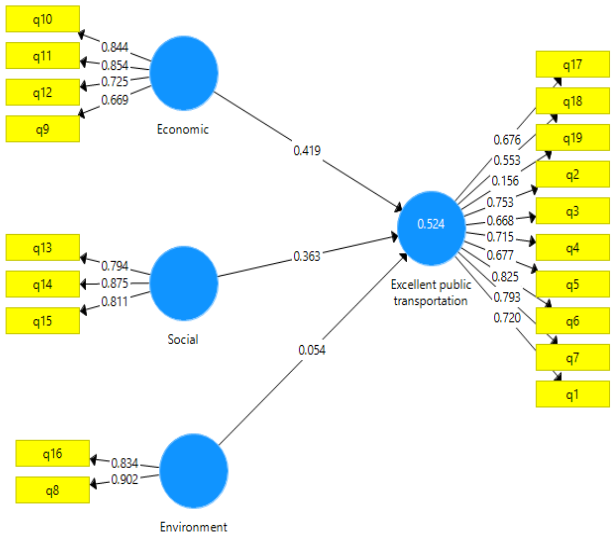


Figure 2: Initial Constructed Model

After the model is constructed, we run the model and get the convergent validity result. Convergent validity is measured based on the correlation among the estimated score component. The value is high if the weight is more than 0.7. In some case for early-stage research, the measurement scale between 0.5 to 0.6 is reliable enough [34]. The result can be seen in Table 6.

Table 6: Initial Outer Loading (Cross-Loading Factor) Result

Code	Economic	Environment	Excellent Public Transportation	Social
Q10	0.844			
Q11	0.854			
Q12	0.725			
Q13				0.794
Q14				0.875
Q15				0.811
Q16		0.834		
Q17			0.676	
Q18			0.553	
Q19			0.156	
Q2			0.753	
Q3			0.668	
Q4			0.715	
Q5			0.677	
Q6			0.825	
Q7			0.793	
Q8		0.902		
Q9	0.669			
Q1			0.720	

According to the theory, the limit of the cross-loading factor is 0.5, and we remove the factor that less than 0.5. it is found that Q19 has a cross-loading factor less than 0.5, or the value is 0.159. So, we eliminate from the constructed model and run the software again. The result is all of the loading factor values are higher than 0.5. It means that fulfilled required convergent validity. Thus, all indicators from the model are appropriate to depict each measured variable (Table 7).

Table 7: Final Outer Loading (Cross Loading Factor) Result

Code	Economic	Environment	Excellent Public Transportation	Social
Q10	0.844			
Q11	0.855			
Q12	0.724			
Q13				0.792
Q14				0.876
Q15				0.812
Q16		0.833		
Q17			0.676	
Q18			0.542	
Q2			0.757	
Q3			0.668	
Q4			0.717	
Q5			0.674	
Q6			0.827	
Q7			0.800	
Q8		0.903		
Q9	0.667			
Q1			0.724	

3.5. Average Variance Extracted (AVE)

The higher AVE value means, the better of correlation among the constructed model. The minimum required values of AVE are 0.5. According to Table 8, all of the variables have AVE values higher than 0.5. So, it can be mentioned that all of the variables are suitable to fulfil minimum requirement.

Table 8: Average Variance Extracted (AVE) Result

Variable	Cronbach's Alpha	rho A	Composite Reliability	Average Variance Extracted (AVE)	R-square
Economic	0.777	0.794	0.858	0.603	
Environment	0.679	0.708	0.860	0.754	
Excellent Public Transportation	0.877	0.884	0.902	0.510	0. 518
Social	0.772	0.801	0.867	0.684	

3.6. Discriminant Validity

Discriminant validity measured how far the differences of a construct with other constructs. High value of discriminant validity can be determined as a special or unique variable and described the measured phenomenon.

According to Table 9 and Table 10, we compare the latent variable covariance with the square root of AVE. It is depicted that all of square root of AVE value are higher than latent variable covariances. So, discriminant validity criteria are fulfilled for all variables.

Table 9: Latent Variable Covariances

Variable	Economic	Environment	Excellent Public Transportation	Social
Economic	**1.000**	0.569	0.640	0.535
Environment	0.569	**1.000**	0.511	0.606
Excellent Public Transportation	0.640	0.511	**1.000**	0.618
Social	0.535	0.606	0.618	**1.000**

Table 10: Square Root of AVE

Variable	**AVE**	**√ (AVE)**	**Discriminant Validity**
Economic	0.603	0.776531	Fulfilled
Environment	0.754	0.868332	Fulfilled
Excellent Public Transportation	0.51	0.714143	Fulfilled
Social	0.684	0.827043	Fulfilled

3.7. Composite Reliability

Reliability test is done by calculating composite reliability. The minimum value of the composite reliability is 0.6. according to Table 8, all of the construct indicator has higher value than 0.6, so it can be defined that the model has high reliability and reliable. Furthermore, Cronbach's Alpha is set an additional parameter. The minimum of Cronbach's Alpha values in the variable to be reliable is 0.7. According to Table 8, all of Cronbach's Alpha values are higher than 0.7, except the environment variable (0.679). So, the reliability of three variables (economic, social, and excellent public transportation) is high, but the environment variable has low reliability.

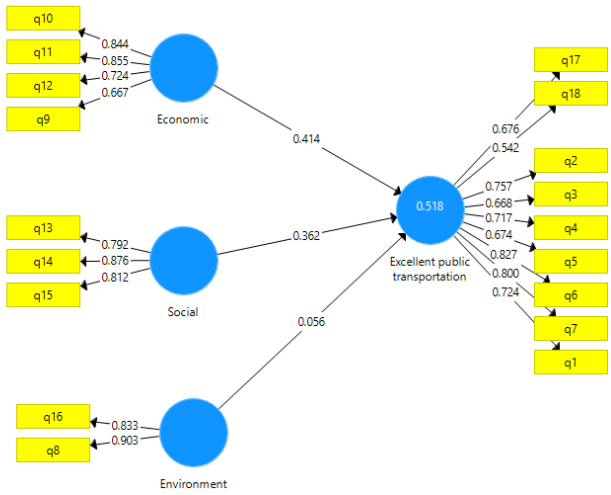


Figure 3: Final Structured Model

3.8. Structural Model Analysis

The subsequent analysis is a structural model. The analysis is conducted to measure the correlation among the constructs or variables, significance, R-squared. According to Figure 3, it is depicted that the number between the yellow rectangle and the blue circle (on the arrow) is cross loading factors value or outer loading value of the model. The description of the value can be referred to Table 7. Furthermore, the values between the blue circle of economic, social, and environment variable and the blue circle of excellent public transportation (on the arrow) is the Original Sample (O) value. Those values can be seen in Table 11. The number on the blue circle of excellent public transportation is R-squared value. The value is 0.518, it means that the expectation of contribution toward BRT execution is at moderate level. The rest is affected by other factors from the outside that are not described in this study.

Table 11: Coefficient Path Structured Model

Variable	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (|O/STDEV|)	P-Values
Economic -> Excellent Public Transportation	0.414	0.422	0.072	5.745	0.000
Environment -> Excellent Public Transportation	0.056	0.056	0.102	0.549	0.583
Social -> Excellent Public Transportation	0.362	0.364	0.078	4.672	0.000

The differences of Figure 2 (initial model) and Figure 3 (final model) is at removing the Q19 indicator. The removed Q19 is caused by the loading factor in convergent validity measurement is less than 0.5. The effect is the final model is more valid than the

initial model, but the R-squared of the initial model (0.524) is a little bit higher than the final model (0.518). Furthermore, the Original Sample (O) value of the initial model is not entirely different from Original Sample (O) value of the final model.

After finishing the model structural analysis, it is processed the bootstrapping. The result can be seen in Table 11. According to the table, economic variables have t-statistics at 5.745, or greater than 1.96 (alfa is 5%) or p-value is less than 0.05 (0.000). It means that the economic variable has a significant factor with excellent public transportation variable. Furthermore, the social variable has t-statistics at 4.672, and the p-value is 0.000. It means that the social variable has a significant factor toward excellent public transportation variable. Meanwhile, the environment variable has t-statistics less than 1.96 (0.549) and p-value is at 0.583 (greater than 0.05). It means that environment variable has no significant factor for excellent public transportation variable. So, only the economic and the social variables which have a significant factor for excellent public transportation variable.

Regarding the Original Sample (O) value, the correlation among those variables shows that the economic and the social variables has a significant correlation with excellent public transportation variables. The value is 0.414 and 0.362. However, for environment variable does not have a significant correlation with excellent public transportation variables and the value is 0.056. So, the economic variable affects the excellent public transportation variable is at 41.4%, and the social variable is at 36.2%. Nevertheless, the environment variable does not affect to excellent public transportation variables.

3.9. Implication and Recommendation

The BRT is new and has been followed the service procedure to fulfil ridership service. The objective of the BRT is to connect ridership and local community or non-ridership. Ridership means people who use the BRT frequently. Non-ridership means the people whether they use the BRT or not, but they get the impact of the BRT execution. In this case, they are defined as marginal community. The role of them is important to support the BRT sustainability and enhance the public engagement.

The demographics result depicts that most non-ridership is non-formal occupations such as street food sellers and pedicab drivers. Furthermore, most of them are never tried the BRT. It is indicated that the BRT implementation or execution belongs to ridership in terms of the service. Indeed, the BRT is set for ridership, but non-ridership shall get the impact at least increasing the opportunity to access, and gain the new transportation mode, economic opportunity, and positive social impact.

According to the survey result of the correlation between demographics with presumptions, living areas of the non-ridership are more frequent, giving significant factors, followed age ratio and gender. The different living areas of non-ridership leads to different presumptions about the BRT schedule, economic productivity, economic opportunity, local mobility, reducing traffic accidents, and low fossil fuel use. Most of the non-ridership in Purbalingga is labour. They rely on public transportation that supported their mobility. So, they need less waiting time at the BRT stop. The BRT provider shall adjust the condition by adding more frequent bus schedules in Purbalingga at peak hours [35].

Economic productivity and opportunity factors depict that Purbalingga has lower economic status than Purwokerto. The economic status is the ability of the non-ridership to access the industry or economic centre [30]. Actually, Purbalingga has more factories, but Purwokerto is more developed than Purbalingga, such as ease of public transportation access and strategic geographical area. So, the BRT shall open new opportunities for non-ridership to open their economic access and develop Purbalingga to be similar economic status with Purwokerto. It can be done by promoting the tourism object, traditional food, museum, and some interesting places, so that increase non-ridership economic opportunities.

Reducing traffic accidents and using fossil fuel less, the BRT can be a pioneer of public transportation that excellent for safety driving and green transportation. It will lead the use of excellent vehicles standard to support both cities to be more eco-friendly mobility and increase local or non-readership healthy.

Meanwhile, the age ratio factor influences different presumption among the non-ridership in terms of traffic congestion and e-payment. In term of traffic congestion, a different age ratio has different background. So, some of non-ridership who use private vehicle trust that the BRT can reduce traffic congestion, especially at peak hours. Furthermore, some non-ridership agree that the BRT shall provide an e-payment system. Actually, the payment system will ease the provider to record the ridership occupancy. So, it can be reliable transport and reach all society segmentation in terms of preparing a new model of the payment system.

The last is the gender that has significant presumption in term of accommodating their destination. Recently, the BRT facilities shall be improved, mainly to attract women. The non-ridership women hope that the BRT can accommodate their destination with appropriate and comfort facilities [36]. So, it will reduce non-ridership women using a private vehicle [37].

Referring to the structured model from SEM-PLS, it is indicated that economic and social variable have significant factors to excellent public transportation variable. It means that the non-ridership is strongly agree that the social and economic variables must be applied with BRT execution. The social and economic variables are perceived by non-ridership, whether direct or indirect effect such as enhance social mobility, reduce traffic accidents and congestion, and opening economic opportunity and access. According to the phenomenon, the BRT provider shall capture the opportunity by enhancing the service onboard and offboard. Onboard service is for ridership and off-board service is non-ridership. Onboard service is universal and the procedure is standardized but off-board service can be set as a sustainable society development program. Environment variable does not have a significant correlation with the excellent public transportation variables. It is indicated that the availability of the BRT is far from the green transportation campaign. Indeed, the primary purpose of the BRT is minimizing traffic congestion and promoting sustainable transportation with low CO2 emissions [38]. However, by implementing green public transportation, the BRT provider shall invest more and sometimes it will not be required recently, especially in developing countries.

According to the structural model from SEM-PLS for excellent public transportation variables, the expectation of contribution toward the BRT execution is 51.8%. It is moderate level and still far from the expectation. However, it needs more empirical study to improve the percentage, so that the phenomenon can be captured wider.

The non-ridership as marginal community is taken to offer significant improvement service by the BRT provider. The BRT provider can improve the journey time, journey time reliability, and realistic alternative mass transportation [29]. By knowing the presumption of the non-ridership, the BRT provider can purpose some program to enhance the BRT occupancy and improve the service. Meanwhile, the social, economic, and environment benefits can be gained to attract more public engagement among the programs. The program can be designed as facilitating economic opportunity program, social development program, and promoting green transportation program.

4. Conclusions

BRT in Purwokerto and Purbalingga is new public transportation that is operated by the Ministry of Transport. As the new public transportation, the support from the non-ridership is a compulsory to maintain the sustainability. According to the analysis, it gives the insightful results;

1) The living area has dominance influenced factors, followed age ratio and gender about the BRT condition. The living area covers the presumption about the BRT schedule, economic opportunity and productivity, local mobility access, reducing traffic accident, and fossil fuel use. For the age ratio, it covers traffic congestion and e-payment system, and the gender covers the destination presumption.
2) The SEM-PLS model shows that social and economic criteria have correlation with excellent public transportation criteria, but not with environment criteria. So, it can be a suggestion to the BRT provider to concern more environmental issue to raise the community awareness and the public engagement.

In the BRT provider side, the suggestion will lead to purpose some programs to improve the service and increase the occupancy. In the non-ridership side, the proposed program will attract the sense of awareness and the public engagement from the BRT execution. Furthermore, the model is demonstrated at the moderate percentage, so the future research in this area can be directed to add more variables to reach excellent public transportation. The added variables such as specific involved communities, political will, regulation, gender equality, economic opportunity, and local culture to capture more public engagement.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

This acknowledgement is given to Ministry of Education and Culture Republic of Indonesia for the funding. We thank so much to the BRT ridership who gave free time to fill our questionnaire and also surveyor team from Industrial Engineering and Information System Department, IT Telkom Purwokerto, Central Java, Indonesia.

References

[1] O. Afolabi, "Commuters Perception and Preferences on The Bus Rapid Transit in Lagos State," JORIND, **14**(2), 2016.
[2] J. Cao, X. Cao, C. Zhang, X. Huang, "The gaps in satisfaction with transit services among BRT, metro, and bus riders: Evidence from Guangzhou," Journal of Transport and Land Use, 97–109, 2015, doi:10.5198/jtlu.2015.592.
[3] D. Darwis, Fatmawati, Nasrulhaq, "Hubungan Kualitas Pelayanan dengan Kepuasan Masyarakat Pengguna Bus Rapid Transit (BRT) Maminasata," Administrasi Publik, **3**, 2017.
[4] J.M. Bunker, "High volume bus stop upstream average waiting time for working capacity and quality of service," Public Transport, **10**(2), 311–333, 2018, doi:10.1007/s12469-018-0179-1.
[5] M. Islam, M. Brussel, A. Grigolon, T. Munshi, "Ridership and the Built-Form Indicators: A Study from Ahmedabad Janmarg Bus Rapid Transit System (BRTS)," Urban Science, **2**(4), 95, 2018, doi:10.3390/urbansci2040095.
[6] D. Oviedo, L. Scholl, M. Innao, L. Pedraza, "Do Bus Rapid Transit Systems improve accessibility to job opportunities for the poor? The case of Lima, Peru," Sustainability (Switzerland), **11**(10), 2019, doi:10.3390/su11102795.
[7] D.A. Rodriguez, C.E. Vergel-Tovar, "Urban development around bus rapid transit stops in seven cities in Latin-America," Journal of Urbanism, **11**(2), 175–201, 2018, doi:10.1080/17549175.2017.1372507.
[8] E. Saputra, C. Widyasmara, "The effect of urban spatial structure on rapid bus transit services in Yogyakarta and Surakarta, Indonesia: A comparative study of the Trans Jogja and the Batik Solo Trans," Geografia - Malaysian Journal of Society and Space, **10**(2), 46–58, 2014.
[9] S. Tao, J. Corcoran, I. Mateo-Babiano, D. Rohde, "Exploring Bus Rapid Transit passenger travel behaviour using big data," Applied Geography, **53**, 90–104, 2014, doi:10.1016/j.apgeog.2014.06.008.
[10] L. Prayogi, "The Influence of Bus Rapid Transit System on Urban Development : An Inquiry to Boston and Seoul BRT Systems," International Journal of Built Environment and Scientic Research, **1**(1), 1–8, 2017.
[11] A. Koling, W.-B. Zhang, K. Zhou, H. Meng, Bus Rapid Transit (BRT) Toolbox : Assessing Person Throughput to Measure Transportation Impacts for BRT Projects, 2018.
[12] S. Tabassum, S. Tanaka, F. Nakamura, "Improving Access considering Commuters ' Perception (A Case Study of Lahore BRT) Keywords :," International Journal of Innovative Research in Engineering and Management (IJIREM), **3**(4), 283–289, 2016.
[13] M. Woldeamanuel, C. Olwert, "The Multimodal Connectivity at Bus Rapid Transit (BRT) Stations and the Impact on Ridership," Journal of the Transportation Research Forum, **55**(1), 87–102, 2016, doi:10.5399/osu/jtrf.55.1.4340.
[14] B. Ferhan, X. Zhou, "Drivers , Barriers and Implementation of Social Sustainable in Supply Chain A qualitative study of SMEs," 1–83, 2016.
[15] S. Nasrin, "Acceptability of Bus Rapid Transit (BRT) to Commuters in Dhaka," 2015.
[16] H.I. Okagbue, M.O. Adamu, S.A. Iyase, E.A. Owoloko, On the Motivations and Challanges Faced by Commuters Using Bus Rapid Transit in Lagos, Nigeria, The Social Sciences, **10**(6), 696–701, 2015, doi:10.3923/sscience.2015.696.701.
[17] J. Dubé, F. Des Rosiers, M. Thériault, P. Dib, "Economic impact of a supply change in mass transit in urban areas: A Canadian example," Transportation Research Part A: Policy and Practice, **45**(1), 46–62, 2011, doi:10.1016/j.tra.2010.09.002.
[18] J.B. Ingvardson, O.A. Nielsen, "Effects of new bus and rail rapid transit systems–an international review," Transport Reviews, **38**(1), 96–116, 2018, doi:10.1080/01441647.2017.1301594.
[19] E. Adewumi, D. Allopi, "An Appropriate Bus Rapid Transit System," International Journal of Science and Technology Volume 3 No. 4, **3**(4), 248–254, 2014.
[20] A.P. Ajayi, "Assessment of the Influence of Lagos Bus Rapid Transit Scheme (BRT-Lite) on Road Traffic Crashes (RTC) on Lagos Mainland-Island Corridor," The Open Transportation Journal, **11**(1), 102–109, 2017, doi:10.2174/1874447801711010102.
[21] A. Nikitas, M. Karlsson, "A Worldwide State-of-the-Art Analysis for Bus Rapid Transit: Looking for the Success Formula," Journal of Public Transportation, **18**(1), 1–33, 2015, doi:10.5038/2375-0901.18.1.3.
[22] A. Mishra, S.A.S.A. Kumar, P. Pradeep, "Study of Bus Rapid Transit system In Respect to Growing Cities of India," Nternational Journal of Engineering Research & Technology (IJERT), **2**(10), 1378–1383, 2013.
[23] M. Poku-Boansi, G. Marsden, "Bus rapid transit systems as a governance reform project," Journal of Transport Geography, **70**(June), 193–202, 2018, doi:10.1016/j.jtrangeo.2018.06.005.
[24] S.R. Majumdar, "The case of public involvement in transportation planning using social media," Case Studies on Transport Policy, **5**(1), 121–133, 2017, doi:10.1016/j.cstp.2016.11.002.
[25] E. Cascetta, F. Pagliara, "Public Engagement for Planning and Designing Transportation Systems," Procedia - Social and Behavioral Sciences, **87**,

103–116, 2013, doi:10.1016/j.sbspro.2013.10.597.
[26] K. Rangarajan, S. Long, A. Tobias, M. Keister, "The role of stakeholder engagement in the development of sustainable rail infrastructure systems," Research in Transportation Business and Management, **7**, 106–113, 2013, doi:10.1016/j.rtbm.2013.03.007.
[27] N. Hassen, P. Kaufman, "Examining the role of urban street design in enhancing community engagement: A literature review," Health and Place, **41**, 119–132, 2016, doi:10.1016/j.healthplace.2016.08.005.
[28] J.R. Gardner, R. Rachlin, H.W.A. Sweeney, A. Sweeny, H.W. Allen, Handbook of Strategic Planning, Wiley, 1986.
[29] D. Mobereola, Lagos Bus Rapid Transit Africa ' s first BRT scheme (The Lagos BRT-Lite System), 2009.
[30] U. Dutta, "A Framework to Identify Economic and Community Benefits of the Proposed BRT System in Southeast Michigan," Current Urban Studies, **05**(02), 164–184, 2017, doi:10.4236/cus.2017.52010.
[31] N. Duduta, C. Adriazola, D. Hidalgo, L.A. Lindau, R. Jaffe, "Understanding road safety impact of high-performance bus rapid transit and busway design features," Transportation Research Record, **2317**, 8–14, 2012, doi:10.3141/2317-02.
[32] V.J. Racehorse, G. Zhang, A. Sussman, A. Jian, T. Parker, "Bus rapid transit system deployment for high quality and cost-effective transit service: A comprehensive review and comparative analysis," IET Intelligent Transport Systems, **9**(2), 175–183, 2015, doi:10.1049/iet-its.2013.0176.
[33] D. Prabowo, BRT Purwokerto Terus Tambah Halte, Radar Banyumas, 2019.
[34] A. Rifai, "Partial Least Square-Structural Equation Modeling (PLS-SEM) untuk mengukur ekspektasi penggunaan repositori lembaga: Pilot studi di UIN Syarif Hidayatullah Jakarta," Al-Maktabah, **14**(1), 56–65, 2015.
[35] P. Starkey, J. Hine, Poverty and Sustainable Transport How Transport Affects Poor People with Policy Implications for Poverty Reduction: A literature review, 2014.
[36] D. Martinez, E. Salgado, P. Yañez-pagans, Connecting to Economic Opportunity ? The Role of Public Transport in Promoting Women ' s Employment in Lima, 2018.
[37] K. Terayama, M. Odani, "Expected role of public transportation services in securing residents ' accessibility to the city center in suburban housing development areas," Transportation Research Procedia, **25**, 4258–4269, 2017, doi:10.1016/j.trpro.2017.05.237.
[38] L. Velazquez, N.E. Munguia, M. Will, A.G. Zavala, S.P. Verdugo, "Sustainable transportation strategies for decoupling road vehicle transport and carbon dioxide emissions," **26**(3), 373–388, 2015, doi:10.1108/MEQ-07-2014-0120.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 805-815 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Mentoring Model in an Active Learning Culture for Undergraduate Projects

Wongpanya Nuankaew[1], Kanakarn Phanniphong[2], Sittichai Bussaman[3], Direk Teeraputon[4], and Pratya Nuankaew[5,*]

[1]*Faculty of Information Technology, Rajabhat Mahasarakham University, Maha Sarakham, 44000, Thailand*

[2]*Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-Ok, Chonburi, 20110, Thailand*

[3]*Faculty of Science and Technology, Rajabhat Mahasarakham University, Maha Sarakham, 44000, Thailand*

[4]*School of Education, University of Phayao, Phayao, 56000, Thailand*

[5]*School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand*

ARTICLE INFO

Article history:
Received: 17 June, 2020
Accepted: 17 August, 2020
Online: 28 August, 2020

Keywords:
Active Learning Culture
Educational Data Mining
Engineering Education
Learning Analytics
Lifelong Learning
Mentoring Model

ABSTRACT

Senior projects allow students to move the learning process from basic knowledge to an interdisciplinary approach. The purpose of this research is (1) to analysis attitude and perception, which is a collaboration between teachers and students to develop a model for clustering of appropriate advisors and advisee who cooperate in senior project, and (2) to develop factors that are significant to predict the right match in senior projects course. Data collection was a questionnaire consisting of 463 samples from 7 administrators, 68 lecturers, 26 staff and 362 students from two institutions: The Rajabhat Mahasarakham University, and the University of Phayao. The research methodology was designed and divided into three sections: preparation, implementation, and conclusion. The result shows that the satisfaction and the overall acceptance level were at a high level (mean = 4.04, S.D. = 0.88). Moreover, the developed model has the highest level of efficiency (accuracy = 98.06%). While the factors that are important for matching recommendations consists of 9 factors: policies of the organization, vision of the organization, mission of the organization, experience and achievements of researchers, qualifications of research team, interest in the research topics, impressions and examples in the past, technology and laboratory support, and budget support. For the future, the researchers are aimed to research on the development of students' academic achievement and aims to promote a learning culture based on the results of this research.

1. Introduction

In the world of learning and educational theories, developing learners to understand research problems and constructing the concepts to create research problems is most difficult and complicated in managing. Learners often have different learning styles, which are classified according to learner characteristics [1–4]. At the same time teachers have skills, abilities, experience, knowledge, expertise, and aptitude in different contexts [1]. Thus, the response that occurs between the student and the teacher is important. It can also be used to develop a structured relationship for collaboration between students and teachers through feedback and perception of both parties [3,4].

The problem is related to the context of the curriculum that has been instructed in the field of information technology and other disciplines, which is concerned with the development of students' knowledge by using senior projects as a knowledge base learning [5]. However, a popular and applied method for developing an effective educational process is proactive learning [6] and proactive learning culture [7].

*Corresponding Author: Pratya Nuankaew, School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand, +66 89 961 4832 & pratya.nu@up.ac.th

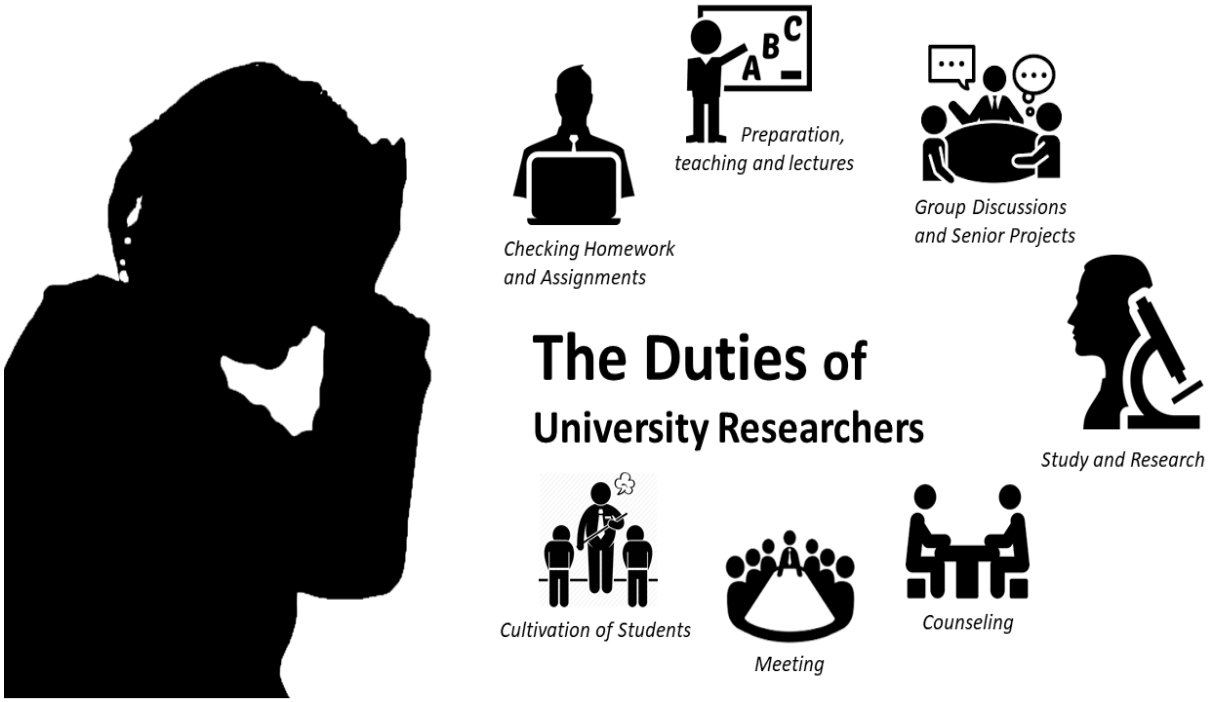


Figure 1: The Duties of University Researchers

Proactive learning or active learning is an essential feature of a person in a learning society that is aware of and responds to many events and stimuli. Having a proactive feature will allow the person to behave appropriately and produce good results. In the active learning perspective, learners play a role in acquiring knowledge and learning in an interactive way until knowledge and understanding are applied, able to analyze, synthesize, evaluate or create things and develop themselves to their full potential.

As well as organizing a learning experience to participate in discussions to practice communication skills, resulting in a high learning result [8-13]. Key aspects of learning management of active learning concepts that emphasize the role and participation of the learners. Knowledge arises from experience, building knowledge, and reviewing learners. In addition, it is a learning management aimed at developing brain potential. Learning management that allows learners to participate in learning The process of creating situations for learners and emphasizing advanced thinking skills [8-10]. It can be concluded that active learning is not only for learner development but also being used for industry development [11-12].

In addition, to creating a learning culture that promotes critical and analytical thinking [8] there needs to be a concrete order for the learner to understand the process of the problem and the solution. According to the intentions of the curriculum for modern education [1,2,9–12], it changed the role of instructors and learners to become colleagues, which is a matter of knowing and cooperating. The instructor is not responsible for educating the learner anymore. The instructor acts as a coach that guides players to play only.

At the same time, the students need to solve a specific problem in the educational problems. It seems that students will need to find solutions to educational problems in their quest to understand their own thinking processes and knowledge.

Moreover, the vast majority of researchers at the university need to perform teaching tasks, which spend most of their time for preparing in learning management and summarizing students' tasks as demonstrated in Figure 1. Teachers also need more time to evaluate the learner outcomes. This corresponds to the problem on students who wish to perform well with their senior project.

Figure 1 presents an overview of the duties and responsibilities of most researchers at universities in Thailand. It consists of teaching, conducting research, student counseling, planning classroom management, meetings, promoting students' intrinsic motivation, and so on. It seems that teachers have little time to support student projects. Whereas the collaboration of teachers and students used in senior project development consists of five phases as shown in Figure 2.

Figure 2 shows the senior project development phase which consists of five phases. Phase 1: The process of discovering senior projects based on the student's interests, relevant to teacher research and teacher interest, and the appropriate competence of the faculty member in the organization. Phase 2: The process of transferring academic results, knowledge, interests, sharing, planning, group discussion, and senior project framework development. Phase 3: Development phase, which will be in a long period of time. Phase 4: Project effectiveness testing and evaluation phase, which will test the project for improvement. Phase 5: The publishing process, which presents knowledge and new discoveries obtained from the project. All five phases need to be completed in two semesters or approximately eight months.

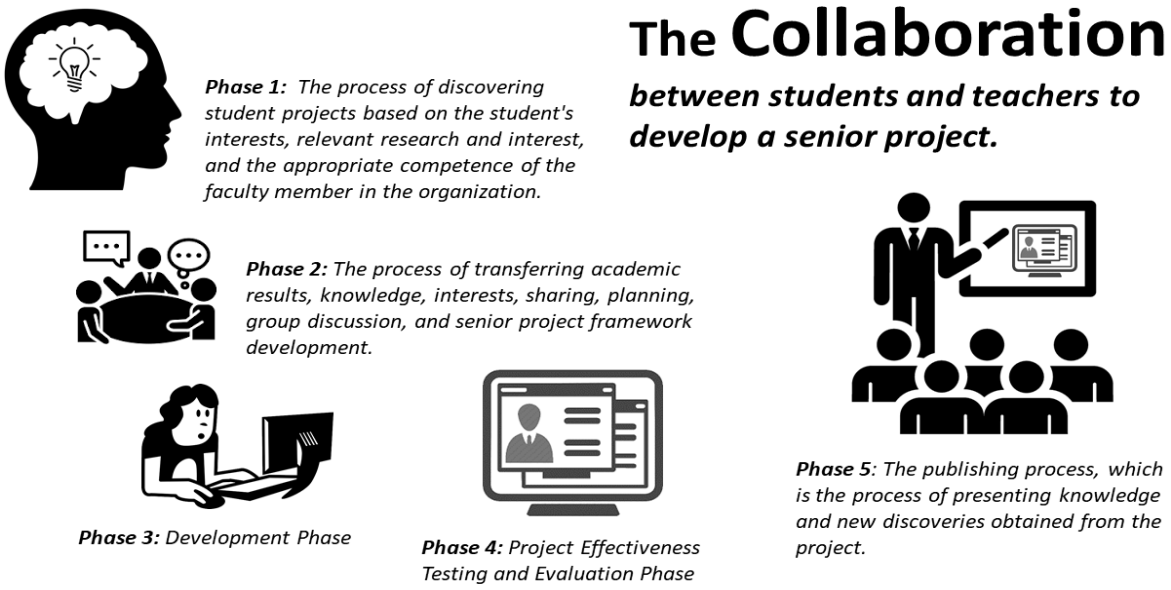


Figure 2: Senior Project Development Process

Considering the abundant and heavy workload of the teacher and the limitations for the development of student projects, adopting proactive learning by creating an appropriate learning culture is important. More than that, the reason for the researchers who emphasized the model development and data analysis in this study was because the research team assembled teams from many institutions, but encountered the same problem. Our common problem is to recommend suitable teachers to students in senior projects. It is important, but difficult. Therefore, this research focused on solving problems proactively using proactive learning strategies to create opportunities for corrective action and reduce impropriety. In addition, the professional group of the researchers is in the information technology field, as researchers try to use technology to help solve the problems. The purpose of this research is (1) to analyze the attitude and perception, which is a collaboration between teachers and students to develop a model for the clustering of appropriate advisors and advisee who cooperate in the senior project, and (2) to develop factors that are significant to predict the right match in senior projects course. Moreover, the major goal of the research team is to develop a student proactive learning culture in educational institutions, which is planned for future research.

The research outline consists of five parts: a summary of the relevant research in Section 2, data collection and research methodology is presented in Section 3, research results and discussion reports are provided in Section 4. Finally, the conclusion is summarized in Section 5, respectively.

2. Literature Reviews

2.1. Cooperative Learning

Cooperative learning model is one of the learning style that aims to communicate and collaborate with both parties on academic achievements. In addtion, cooperative learning is a learning activity technique that allows students to learn together in small groups, each made up of members with different abilities. Each person is truly involved in learning [17-19]. In the group's learning and success, where the group exchanges views, shares resources, encourages one another, group members are not only responsible for their own grades but are also responsible for their own performance. Learn of all fellow members of the group, it can be called individual success, group success.

Examples of research work related to the concept of cooperative learning such as eL-PCDA model that management cooperative learning in e-Learning, based on learner's behavioral environment, scheduling, and monitoring of learning activities for cooperative learning [18]. Anothor example is the multi-agent system (MAS) that instead of single-agent learning to handle the performance enhancement of cooperative learning for decision making [18]. The last example is the multi-user streaming feedback system for cooperative learning application whereas teachers can monitor their students and problems in real-time, and the students have more willingness towards learning [13]. Obviously, the demonstration shows that cooperative learning can be a learning achievement, which is important in including goals that both parties emphasize.

2.2. Practical Learning

Practical learning is a way of learning from experience. Because work is more problematic than problem solving, it requires more than opening the textbook. It's important for learners to find a way to solve the problem appropriately be correct and reasonable, as well as being useful and applicable [20-23].

Practical learning is learning through experience, the theories used in learning management are Kolb's experiential leaning theory [1-2]. Kolb defines learning as a process in which knowledge is created through the transformation of experience. Learning can be explained in four main steps: Concrete Experience (CE), Reflection Observation (RO), Abstract Conceptualization

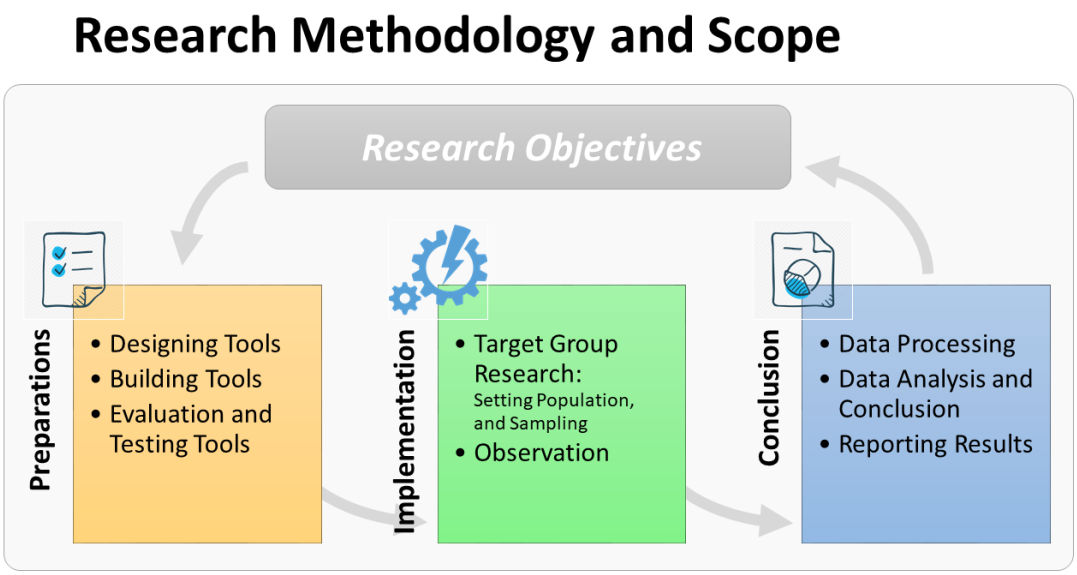


Figure 3: Research Methodology

(AC), and Active Experimentation (AE). Which is evident that learning from practice requires the use of knowledge and experience, which will help develop the learners' attitudes and behavior towards an important goal of learning.

However, practical learning is not just a matter of practice, it must Research Methodology

The research methodology is designed and shown in Figure 3.

2.3. Research Objectives

The purpose of the research is (1) to make an analysis on the attitude and perception, which is a collaboration between teachers and students to develop a model for the clustering of appropriate advisors and advisee who cooperate in the senior project, and (2) to develop factors that are significant to predict the right match in senior projects course.

2.4. Research Preparation

1) Designing Tools

At the stage of designing tools, the researchers conducted a preliminary study of the research process used for finding attitudes and satisfaction, which found that the appropriate and interesting tool was a questionnaire. Researchers examined questionnaires through the researcher's organization according to procedures and standards of the research process. Questionnaires are routinely used to assess teaching and learning. It is conducted on an annual basis. In addition, the questionnaire quality assessment is described in the review stage section.

At the same time, the data used in the analysis and design of the questionnaire were used by interviewers, including teachers who had received research grants, students who presented and published senior projects, and university administrators.

2) Building Tools

In the process of building tools, the questionnaire was designed into 3 parts, including, Part 1: General information of the respondents, Part 2: Satisfaction and Attitude Towards Senior Projects, and Part 3: Suggestions for Senior Projects. While the main topic of part 2 is shown in Table 1.

Table 1: Issues and Essence on Senior Projects

Main Essence	Dimensions Related to Essence: Sub-Stages
1. Organization	1.1 The policies of the organization 1.2 The vision of the organization 1.3 The mission of the organization 1.4 Management and support of the organization
2. Researcher	2.1 Experience and achievements of researchers 2.2 Aptitude and expertise of researchers 2.3 Qualifications of Research Team
3. Students	3.1 Knowledge and ability of learners 3.2 Interest in the research topics 3.3 Impressions and examples in the past
4. Project Support	4.1 Technology and laboratory support 4.2 Staff support 4.3 Budget support

The essence of Part 2 is to find out the level of opinion and attitude on the factor that encourage learners and teachers to collaborate on research, which is comprised of four main areas: organization, researcher, students and project support.

3) Review Stage

After designing and constructing the tool, it was tested for confidence and accuracy by 3 experts from the Department of Information Technology, Faculty of Information Technology, Rajabhat Mahasarakham University, Thailand.

2.5. Evaluation and Testing Tools

The research actions were divided into two phases: 1) Target group research, which describe and define the scope of data collection, 2) Research observation, which shows the process of conducting research and gathering data.

1) Target Group Research

a) Population

Data gathering is done in all dimensions, which was limited to the area from two institutions: The Rajabhat Mahasarakham University, and the University of Phayao. The population was defined by four groups: administrators, lecturer, staff and students.

b) Sampling

Sampling was randomly selected from 463 samples, which had been divided into four groups: 7 administrators, 68 lecturers, 26 staff and 362 students.

2) Research Observation

a) Respond to the objectives

After completing the data collection the data sets were already considered as information which provided covers for the purpose of the research.

b) Meet the research framework

The data set responds to the research based on the research framework and is used to test the hypothesis in its entirety.

c) Proceed with caution

In gathering information, it will only attract the attention of the respondents and collected from the reality to get the information according to the actual situation.

2.6. Research Conclusion

1) Data Processing

At this stage, the instrument used for data analysis was selected, consisting of two parts: standard deviation (S.D.) and mean. Analysis of data was divided into groups and topics has been determined and summarized in the next section.

2) Data Analysis and Conclusion

The sample set from the 463 samples was analyzed and summarized as shown in Table 2 and Table 3.

Table 2: Data collection classified by gender

Gender	Samples (n = 463)				
	Administrators (s_1)	Lecturers (s_2)	Staffs* (s_3)	Students (s_4)	Total
Male	4 (0.86%)	27 (5.83%)	13 (2.81%)	141 (30.45%)	185 (39.96%)
Female	3 (0.65%)	41 (8.86%)	13 (2.81%)	221 (47.73%)	278 (60.04%)
Total:	7 (1.51%)	68 (14.69%)	26 (5.62%)	362 (78.19%)	463 (100%)

* staffs = Practitioners in the organization but not performing teaching duties

Table 2 shows that most of the respondents are female, with 278 people representing 60.04 percent of all data providers. The researchers found that it was consistent with the students at the University of Phayao and Rajabhat Mahasarakham University, which had mostly female students.

Table 3: Data collection classified by age

Age	Samples (n = 463)				
	Administrators (s_1)	Lecturers (s_2)	Staffs (s_3)	Students (s_4)	Total
20 - 30 years old	1 (0.22%)	24 (5.18%)	12 (2.59%)	356 (76.89%)	393 (84.88%)
31 – 40 years old	1 (0.22%)	30 (6.48%)	9 (1.94%)	6 (1.30%)	46 (9.94%)
41 – 50 years old	1 (0.22%)	11 (2.38%)	44 (0.86%)	0 (0.00%)	16 (3.46%)
51 – 60 years old	4 (0.86%)	3 (0.65%)	1 (0.22%)	0 (0.00%)	8 (1.73%)
Total :	7 (1.51%)	68 (14.69%)	26 (5.62%)	362 (78.19%)	463 (100%)

Table 3 shows that the majority of data providers are between the ages of 20-30, with 393 people representing 84.88 percent of all data providers.

The researchers found that most of the respondents were between 20 and 30 years old, as it was the main target group the research focused on. However, other age groups are consistent with the target group, as it is the mentor group who provides consultation to the senior project.

Table 4: Data collection classified by education

Education	Samples (n = 463)				
	Administrators (s_1)	Lecturers (s_2)	Staffs (s_3)	Students (s_4)	Total
Doctorate	6 (1.30%)	9 (1.94%)	0 (0.00%)	0 (0.00%)	15 (3.24%)
Master	1 (0.22%)	55 (11.88%)	14 (3.02%)	6 (1.30%)	76 (16.41%)
Bachelor	0 (0.00%)	4 (0.86%)	12 (2.59%)	356 (76.89%)	372 (80.35%)
Total :	7 (1.51%)	68 (14.69%)	26 (5.62%)	362 (78.19%)	463 (100%)

Table 4 shows that the majority of the data providers have a bachelor's degree, with 372 people representing 80.35 percent of all data providers. The researchers found that most of the respondents had a bachelor's degree because they were students of the primary target group. In addition, respondents with a master's and doctoral degree show the proportion of mentors.

In interpreting the data according to the characterization criteria, the interpretation is based on a five-level interpretation method by comparing it with the criteria that divides the level estimation into five equal levels, as followed in Equation (1). The result of the calculation is shown in Equation (2).

$$Width\ of\ the\ level\ = \frac{Maximum - Minimum}{Number\ of\ levels} \quad (1)$$

$$Width\ of\ the\ level\ = \frac{5-1}{5} = 0.8 \quad (2)$$

From the calculation results in Equation (2), the interpretation results can be specified as shown in Table 5.

Table 5: Interpretation and Meaning

Width of the level	**Interpretation**	**Meaning**
1.00 – 1.80	Highly Unacceptable	The lowest level of satisfaction
1.81 – 2.60	Unacceptable	Low level of satisfaction
2.61 – 3.40	Acceptable	Satisfaction

Width of the level	Interpretation	Meaning
3.41 – 4.20	Highly Acceptable	High level of satisfaction
4.21 – 5.00	Maximum Acceptable	The highest level of satisfaction

Table 6: Satisfaction and acceptance toward the factors

Stages	Satisfaction and Acceptance (n = 463)						
	S1	*S2*	*S3*	*S4*	*Total Mean*	*Total S.D.*	*Interpretation*
Stage 1 : Organization							
Stage 1.1	3.71	2.87	3.04	3.69	3.53	1.12	Highly Acceptable
Stage 1.2	3.86	3.43	3.39	3.98	3.86	0.92	Highly Acceptable
Stage 1.3	3.57	3.57	3.42	3.96	3.87	0.87	Highly Acceptable
Stage 1.4	4.42	4.24	3.89	4.16	4.16	0.82	Highly Acceptable
Average	3.89	3.53	3.44	3.95	3.86	0.93	Highly Acceptable
Stage 2 : Researcher							
Stage 2.1	4.29	3.77	3.65	4.06	4.00	0.82	Highly Acceptable
Stage 2.2	4.00	4.10	4.08	4.17	4.15	0.78	Highly Acceptable
Stage 2.3	4.27	4.04	4.12	4.17	4.15	0.85	Highly Acceptable
Average	4.19	3.97	3.95	4.13	4.10	0.82	Highly Acceptable
Stage 3 : Students							
Stage 3.1	3.86	3.88	3.58	4.03	3.98	0.93	Highly Acceptable
Stage 3.2	3.71	3.50	3.39	3.91	3.82	0.89	Highly Acceptable
Stage 3.3	3.71	3.37	3.35	3.96	3.84	0.89	Highly Acceptable
Average	3.76	3.58	3.44	3.97	3.88	0.90	Highly Acceptable
Stage 4 : Project Support							
Stage 4.1	4.86	4.75	4.58	4.27	4.37	0.93	Highly Acceptable
Stage 4.2	4.71	4.71	4.65	4.30	4.38	0.85	Highly Acceptable
Stage 4.3	4.43	4.71	4.54	4.36	4.42	0.83	Highly Acceptable
Average	4.67	4.72	4.59	4.31	4.39	0.87	Highly Acceptable
Total Average	4.11	3.92	3.82	4.08	4.04	0.88	Highly Acceptable

From the analysis of the overall level of satisfaction and acceptance of factors, it was found at a high level of agreement (mean = 4.04, S.D. = 0.88). Moreover, the level of satisfaction with the project is the highest (mean = 4.39, S.D. = 0.87). Therefore, it can be concluded that the respondents agreed and accepted the teaching and learning as a senior project. From the data obtained from the respondents, the researchers grouped the data by clustering through the use of machine learning tools and data mining techniques, which consisted of three parts:

a) K-Means

The k-means is the most common algorithm used as an iterative refinement technique. It is also called Lloyd's algorithm [14], especially in the computer science community. The k-mean algorithm is performed by switching between two steps: (1) the assignment step, which assigns each observation to the cluster with the closest mean. (2) updates the procedure which calculated the new means to be a centroid of observations in a cluster.

b) Decision Tree

The decision tree is one of the learning methods used in statistics, machine learning, and data mining. It works by determining the data from observations and separating data for use in data consideration and finding predictive results. The benefit is on getting important factors, which are caused by the nodes or the decision-making part of the model [15–16].

c) Model Performance

The goal of the model performance is to assess the results and review the process [15–16]. The tools are used in the research, including cross-validation methods as shown in Figure 4, and confusion matrix as mentioned in Figure 5.

Figure 4 displays the separation of data for evaluating the model. The cross-validation method divides the data into two parts. The first part is used for modeling and the remainder is applied to test the model. In addition, model evaluation requires a tool called a confusion matrix [16] to test the model's performance, with the principles shown in Figure 6.

Figure 5 presents the composition of the confusion matrix performance, which is composed of the actual class and the predicted class. An important benefit of the performance of the confusion matrix is the ability to determine the model's ability to predict results, such as the predictive ability or accuracy, model precision, model sensitivity, and model specificity (recall measurement). These values are used to determine the actual performance model. Moreover, Figure 5 also demonstrates the formulas and methods for calculating the various performance parameters in detail.

3) Reporting Results

From Table 6, the most attention toward the factors is the stage 4.3 budget support, which is the highest level of satisfaction (mean = 4.42, S.D. = 0.83). The second is the stage 4.2 staff support, which has a high level of satisfaction (mean = 4.38, S.D. = 0.85). The lowest is the stage 1.1 organization policies, whereas the level of satisfaction is at the middle level (mean = 3.53, S.D. = 1.12).

Interesting suggestions include that the budget is delayed and insufficient. Researchers are not very interested because they think it is difficult and not worth the money. Researchers who are truly interested in joining are few, most of whom participate because of their duties. There should be incentives to motivate participants and to develop research in the organization.

3. Research Results and Discussion

The research results classified the research report into four topics which are modelling results, model testing results, model applying results, and summarizing the significant factors.

3.1. Modelling Results

Modelling results are the reports of the various models on different criteria, including the depth of the decision tree, and types of cross-validation method tests, where the different results are shown in Table 7.

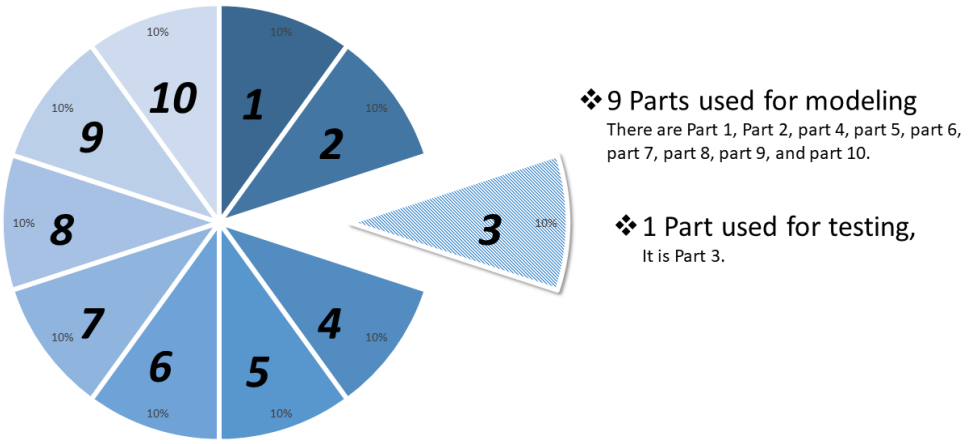


Figure 4: Cross-Validation Methods

Accuracy : $\frac{TP+TN}{TP+TN+FP+FN}$		Actual Class: *Positive*	Actual Class: *Negative*	Precision
Predicted Class	*Positive*	True Positive : TP	False Positive: FP (Type 1 Error)	***Positive Predictive Value :*** $\frac{TP}{TP+FP}$
Predicted Class	*Negative*	False Negative : FN (Type 2 Error)	True Negative: TN	***Negative Predictive Value :*** $\frac{TN}{TN+FN}$
Recall		***Sensitivity :*** $\frac{TP}{TP+FN}$	***Specificity :*** $\frac{TN}{TN+FP}$	

Figure 5: Confusion Matrix Performance

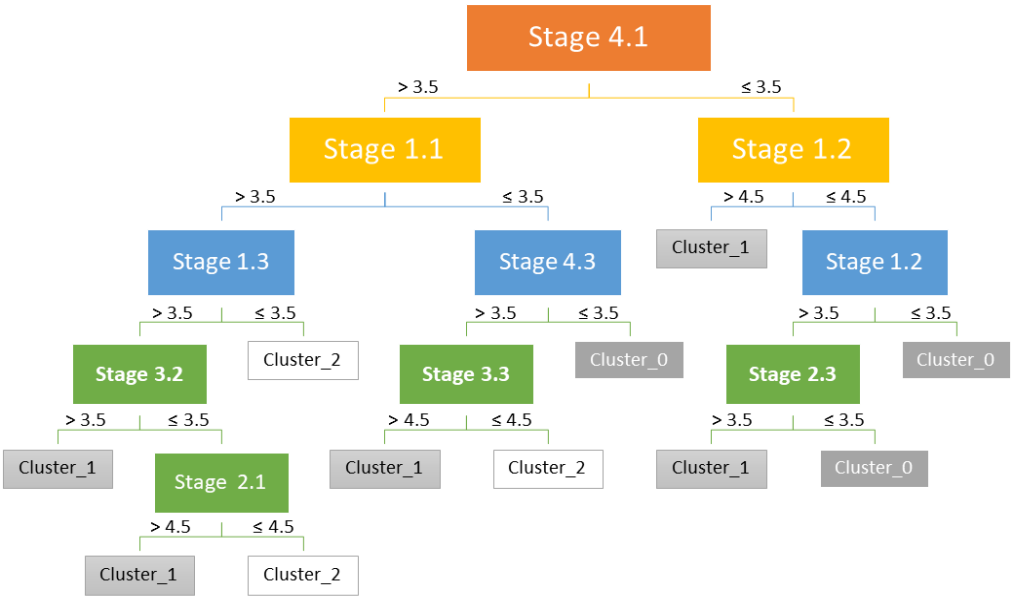


Figure 6: The decision tree model

Table 7: Model results from k-means clustering

Cluster Number	Depth of Decision Tree	Type of Cross-Validation Methods		
		5-Fold	*10-Fold*	*Leave-one-out*
3 Clusters*	Level 3	90.51%	92.21%	92.66%
	Level 4	95.91%	95.88%	96.33%
	Level 5	95.90%	96.53%	96.98%
	Level 6*	96.12%	96.75%	98.06%*
	Level 7	95.90%	96.75%	97.62%
	Level 8	95.68%	96.32%	96.76%
	Level 9	95.25%	96.32%	96.76%
4 Clusters	Level 3	71.30%	76.01%	81.21%
	Level 4	87.25%	86.63%	82.51%
	Level 5	88.33%	89.00%	84.88%
	Level 6	89.62%	89.65%	86.83%
	Level 7	89.41%	90.09%	88.12%
	Level 8*	89.85%	90.52%*	88.12%
	Level 9	90.28%	90.51%	89.20%
5 Clusters	Level 3	61.99%	61.97%	61.12%
	Level 4	77.73%	78.83%	81.64%
	Level 5	87.69%	87.25%	85.96%
	Level 6	87.04%	87.90%	85.53%
	Level 7	87.70%	88.33%	87.26%
	Level 8	88.78%	86.83%	87.26%
	Level 9*	89.21%*	89.21%*	87.69%

Table 7 shows that the k-means model with the highest accuracy is the decision tree model that is classified into 3 clusters by dividing the model testing into the leave-one-out cross-validation method with a depth of 6 levels of the decision tree model and has an accuracy of 98.06%

However, the test results classified by other clusters yield a lower accuracy. For example, the 4 clusters with the highest results are 90.52% and the 5 clusters have the highest results of 89.21%.

3.2. Model Testing Results

From the results of the prototype model development, it can be concluded that the model with the highest accuracy is the development of the model from k-means clustering, with the appropriate number of 3 clusters and there is a leave-one-out cross-validation result with an accuracy of 98.06%. Details of the developed model are shown in Table 8.

Table 8: The model testing results

Accuracy: 98.06% (+/- 13.82%)		Actual Class			Precision Class
		True Cluster_1	*True Cluster_2*	*True Cluster_3*	
Predicted Class	*Pred. Cluster_1*	220	3	1	98.21%
	Pred. Cluster_2	4	179	1	97.28%
	Pred. Cluster_3	0	0	55	100.00%
Recall Class		98.21%	98.35%	96.49%	

3.3. Model Applying Results

From the model that has been selected and demonstrated the performance, this section shows the decision tree model in Figure 6 as displayed in the end of paper. The decision tree rules for self-testing are shown in the test results in Table 9, and the centroid of each cluster is shown in Table 10.

Table 9: The decision tree rules for self-testing

Rule	Condition (If)	Prediction (Then)
1	If Stage 4.1 > 3.500 and Stage 1.1 > 3.500 and Stage 1.3 > 3.500 and Stage 3.2 > 3.500	Then, suitable for cluster_1 = 99.52%, and suitable for cluster_2 = 0.48%.
2	If Stage 4.1 > 3.500 and Stage 1.1 > 3.500 and Stage 1.3 > 3.500 and Stage 3.2 ≤ 3.500 and Stage 2.1 > 4.500	Then, suitable for cluster_1 = 100%.
3	If Stage 4.1 > 3.500 and Stage 1.1 > 3.500 and Stage 1.3 > 3.500 and Stage 3.2 ≤ 3.500 and Stage 2.1 ≤ 4.500	Then, suitable for cluster_1 = 12.50%, and suitable for cluster_2 = 87.50%.
4	If Stage 4.1 > 3.500 and Stage 1.1 > 3.500 and Stage 1.3 ≤ 3.500	Then, suitable for cluster_2 = 100%.
5	If Stage 4.1 > 3.500 and Stage 1.1 ≤ 3.500 and Stage 4.3 > 3.500 and Stage 3.3 > 4.500	Then, suitable for cluster_1 = 60%, and suitable for cluster_2 = 40%.
6	If Stage 4.1 > 3.500 and Stage 1.1 ≤ 3.500 and Stage 4.3 > 3.500 and Stage 3.3 ≤ 4.500	Then, suitable for cluster_1 = 1.23%, suitable for cluster_2 = 98.15%, and suitable for cluster_3 = 0.62%.
7	If Stage 4.1 > 3.500 and Stage 1.1 ≤ 3.500 and Stage 4.3 ≤ 3.500	Then, suitable for cluster_3 = 100%.
8	If Stage 4.1 ≤ 3.500 and Stage 1.2 > 4.500	Then, suitable for cluster_1 = 100%.
9	If Stage 4.1 ≤ 3.500 and Stage 1.2 ≤ 4.500 and Stage 1.2 > 3.500 and Stage 2.3 > 3.500	Then, suitable for cluster_1 = 100%.
10	If Stage 4.1 ≤ 3.500 and Stage 1.2 ≤ 4.500 and Stage 1.2 > 3.500 and Stage 2.3 ≤ 3.500	Then, suitable for cluster_3 = 100%.
11	If Stage 4.1 ≤ 3.500 and Stage 1.2 ≤ 4.500 and Stage 1.2 ≤ 3.500	Then, suitable for cluster_3 = 100%.
		Correct: 456 out of 463 training examples (98.49%).

Table 10: The average within centroid of each cluster

Stage and Cluster			
Stages	*Cluster_1*	*Cluster_2*	*Cluster_3*
Stage 1. Organization			
Stage 1.1	4.43	2.68	2.74
Stage 1.2	4.49	3.49	2.60
Stage 1.3	4.48	3.48	2.70
Stage 1.4	4.43	4.24	2.81
Stage 2. Researcher			
Stage 2.1	4.48	3.82	2.68
Stage 2.2	4.45	4.19	2.84
Stage 2.3	4.51	4.12	2.84
Stage 3. Students			
Stage 3.1	4.50	3.70	2.86
Stage 3.2	4.44	3.34	2.93
Stage 3.3	4.49	3.34	2.88

Stage 4. Project Support			
Stage 4.1	4.44	4.91	2.37
Stage 4.2	4.41	4.87	2.74
Stage 4.3	4.45	4.87	2.88
Stage 4.4	4.44	4.91	2.37
Total Number of Items: 463	224 items	182 items	57 items

3.4. Summarize the Significant Factors

From the data collection, there were 463 people who provided attitudes and satisfaction to the research. It can be concluded that the overall level of satisfaction is accepted (mean = 4.04, S.D. = 0.88), as shown in Table V. Therefore, it can be concluded that the respondents agreed and accepted the teaching and learning as a senior project. In addition, the prototype model has been developed for predicting the highest accuracy levels (98.06%): shown in Table VI and Table VII. Moreover, the results of the self-test data model are very accurate at the highest level, which is equal to 98.49% (Correct: 456 out of 463 training examples) as shown in Table VIII. Therefore, it can be concluded that this study was successful. It can develop highly accurate models and also have high performance model testers.

Finally, the significant factors from this research are composed of nine factors as follows: Stage 1.1 The policies of the organization, Stage 1.2 The vision of the organization, Stage 1.3 The mission of the organization, Stage 2.1 Experience and achievements of researchers, Stage 2.3 Qualifications of Research Team, Stage 3.2 Interest in the research topics, Stage 3.3 Impressions and examples in the past, Stage 4.1 Technology and laboratory support, and Stage 4.3 Budget support.

From clustering and discovering these significant factors, the researcher can use the research results to develop a program for advising advisors in a serious project course in order to effectively pair with students having a common attitude.

3.5. Research Discussions

The research discussions are divided into three points: (1) A proactive learning culture and learning style model development, (2) Scientific aspects of educational data mining, and (3) Utilization of research activities and research results.

1) A proactive learning culture and learning style model development

The reason for the researchers emphasizing on the model development and data analysis was because the research team was assembled from many institutions but encountered the same problem. Our common problem is to recommend suitable teachers to students in senior projects. Therefore, this research focused on solving problems proactively using proactive learning strategies to create opportunities for corrective action and reducing impropriety. In addition, the professional group of the researchers is in the information technology field where the effort is on using technology to help solve problems.

2) Scientific aspects of educational data mining

In the process of developing research instruments tool, researchers examined questionnaires through the researchers' organization according to procedures and standards of the research process. Questionnaires are routinely used to assess teaching and learning. It is conducted on an annual basis. The past assessments are only through statistical analysis, not analyzed for improvement and problem solving. In this research, the researchers analyzed the results of stakeholder assessments in order to form a correlation group and present the stakeholders on further improving the curriculum, and teaching and learning management.

In the conclusion of scientific aspects of educational data mining, the use of available information is important and necessary.

3) Utilization of research activities and research results

From the research process and routine operations of the organization, the researcher found that the organizational management and its members realized the importance of conducting this research. In addition, the models in Figure 6, the decision tree model, and Table 9 decision tree rules for self-testing can be further developed into computer programs, applications, and technological applications.

One of the obvious benefits of research is engagement within the organization. It makes the people in the organization aware of the importance of applying technology. In addition, students who participate in the project can also apply the knowledge gained in future students' lives and work.

4. Conclusion

From this research, the researcher can summarize according to the research objectives as follows: (1) to survey attitude and perception, which is a collaboration between researchers and students, (2) to develop a model for clustering of advisors and students, and (3) to develop factors that are significant to predict the right match in senior projects course. The results of the survey attitude and perception, which is a collaboration between researchers and students is shown in Table V. The result is that respondents have a high level of overall satisfaction (mean = 4.04, S.D. = 0.88). Thus, it can be concluded that the respondents agreed and accepted the teaching and learning as a senior project.

The result of model development of the clustering model is for matching advisors and students. The result is that the model has the highest level of accuracy (98.06%) as shown in Table VI and Table VII. Moreover, the results of the self-test data model are very accurate at the highest level, which is equal to 98.49% (Correct: 456 out of 463 training examples) as shown in Table VIII.

The result with the development of the factors are significant to predict the right match in senior projects course. The result is an important factor consisting of nine factors, which are stage 1.1 the policies of the organization, stage 1.2 the vision of the organization, stage 1.3 the mission of the organization, stage 2.1 experience and achievements of researchers, stage 2.3 qualifications of research team, stage 3.2 interest in the research topics, stage 3.3 Impressions and examples in the past, stage 4.1 technology and laboratory support, and stage 4.3 budget support.

Therefore, from the research study on collecting data Rajabhat Mahasarakham University, the Maha Sarakham University, and the University of Phayao with a total of 463 samples, this work concludes that the objectives have been achieved with the four machine learning tools and data mining techniques: k-means, decision tree, cross-validation methods, and confusion matrix. For

future research projects, the researchers are committed to further studying on the development of learners' achievement and aims to promote a learning culture based on the results of this research and active learning of educational institutions.

Author Contributions

Wongpanya Nuankaew, Kanakarn Phanniphong and Sittichai Bussaman designed the research problems and research framework; Wongpanya Nuankaew and Pratya Nuankaew collected the data; Pratya Nuankaew and Direk Teeraputon analyzed the data and constructed the clustering models; Pratya Nuankaew wrote the paper and all authors had approved the final version; Pratya Nuankaew is responsible for communicating and delivering the research to the journals.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is supported by the five organizations: (1) Faculty of Information Technology, Rajabhat Mahasarakham University, Maha Sarakham, 44000, Thailand. (2) Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-Ok, Chonburi, 20110, Thailand. (3) Faculty of Science and Technology, Rajabhat Mahasarakham University, Maha Sarakham, 44000, Thailand. (4) School of Education, University of Phayao, Phayao, 56000, Thailand. (5) School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand. The researchers would like to thank the advisor, lecturers, students, technicians, and all respondents for their entire support.

References

[1] P. Nuankaew, P. Temdee, "Matching of compatible different attributes for compatibility of members and groups" Int. J. Mob. Learn. Organ., **13**(1), 4–29, 2019. https://doi.org/10.1504/IJMLO.2019.096469.

[2] P. Nuankaew, W. Nuankaew, K. Phanniphong, S. Imwut, S. Bussaman, "Students Model in Different Learning Styles of Academic Achievement at the University of Phayao, Thailand" Int. J. Emerg. Technol. Learn. IJET, **14**(12), 133–157, 2019. https://doi.org/10.3991/ijet.v14i12.10352.

[3] F. M. van der Kleij, "Comparison of teacher and student perceptions of formative assessment feedback practices and association with individual student characteristics" Teach. Teach. Educ., **85**, 175–189, Oct. 2019. http://doi.org/10.1016/j.tate.2019.06.010.

[4] C. Brandmiller, H. Dumont, and M. Becker, "Teacher Perceptions of Learning Motivation and Classroom Behavior: The Role of Student Characteristics" Contemp. Educ. Psychol., 101893, 2020. http://doi.org/10.1016/j.cedpsych.2020.101893.

[5] C.-Y. Chen and P. P. Chong, "Software engineering education: A study on conducting collaborative senior project development" J. Syst. Softw., **84**(3), 479–491, 2011. http://doi.org/10.1016/j.jss.2010.10.042.

[6] D. A. Morley, "Enhancing networking and proactive learning skills in the first year university experience through the use of wikis" Nurse Educ. Today, **32**(3), 261–266, 2012. http://doi.org/10.1016/j.nedt.2011.03.007.

[7] I. Gnizy, E. Baker William, and Grinstein Amir, "Proactive learning culture: A dynamic capability and key success factor for SMEs entering foreign markets," *International Marketing Review*, **31**(5), 477–505, 2014. https://doi.org/10.1108/IMR-10-2013-0246.

[8] B.-K. Joo and T. Lim, "The Effects of Organizational Learning Culture, Perceived Job Complexity, and Proactive Personality on Organizational Commitment and Intrinsic Motivation," *Journal of Leadership & Organizational Studies*, **16**(1), 48–60, Aug. 2009. https://doi.org/10.1177/1548051809334195.

[9] A. Ruiz-Iniesta, G. Jimenez-Diaz, M. Gomez-Albarran, "Recommendation in Repositories of Learning Objects: A Proactive Approach that Exploits Diversity and Navigation-by-Proposing" in 2009 Ninth IEEE International Conference on Advanced Learning Technologies, 543–545, 2009. http://doi.org/10.1109/ICALT.2009.23.

[10] D. Gallego, E. Barra, S. Aguirre, G. Huecas, "A model for generating proactive context-aware recommendations in e-Learning systems" in 2012 Frontiers in Education Conference Proceedings, 1–62012, 2012. http://doi.org/10.1109/FIE.2012.6462246.

[11] M. Reuter et al., "Learning Factories' Trainings as an Enabler of Proactive Workers' Participation Regarding Industrie 4.0" Procedia Manuf., **9**, 354–360, 2017. http://doi.org/10.1016/j.promfg.2017.04.020.

[12] S. Moon, J. G. Carbonell, "Proactive learning with multiple class-sensitive labelers" in 2014 International Conference on Data Science and Advanced Analytics (DSAA), 32–38, 2014. http://doi.org/10.1109/DSAA.2014.7058048.

[13] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, "Promoting reading comprehension and critical–analytic thinking: A comparison of three approaches with fourth and fifth graders" Contemp. Educ. Psychol., **46**, 101–115, 2016. http://doi.org/10.1016/j.cedpsych.2016.05.002.

[14] P. Nuankaew, P. Temdee, "Of online community: Identifying mentor and mentee with compatible different attributes and decision tree" in 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 1–6, 2015. http://doi.org/10.1109/ECTICon.2015.7207130.

[15] P. Nuankaew, P. Temdee, "Determining of compatible different attributes for online mentoring model" in 2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace Electronic Systems (VITAE), 1–5, 2014. http://doi.org/10.1109/VITAE.2014.6934434.

[16] P. Nuankaew, W. Nuankaew, T. Thamma, "The Recommended System for the Relationship between Educational Programs and Stu-dents' Interests" in International Conference on Digital Arts, Media and Technology, 2-3, 2016.

[17] W. Tan, A. Tang, T. Wang, W. Tan, L. Li, Z. Zhang, "An Overall Lifecycle Modeling Method for E-Learning Services and the Cooperative Mechanism Based on PCDA" in 2013 IEEE International Conference on Systems, Man, and Cybernetics, 669–674, 2013. http://doi.org/10.1109/SMC.2013.119.

[18] D. A. Vidhate, P. Kulkarni, "Performance enhancement of cooperative learning algorithms by improved decision making for context based application" in 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), 246–252, 2016. http://doi.org/10.1109/ICACDOT.2016.7877588.

[19] C.-T. Chang, C.-Y. Tsai, S.-E. Peng, H.-H. Tsai, P.-T. Yu, "On the Design of Multi-User Streaming Feedback System for Application of Cooperative Learning" in 2016 International Computer Symposium (ICS), 656–659, 2016. http://doi.org/10.1109/ICS.2016.0134.

[20] P. Maresca, E. Gómez, J. Caja, C. Barajas, R. Ledesma, "Academic Learning Platform for Practical Classes: A Learning Model in Manufacturing Engineering" Procedia Eng., **132**, 205–212, 2015. http://doi.org/10.1016/j.proeng.2015.12.471.

[21] A. Bouabid, P. Vidal, J. Broisin, "Integrating Learning Management Systems and Practical Learning Activities: The Case of Computer and Network Experiments" in 2009 Ninth IEEE International Conference on Advanced Learning Technologies, 398–402, 2009. http://doi.org/10.1109/ICALT.2009.181.

[22] A. Mathew, V. Pangracious, A. Varghese, A. Thomas, D. Mathew, J. George, "PROJECT HAWK: An innovative introduction of practical learning and entrepreneurship in engineering education" in 2012 2nd Interdisciplinary Engineering Design Education Conference (IEDEC), 34–40, 2012. http://doi.org/10.1109/IEDEC.2012.6186919.

[23] A. Smith, C. Bluck, "Multiuser Collaborative Practical Learning Using Packet Tracer" in 2010 Sixth International Conference on Networking and Services, 356–362, 2010. http://doi.org/10.1109/ICNS.2010.56.

[24] G. Hamerly J. Drake, "Accelerating Lloyd's algorithm for k-means clustering" in Partitional clustering algorithms, Springer, 41–78, 2015. https://doi.org/10.1007/978-3-319-09259-1_2.

[25] P. Nuankaew, W. Nuankaew, P. Temdee, "Institution recommendation using relationship optimisation between program and student context" Int. J. High. Educ. Sustain., **2**(4), 279–302, 2019. https://doi.org/10.1504/IJHES.2019.103365.

[26] P. Nuankaew, W. Nuankaew, K. Phanniphong, R. Fooprateepsiri, S. Bussaman, "Analysis Dropout Situation of Business Computer Students at University of Phayao" in The Impact of the 4th Industrial Revolution on

Engineering Education, Cham, 419–432, 2020. http://doi.org/10.1007/978-3-030-40274-7_42.

[27] P. Nuankaew, "Dropout Situation of Business Computer Students, University of Phayao" Int. J. Emerg. Technol. Learn. IJET, **14**(19), 2019. https://doi.org/10.3991/ijet.v14i19.11177.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 808-819 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Innovative Solution for Parking-Sharing of Private Institutions Using Various Occupancy Tracking Methods

Adrian Florea*, Valentin Fleaca, Simona Daniela Marcu

Lucian Blaga University of Sibiu, Department of Computer Science and Electrical Engineering, Sibiu, 550025, Romania

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 28 September, 2020
Online: 12 October, 2020

Keywords:
parking sharing
embedded system
web application
smart parking
image processing

ABSTRACT

This work presents an innovative solution for parking-sharing of private institutions based on daily occupancy patterns and using different real time tracking methods of vacant parking slots. The research objective consists in finding the most accurate cars detection method, for determining of vacant parking slots and updating them on application web page. Beside the technical innovation represented by image processing algorithms used, this paper promotes the concept of sharing economy with many social benefits like car flow optimization, reducing fuel, pollution, loss of time and creating financial advantages for parking owners. The main software component is a web application which is connected with Raspberry Pi microcontroller, 2 Pi cameras and one fix camera for parking management. It facilitates reserving a place, opening the barrier and allows entering, exiting and revising the number of vacant slots and synchronization with the web application and the supporting database. The web application provides the following facilities: real time parking status view, reservation on a specific time by license plate number, administration module that includes payment system and updates about users and prices, implementation of the gamification concept in the management of parking spaces. The solution was piloted at Lucian Blaga University of Sibiu (LBUS) Romania. The developed solution is flexible, extensible and applicable to crowded university cities, but also to other private organizations that have inefficiently operated parking slots.

1. Introduction

This paper extends the previous work that was presented at the 2018 Thirteenth International Conference on Digital Information Management (ICDIM) [1], by presenting an innovative solution for parking-sharing of private institutions based on daily occupancy patterns and using different real time tracking methods of vacant parking slots.

1.1. The need of innovative smart parking solutions

The general-purpose objective of smart cities is finding smart solutions for increasing the citizens' welfare and comfort by providing high quality services like smart mobility, smart health, and smart environment. Unfortunately, the shortage of parking spaces and, the continuous infrastructure development generates city crowding [2] and traffic congestion. The constant stopping and starting in traffic jams burn the fuel at higher rate in comparison to when someone travels smoothly through the highway. Thus, the congested areas have a higher emission of Carbon Dioxide than other areas [3]. A solution that some countries tried was to put a number of traffic policemen to help manage the traffic in a better way [4]. The problem with this was that a one out of three-traffic policeman suffers from either one or more respiratory morbidities like allergic rhinitis, chest symptoms (cough, wheeze, and breathing difficulty/chest tightness), allergic eye symptoms, sinusitis [5].

Especially in touristic cities, at seaside resorts or others with many visitors yearly, the lack of parking spaces is emphasized by unpopular local authorities' decisions with short or long term effect. For example, in Sibiu, during the theatre festival or international events like auto-moto rallies, the municipality restricts parking places, providing these for guests or organizers. In 2019 there were 700,000 people who visited or participated to shows, more than 4 times the population of Sibiu that produces a further pressure on local drivers. Finding with difficulty a vacant parking place became ordinary activity for most of the drivers.

*Corresponding Author: Adrian Florea, Lucian Blaga University of Sibiu, Department of Computer Science and Electrical Engineering, Sibiu 550025, Romania, Emil Cioran Street, No. 4, Email adrian.florea@ulbsibiu.ro

Studies developed by researchers from Boston University [6] reveals that over 30% of the drivers lost around 8 minutes looking for a free space around the zone where they must stop their car. According to [7], the drivers spend in medium 17 hours yearly looking for vacant places on streets, in lots, or in garages generating tremendous financial losses (tens of billions of dollars) to economy. From our experience in Sibiu, Romania, finding a parking place near the Engineering Faculty takes every morning minimum 10 minutes of searching, beside the fuel costs. Arriving to Faculty earlier with around 20 minutes before starting the classes might solve sometimes this problem. Implementing a smart parking system would shorten the search time for a parking space, reduces CO2 emissions – a key aspect of smart crowded cities, optimizing the car traffic, increasing safety and sometimes even avoid fines caused by illegal parking.

An innovative smart parking system incorporates several important features [8]:

- It helps the business growth and economic development by different solutions like providing incentives for parking or sharing excess spaces of private institutions.
- It helps drivers and municipality to efficiently exploit citywide parking space or providing unexpected (vacant) spaces in real time from private parking which share their spaces.
- It analyses the daily occupancy patterns based on institution / company schedule to understand driver patterns and to predict spaces to be released.

In our smart parking solution, beside innovative technical implementation, the emphasize is on the problem of sharing economy (until the municipalities make new parking lots – a very difficult and time-consuming process, the sharing empty parking places is an acceptable solution with positive economic impact. Especially in big cities finding a parking place represents one of the major challenge of a drivers, and smart parking would represent even urban smart mobility solution for traffic congestion control using parking charges differentiated by time and area. In busy days when most of people are running to work / schools and Saturday afternoon when everybody is engaged in the social and commercial life of the city or during major events when traffic congestion is high, the municipalities' decision factors could coordinate with traffic planners the city mobility to update parking lot status and maximize use of public transit, and alternative transportation options. However, this would be applied by local municipalities. Our proposal is dedicated for private organizations who may share their excess parking places with ordinary traffic drivers without affecting their own employees.

1.2. Scientific motivation based on bibliometric study

As motivation of our research we also present the following bibliometric study, developed by us at the end of June 2020. 812 articles containing the phrase "*smart parking*" in the "Title", "Abstract" or "Keywords" and that appeared at least 25 times in the article's content were selected. The charts were developed using Microsoft Excel and VOSviewer Software (https://www.vosviewer.com/).

Analyzing the period 2011-2020, there is an increasing evolution of the number of articles published in the SCOPUS database, which refers to the phrase "smart parking". The statistics (see Figure 1) were compiled at the end of June 2020 and the number of articles in 2017 has already been reached, despite the reduction in the number of conferences (and implicitly of articles) caused by the COVID-19 pandemic.

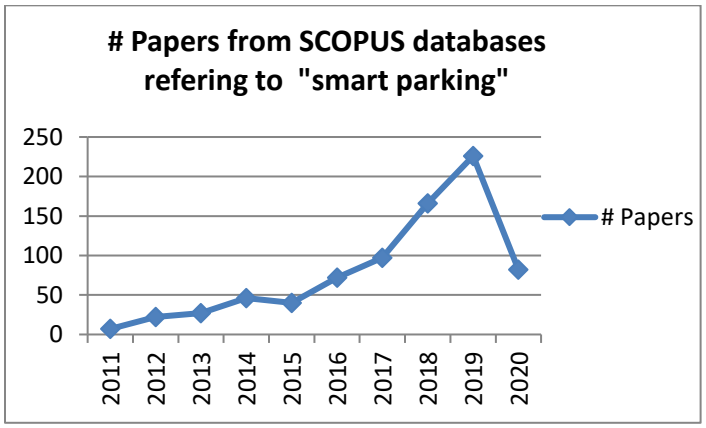


Figure 1. Number of publications indexed in the Scopus databases referring to the concept of smart parking (2011–2020)

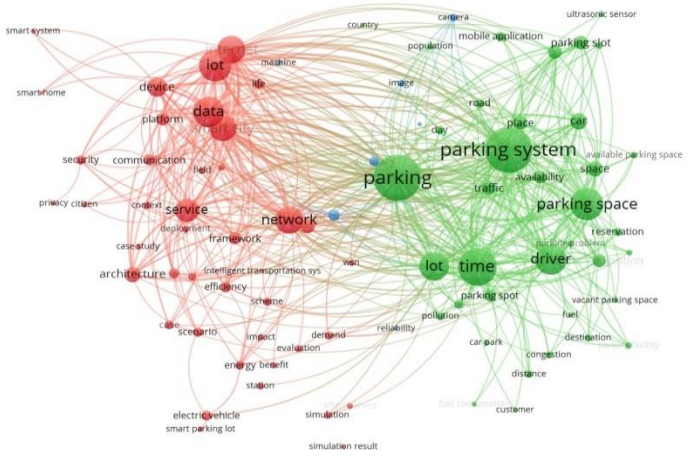


Figure 2. Number of publications indexed in the Scopus databases referring to the concept of smart parking (2011–2020)

By applying clustering algorithms on the selected papers resulted in 87 keywords grouped into three clusters. In Figure 2 is drawn with different colors a map displaying relations between the analyzed keywords. The frequently key terms are presented in the central part of the map. The style information like the size and font of point specific to a certain keyword express its frequency (higher or lower). First, the resulting network is characterized by numerous links and looks rather dense. The first cluster was located around the term "parking" including the most terms defining the problem of smart parking: parking system/lot/space/slot/spot, reservation, road, fuel, congestion, pollution, availability, etc. In the second cluster terms related to solution (most of them technical) can be observed: network, Internet, IoT (of Things), Communication, device, platform, framework, architecture, service, smart system/home/city, etc. The third is the shortest and is strict related to some tools and their performance: image, camera, accuracy, reliability, etc. However, the three clusters do not contain keywords such as "sharing", "economy", and very few or not at all contain "image processing", "machine learning algorithms", "thermo vision camera", etc.

The conclusions of this bibliometric study show us that very few researchers consider the economic view of smart parking and especially the sharing concept in order to create benefits.

Other studies show that there are a lot of startups worldwide that target different aspects of a smart parking: payment, access monitoring or even vacancies but do not analyse the problem from perspective of sharing economy and private parking lots and also does not present technical solutions for detection of free parking spaces [9].

1.3. Premises for success. Brief description of solution proposed

In our previous work [10] we predicted "the occupancy rate of a building" based on embedded system which retrieves data from sensors and, using Multi-Layer Perceptrons we predicted the energy consumption in smart buildings knowing alternative data sources, like the number of cars from parking space. It was computed the correlation between the number of vehicles over time and time-related parameters, like day of week, hour of day, and type of day, for validating the machine learning model of parking lot occupancy. This paper describes in detail the smart parking system implemented at Engineering Faculty of Lucian Blaga University of Sibiu (LBUS), Romania. The occupancy pattern during one year includes around 40 weeks, in rest being holidays. Engineering Faculty contains about 120 parking slots. The institution time scheduler shows that usually after 5 p.m. are much less classes than during the morning. In conclusion, there is a quite high degree of vacancy in the parking which could generate additional income by sharing them.

The proposed solution includes the hardware components (Raspberry Pi Model 3B with a Pi camera) for license plate recognition, the opening / closing the barrier, three led that form the traffic light, and software application for real time detection of each empty parking space based on Motion Detection algorithm; the reserving a parking space for a time interval is done through a web application. Raspberry Pi microcontroller and the WEB application communicate through the Node-RED program that use the MQTT services for transmitting real-time information to the application without the need to refresh the webpage ensuring real time synchronization of web application and database with existing vacant parking spaces.

Our solution provides the next benefits:

- Innovate by relying on the principle of sharing economy / parking sharing of private institutions: analyses the problem from the perspective of creating financial advantages for the owner in conjunction with reducing the disadvantages mentioned above, loss of time, pollution, etc.
- It is based on the profile / schedule of the institution (daily and / or on much longer periods – monthly, annually).
- Innovate by applying image processing methods for monitoring vacancies from parking lot based on daily occupancy patterns and real time tracking of vacant parking slots and updating in real time the web application and the supporting database.
- Provides a multidisciplinary character combining different disciplines of Computer Engineering ranging from Image Processing and classification algorithms specific to Artificial Intelligence, to Embedded Systems and IoT, Web Design, Information Security (payment), targeting social welfare, improving the citizens' life quality.

The rest of the paper is organized into six sections. Section 2 analyses the related works. Section 3 introduces our smart parking solution, its advantages as well as the analysed use-case scenarios. Section 4 presents the application's hardware and software architecture, whilst the Section 5 describes the image processing methods used for finding the vacant slots in the parking space. The Section 6 illustrates some results obtained after software implementation of methods from fifth section. Section 7 emphasizes the paper's conclusions and proposes further work research.

2. Related Work

The reservation-based parking system developed by researchers from Nebraska University, Lincoln, USA [11] focus on parking spaces managing and real-time monitoring. Their solution is based on reservation and the selection of the vacant slot for parking is made based on the walking distance to the destination and the cost. The main drawback of their solution consists in the synchronization problem when more than one motorists apply simultaneously for reservation. A "bottleneck" occur within the system and increases the update time out of the need to serially treat the requests. In our work we solve this problem at the session level of the OSI reference model [12] of communications protocols. Different names for every session are used, and the messages are received and sent only to the specific session.

In [13] the researchers from University of South Florida developed an embedded system for a real-time parking aiming to inform the drivers using a mobile application, about parking spaces availability. Detecting the vacant spaces was done with ultrasonic sensors. When a car is detected in front of the parking lot, the ultrasonic sensor reads and refreshes the number of cars on the entrance display and on the smart phone application. The accuracy suffers in case of golf carts or people because the sensors consider them as cars. Unlike them, our solution of entering and leaving is based on image recognition of license plate number.

The Indian researchers from Pune [14] implemented an intelligent reservation parking system using QR code technology and a camera that reads at parking entrance the QR code and checks its correspondence in database. In our case, the system includes newer and different hardware components than those used in [14] and besides that we implemented image recognition for license plate instead of QR technology, we also use image processing for updating the web application according to cars presence or not from the parking.

Finding vacant parking slot by the drivers becomes an NP-hard problem with various constraints. In [15] a genetic algorithm is used for solving scheduling problems and finding the best parking region into shopping malls and the slot to park the vehicle without affecting medium waiting time. The kernel of their system is a number of autonomous trolleys. These are used to take the vehicles from the entrance to best parking slot identified by the system. The customer can leave their vehicle at the entrance and enjoy the

shopping. The trolley captures the image of the surrounding after the vehicle is successfully parked in the slot. The image and location map is send to the customer mobile application. The main disadvantage is the cost of implementation and the necessity that the customer to have a WiFi connection.

In [16] the author reviews and presents practical examples based on smart technologies for guiding drivers to parking lot in urban areas. There are four categories (from simple to complex) depending on the equipment used and the services provided by parking lot: typical, semi-automatic, automatic and multi-storey. The simplest (typical) involves no equipment and services, the semi-automatic allows payment at the cashier and the operator who operates the barriers, the automatic is fully equipped with an automatic parking equipment and payment on terminals. The disadvantage of first three categories is that these parking systems are not providing information about the occupancy of car parks (free parking spaces) except for the direction and possibly the distance to the destination. The most complex, multi-storey car parks are fully automatic and equipped with smart technologies. Unfortunately, from the administrator's viewpoint their purchasing and correct implementation is costly and from the driver's point of view, the disadvantages are their reliability and detection accuracy (most of the time different of 100%).

Smart parking alternative solutions based on magnetic sensors and IoT systems are revealed by engineering companies. Magnetic sensors, placed on the side of road or parking spot, on the road surface, or in the ground in the form of Inductive-loop detectors, can be used to check the occupancy of the parking lot, determining the number of free spaces left every time a car passes over them and disrupts the magnetic field [17]. The major disadvantages of this solution are the high installation cost and time and require civil engineering.

In [18] are briefly shown examples from Siemens and Huawei IoT-integrated parking management systems. These provide drivers information about the availability of parking lots and connect with municipalities decision factors. The IoT system consists in environmental sensors, a cloud-based web platform enabling also the automated parking payment. Our solution is very close to this system but furthermore promote the idea of parking sharing by private institutions.

In the European project proposal "Unlock the sharing economy for the parking sector" [19] the authors promote the parking sharing concept proposing integration of parking sharing platform into connected and autonomous transportation systems in smart cities. Based on mathematical model and simulation the authors tests how motivated are the drivers to use the platform which includes private parking shared for public. Furthermore they reveal the need for collaboration between academic and industrial partners as co-developers and city councils and local communities as decision makers and private companies as parking sharers.

3. The smart parking solution

This work aims to create an embedded system that solves the parking problem described above. The main software component is a web application which is connected with Raspberry Pi for parking management: reserving a slot and authorizing the entering and exiting and revising the number of vacant slots, real time detection of each empty parking space based on Motion Detection algorithm and synchronization with the web application and the supporting database. Of course the technical implementation is complementary to the sharing economy concept that in order to be efficiently applied requires appropriate understanding the parking occupancy patterns of each private institution. The project development stages are:

- Implementation of hardware-software application for identifying car license plate number and real time detection of each parking space (to determine if they are vacant or not)
- Development of the web application
- Real time synchronization with the web application and database

The web application provides the following facilities:

- Real time parking status view
- Reservation on a specific time by license plate number
- Contact parking management by email
- Administration module that includes updates about users and prices
- Implementation of the gamification concept in the management of parking spaces.

In order to make drivers estimate an exact time at which their car will be parked on that place, is introduced the concept of "Gamification". In this case, this concept aims at reserving parking spaces for as accurate a period of time as possible, in order to better manage vacant spaces and in addition to this, it also has the purpose of reducing fuel, driving time and streamlining traffic at peak hours. This is done using the web application and encourages drivers to make as accurate estimation of time as possible by rewarding them with points that can be used later for parking payment.

Scalability issues and proposal for solving are treated in [1].

3.1. The use case diagrams

In Figure 3 is presented the possible use cases of the application: the involved actors and the actions performed. There are three types of actors: the administrator, a previously enrolled user and the new user. The administrator manages the parking lots and their status; adds new locations and parks; manages new users; allows sign-in of new users; allows and control making the new reservations and set the price for parking. A registered user may see the parking status, may reserve a parking slot into a location, and obviously the sign-in process. The reservation process supposes: choosing the date and time for access, the parking location and slot and providing the license plate. The new user can see the parking status, the contact information and the geographical area of parking location, and he may register.

3.2. Reservation

The smart parking system is reservations-based. Any registered user may reserve a parking slot in case exists at least one vacant. Reservation includes the parking name, parking space (selected by user), hourly rates, the estimated date and time of arrival and

departure. The parking spaces reservation is made from web application. This reservation can only be made if the user already has an account (as new user performed the registration process and then as registered user make the sign-in process).

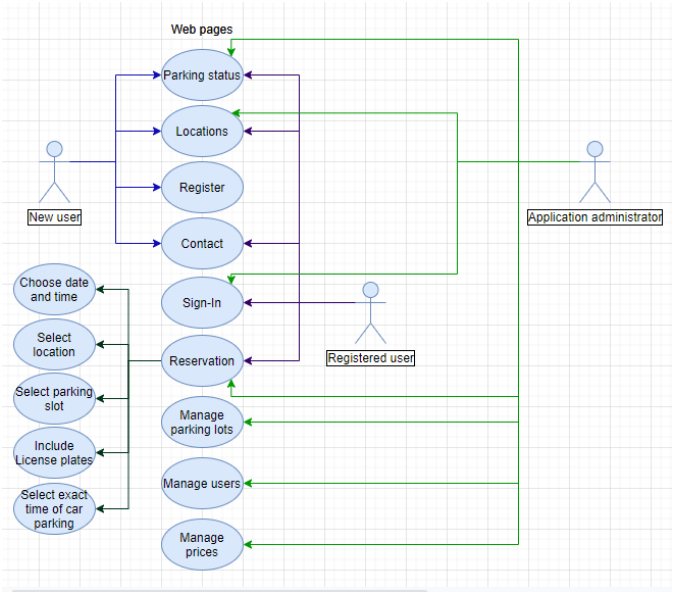


Figure 3. The application use case diagram

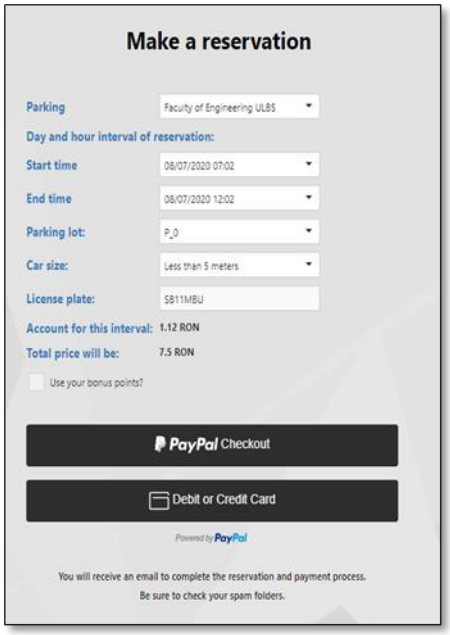


Figure 4. Reservation form

Extending the original work [1], in the Reservation page (see Figure 4) we added hourly payment implementation and the type of parking desired based on car's size (less than 5 meters, between 5-10 meters and higher that 10 meters). This constraint has been implemented to ensure an appropriate number of parking spaces according to the size of the car, otherwise there is a risk that the car is too large compared to the parking space and then two places must be invalidated even if only one car is entered. To change the type of parking is required to access the configuration page, and for each parking space the type is changed by the administrator. Payment implementation buttons have also been added using PayPal technology, which currently perfectly simulates a real payment. This process requires bank protocols which are not finalized yet.

3.3. Real time parking status view

As Figure 5 illustrates, the parking degree of occupancy can view in real time by anyone, and, for the car parked, which are their license plate numbers. The busy spaces are drawn with Red color, and vacant spaces are drawn with Green color. In "Parking status" page can be seen in real time the occupancy of the parking. The parking slots are updated by reloading the page when a message is received from the Node-Red with the parameter "Access". All parking lots are made as dynamic buttons. All parking coordinates together with id and color are taken from database into web application and displayed as buttons of different sizes depending on the parking type. For each set of 4 points that represent a parking space has been made some calculations to determine the parking type. The types of parking can be as follow: normal parking, diagonally parking on the right, diagonally parking on the left and lateral parking (see Figure 5).

In addition to the original work [1], in the Parking Status page all the coordinates corresponding to each parking space are taken and it is determined for each place whether it is straight, lateral or diagonal. These places are displayed on the page according to the coordinates received in the database (x, y) and the corresponding color in real time.

The management of parking in the web application is done in the following way: when a new parking lot is added, the "start" image is taken from the moment the fixed camera is installed in the parking lot, it is sent to Raspberry Pi for processing (is checked if there are parking spaces in the image , how many, and possibly the two-dimensional coordinates x, y for each parking space). This image is added as a background on the real time parking view page and is colored according to coordinates which is occupied in red and the rest in green (similar to Figure 5).

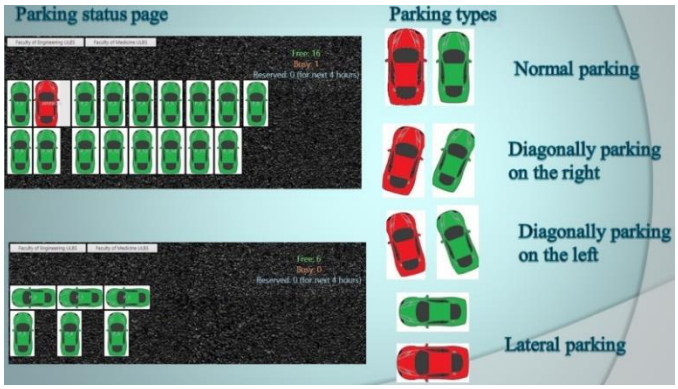


Figure 5. Real time status parking view

3.4. License plate recognition

At the entrance into the parking is checked the car license plate number. When reach in the front of the barrier, the driver push <*I arrive*> button from "Home" page to send instruction to camera for capturing and checking license plate number. If the drivers have a reservation that is +/- 15 minutes around estimated arrived date and time he has access into the parking. Otherwise, it is

checked if exist vacant slots, that are not reserved in the next 4 hours. In the affirmative case the car is authorize to enter in the parking space, otherwise not.

4. The application architecture, tools and programming environment used

This embedded system consists in the following main components:

- Hardware component which contains a Raspberry Pi Model 3B with a two Pi cameras (one for entering and one for exiting from parking lot), a Hikvision camera from the category "Bullet Series Thermal", stepper motor and some LEDs. Raspberry Pi is used together with the library that includes License plate recognition algorithm (OpenALPR algorithm from OpenCV library) that checks the number of license plate with the previously made reservation and opening the gate, parking lot and occupancy detection algorithm, Pi camera used for taking picture of each car arrived on parking entrance to apply license plate recognition algorithm, python script to manage barrier using a stepper motor and providing rotation angles, and another scripts that commands LEDs lighting for semaphore simulation. The Hikvision camera is used for parking monitoring and providing inputs for Motion Detection algorithm.
- ASP .NET web application developed using DevExpress controls. Website structure it consists of several modules who communicate with each other for a better functionality of application. MQTT (Message Queuing Telemetry Transport) and Node-Red are used for communication between web application and Raspberry Pi. The main web components are pre-setting the parking configuration based on the architecture of the location, user registration, parking slot reservation, gamification concept, real time parking visualization with counter for each vacant, occupied and reserved (for next two hours) parking slots, administration module to manage users, parking and prices for each zone.
- MySQL database. The database is essential and structured in such a way as to allow a much better flexibility, efficiency and dynamism so that the web application can be applied much more easily from one parking to another. The MySQL database contains few tables with foreign keys and records in database all reservations, user information, prices, history and the number of free parking spaces.

In Figure 6 is presented the application architecture together with all project stages from online reservation to entering car into a parking space. The first step is Motion Detection algorithm (1) in order to set the parking configurations and saving the coordinates in database (2). When the camera is installed and fixed, the parking spaces in an image captured in real time are delimited with the help of a Python script. The delimited places are saved in the database together with the coordinates and the id corresponding to each place, and the initial colour is green (free places). The Motion Detection algorithm is in real time applied and used for updating the colour corresponding to each place (red - occupied, green - free) in the database and web application. Then, it follows the online reservation (3) and inserting a record in database. When the driver arrives in front of the barrier, he has to push the <I arrive> button from Home page. After all this, the Node-Red is sending using MQTT, messages and instructions from/to application. In step (4), after pushing the *<I arrive>* button the web application through Nod-Red is sending to Raspberry Pi the instruction to take a picture. Then the checking access in parking is start calling OpenALPR algorithm (5) which provide the license plate number (6) to web application. So, if the car that arrives in the front of barrier has a valid reservation (the license plate number is successfully found in database (7) and the result is sent to web application (8)) then the latest grants the permission to enter in the parking space (9) and send this message through Node-Red to Raspberry Pi, the green led is turning on (10), and the barrier opens (11). Otherwise, the red led is turning on. After verification of license plate number, web application sends instructions to turn on / off the lights and open the barrier or not. While all this checks are made, yellow light is on. Finally, if the access is granted, green light is turning on, and the barrier opens. Otherwise, red light is turning on, and the barrier does not open.

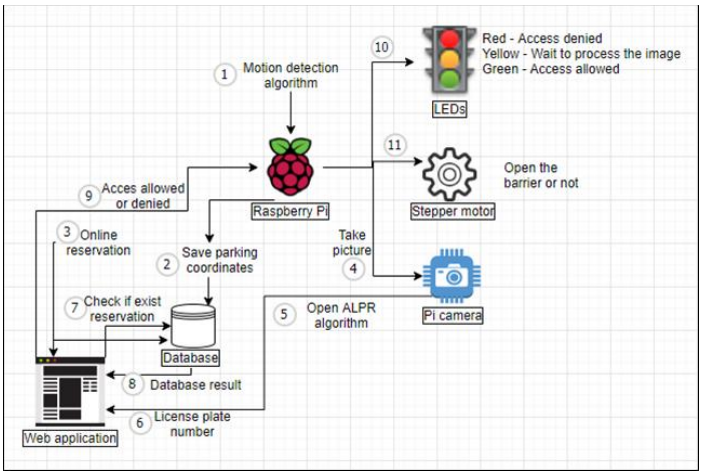


Figure 6. Application hardware-software architecture

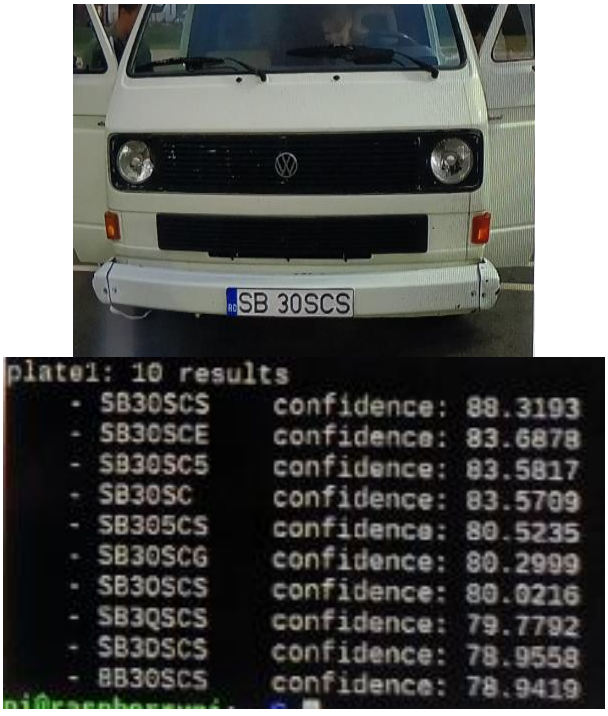


Figure 7. ALPR algorithm (up - image source, down - the result)

For updating the web application and database with the correct number of slots when the car leaves the parking is used the second Pi camera. So, if a car intends to exit from parking the same license plate recognition algorithm is applied. In the web application we check if that license plate number is inside the parking or not. In the affirmative case, is change accordingly the parking slots and opens the barrier and turn on the green led. The contrary case means that the car intends to enter in the parking. Ultrasonic sensors could be alternative solution for detecting the presence or absence of the car. When the car leaves, sensors detect this thing and mark this thing in the database. We opted in this project to use the second Pi camera instead of sensors from simplicity reason.

4.1. Software tools

1) OpenALPR: is an open source library written in C++, Java, Node.js, and Python that analyses video images and streams to identify the license plate numbers [20]. The Computer Vision algorithm is in Python implemented. The output returned is the text represented by the characters of the license plate (see Figure 7).

OpenALPR return license plate number with the higher confidence from ten possible results. The algorithm can be configured to work both with European and American license plate numbers. This library is portable and can be compiled and run on Linux, Mac and Windows. OpenALPR needs the next additional libraries:

- Tesseract OCR: it is one of the most accurate optical character recognition (OCR) engine for various operating systems.
- OpenCV: it is a Computer Vision library enhanced with machine learning algorithms.
- Leptonica: contains software dedicated for image processing and analysis applications. The license plate region detector uses the Local Binary Pattern (LBP) algorithm [21]. LBP is a simple but effective texture operator that labels the pixels of an image by comparing it with neighboring pixels, and considers the result to be a binary number. The advantage of LBP operator consists in its resistance to changes in the level of gray generated mainly by the variation of light.

2) MySQL: is used for database. In Figure 8 can be viewed the database architecture, the structure of the tables, the primary keys of each table and their links to other tables as foreign keys, which allow adjacent selections from several tables (joins), but also allow faster retrieval data in this way.

The Users table contains information about users and their type such as “User”, “Visitor” or “Admin”. The Reservations table contains all reservations made using the online reservation such as parking zone, car size, parking spot where driver should park, time of arrival and departure, license plate number, account that need to be paid before, total price and user id. The History table keeps the temporary information retaining only the vehicles that has entered in the parking and are still there. Here are stored the exactly time of arrival and departure of each car that had access to parking. Price table stores the hourly rate for every parking and ratio for account payment for each parking zone and each parking type. The Parkingzone table stored the name of every parking zone (for the moment “Faculty of Engineering ULBS”, “Faculty of Medicine ULBS”) and in Parking_types table are stored the car size types such as “Less than 5 meters”, “Between 5 and 10 meters” and “Higher than 10 meters”. In table Parking_coordinates are stored information about each parking slot. Every row from this table contains a set of 4 points which determines a parking slot, id from image file and colour in real time (red or green depending if parking slot is vacant or not).

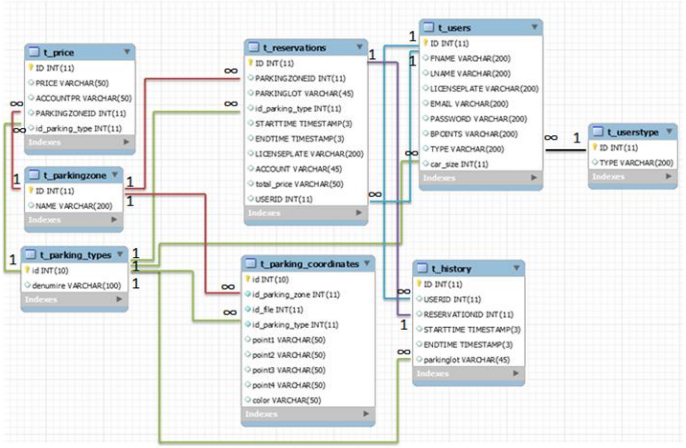


Figure 8. Database architecture

3) Node-Red: represents a programming tool used in general for IoT programming that connects hardware components with API’s and online services in new and interesting ways. To connect the Raspberry Pi, the Pi camera, the LEDs and the stepper motor of the web application is used Node-Red programming tool.

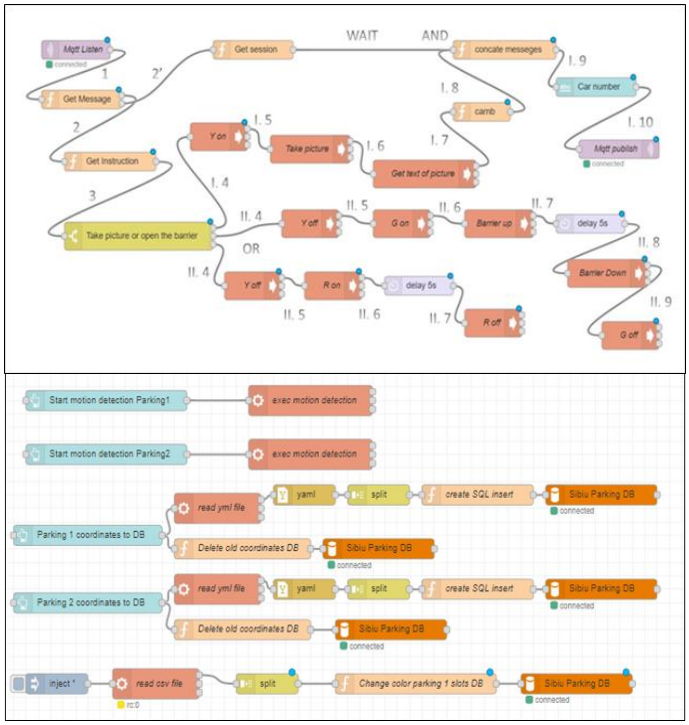


Figure 9. Implementation in Nod-Red of parking access (up) and Motion Detection algorithm (down)

In the upper side of Figure 9 is presented the whole flow of Node-Red. The purple nodes are MQTT nodes. “MQTT in” is waiting for messages from web application while “MQTT out” is sending messages to web application. The nodes with an “f” are functions, the yellow/green node is a switch node, the nodes with

an arrow are execution nodes, delay nodes represent a delay of 5 seconds, and the blue node represents a text output that displays the final message that has to be sent. When a message arrives, using function nodes we get the session and instruction from message. The session is used to know from which web session had been called the instruction, to send back the message only to that session. If instruction is "new Car" we go on the first value of switch. This value generates the execution on commands like turn yellow led on, take a picture, get the license plate number from picture, concatenate with session and send message. The second value of switch determine the fact that the car has access in parking and send the commands to turn off yellow led, turn on the green one, open the barrier, wait 5 seconds, close the barrier and turn off the green led. Last value of switch represents the fact that the car doesn't have access and turns off yellow led, turn on red led, wait 5 second and turn off the red led.

In the down side of Figure 9 are the scripts corresponding to saving coordinates for each parking slot from real time image captured with Hikvision camera when this is installed for the first time; Command to apply Motion Detection algorithm and saving real-time colors to database at each 1 second.

4) DevExpress: provides various controls for ASP .NET web applications. The web application was splited into several modules with different functionalities combined to achieve the goal, and easy understanding from the user perspective. There are implemented the 3 menu types (for Visitor, Users and Administrator) according to the application use case diagram (see Figure 3). The Register form (see Figure 10) is completely made from DevExpress controls. For the web application in order to have similar format for all pages as possible (the only part what differs in pages is the content) a page called "master" is used, in which CSS or JavaScript files can be imported, elements can be added that are desired in the future in all the pages created later, such as menu, footnote, or any other elements whose presence is desired in more than one page. This page is responsible for the layout of the application.

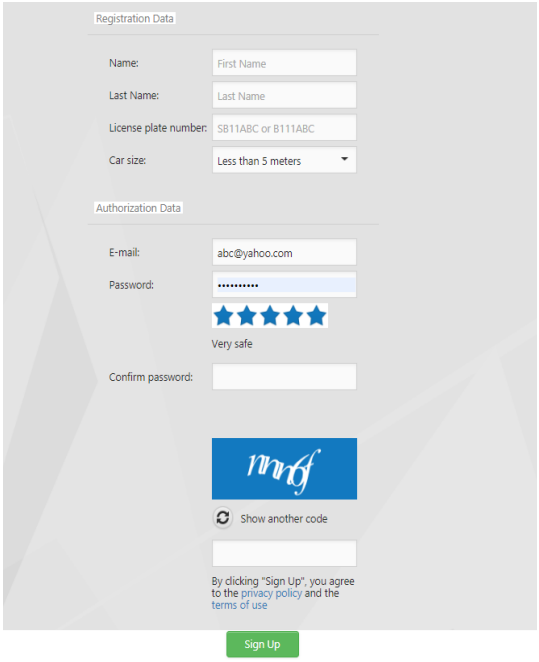


Figure 10. Register form

5) MQTT is a machine-to-machine (M2M) connectivity protocol. The messages are sent and received using an online MQTT broker. The Node-Red receives the message using the MQTT in node. The implementation of MQTT in the web application is done by installing an M2Mqtt service, which uses a PAHO client from Eclipse, the programming being done in JavaScript.

5. Finding the vacant parking slot using Image Processing methods

Finding vacant spaces can be a challenge for large parking lots and especially at rush hours. The technical solution proposed for implementing a smart parking system in this work is based on Image Processing. Different methods was tackled in order to set the best one. The Hikvision camera is positioned on a considerable height (3 meters above the ground), so that it could see the entire parking lot.

In the following are provided brief details about the methods tried and will be discussed in detail the most accurate that is used in the implementation of the smart parking system.

5.1. Using the Hough Transform

First was tried the detection of the parking lot grid using Hough Transform (dedicated to feature extraction and object detection) and Hough Probabilistic for lines [22]. The potential challenge referred to the fact that the parking lot grid is not clear/visible enough from the camera position in case of fog/snow/rain/old paint on pavement/fallen leaves. The approached solutions were:

i) Manual parking lot grid initialization

- Reading the image
- Convert image to grayscale to reduce the info in the photo
- Apply Gaussian Blur to remove even more unnecessary noise
- Detect the edges with Canny detector
- Detect lines with Hough Transform

ii) Machine-Learning algorithm for detecting the grid (fed with multiple photos from different days/weeks/months) which generates as output the parking slots.

On web application, first was tried the rendering of background image every 1 second (and after every 5 seconds) with real time picture from camera. The rendering takes too long. We tried all this on a separated asynchronous thread, but with no success. After that we were thinking about drawing each set of points on a background image from web application. This was a good idea, but wasn't good looking. The main drawback was that other lines appear besides the desired ones and the hard part was that we need to identify the useful lines and remove the rest. We tried to selectively analyze only of the useful part and not of the whole image. Unfortunately the results was not satisfactory. These are presented in the 6.1 section of the paper.

5.2. Car detection via static thermovision camera

The Hikvision camera allows acquiring images both from infrared and visible specter. Thus, the second method aims car detection via static thermovision camera by tracking the engine of the car

(considering that it will dissipate the highest amount of heat). In the database is saved the slot with highest temperature produced by the engine of a new car that is coming from outside the parking. We update as an occupied slot, in the database, and in web application, whenever a new hotspot reaches to a previously vacant slot. A challenge that we faced was finding the correct threshold for the heat level. This is because in the sunny days even the pavement/asphalt can reach the temperature of running engine. To overcome this, we ignore not moving heat spots. This method could be the best for using at night consumes moderate computational power but provides moderate accuracy. The main disadvantage is the cost of camera which is relatively high.

However, were made without success, temperature comparisons relative to the cars that enter the parking lot, against to those that stay and have the engine "cold" or "colder" according to the previous idea that different temperature of the cars represents a concrete indicator regarding a vacant or occupied place (see Figure 13).

5.3. Using Motion Detection algorithm

Detecting a free parking slot based on car motion can be done in multiple ways:

- Car tracking via motion detection – from the moment the barrier is open, is known that only a car would move near the parking entrance and it can be track it up to a point where it parks. The location is saved in the database as occupied. The challenge in this case is that if the driver, on his way to a slot, will stop too much time on the road the tracking will fail (a threshold variable for the motion detection has been reached). Thus, is need variable thresholding time for the tracking (tracking the car until it reaches exactly on a parking spot by calculate the distances towards the nearest available slot's grid lines). It has low computational impact but may not be accurate enough and it does not work well during night
- The second way suppose no car detection or tracking, actually the motion is indirectly determined. It is just detected if a slot is occupied or not by calculating the average of pixels between 4 grid lines. An empty slot has to have a lot of darker pixels (empty slot case) than any other colored pixels (the case when car is parked). The points method (based of pixels between 4 grid lines) provides the best accuracy and is easy integrated in the web application. Basically, in the pre-processing stage (pre-setup), the image is analyzed and is configured the 2D map of the parking lot that is loaded in the web page. Subsequently, the stored information related to the spatial position of the existing parking spaces is kept in database. This motion detector algorithm is based on some coordinates given by user. Thus, user has to give position for each of the parking spots by clicking on the initial image. This process is made only once, when the camera is fixing (pre-setup stage). Each set of 4 points that made up a parking space for all of the spaces in the lot are stored in an *.yml* file. Clicking edges of each parking slot is made using a mouse as a "paintbrush". Then some calculations are made to find out the center of the rectangle (to label each space with a number, used later on web application). After drawing the rectangles, all there was left to do was examine the area of each rectangle to see if there was a car in or not. By taking each rectangle (grayed and blurred), determining the area and doing an average on the pixels, we can tell that there was not a car in the spot if the average was high (more dark pixels) and the rectangle color is changing. One challenge that could arise is if car has a close color to the pavement. In this case, even if a car has the same color as the slot when it is empty, it will still reflect light from the windshields so will result different average color of pixels than an empty slot.

The most benefits of this method are that provide high accuracy and requires moderate to low computational power. By this indirect Motion Detection algorithm is known exactly that a new car which entered in the parking lot cover a previously empty place (however it is not 100% sure that is the reserved place, it would have been useful to combine the algorithm with the above solution that follows the car from the barrier to the parking slot). The introduction of the type of car is necessary and efficient. It allows the display of parking on the web page in a way as close as possible to reality and the parking spaces are updated in real time depending on what colors the Motion Detection algorithm returns for each parking slot (see Figure 14).

Next, we summarize the strengths and the weakness of car detection methods:

Table 1: Algorithms Summary

Algorithm	Advantages	Disadvantages
Hough Transform	• Handle missing data • Express a high level of detail	• Due to the heterogeneity of the objects in the image, the algorithm produces as a result lines that are not part of the parking lot (poles, houses, etc.). • Sensitive to fog / snow / rain / fallen leaves, etc.
via static thermovision camera	• May be successfully used at night • Consumes moderate computational power	• Moderate accuracy • Relative high costs
Motion Detection	• High accuracy • Low computational power	• Limited by the relation between cars color and pavement

6. Comparative results between algorithms

6.1. Results after using Hough Transform for detection of the parking lot grid

In Figure 11 and Figure 12 are presented the original images (two parking lots) and the results respectively after detecting the lines in parking using Hough Transform. OpenCV encapsulates the math of the Hough Transform into HoughLines(). This approach was tested on different videos and real time images but without success.

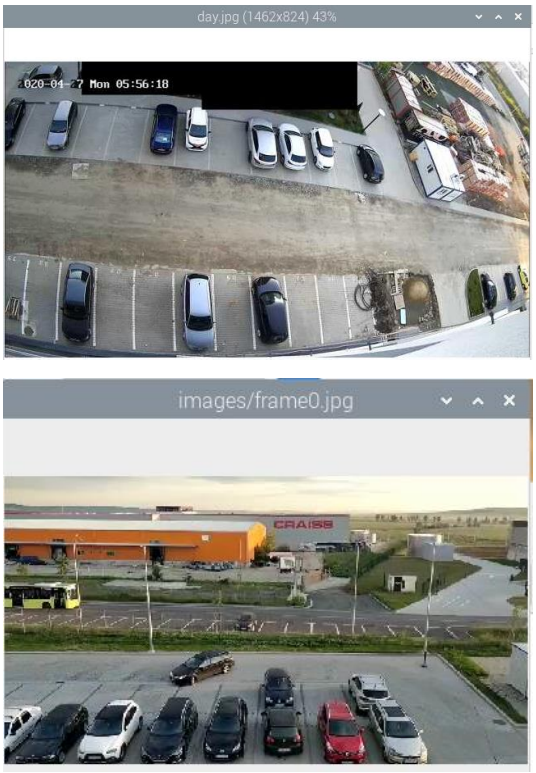


Figure 11. Original images

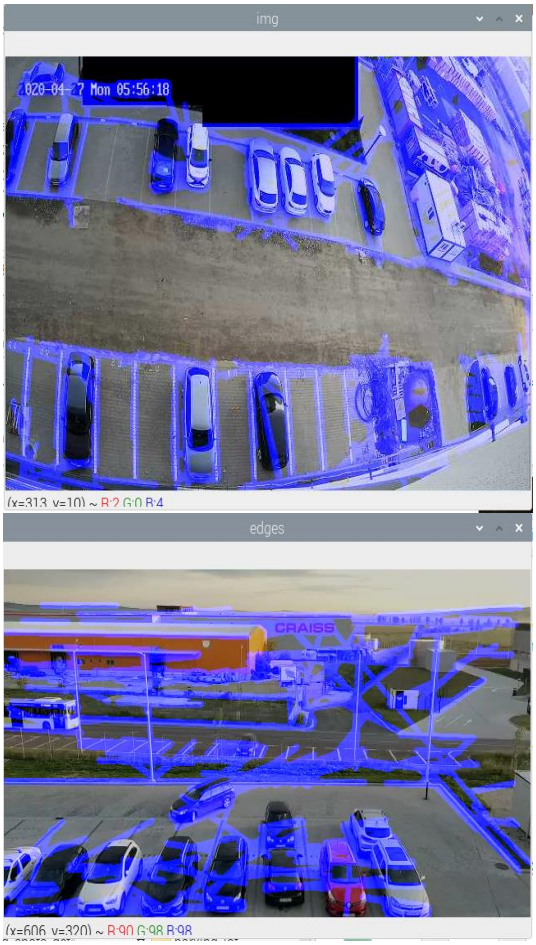


Figure 12. Final result

6.2. Results from thermovision camera

One explanation for the disappointing results may be that the car body insulates the engine from the outside and its high temperature is not captured by the thermovision camera. Also, heated asphalt produces a higher temperature than cars. The Figure 13 illustrates the image of the parking lot captured with the camera both in the infrared and in the visible spectrum. In the Figure 14 it can be seen how the driver getting out of the car has the highest temperature in the image and not the engine. Unfortunately, the method did not help us to accurately identify free parking spaces.

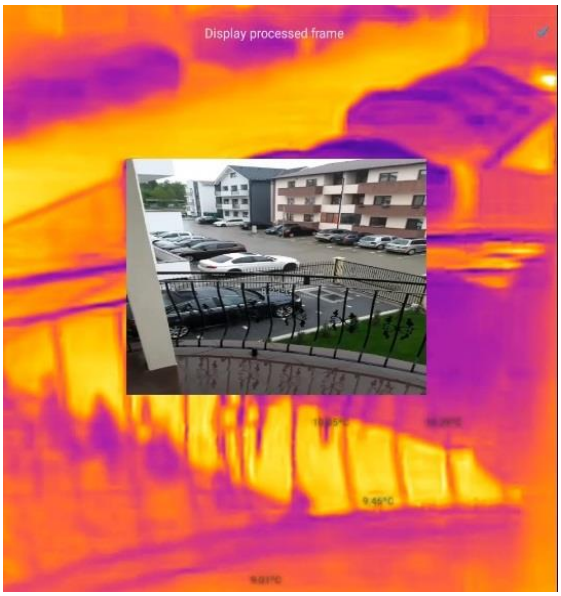


Figure 13. The parking lot seen in both spectrum: infrared and visible

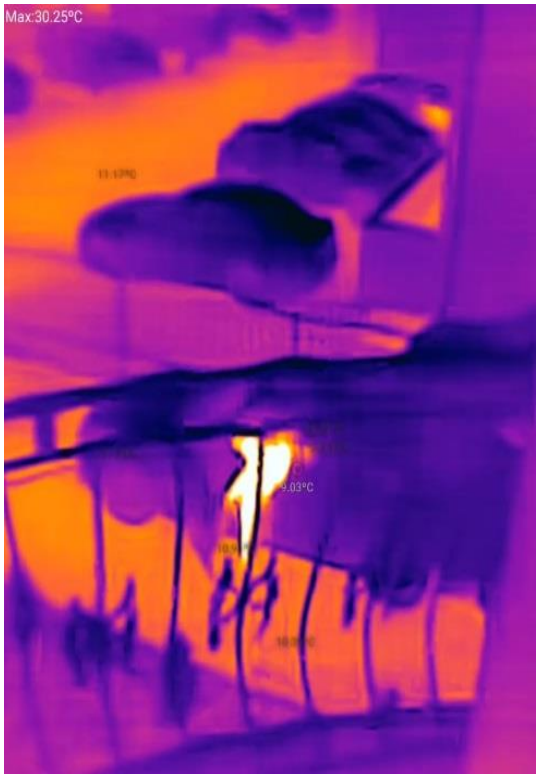


Figure 14. Driver exiting from the car parked seen with thermovision camera

6.3. Results after using Motion Detection algorithm

Looking both to Figure 15 and to Figure 5 it can be seen that the indirect Motion Detection algorithm helps to create a more real 2D structure of the parking lot and using it we can determine the occupancy of the parking space (to know which slots are vacant or not). In the upper side of the Figure 15 is capture the pre-setup stage when are identified the slots (with Blue colour) of the parking, in the middle it can be seen a scenario with two empty spaces and the down side image shows the previous places that are now occupied. It should be mentioned that for a better exemplifying, only the bottom of the test image was treated (since the pre-setup phase) without taking into account the cars and the places from the top of each of the three images. Once the algorithm is implemented, the degree of parking occupancy can be accurately determined at any given time.



Figure 15. Motion Detection algorithm results

7. Conclusions and further work

This paper presents a smart parking embedded system with focus on image processing algorithms as a solution for smart mobility. We implemented and comparatively presented the results of various image processing algorithms like Hough Transform, with thermovision camera and based on motion detection for real time tracking of vacant parking slots or OpenALPR for identifying car license plate number. The accuracy of the identification system of license plate numbers depends on the quality of the camera used. For example, in the first tests we noticed that for the number SB11MBU the number 11 was not read, the result being SBMBU, probably due to the confusion with the letter 'i' or 'I'. Another problem would be the brightness; the 5MP resolution Pi camera is not efficient enough to take clear pictures when the brightness is low. Anyway, with this cost-effective Pi camera we perfectly detected 188 from 204 cars meaning 92.15% accuracy (a reasonable performance).

The main software component of the system is the web application that implements the payment system with different rates depending on the number of hours or days of parking and integrates the gamification concept rewarding the driver with bonuses if he exactly respects the parking period for which he made the reservation. When booking a parking slot, an account is paid to make sure that the slot is available on arrival thus avoiding the queue for payment and the need to change money.

This work brings social, economic and scientific arguments to promote innovation in smart parking and especially promoting the parking sharing concept. Even if we developed a prototype tested at LBUS Sibiu (a town with around two hundred thousand inhabitants) we appreciate that the smart parking system proposed is functional, scalable [1] at least to moderate sized cities and extendable to many private institutions which might share their excessive parking places according to the daily occupancy patterns of their employees. Excess parking spaces of the parking lots of different faculties / universities, but also of private companies can be shared to users in need! If a power outage occurs or the barrier does not work, to avoid creating a long queue at the entrance that could block traffic then the implemented technical solution involves the automatic lifting of the barrier and allowing free entry and exit (without scanning the license plate number) of cars depending on the remaining available space. Even in these conditions, the Motion Detection algorithm and the takeover of the occupied spaces in the web application works, announcing the exact number of free spaces in the parking lot.

There are at least two work directions we will consider in the near future:

- The web application should provide an indication of what would be the best place to park depending on the size of the car based on genetic algorithms.
- With the help of Google Earth or Google Street applications, we will take over the spatial arrangement of the parking lots and we will generate the parking architecture in the web page (see Figure 5), thus offering a greater flexibility for application.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was partially developed under the ERASMUS+ KA2 project "THE FOF-DESIGNER: DIGITAL DESIGN SKILLS FOR FACTORIES OF THE FUTURE", financing contract no.

2018-2553 / 001-001, project number 601089-EPP-1-2018-1-RO-EPPKA2-KA, web: https://www.digifof.eu/

References

[1] S. D. Marcu, A. Florea, "Smart parking system-another way of sharing economy provided by private institutions". In 2018 IEEE Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 2018. https://doi.org/10.1109/ICDIM.2018.8846986

[2] E. J. Tomaszewska, A. Florea, "Urban smart mobility in the scientific literature - bibliometric analysis", Engineering Management in Production and Services, **10**(2), 41-55, 2018, https://doi.org/10.2478/emj-2018-0010

[3] A. Florea, L. Berntzen, M. R. Johannessen, D. Stoica, I.S. Naicu, & V. Cazan, "Low Cost Mobile Embedded System for Air Quality Monitoring", In Proceedings of the Sixth International Conference on Smart Cities, Systems, Devices and Technologies (SMART), Venice, Italy (25-29), 2017.

[4] N. Zhao, & Z. Li, "Optimize traffic police arrangement in easy congested area based on improved particle swarm optimization". Procedia-Social and Behavioral Sciences, **138**, 408-417, 2014. https://doi.org/10.1016/j.sbspro.2014.07.219

[5] G. Gowda, & R. Thenambigai, "A Study on Respiratory Morbidities and Pulmonary Functions among Traffic Policemen in Bengaluru City". *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, ***45***(1), 23–26, 2020. https://doi.org/10.4103/ijcm.IJCM_102_19

[6] Y. Geng, C. G. Cassandras, "A new smart parking system infrastructure and implementation". Procedia - Social and Behavioral Sciences, **54**, 1278–1287, 2012, https://doi.org/10.1016/j.sbspro.2012.09.842

[7] K. McCoy, "Drivers spend an average of 17 hours a year searching for parking spots", USA TODAY, December 15, 2019, https://eu.usatoday.com/story/money/2017/07/12/parking-pain-causes-financial-and-personal-strain/467637001/, last accessed 25.08.2020.

[8] B. Chan, "Smart parking is innovative, but what it enables is even more innovative", Digital Transformation, Innovation Management, Internet of Things, IoT, Smart City, September 21, 2017, https://strategyofthings.io/smart-parking-innovation, last accessed 25.08.2020.

[9] I.T. Cervera, "Internet of Things & startups: state of the art and emerging trends", Master's thesis, Universitat Politècnica de Catalunya, 2020.

[10] J. A. Oliveira-Lima, R. Morais, J.F. Martins, A. Florea, C. Lima, "Load forecast on intelligent buildings based on temporary occupancy monitoring", Energy and Buildings, **116**, 512-521, 2016, https://doi.org/10.1016/j.enbuild.2016.01.028

[11] H. Wang, "A reservation-based Smart Parking System", Ph.D Thesis, University of Nebraska-Lincoln, 2011.

[12] International Telecommunication Union ITU-T X.200, "Information technology – Open Systems Interconnection – Basic Reference Model: The basic model", July 1994.

[13] O. Dokur, "Embedded System Design of a Real-time Parking Guidance System", MSc Thesis, University of South Florida, October 2015.

[14] F. I. Shaikh, P. N. Jadhav, S. P. Bandarkar, O. P. Kulkarni, N. B. Shardoor, "Smart Parking System Based on Embedded System and Sensor Network", International Journal of Computer Applications in Technology **140**(12):45-51, 2016, https://doi.org/10.5120/ijca2016909532

[15] D. Thomas, B. C. Kovoor, "A genetic algorithm approach to autonomous smart vehicle parking system", Procedia Computer Science, 125, 68-76. 2018, https://doi.org/10.1016/j.procs.2017.12.011

[16] J. Hanzl, "Parking Information Guidance Systems and Smart Technologies Application Used in Urban Areas and Multi-storey Car Parks", Transportation Research Procedia, **44**, 361-368, 2020, https://doi.org/10.1016/j.trpro.2020.02.030

[17] M. Pietruch, "Innovative Technology for Quick and Easy Smart City Parking Solutions", Comarch Telecommunications Blog, May 08 2019, https://www.comarch.com/telecommunications/blog/innovative-technology-for-quick-and-easy-smart-city-parking-solutions/, last accessed 25.08.2020.

[18] Digiteum Team, "Parking IoT Technologies: About Smart Parking Systems", Smart Parking Solutions, November 27, 2019, https://www.digiteum.com/iot-smart-parking-solutions, last accessed 25.08.2020.

[19] W, Liu; H. Haitao, "Unlock the sharing economy for the parking sector", ETH Zurich Research Collection, European Transport Innovation Challenge: Creating Mobility Solutions, 2017, https://doi.org/10.3929/ethz-b-000171167

[20] https://github.com/openalpr/openalpr, Retrieved at 25.08.2020.

[21] D. C. He, L. Wang, "Texture Unit, Texture Spectrum, and Texture Analysis", IEEE Transactions on Geoscience and Remote Sensing, **28**, 509-512, 1990, https://doi.org/10.1109/TGRS.1990.572934

[22] J. Illingworth, J. Kittler, "A survey of the Hough transform", Computer vision, graphics, and image processing, **44**(1), 87-116, 1988, https://doi.org/10.1016/S0734-189X(88)80033-1

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 1, 810-818 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Ferromagnetic Core Reactor Modeling and Design Optimization

Subash Pokharel*, Aleksandar Dimitrovski

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, 32816, USA

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 13 January, 2021
Online: 05 February, 2021

Keywords:
Magnetic equivalent circuit (MEC)
Leakage permeance
Finite element analysis (FEA)
Multi-Objective (MO) Optimization
Genetic algorithm (GA)

ABSTRACT

This article presents an analytical model of a single-phase ferromagnetic core power inductor (reactor) based on a magnetic equivalent circuit (MEC). The MEC model consists of magnetomotive forces (MMFs) and reluctances for all flux paths: magnetic core and leakage flux paths. The MEC elements are found established on the characteristics of the ferromagnetic material and the reactor's dimensions. Then the inductances of the reactor are determined using the MEC for a group of Silicon Steel (Si-Fe) sheets, which are juxtaposed with the inductances obtained using the finite element analysis (FEA) method. This comparison corroborates the MEC presented here. Furthermore, for the unsaturated (linear) region of the B-H characteristics of the same sets of Si-Fe materials, the inductive reactance closed-form solution of the reactor is obtained as a function of the design parameters. As an application example of the presented analytical model, reactor design optimizations (single-objective and multi-objective) are formulated and solved based on the derived closed-form reactance expression resulting in a reactor not only optimized in reactance but also in terms of material use and size.

1 Introduction

This article augments the underlying work initially introduced in the 2019 North American Power Symposium (NAPS) [1].

Inductors (or the reactors in power engineering) are fundamental and most uncomplicated circuit elements that provide economical, sturdy, and efficient solutions to a spectrum of power system intricacies. The inductors have been used to limit fault and load current, compensate reactive powers, filter harmonics, damp the transients, and balance the loads, to mention a few applications in the electric power system. As a result of the significant expense and restricted usefulness for sufficient and optimal AC flow controls with FACTS devices, inductors are considered a superior alternative in power systems [2]. The use of reactors can be in conjunction with improving electric power quality, especially in the ever-changing contemporary power networks. The reactors can be connected either in series or parallel with the network based on a particular application. Contingent upon the voltage and power ratings, they can be either dry-type or oil-submerged sort. Usually, dry-type series reactors are made with air cores, and oil-immersed shunt types are made with ferromagnetic cores. Powdered core reactors, which are manufactured by the compression of the very fine particles of the magnetic materials, are preferred for high-reliability military and space applications because they are robust against shocks, vibrations, and nuclear radiations. Also, flux containment is better in powdered core compared to the ferromagnetic cores [3]. A simple schematic of a single-phase ($1-\Phi$) ferromagnetic core reactor for use in power systems is shown in Figure 1.

The inductor is symmetrical with non-ferromagnetic (air) gaps in the center of the inner leg. Over the gaps, there is AC winding connecting to the AC circuit. The gaps are used to limit the magnetic flux in order to prevent reactor's saturation from the rated operating supply. The core of ferromagnetic reactors is made of a material with high permeability that acts as an excellent conductor for magnetic flux. The ferromagnetic core's ability to conduct and concentrate the magnetic flux flow and, consequently, the reactor's inductance is dictated by its permeability. The relationship between the flux flow in the ferromagnetic core and the current supply through the coils of an electromagnetic device (like the reactor) is nonlinear. The rate-of-change (ROC) of flux for the field strength shifts from moderate to rapid to low when the operating condition shifts from the inception to linear to the saturated region. Generally, magnetic power devices' operation is not recommended in the saturation region, where more increment in MMF supply produces a minuscule change in flux. The operation of such devices in the saturation region increases the magnetizing current, responsible for

*Corresponding Author: Subash Pokharel; Email: pokharelsbs@ieee.org

increased losses and possible hazardous operating conditions. For some applications which work on the principle of the magnetic amplifier (MA), like continuously varying series reactor (CVSR) and fault current limiters (FCL), the inductance of the device is changed without any moving parts. The inductance variation is achieved with the help of DC bias supply by controlling the permeability and saturation of the ferromagnetic core as described in [2, 4]. The use of the laminated ferromagnetic core is the general practice to lower core losses; the ferromagnetic core in Figure 1 is shown as a solid rectangular core for simplicity only.Placing non-ferromagnetic gaps with magnetic properties similar to free space in the ferromagnetic core is also common. The effective permeability and the inductance of the device get reduced by such non-ferromagnetic gaps. However, the device's usable range of operation gets increased with dominant reluctance in the magnetic circuit. Likewise, the non-ferromagnetic gaps assume a crucial part in making the flux flow sensitivity lower to external conditions like temperature. The adequate selection of the air gap dimension is an essential job of the reactor designer because there is a need to struck an equilibrium between avoiding core saturation with a significant gap and imposed constraints to achieve the desired value of the inductance.

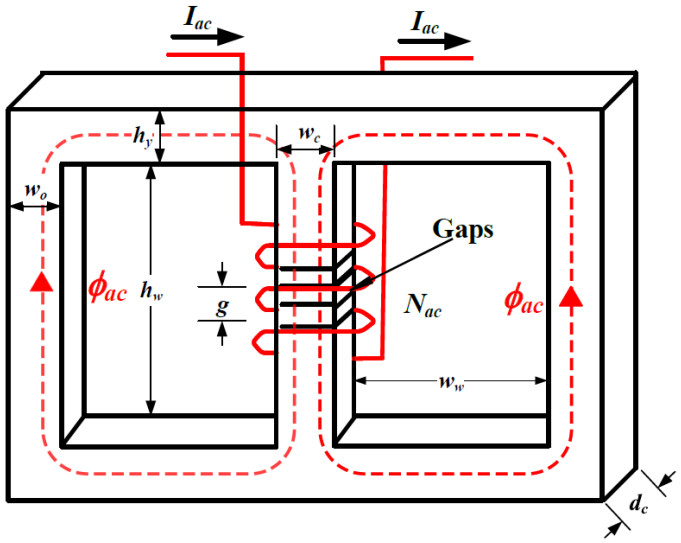


Figure 1: Ferromagnetic core reactor and its dimensions

FEA based methods are usually adopted for accurate parameter calculations and characterization of the electromagnetic devices. This numerical method can be tedious with a tremendous computational burden. It also lacks closed-form solutions. Therefore, it is not suitable for the repetitive design processes and dynamic analyses of magnetic power devices [5, 6]. In [7] and [8], the authors introduced the magnetic equivalent circuit (MEC) based method a long time ago. MEC is an alternative magnetic field modeling method for electromagnetic devices with a considerably reduced computational resources requirement. The MEC approach is based on the design inputs, material characteristics, and physical device geometry. It represents such devices with lumped magnetic circuit components: magnetomotive forces (MMFs) and reluctances. The MEC representation is considered coarser than the FEA but more acceptable than the electrical-equivalent lumped parameter models, which are based on fundamental equations. Hence, MEC is considered a bargain between those approaches in terms of computational resource need and accuracy [9, 10]. MEC approach is based on the design inputs, material specifications, and geometry of the device. The inductance calculation of a device is straightforward with this approach. MEC has the ability to accommodate local saturation effects; however, because of the eddy current and skin effects, the MEC representation becomes challenging. A comprehensive understanding of the device in consideration and a significant engineering judgment is required to generate a reluctance network with spatial discretization [9]–[13].

A great deal of engineering time and experience can be avoided with the optimization-based design process, which can produce a better design with less engineering effort [14]. A reactor's design can involve many performance targets, some more important than others, with a set of constraints. With multi/many-objective design optimization, the feasible solution space might be discontinuous and non-convex. The evolutionary algorithms are typically chosen for such problems. A large number of design evaluations are necessary to produce a Pareto optimal front, and approximate analytical models make such computations efficient. In this article, the constrained single objective (SO) and multi-objective (MO) optimization design formulations, coupled with the detailed MEC-based analytical model that captures the reactor's electromagnetic behavior, are presented. These design examples signify the importance of computationally efficient and accurate analytical MEC models for the reactor's design optimization.

The article is structured as follows. Section 2 proposes an adequately precise yet simple MEC for a gapped ferromagnetic core reactor. The inclusion of the fringing fluxes in the MEC proposed in the previous section improves its accuracy. The methodology for the confirmation of the MEC model is presented in Section 3. The next section follows the methodology introduced with a case study for a set of Silicon Steel (Si-Fe) ferromagnetic materials. The design optimization examples are presented in Section 5, facilitated by the reactance expressions obtained from the validated MEC. Based on the proposed MEC, its validation, and optimization examples, conclusions are drawn in Section 6.

2 Magnetic Equivalent Circuit

The dimensions of the simple single-phase gapped ferromagnetic core reactor are shown in Figure 1. The gaps (air) in the core make the leakage flux not influential for the reactor's magnetostatic characterization. For this reason, an arbitrary winding configuration choice is made for this device. The reactor features two windows of width w_w and height h_w. w_o and w_c are the widths of the exterior and middle leg, respectively. The central leg has non-ferromagnetic gaps of elevation g which stretches d_c into the surface, which is the same throughout the reactor. Both top and bottom yokes have the same height of h_y.

Figure 2 represents one of the simplest possible MECs of the reactor. The reluctances (resembling resistors from electric circuits) in the proposed MEC are positioned to interpret the accurate reactor physical structure. Similarly, the MMF source position (resembling voltage source from the electric circuit) in the MEC indicates the

winding position in the reactor. The ferromagnetic core reluctances are denoted by $R_1 - R_7$, dependent on the magnetic flux through the corresponding elements. In MEC, these reluctance are shown to be a function of flux.

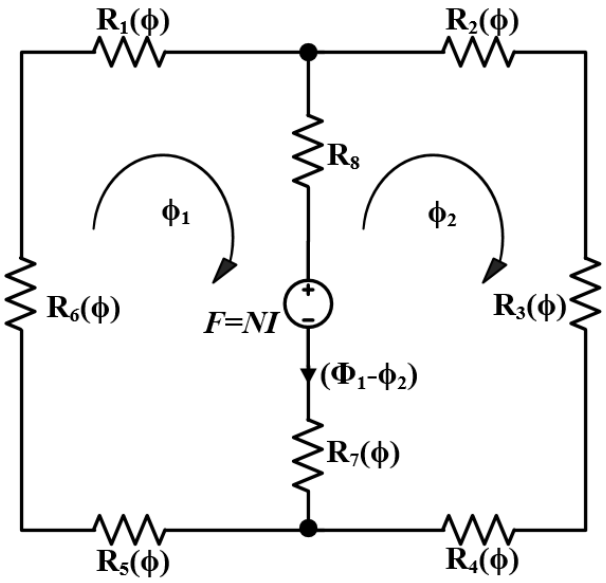


Figure 2: Magnetic equivalent circuit

However, the non-ferromagnetic air gap reluctance (R_8 in Figure 2) is fixed and does not depend on the corresponding flux. The MMF source in the middle leg of the proposed MEC is the product of the number of winding turns (N) and current supply through it (I), i.e., $F = NI$. In Figure 2, loop fluxes are represented by ϕ_1 and ϕ_2, making the inner leg branch flux to be $(\phi_1 - \phi_2)$ in the specified direction.

For a representative branch k, the generalized reluctance is:

$$R_k = \frac{l_k}{A_k \mu} \tag{1}$$

In equation (1), A_k and l_k are cross-sectional area normal to the flux flow direction and length of branch element, respectively, and μ is the permeability (absolute) of the material used. The absolute permeability depends on the type of flux path, based on which the reluctances can be categorized into:

2.1 Core Piece Reluctance

The reluctances of ferromagnetic core pieces are intrinsically nonlinear. The reluctance of each ferromagnetic element in the MEC depends on the material property, physical geometry, and supply through the winding around the core. The core piece's absolute permeability can be demonstrated as a function of either flux density (B) or flux intensity (H). Considering the symmetry of the reactor, based on (1), the characteristic core piece reluctances can be written as:

$$R_1(\phi) = \frac{2w_w + w_o + w_c}{2d_c h_y \mu_B(\frac{\phi}{d_c h_y})} \tag{2a}$$

$$R_3(\phi) = \frac{h_y + h_w}{d_c w_o \mu_B(\frac{\phi}{d_c w_o})} \tag{2b}$$

$$R_7(\phi) = \frac{h_y + h_w - g}{d_c w_c \mu_B(\frac{\phi}{d_c w_c})} \tag{2c}$$

Also, $R_1(\phi) = R_2(\phi) = R_4(\phi) = R_5(\phi)$ and $R_3(\phi) = R_6(\phi)$. Each branch reluctance in (2) has absolute permeability represented as the functions of flux density through them. In [15], the author describes the process of finding these absolute permeabilities, which can be summarized by the following expressions.

$$\mu_B(B) = \mu_0 \frac{A(B)}{A(B) - 1} \tag{3}$$

where,

$$A(B) = \frac{\mu_r}{1 - \mu_r} + \sum_{k=1}^{K} \left\{ \alpha_k |B| + \frac{\alpha_k}{\beta_k} log\left(\frac{e^{-\beta_k \gamma_k} + e^{-\beta_k |B|}}{1 + e^{-\beta_k \gamma_k}} \right) \right\} \tag{4}$$

In equation (4), B represents the magnetic flux density; μ_r and μ_0 are the relative and absolute permeabilities, respectively. Here, anhysteretic characteristics are considered with the reactor operating in the linear region. Moreover, α, β, and γ are the permeability function parameters for select core materials.

An iterative numerical search is needed for each branch of the MEC corresponding to the core piece ferromagnetic components to find a value of reluctance, given some initial values for the flux density. Figure 3 shows the flowchart summarizing this iterative process. The absolute permeability calculation is the first step of the iterative process. For mesh circuit analysis, permeability as a function of flux density is preferred. Using the absolute permeability ($\mu \equiv \mu_B(B)$), the reluctance of each piece can be calculated using (1). In the next step, each branch is represented in the standard general form: a series combination of an MMF source and the reluctance. Based on the complete circuit element formation, mesh analysis is carried out to find the fluxes through each loop. Except for the first iteration, as indicated in the flowchart, the stopping criteria for the iterative process are the maximum number of iterations and branch flux relative error between consecutive iterations.

2.2 Air Gap Reluctance

In the MEC model of Figure 2, the reluctance R_8 is the air gap reluctance. It is a constant reluctance without any dependence on the flux passing through it. Following (1), the reluctance of the air gap is given by:

$$R_g = \frac{g}{d_w w_c \mu_0} \tag{5}$$

While getting reluctance in (5), it is assumed that the cross sectional flux flow between the top and bottom of the air gap is streamlined. However, there exists a portion of a flux that flows in the vicinity of the air-gap corners spreading in all directions,

which is known as the fringing flux. Because of this flux, the effective cross-sectional area of the air gap is increased, resulting in decreased air gap reluctance. The fringing fluxes should be incorporated into the air gap reluctance computation to make the MEC highly accurate.

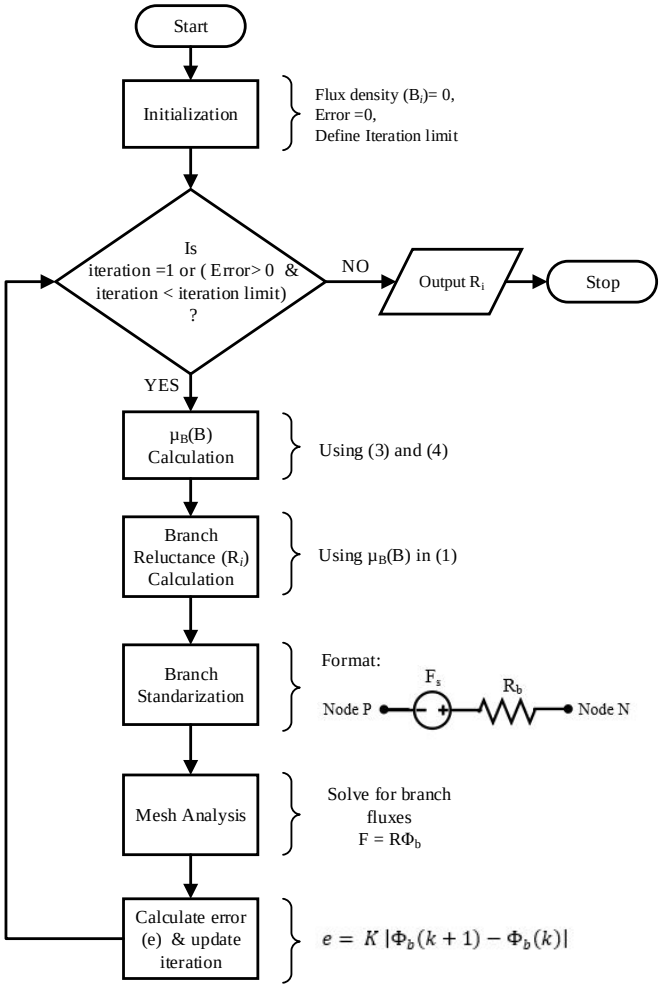


Figure 3: Iterative non-linear reluctance calculation flowchart

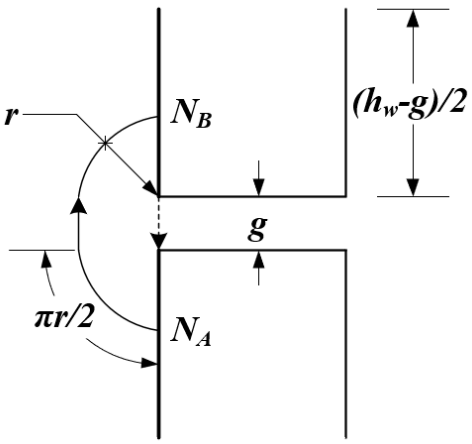


Figure 4: Fringing permeance calculation

The Ampere's circuital law, which relates the current to the associated magnetic field, can be described by (6) representing the magnetic field intensity H(A/m) over a closed path c to be equivalent to the current enclosed by that path ($I_{enclosed}$.)

$$\oint_c H \cdot dl = I_{enclosed} \tag{6}$$

The fringing effect can be captured by applying Ampere's law for the uniformly flowing flux from surface node N_A to N_B along the expanse of the flux route shown in Figure 4 as:

$$H(g + \pi r) = F \tag{7}$$

In (7), H and $F(= NI)$ are the magnetic field strength (intensity) and the MMF source respectively. Based on the relationship with H, the magnetic flux density ($B = \mu H$) can be represented as:

$$B = \frac{\mu_0 F}{(g + \pi r)} \tag{8}$$

The surface integration of the flux density gives the flux coming out of surface node N_A, given by (9).

$$\phi = \int BdS = \int_0^{(h_w - g)/2} \frac{\mu_0 F}{(g + \pi r)} d_c dr \tag{9}$$

Also, the fringing flux incorporated with the designated path can be linked to fringing permeance (P_{fring}) and MMF source as:

$$\phi = P_{fring} F \tag{10}$$

From equations (9) and (10):

$$P_{fring} = \frac{\mu_0 d_c}{\pi} log\left\{1 + \frac{\pi(h_w - g)}{2g}\right\} \tag{11}$$

For the central leg with air gap, the depth and the width are the same; therefore, the total fringing permeance is four times the P_{fring}. Hence, the air gap reluctance can be given by:

$$R_{air-gap} = R_8 = R_g \| \frac{1}{(4P_{fring})} = \frac{R_g}{1 + 4R_g P_{fring}} \tag{12}$$

3 MEC Validation Methodology

A meshed circuit analysis based on Kirchhoff's laws from electric circuits can be adopted to solve the proposed MEC once each MEC element is determined. This analysis gives flux through each branch and loop of the MEC. The physical law of energy conversation holds for the MEC. Therefore, a balance between the energy generated by the MMF source(s) and energy dissipated into the reluctances always exists. The energy balance principle can be exploited to find the overall equivalent inductance of the inductor directly. From [16] the reactor's inductance is:

$$L_{MEC} = \frac{1}{I^2} \sum_{k=1}^{n} \phi_k^2 R_k \tag{13}$$

Here, $1 \leq k \leq n$ is the index representing each branch in the MEC with a total number of branches being n; while R_k and ϕ_k are the reluctance and flux through the branch k respectively. Furthermore, the winding coil has current I following through it.

By performing different analyses (magnetostatic/transient) using the software packages based on FEA can give the inductance of a reactor. A three-dimensional (3-D) model of an inductor with minimal error tolerance set up can compute the inductance (apparent/incremental) accurately in the FEA software package.

The proposed MEC of the reactor is verified by comparing the equivalent inductances obtained from MEC as described in (13) using MATLAB®[1], and the FEA based approach using ANSYS Maxwell®[2]. The following case study presents the comparative analysis.

4 Case Study

A contextual analysis for the affirmation of the presented analytical model is carried out in this section, followed by the closed-form reactance calculation based on the validated model.

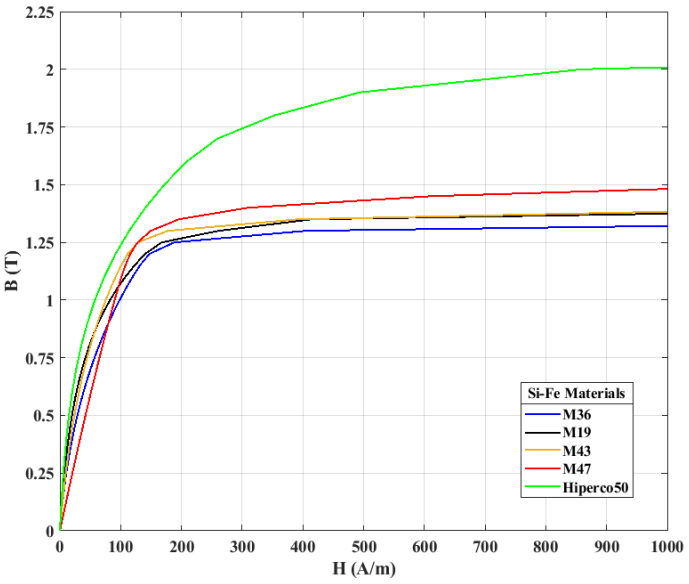


Figure 5: Candidate ferromagnetic material characteristics

4.1 MEC Validation

A single-phase reactor considered for the MEC validation has the parameters, as shown in Table 1.

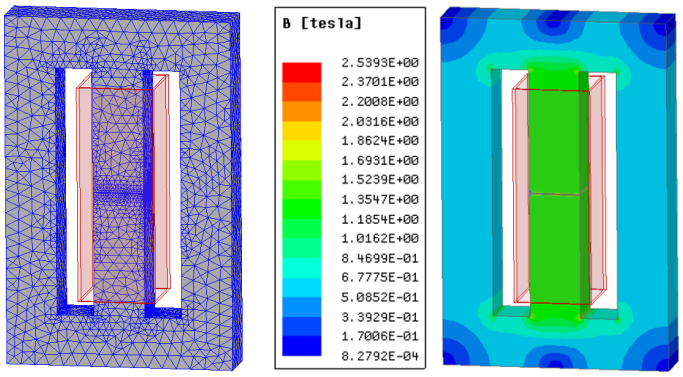
Table 1: Gapped core reactor parameters

Parameters	Symbol	Value
Window height	h_w (m)	0.374
Window width	w_w (m)	0.056
Leg width	$w_o = w_c$ (m)	0.076
Yoke/base height	h_y (m)	0.076
Core depth	d_c (m)	0.076
Number of turns	N	39
Rated current	i (A)	$25\sqrt{2}$

[1]https://www.mathworks.com/products/matlab.html
[2]https://www.ansys.com/products/electronics/ansys-maxwell

A set of magnetic core materials (Si-Fe) is considered with accompanying qualities (relative permeability, permeability function parameters, and maximum flux densities) extracted from [17]. The magnetic properties (B-H curves) for these materials generated according to [18] are shown in Figure 5. The two approaches described in the previous section are applied to calculate and compare the inductance of the reactor made up of each of the given materials.

For both linear and non-linear regions of the magnetization characteristics of Si-Fe, it is clear that the anhysteretic curve of Hiperco50 is considerably different from the others, as can be seen in Figure 5.



(a) Adaptive mesh plot (b) Flux density plot

Figure 6: FEA simulation Results for Si-Fe Hiperco50 core

Figure 6 shows 3-D FEA results for the reactor with the Si-Fe Hiperco50 core with the parameters from Table 1. The adaptive meshing has been applied to the model to make the simulation more accurate. The figure also shows the flux density distribution throughout the core, and it can be seen that the middle leg has the highest flux density.

For every core materials from the group, the inductances using both approaches and relative error between them are summarized in Table 2. The errors are within 1% for each of the materials. This minimal range of errors across a group of Si-Fe materials confirms the accuracy of the proposed MEC.

Table 2: Inductance comparison

Si-Fe Materials	Inductance (mH)		% Error
	MEC	FEA	
M19	7.853	7.880	0.348
M36	7.905	7.871	-0.427
M43	7.910	7.935	0.310
M47	7.831	7.904	0.925
Hiperco50	8.012	8.081	0.850

4.2 Reactance Calculation

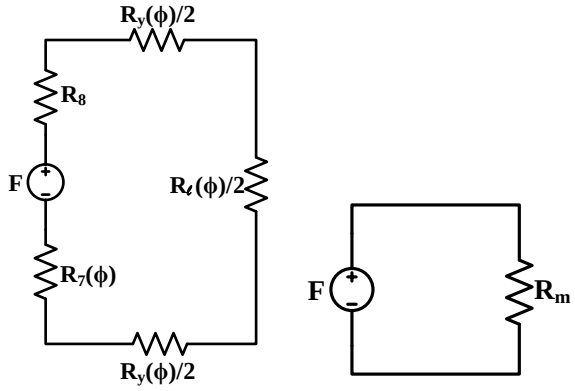
For a 60 Hz system, a single-phase inductor's reactance can be expressed as:

$$X_L = 120\pi L \tag{14}$$

where,

$$L = \frac{N^2}{\mathcal{R}_m} \tag{15}$$

here, $\mathcal{R}_m$ is the equivalent reluctance of the reactor. A circuit reduction on MEC can be applied to obtain it, as shown in Figure 7.



(a) Step I (b) Step II

Figure 7: MEC reduction

As the design of the reactor is symmetrical, a circuit simplification by the parallel combinations of the reluctances corresponding to the yoke and vertical leg transforms the MEC from Figure 2 to Figure 7a. Furthermore, the series reluctances in a single loop MEC in Figure 7a can be combined to achieve the simplest MEC as in Figure 7b. Accordingly:

$$\mathcal{R}_m = R_y(\phi) + 0.5R_l(\phi) + R_7(\phi) + R_8 \tag{16}$$

In Figure 7, $R_y \equiv \{R_1 = R_2 = R_4 = R_5\}$ and $R_l \equiv \{R_3 = R_6\}$. A further break down of (16) for a linear operation region (in the magnetization characteristic curve) returns:

$$\mathcal{R}_m = \frac{6.25 \times 10^5 g}{0.7854 d_c w_o + g d_c log\left(-0.5708 + \frac{1.5708 h_w}{g}\right)} + \frac{X\left\{-h_y g + w_o(w_o + w_w)\right\} + Y h_y(h_y + h_w)}{d_c h_y w_o} \tag{17}$$

A symbolic mathematical computation program called MATHEMATICA®[3] has been used to obtain this simplified expression. For different core materials, the coefficients X and Y in (17) are summarized in Table 3.

[3]https://www. wolfram.com/mathematica/

Table 3: Reactance coefficients

Si-Fe Materials	Coefficients	
	X	Y
M19	44.324	66.486
M36	58.996	88.494
M43	51.342	77.012
M47	86.369	129.554
Hiperco50	36.473	54.710

5 Design Optimization

5.1 Single Objective Optimization

An example of a single objective optimization problem is set up using the expression for the reactance of the reactor. The use of minimal material for the core is the objective while complying with several design constraints. Here, it is assumed that the reactor core is made up of Si-Fe M36 material. The design constraints include the target reactance, flux density limits for each element of MEC ($n = 8$), and minimum and maximum limits for the design parameters. The design parameter vector (19) consists of all design parameters. Their ranges and the initial points are summarized in Table 4.

$$\begin{aligned} \min_{\mathbf{x}} \quad & V = d_c\{2h_w w_o + w_c(h_w - g) + 2h_y(2w_w + 2w_o + w_c) \\ \text{s.t.} \quad & X_L = x_{design}, \\ & B_k \le B_{max}, \ \forall k \in n, \\ & \mathbf{x}_{min} \le \mathbf{x} \le \mathbf{x}_{max} \end{aligned} \tag{18}$$

In (18), $B_{max} = 1.25\ T$ is the maximum flux density in the linear region of the B-H curve for Si-Fe M36, and $x_{design} = 2\ \Omega$.

Table 4: Domain of design parameters

Parameters	$\mathbf{x}_{min}$ (m)	$\mathbf{x}_{max}$ (m)	$\mathbf{x}_{initial}$ (m)
w_o	0.0762	0.1016	0.0889
w_w	0.0559	0.0762	0.0635
h_y	0.0762	0.1016	0.0889
h_w	0.3739	0.5080	0.4572
d_c	0.0762	0.1016	0.0889
g	0.0015	0.0023	0.0020
w_c	0.0762	0.1016	0.0889

$$\mathbf{x} = [w_o \quad w_w \quad h_y \quad h_w \quad d_c \quad g \quad w_c]^T \tag{19}$$

A non-linear solver called 'Knitro' with YALMIP toolbox from MATLAB® has been used to solve this non-linear optimization problem. Table 5 summarizes the optimization results.

Table 5: SO optimization results

Parameters	Value (m)	Parameters	Value (m)
w_w	0.0559	$w_o = w_c$	0.0762
h_y	0.0762	h_w	0.3739
g	2.286×10^{-3}	d_c	0.0762

The optimal (minimum) ferromagnetic core material volume is found to be $1.0452 \times 10^{-2} m^3$. An approximate 'back-of-the-envelope' calculation for the volumetric minimization of the reactor with some educated guess regarding the constraints verifies the presented results. The air-gap (g) is very sensitive towards the reactance and the flux density in the core; however, it has a minimal impact on the reactor's total ferromagnetic volume. Therefore, the constraint set $\mathbf{x}$ other than g can be guessed towards the minimum limit to get the minimum ferromagnetic volume. If $\mathbf{x}$ is chosen to be x_{min}, the ferromagnetic volume would be $1.0457 \times 10^{-3} m^3$.

The optimal ferromagnetic volumes for different Si-Fe materials are summarized in Figure 8. In the figure, the ferromagnetic volumes are very near each other; however, the price difference between the materials will impact material selection.

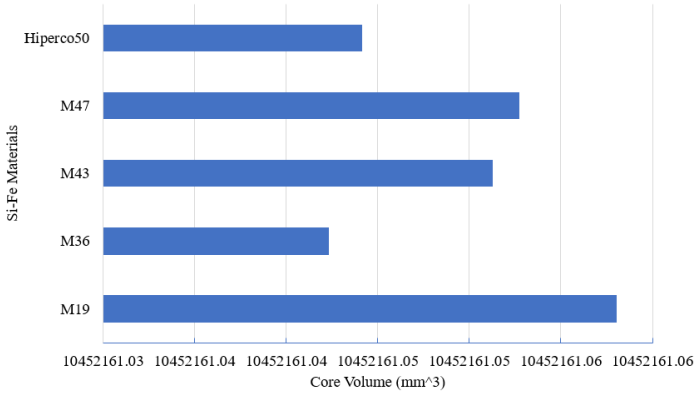


Figure 8: Optimal core volume comparison

5.2 Multi-Objective Optimization

Reactance is the most important characteristic of the ferromagnetic core inductor. In addition to the core material volume minimization, an additional objective of reactance maximization is introduced here. The multi-objective (MO) optimization searches a vast design space and is quite effective in finding optimal machine designs. An evolutionary population-based genetic algorithm [19] is implemented to solve a MO reactor optimization problem. A genetic algorithm (GA) can find multiple solutions from a population of solution candidates in one execution, which is not possible with classical optimization [14].

The goal of this design example is to come up with a single-phase power reactor design that has a reactance of at least X_{min}, and maximum ferromagnetic core flux density below B_{max}. It is desirable to minimize the reactor's ferromagnetic core material volume and maximize the reactance at the rated system conditions.

$$\begin{aligned} \min_{\mathbf{x}} \quad & \begin{bmatrix} V & \frac{1}{X_L} \end{bmatrix} \\ \text{s.t.} \quad & X_L \geq X_{Lmin}, \\ & B_m \leq B_{max}, \\ & \mathbf{x}_{min} \leq \mathbf{x} \leq \mathbf{x}_{max} \end{aligned} \tag{20}$$

The ferromagnetic core material used in this optimization problem is the same material used for single objective optimization problem described in the previous section (Si-Fe M36). The optimal reactor would have a reactance of at least 1.5 Ω (X_{min}) and the flux density of the 1.25 T (B_{max}) so that the operation of the reactor is within the linear region of the characteristics curve of the selected core material for the rated supply. In 20, the flux density has been bounded only for the middle leg flux density because it is the dominant flux density region. The gapped middle leg flux density can be approximated by:

$$B_m = \frac{\phi_m}{A} = \frac{Ni}{\mathcal{R}_m A} \approx \frac{\mu_0 Ni}{g} \tag{21}$$

The parameter bounds are the same as in the previous section, given in Table 4. The free parameter vector is represented by (19). The symbols in parameter vector are as defined in Table 1. It is assumed that the window spaces are sufficient for the winding.

The fitness function for the MO optimization problem can be defined as:

$$f = \begin{cases} \varepsilon(\bar{c} - 1)[1 \quad 1]^T, & \bar{c} < 1 \\ \begin{bmatrix} \frac{1}{V} & X_L \end{bmatrix}^T, & \bar{c} = 1 \end{cases} \tag{22}$$

In (22), ε–a small positive number (in order of 10^{-10})– does not have an influence on the optimization outcomes but is appropriate for the observation of the optimization progress.

And, $\bar{c}$ is the aggregate constraint of the MO problem, which is:

$$\bar{c} = \frac{1}{n_c} \sum_{i=1}^{n_c} c_i = \frac{c_1 + c_2}{2} \tag{23}$$

where, n_c is constraint number (here, $n_c = 2$). Once X_L and B_m have been evaluated using (14) and (21), constraint functions can be put together as (24):

$$c_1 = gte(X_L, X_{L_{min}}) \tag{24a}$$

$$c_2 = lte(B_m, B_{max}) \tag{24b}$$

where, $gte()$ and $lte()$ are greater-than-or-equal-to and less-than-or-equal-to functions, respectively. These functions are defined as:

$$\text{lte}(x, x_{max}) = \begin{cases} 1, & x \leq x_{max} \\ \frac{1}{1 + x - x_{max}}, & x > x_{max} \end{cases} \tag{25a}$$

$$\text{gte}(x, x_{min}) = \begin{cases} 1, & x \geq x_{min} \\ \frac{1}{1 + x_{min} - x}, & x < x_{min} \end{cases} \tag{25b}$$

The MO optimization is carried out using GOSET [20], a MATLAB-based genetic optimization toolbox. The optimization

has been performed with a population size of 1500 over 2000 generations with the specifications described by Table 6, and the parameter bounds given in Table 4.

Table 6: Genetic algorithm parameters

Parameters	w_o	w_w	h_y	h_w	d_c	g	w_c
Encoding	log	log	log	log	log	log	log
Chromosome	1	1	1	1	1	1	1

Table 7 summarizes the GA options carried out in the different steps of the MO optimization.

Table 7: Summary of GA options implemented

GA	Elitist nondominated sorting GA(NSGA-II) [21]
Selection	Tournament selection
Death	Random replacement
Gene repair	Hard limiting
Scaling	Offset scaling

Figure 9 represents the objective space at the end of the optimization, where each point represents the objectives of the corresponding design. The objective space plot is according to the fitness function definition. Therefore, reactance is plotted against the reciprocal of the ferromagnetic core material volume. The design points can be divided into nonviable, dominated, and nondominated design sets. The nonviable designs are near the origin and the negative axes of the objective space. The figure indicates the distinct design sets which are viable but dominated or nondominated.

Figure 10 shows only the nondominated designs with the axes representing the reactance and the material volume. This plot clearly shows the tradeoff between the objectives. A sample design on the Pareto optimal front is also indicated. The parameters representing the sample design are summarized in Table 8.

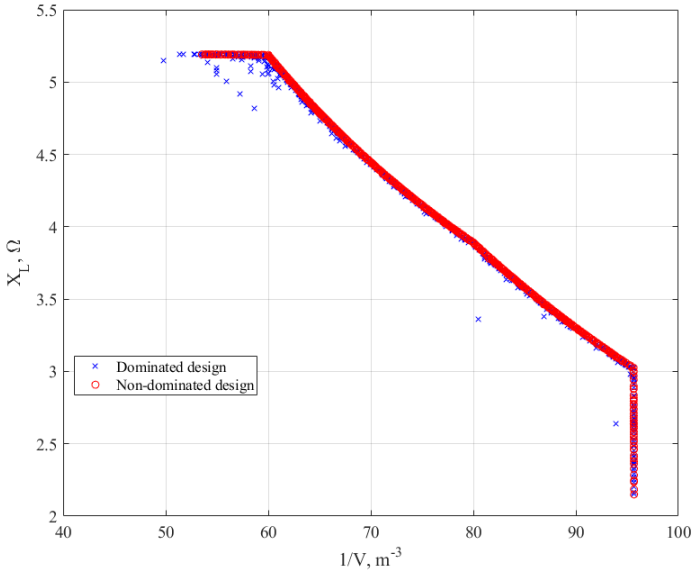


Figure 9: Objective design space

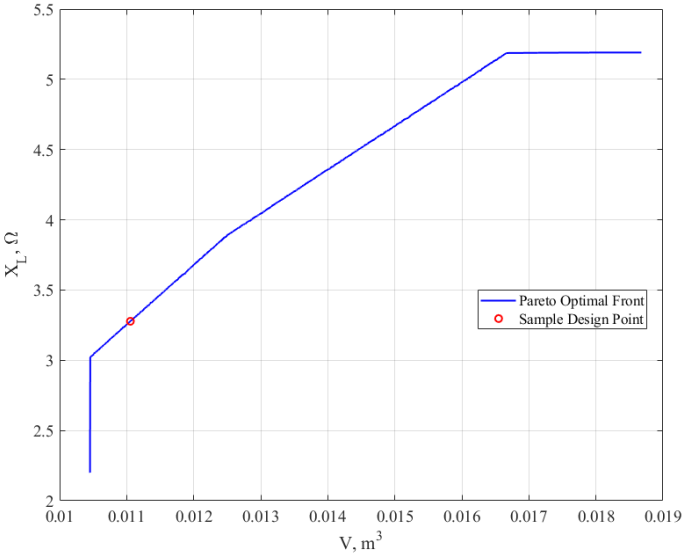


Figure 10: Pareto frontier boundary

Table 8: Sample design results

Parameters	Values (m)	Parameters	Values (m)
w_o	0.085	w_w	0.05589
h_y	0.076233	h_w	0.3739
d_c	0.0762	g	0.001524
w_c	0.0762		

For the sample design, the ferromagnetic material volume is 0.011176 m^3, and the reactance of the inductor is 3.32 Ω. This reactance is above the minimum limit X_{Lmin}, and the flux density is found to be 1.137 T, which is less than the flux density limit B_{max}.

To check the validity of the sample design obtained above, a 3-D FEA model is created again with the parameters from Table 8. The magnetostatic analysis from FEA shows that the reactance of the sample design is 2.96 Ω, which is ≈ 12% off from the MO optimization results. It is a fairly good result considering the simplicity of the model used, and more accurate than a spreadsheet-based design typically used by manufacturers.

6 Conclusions

In this article, a systematic analytical representation was proposed for a 1-Φ ferromagnetic core air-gapped inductor based on MEC. For a set of Si-Fe materials for core, the proposed MEC was substantiated after collating the inductances obtained by using the 3-D FEA method. The minuscule error of the inductance values throughout the entire group of materials validated the proposed model. Furthermore, when the ferromagnetic core is in the unsaturated (linear) region of the operation, an inductor reactance expression was derived as a function of design parameters. To illustrate one of the applications of the suggested systematic model, both SO and MO optimization problems were formulated and solved. They prove the MEC's applicability for obtaining optimal reactor designs for a specific application and highlight its efficiency and importance.

The future work will include reactor design improvements with more complex models, including more detailed flux leakages, and analyses (loss and thermal), as well as data from tests on the physical device with comparison of computational and experimental results.

Conflict of Interest The authors declare no conflict of interest.

References

[1] S. Pokharel, A. Dimitrovski, "Analytical Modeling of A Ferromagnetic Core Reactor," in 2019 North American Power Symposium (NAPS), 1–6, 2019, doi:10.1109/NAPS46351.2019.9000352.

[2] A. Dimitrovski, Z. Li, B. Ozpineci, "Magnetic Amplifier-Based Power-Flow Controller," IEEE Transactions on Power Delivery, **30**(4), 1708–1714, 2015, doi:10.1109/TPWRD.2015.2400137.

[3] C. W. T. McLyman, Transformer and inductor design handbook, CRC press, 2017.

[4] A. Dimitrovski, Z. Li, B. Ozpineci, "Applications of saturable-core reactors (SCR) in power systems," in 2014 IEEE PES T&D Conference and Exposition, 1–5, 2014, doi:10.1109/TDC.2014.6863404.

[5] A. Deihimi, "Improved model of saturated regions in Magnetic Equivalent Circuits of highly saturated electromagnetic devices," in 2008 11th International Conference on Optimization of Electrical and Electronic Equipment, 45–50, 2008, doi:10.1109/OPTIM.2008.4602385.

[6] M. Moallem, G. E. Dawson, "An improved magnetic equivalent circuit method for predicting the characteristics of highly saturated electromagnetic devices," IEEE Transactions on Magnetics, **34**(5), 3632–3635, 1998, doi:10.1109/20.717858.

[7] E. Lwithwaite, "Magnetic equivalent circuits for electrical machines," in Proceedings of the Institution of Electrical Engineers, volume 114, 1805–1809, IET, 1967.

[8] C. Carpenter, "Magnetic equivalent circuits," in Proceedings of the Institution of Electrical Engineers, volume 115, 1503–1511, IET, 1968.

[9] M. Amrhein, P. T. Krein, "3-D Magnetic Equivalent Circuit Framework for Modeling Electromechanical Devices," IEEE Transactions on Energy Conversion, **24**(2), 397–405, 2009, doi:10.1109/TEC.2009.2016134.

[10] S. D. Sudhoff, B. T. Kuhn, K. A. Corzine, B. T. Branecky, "Magnetic Equivalent Circuit Modeling of Induction Motors," IEEE Transactions on Energy Conversion, **22**(2), 259–270, 2007, doi:10.1109/TEC.2006.875471.

[11] J. Cale, S. D. Sudhoff, Li-Quan Tan, "Accurately modeling EI core inductors using a high-fidelity magnetic equivalent circuit approach," IEEE Transactions on Magnetics, **42**(1), 40–46, 2006, doi:10.1109/TMAG.2005.859439.

[12] B. N. Cassimere, S. D. Sudhoff, D. H. Sudhoff, "Analytical Design Model for Surface-Mounted Permanent-Magnet Synchronous Machines," IEEE Transactions on Energy Conversion, **24**(2), 347–357, 2009, doi:10.1109/TEC.2009.2016139.

[13] A. Taher, S. Sudhoff, S. Pekarek, "Calculation of a Tape-Wound Transformer Leakage Inductance Using the MEC Model," IEEE Transactions on Energy Conversion, **30**(2), 541–549, 2015, doi:10.1109/TEC.2015.2390260.

[14] S. D. Sudhoff, Optimization-Based Design, 1–44, Wiley-IEEE Press, 2014, doi:10.1002/9781118824603.ch01.

[15] S. D. Sudhoff, Magnetics and Magnetic Equivalent Circuits, 45–112, Wiley-IEEE Press, 2014, doi:10.1002/9781118824603.ch02.

[16] S. Pokharel, A. Dimitrovski, "A Gapless Ferromagnetic Core Reactor - Magnetic Equivalent Circuit and Inductance," in 2019 IEEE Power Energy Society General Meeting (PESGM), 1–5, 2019, doi:10.1109/PESGM40551.2019.8973400.

[17] S. D. Sudhoff, Selected Magnetic Steel Data, 443–444, Wiley-IEEE Press, 2014, doi:10.1002/9781118824603.app3.

[18] G. M. Shane, S. D. Sudhoff, "Refinements in Anhysteretic Characterization and Permeability Modeling," IEEE Transactions on Magnetics, **46**(11), 3834–3843, 2010, doi:10.1109/TMAG.2010.2064781.

[19] D. S. Weile, E. Michielssen, "Genetic algorithm optimization applied to electromagnetics: a review," IEEE Transactions on Antennas and Propagation, **45**(3), 343–353, 1997, doi:10.1109/8.558650.

[20] S. Sudhoff, Y. Lee, "GOSET Manual Version 1.03," Available from Scott Sudhoff, Purdue University School of Electrical and Computer Engineering, 2004.

[21] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, **6**(2), 182–197, 2002, doi:10.1109/4235.996017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 816-823 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Performance Analysis of Grid-Connected PV Rooftop, at Sakon Nakhon Province, Thailand

Supalak Sathiracheewin[1,*], Patamaporn Sripadungtham[2], Settakorn Kamuang[1]

[1]*Faculty of Science and Engineering, Kasetsart University Chalermphakiat Sakon Nakhon campus, Sakol Nakhon, 47000, Thailand*

[2]*Faculty of Engineering, Kasetsart University, Bangkok, 10900, Thailand*

ARTICLE INFO

Article history:
Received: 03 July, 2020
Accepted: 23 August, 2020
Online: 28 August, 2020

Keywords:
Weather-Corrected Performance
PV Rooftop
IEC61724

ABSTRACT

The performance ratio (PR) based on IEC 61724 standard is calculated under the influence of seasonal variations and the capability of the system. On the other hand, The National Renewable Energy Laboratory (NREL) of the U.S. Department of Energy proposed the weather-corrected performance ratio (PRcorr). This PRcorr index calculates the performance of the PV system which has taken the weather variations that influence the cell temperature into account. Both techniques were compared with the simulation program to analyze the PR index according to system conditions. The study site is the on-grid PV rooftop at Kasetsart University Chalermphakiat Sakon Nakhon (CSC). The PV panel is oriented to the southwest, 215 degrees azimuth. The angle of the inclination of the panel is 17 degrees. The result shows the trend of monthly PRcorr with little variations. At the same time, the PR value over the year has a lot of variations. The variation of PRcorr and PR is 2.39 and 5.07, respectively. PRcorr has low variability due to the correction of weather factors. The average cell temperature is an important variable. To calculate the average temperature of the panels, one year of data is needed to filter out distorted information. The system available condition is an important factor for the on-grid PV system at low voltage.

1. Introduction

Photovoltaic systems are currently being installed worldwide. The power received was estimated before the PV system was set up. And after installation, the energy generation potential of the PV system is measured. Due to the radiation and ambient temperature are different in each area. This paper is an extension of work originally presented in 16th ECTI-CON [1]. This paper examines the efficiency of the PV system in northeastern Thailand. The result will be a database of the energy generation potential of the PV plant in Sakon Nakhon province. According to the installation, the price per unit is much cheaper and at the same time the price of electricity is increasing. Manufacturers of equipment for the PV plant have increased, which has caused competition in the market. The amount of PV systems installation in Thailand has increased a lot, which is following the energy plan policy of the Thai government [2]. The installation of the PV system has spread throughout Thailand, both on the ground and rooftop. Therefore, many researches demonstrate the efficiency of the PV system, which influences people to install a clean energy plant.

The commonly used to assess the performance of a PV plant that is commonly used is the IEC 61724 standard. Following the IEC 61724 standard, the performance ratio (PR) is calculated by taking into account the radiation intensity, ambient temperature, wind speed, humidity, solar panels, and balance of system (BOS). The performance of the PV system depends on seasonal changes. Chong Li presented the calculation of PR with different PV types over the year in China [3]. The result shows that the performance ratio of poly-silicon PV increases from January to May and monthly array losses in summer are high. Then, the performance ratio of the PV system depends on location, PV module type, and the climate conditions. In 2019, Dhimish et.al. proposed the direction of the PV module for solar radiation that is important for energy production [4]. In addition, the azimuth angle is shown, which influences the maximum energy production. The suitable azimuth angle for the PV installation in the UK is -4 to 2 ºC (the azimuth angle in the south is 0 ºC). The data relation between irradiation and output power is to fitted regression. The R-square value is used to explain the variability data. High relative to R-squared, which is close to 1, means that the PV system generates the most electrical energy. Meanwhile, Singh demonstrated the

*Sathiracheewin.S, Sakon Nakhon, +66(0)42 725 033, Email: supalak.sat@ku.th

performance ratio obtained by PVsyst at Naresuan University, Thailand [5]. The performance ratio in each month is quite stable, which differs from the actual operation of this variation every month. Therefore, the weather variations employ to calculate the PR index based on the IEC 61724 standard. The different climatic conditions in each month affect the PR value.

The weather-corrected performance ratio (PRcorr) is presented by the national renewable energy laboratory (NREL) of the U.S. Department of Energy [6]. This report suggests that the calculation of PR values according to climate change gives a low PR value and a fluctuation value. The weather variation should calculate the only factors that affect the cell temperature. According to research by Basson, the objective of this research is to reduce the sensitivity to climate data [7]. The result shows that the temperature correction has a significant effect to reduce the seasonal changes. Compare to the PR method, its value has significant variation. Syahputra proposed the PRcorr calculation in the real operation in the tropics and subtropics compare to the PR technique of IEC 61724 standard [8]. Also, the temperature coefficient for power is presented in the real operating condition as well. The value of the temperature coefficient varies in different irradiation. As a result, the range value of PRcorr is greater than PR. Therefore, the PRcorr index defined by NREL is also useful for benefit financing.

The design of the PV system involves an analysis of production capacity, often through a simulation program. PVsyst is another simulation program that helps the designer to define from PV sizing, the capacity of the inverter, to energy production estimation [9]. For the simulation, NASA or PVGIS weather data was always used [10]-[11]. To calculate the power ratio correctly, the weather data in the installation area should be measured. Using the estimation data in the simulation model may cause the result to deviate from the actual value [12].

The northeastern region of Thailand is in an area that receives good solar radiation. From the data of the solar map potential improvement project for Thailand 2017, it is apparent that the average monthly daily radiation intensity is 20 to 22 MJ / m^2 per day [13]. In particular, the Kasetsart University Chalermphakiat Sakon Nakhon (CSC) is a campus of Kasetsart University in the province of Sakon Nakhon located in the northeastern part of Thailand with 17.28661 latitude and 104.1061 degrees longitude, as shown in Figure 1. The average solar energy is 17.91 MJ / m^2 per day, and in April the average solar energy is up to 20.84 MJ / m^2 per day. Topograpgy of Sakon Nakhon city is a flat plain with no tall buildings, and there is no large industry that causes air pollution. Therefore, the province of Sakon Nakhon is suitable for the installation of the PV system. Currently, both the public and private sectors in Sakon Nakhon Province are interested and have installed more PV systems. The production of electric power by solar cells, not only reduces the dependence on the main grid but also responses to the government policy. But many people want to know the energy efficiency of the PV system, which has different weather conditions in each area. Appropriate assessment of energy production potential to support decisions made by persons interested in the PV system.

Therefore, this research proposes an efficiency evaluation of a PV production system, based on the IEC 61724 standard and the weather-correct performance ratio of the NREL concept. The result from both concepts were compare to the PVsyst analysis result. The measure data in the study area is input to analyze the significant factors that influence the energy production of the on-grid PV system at low voltage.



Figure 1: Location of Sakon Nakhon, Thailand on Google map.

2. Solar Rooftop at CSC

CSC in Sakon Nakhon covers an area of 6.4 square kilometers. The Faculty of Science and Engineering is part of the University that teaches and provides useful research results to society. One of the research sections conducts clean energy research such as on-grid PV rooftop. The schematic diagram of the grid-connected PV system is shown in Figure 2. The PV system consists of 10 modules of the series, a single-phase 3.6 kW inverter and other equipment used for the installation of the PV system. For installation details, the PV module faces southwest, 215 degrees solar azimuth (north is 0 degrees angle). The angle of inclination of the plate is 17 degrees. The energy output from the PV generation flows at low voltage into the load center of the building. The rated voltage in Thailand is 230 volts.

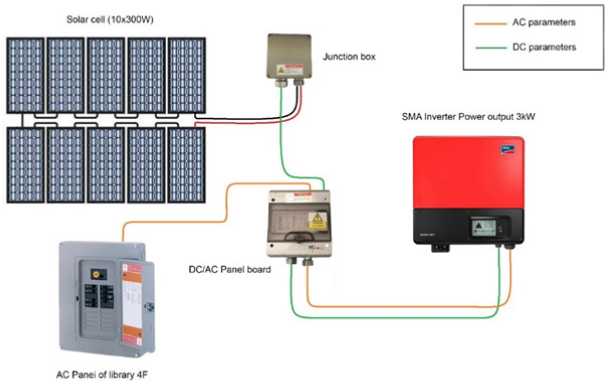


Figure 2: The schematic diagram of 3 kWp PV system

The specification for PV module and inverter is listed in Table 1. The cell type of PV module is poly-crystalline and the total area of PV panels is approximately 19.3 m^2. PV panels send DC power to the junction box via a cable. The cable made of aluminum with a cross-sectional area of 4 mm^2. The junction box contains a DC fuse (20 kA, 2 POLES), after which the DC power flows into the inverter and the AC power is obtained. The AC power flows back into the AC fuse in the junction box and then goes to the load

center. The inverter is a grid-tie inverter of SMA type, Sunny Boy 3600TL-21.

Table 1: PV Module and Inverter Specifications

PV module	Spec.	Inverter	Spec.
Max. Power (Wp)	300	Rated AC Output Power (kW)	36.8
Open-circuit Voltage (V)	45.14	Maximum Current (A)	16
Short-circuit Current (A)	8.74	AC Nominal Voltage (V)	230
Max. Voltage (V)	37.01	Max. Input Current	15
Max. Current (A)	8.12	Input Voltage (V)	175-500
Temp. Coefficient	-0.4574 (%/°C)	Maximum Efficiency (%)	97

According to the SMA inverter, information about electrical energy on the inverter can be found on the website, as the data is transmitted via the university's intranet. The information on this website contains only some information, this is not enough to analyze the performance of the PV system. Consequently, sensors were installed to measure relevant data such as the radiant level, ambient temperature, module temperature, AC power and DC power being recorded at the same reference time. Large amounts of data may contain data discrepancies. Therefore, both the measuring instruments as well as the data acquisition tools are required to have low tolerances [14]-[15].

3. Performance Index of PV System

3.1. IEC 61724 Methodology

For analysis of system performance under standard IEC 61724. This standard has been used as a reference in many studies [16]. The indices specified in the standards are used to measure system performance, including assessing the losses that occur in the system.

Reference Yield (Y_R) is referred to as the reference irradiance generate power (H_t) under standard test condition (G_{STC}) (kwh/m^2), usually 1 kW/m^2. It is the calculation of the energy that is generated by this light intensity. The relationship is shown in equation (1).

$$Y_R = \frac{H_t}{G_{STC}} \qquad \left(kWh.kW_p^{-1}\right) \tag{1}$$

Array yield (Y_A) is the ratio of daily energy generation from PV array (E_{DC}) divided the kW of installed PV array (P_0). This parameter calculates the energy output per day and uses this value to calculate the average per month. Therefore, the average power of the DC part per day of each month is obtained. Y_A value is given by equation (2).

$$Y_A = \frac{E_{DC}}{P_0} \qquad \left(kWh.kWp^{-1}\right) \tag{2}$$

Final Yield (Y_f) is defined as the daily energy generation from PV plant (E_{AC}) divided by the total kW of installed PV array (P_0). This energy refers to the energy gained after passing the inverter or the AC power part. Y_f is calculated using equation (3)

$$Y_F = \frac{E_{AC}}{P_0} \qquad \left(kWh.kWp^{-1}\right) \tag{3}$$

Performance ratio (PR) indicates the actual energy production and the energy generated by the incident irradiation. Therefore, PR is the ratio of Final Yield (Y_f) to Reference Yield (Y_R). This parameter shows the efficiency of the system during solar conversion until the final electrical power is reached. Therefore, the PR index is influenced by all environmental factors and including the efficiency of the PV system. PR calculation is given by equation (4).

$$PR = \frac{Y_F}{Y_R} \times 100\% \tag{4}$$

According to the IEC 61724 standard, the system loss also occurring in the PV system is calculated. The loss is caused by the equipment in the system, such as the PV panel, cable wire, inverter, and other factors that cause the loss to increase.

The standard IEC 61724 also shows the loss value in the system as in equation (5). Array Capture Loss (L_A) is the loss from PV array. It the different value between Y_R and Y_A.

$$L_A = Y_R - Y_A \qquad \left(kWh.kW_p^{-1}\right) \tag{5}$$

Another loss value is system loss (L_S) that depended on equipment effect in the system such as inverter, wire and PV panels. It is calculated in equation (6).

$$L_S = Y_A - Y_F \qquad \left(kWh.kW_p^{-1}\right) \tag{6}$$

3.2. Weather Corrected Performance Ratio

According to the technical report of the national renewable energy laboratory (NREL) [6], the weather-corrected performance ratio (PRcorr) was presented. PRcorr is modified from the PR concept of IEC 61724. PRcorr was created by reducing the impact of fluctuating seasonal conditions, which also causes the PR value to fluctuate. Therefore, PRcorr has improved the formula in IEC61724. PRcorr focuses on the technique to obtain the cell temperature, which affects efficient PV generation. The formula PRcorr calculates the cell temperature based on irradiation and ambient temperature. The equation is shown in (7).

$$PR_{corr} = \frac{\sum_i EN_{AC\_i}}{\sum_i \left[P_{STC}\left(\frac{G_{POA\_i}}{G_{STC}}\right)\left(1 - \frac{\delta}{100}\left(T_{cell\_typ\_avg} - T_{cell\_i}\right)\right)\right]} \tag{7}$$

where

PR_{corr}: corrected performance ratio (%)

EN_{AC}: measured AC power generation (kW)

P_{STC}: total of installed modules' power rating at standard test condition (STC)

G_{POA}: the radiation measurement at plane of array (POA) (kW/m^2)

G_{STC}: irradiance at STC (1,000 W/m^2)

T_{cell}: cell temperature computed from measured meteorological data (°C)

$T_{cell\_typ\_avg}$: average cell temperature computed from one year of weather data using the project

weather file (°C)

δ: temperature coefficient for power (%/°C, negative in sign) that corresponds to the installed modules

i: a given point in time (day)

NREL shows that the use of an adapted reference cell to measure irradiance avoids the need to correct spectral variations as well. In the meantime, other weather effects such as snow losses, environmental pollution losses or the effects of variable irradiance on efficiency are not corrected. Corrections for these additional weather effects could lead to more consistent results.

3.3. Performance Ratio by PVsyst Analysis

In general, the technique and analysis of the performance of the PV system needs to be calculated before the actual installation of the system. Therefore, the use of the simulation programs to analyze a process is necessary, which leads to losses in the system, energy and production performance. The PVsyst simulation program can analyze these parameters and generate solar radiation and other weather data in the study site.

Therefore, when analyzing the performance of the PV system, the correct information needs to be entered compared to the actual value, including cable line, the inverter and PV module type. The installation properties are the actual value in the installation area, such as the azimuth angle and the tile angle of the module. The accuracy of the weather data entered into the simulation program must come as close as possible to the actual conditions and then the high accuracy output will be obtained. The analysis with this simulation program also analyzes the weather-corrected performance ratio [17].

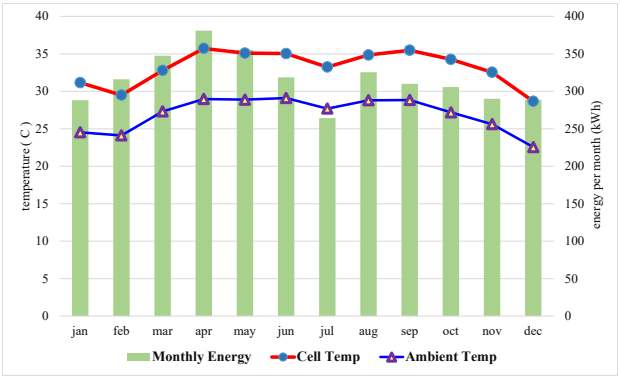
4. Result and Discussion

The results of this experiment are based on annual data for 2017. Figure 3 is the 3 kWp PV plant with the temperature sensor and radiation sensor. All data are collected to data logger. An energy meter is installed for the DC and AC power measurement of the PV system.

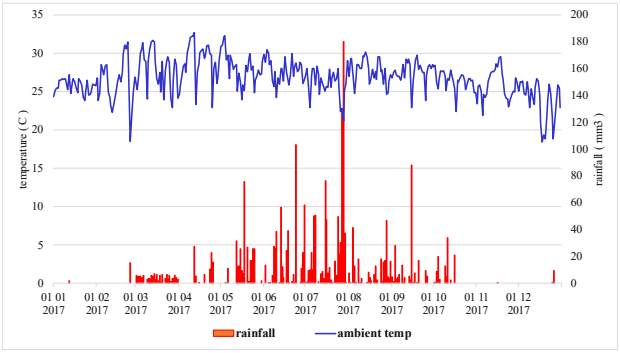


Figure 3: Plant Study at CSC, Sakol Nakhon

Figure 4 (a) shows the average cell temperature, the average ambient temperature, and the monthly energy of the PV system. The average monthly temperature data clearly shows that the panel temperature is higher than the ambient temperature. In 2017, the average ambient temperature is 26.97 °C and the average cell temperature is 33.19 °C. Figure 4 (b) shows the amount of rainfall every day. Note that the heavy rains occur in June, July, and August. As a result, the average ambient temperature also drops. In September and October, the rainfall begins to decrease and then the average ambient temperature gradually increases. In Sakon Nakhon province, from 28 to 29 July 2017, precipitation for both days was up to 310 mm^3, heavy rain caused a sudden flood. The result is a widespread power outage in Sakon Nakhon province. A major power outage meant that the inverter was also disconnected from the PV system. Therefore, the grid-connected PV system has dropped out of the main grid and does not need to generate power for a while.



(a) Cell Temperature and Ambient Temperature Compare to Monthly Energy

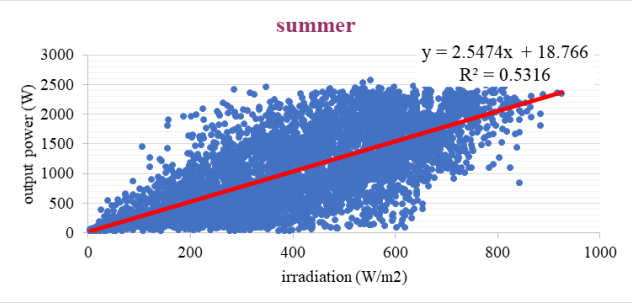


(b) Ambient Temperature and the Rainfall

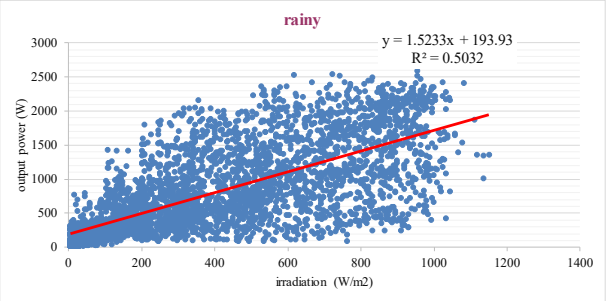
Figure 4: The Weater Factors and Monthly Energy

Because of the PV module faces southwest, 215 degrees solar azimuth, the irradiation data are grouped in the solar azimuth or sun path throughout the year. The sun path refers to the sun that appears daily in the sky. The path of the sun has different values (degrees) every day and causes seasons. From the Sun Orbit database at CSC, each month was grouped with similar orbits. It refers to the sunrise at 6 a.m. and the sunset at 6 p.m. The summer group consists of February, March, and April with an azimuth of the sun between 76 to 97 degrees in the east and 261 to 284 degrees in the west. Rainy season groups are May, June, July, August and September. The sun orbits of this group are between 71 and 80 degrees in the east and between 277 and 292 degrees in the west. The winter group with similar approaches to the movement of the sun are January, October, November, and December with an azimuth angle between 106 and 112 in the east and between 244 and 256 in the west. Taking into account the average movement direction of the sun per month, it can be divided into three types as described above. Each group has daily routes orbit that are close to

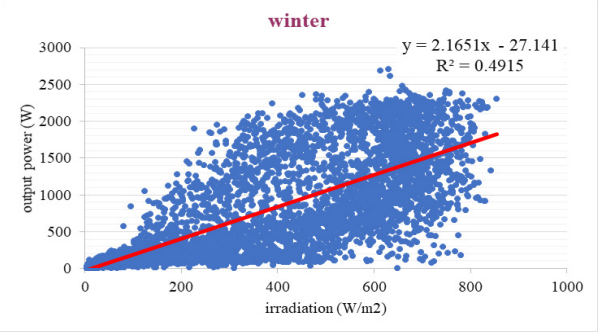
each other. It should be obtained a similar sun intensity. The three groups season relate to the irradiation, shown in Figure 5.



(a) The Summer Group



(b) The Rainy Group



(c) The Winter Group

Figure 5: The Relationship between Solar Irradiation and Output Power

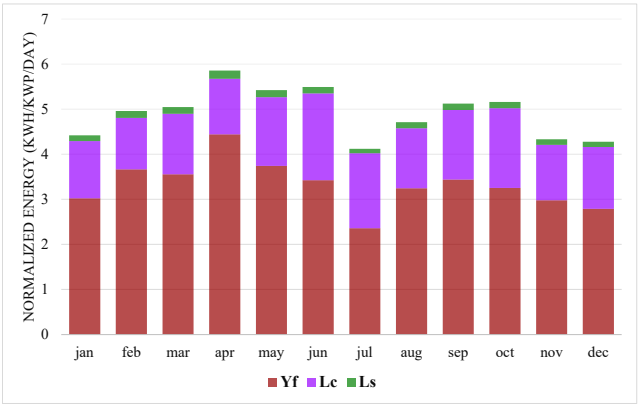
Figure 5 shows the relationship between solar irradiation (W/m^2) and electric power produced (W) according to the monthly groupings mentioned above. Curve fitting is a statistical method used to analyze the data. The R-square indicates the degree of proximity of the data. As the R-squared value approaches 1 or 100%, it means that all data points would fall on the adjusted regression line. According to Figure 5(a), the R-squared of the summer group is higher than the others. The R-squared of Figure 5(b) and Figure 5(c) is 0.5032 and 0.4915, respectively. The winter group has a minimum of R-square value. Since the energy output is approximately linear for irradiation in the study area. This means that the electrical energy generated by the PV system at this time is most valuable.

According to the monthly energy of CSC plant, the electrical energy generated by the PV system is compared with the R-square in Table 2. The summer group has the highest average monthly production capacity of about 348 kWh, while the winter group has the smallest R-square. The average energy during the winter group generated is 293 kWh/month. Therefore, the trend of maximum amount of energy will occur in Sakon Nakhon this month.

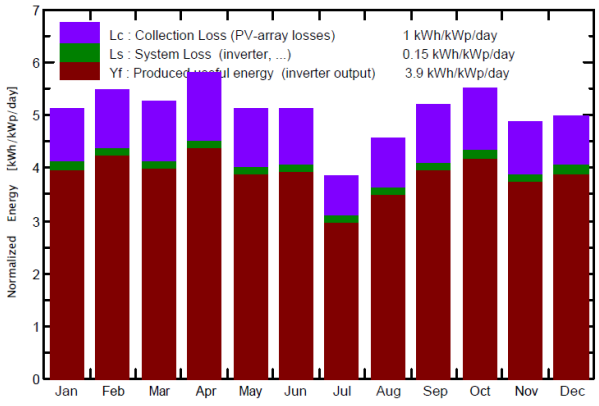
Table 2: Comparative Average Monthly Energy and R-square

Season group	average monthly energy (kWh)	R-square
Summer	348	0.5316
Rainy	315	0.5032
Winter	293	0.4915

For the performance analysis according to IEC 61724 standard, Figure 6 (a) shows the average monthly value of Y_f, L_c and L_s which is 3.33, 1.45 and 0.14 respectively. Compared to the result of PVsyst as shown in Figure 6 (b), Y_f, L_c and L_s which is 3.9, 1.0 and 0.15 respectively. Both techniques are similar energy generation such as the low Y_f is in July. The maximum system loss occurred in April, the IEC method shows 0.19 kWh/kWp/day and PVsyst is 0.169 kWh/kWp/day. The trend of the energy value obtained from the PVsyst program is higher than in the IEC technique, where the actual value is used for the calculation. Therefore, using the prediction program, the energy estimate is inflated. Overall, the amount of energy generation from PVsyst is high. However, the PV array loss is less than with the IEC technique, possibly the dust on the PV module. The system has been installed for almost 3 years and the panel has not been cleaned. Estimating these parameters will input all data as monthly averages. The result of the analysis may be inaccurate or deviate from the actual value.



(a) Y_f, L_c and L_s based on measure data



(b) Y_f, L_c and L_s from PVsyst analysis

Figure 6: Result from the measure data compare to PVsyst analysis

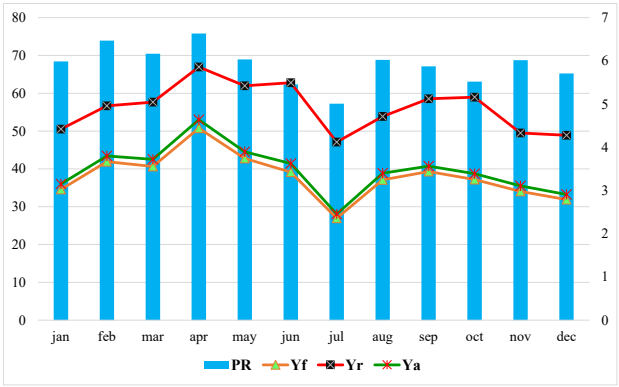


Figure 7: Performance of PV System at CSC

Figure 7 shows the calculation value of the PR index based on Y_F, Y_A, and Y_R data. The graphic shows a comparison of the performances of the grid-connected PV system. For each month of Y_F, Y_A, and Y_R, the line curves are similar. Y_F and Y_A are closed values, as this depends on the efficiency of the inverter and loss in cable. As a result, the average efficiency of the inverter is 96%. The maximum Y_R in April is 5.86 kWh.kW^{-1}, which according to the monthly energy value is 381 kWh. April is considered a hot season. The average temperature is generally high and most skies are cloudless. While July which in the rainy season has a low of Y_R and similar to monthly energy as well. The average Y_R in July is 4.12 and the monthly energy is 264 kWh. On the other hand, Y_F and Y_R lines are separated from each other. It means Y_R can covert to Y_F depend on the system efficiency. It also represents the amount of loss that has occurred in the PV system. The loss consists of the loss of the PV array loss, cable loss and the system loss caused by the efficiency of the inverter.

Every month the average power ratio is calculated. The maximum PR in April, February, and March is 75.8%, 73.94%, and 70.47%, respectively. On the other side, the minimum PR in July is 57.23%. At the end of the year, which is the winter season, monthly energy production tends to decline. Note that the low PR period is during rainy and winter season. PR parameter reflects the performance of the system for converting solar energy into electricity. A high PR also indicates a small loss in the system.

Based on IEC 61724 standard, the PR calculation is caused by the effects of all environmental factors on the PV system. As a result, the PR value fluctuates quickly. Therefore, some researchers do not think about these effects. The PR index should have suitable values to assess the potential of the system. The weather-correct performance is presented by NREL, which corrects the cell temperature. The irradiation and ambient temperature data are used to recalculate the performance of a PV system that directly affects the cell temperature. At the same time, the PVsyst has been used to calculate PR values as well. By creating a simulation model with system unavailability conditions of the main grid system. It indicates the percentage of the period in which the system is unavailable, consisting of 0%, 5%, and 10%. According to the measurement data, there are outage events in a few months. For January, September, October, November, and December are defined to the month of system unavailability condition, as shown in Table 3.

Table 3 shows the PRcorr, PR index based on measured data and PR index analysis by PVsyst. The maximum PRcorr in April is 82.05%, as PR is 75.8%. But in PVsyst calculation, the maximum PR with system unavailability 0%, 5%, and 10% are 78.9% in December, 75.7% in January and 44.7% in February and July respectively. For average PRcorr, PR and PR of PVsyst at 0% are 77.34%, 67.52%, and 77.15% respectively. The variation calculation value is based on statistical techniques. Therefore, the standard deviation is chosen. For standard deviation analysis, the PR and PRcorr are 2.39 and 5.07 respectively.

Table 3: Comparative Average Monthly Performance Ratio (%)

Month	PRcorr	PR	System Unavailability (%)		
			0	5	10
January	74.51	68.42	78.10	75.70	57.40
February	76.37	73.94	77.70	75.30	77.70
March	77.18	70.47	76.10	64.70	76.10
April	82.09	75.80	75.90	73.60	75.90
May	77.96	68.97	76.50	74.10	76.50
June	79.86	62.36	77.20	64.30	77.20
July	73.86	57.23	77.70	66.70	77.70
August	79.67	68.85	77.50	75.00	77.50
September	77.82	67.11	76.50	74.10	57.90
October	76.20	63.07	76.50	74.20	58.50
November	77.49	68.74	77.20	63.80	59.30
December	75.01	65.24	78.90	67.30	61.00
Average	77.34	67.52	77.15	70.73	69.39
Std.	2.39	5.07	0.89	4.87	9.39

PRcorr is significantly higher than PR and the line trend is quite stable, which does not change significantly during the year. While the standard deviation value of PVsyst is tended to increase with the strong value of system unavailability. The PRcorr value varies slightly because the cell temperature was influenced by radiation and ambient temperature. While other weather factors based on the NREL concept are un-correct such as pollution, snow or shading. PRcorr is therefore very valuable and encourages people to be more interested in PV production.

For the other hand, PVsyst analysis, the average PR with 0% system unavailability is similar to PRcorr, but the standard deviation value has very little variation. The standard deviation increases with high system unavailability because the energy output from the PV system cannot flow to electrical load, such as fault event or temporary power from utility. Therefore, the PR with 0% system unavailability is calculated based on the weather corrected condition.

The PR value is a performance indicator that combines the effects of various factors such as the intensity of light falling on the PV module, cell temperature, shading, or device malfunctions in the system. When analyzing these factors, some factors are not related to the weather. However, when weather factors are used to calculate the PR value, it is low and changes rapidly.

Use the PVsyst to determine the monthly energy. The monthly energy output from the PV system is compared with the result of PVsyst, shown in Table 4. The maximum energy in April is 381 kWh, similar to PVsyst, the maximum output in April is 395 kWh. When calculating deviations from the actual values, the maximum is 19% and the minimum is 13%. For average monthly with 10% system unavailability is closely the true value. Because it is stipulated that the PV system will not be able to supply electricity

according to the actual incident (fault in the utility grid). Therefore, a high energy forecast is created for electricity from a PV system that can deliver at any time. Therefore, the use of the analysis program must take these events into account. These events often occur in areas with poor electrical reliability.

Table 4: Comparative Monthly Energy (kWh) with PVsyst Prediction

Month	Monthly Energy	System Unavailability (%)		
		0	5	10
January	288	371	330	273
February	316	358	358	358
March	347	372	372	372
April	381	395	395	395
May	355	363	363	363
June	319	355	355	355
July	264	278	278	278
August	326	327	327	327
September	310	357	311	271
October	306	391	345	299
November	290	338	298	259
December	288	365	321	282
Total (kWh)	3790	4269	4054	3831
Avg. per month	316	356	338	319
Error (%)	-	19	15	13

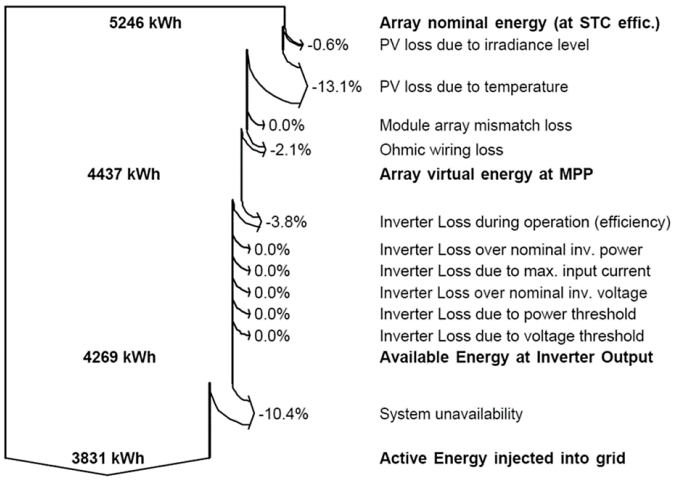


Figure 8: output from PVsyst analysis

The result from PVsyst,as shown in Figure 8, power loss calculation occurs in the PV system that has various types of loss. Nearly 13.1% power loss due to temperature. The high temperature leads to an increase in power loss and then to a reduction in the output power. The parameters used to analyze the PRcorr are derived from the device specification and the actual data in the experimental area. The important parameter is the average cell temperature (Tcell_typ_avg) each month which for average annual is 38.84 ºC. The temperature coefficient for maximum power (δ) is -0.004574 which applies for over the year.

For the on-grid PV system inclement weather such as rain, strong wind will cause the main power system to always fail, such as voltage drop or momentary power outage. These unusual weather conditions cause the inverter protection to work. The inverter's function is to disconnect the circuit so that the energy output of the inverter is zero. Figure 9 demonstrates the real power (P) value with interval time (15 minutes). The red circle is the point of the momentary power outage. Sometimes a power failure occurs for a very short period, such as 2 or 5 minutes. The inverter cuts off the circuit and then turns it on when the main grid can be re-energized. Therefore, the average AC data, for interval time 15 minutes, is reduced. At the mentioned point, the real power (kW) decreases abnormally. These unusual weather conditions cause the inverter protection system to work. Therefore, some monthly energies are reduced even though the cell temperature is low or irradiation is high.

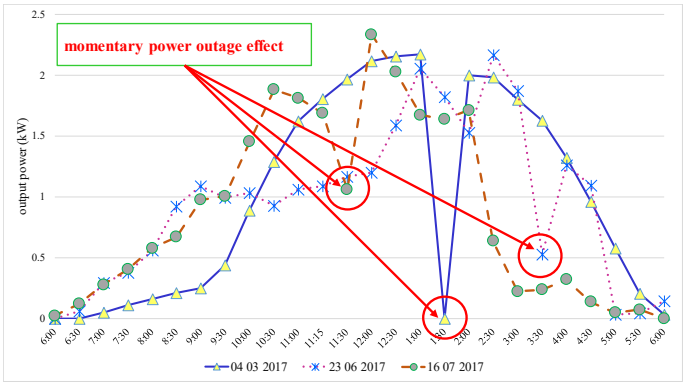


Figure 9: The monentary power outage effect in the on-grid PV system

5. Conclusion

This article presents the technical calculation of the conversion potential of solar energy up to the final electrical power. The technique used is the IEC 61724 standard and the index proposed by the NERL. PVsyst is a simulation model to estimate the energy and the performance ratio with the real conditions. The study site is Kasetsart University Chalermphakiat Sakon Nakhon (CSC). The PV equipment consists of the total PV installation is 3 kWp and 3.6 kW single phase of grid-connected inverter. The energy generation from the PV system flow to the load center of the building at CSC.

The irradiation data show that the R-squared distribution is high, about 0.5316, in February, March, and April, which means high energy production. In April, the maximum monthly energy is 381 kWh. The minimum energy generation occurs in July, about 264 kWh, because of heavy rain and no more the irradiation. For the average of Y_R, Y_A and Y_F is 4.91%, 3.47%, and 3.33%, respectively. The maximum performance ratio of the PV system is 75.80% in April. The average PR is 67.52%, which is influenced by weather factors. Compare to the PRcorr index, which is based on the NREL concept, the maximum PRcorr in April is like the PR index at 82.09% and the average PRcorr at 77.34%. For PVsyst calculation, the average PR with 0% system unavailability is close to the PRcorr and the average PR is decrease with more the system unavailability.

The PRcorr value tends to be less variable over the year. Therefore, PRcorr is usually higher than PR due to the weather-related effects. The effect refers to the cell temperature caused by ambient temperature and irradiation. On the other hand, the average ambient temperature is 26.97 ºC and the average cell

temperature, which is normally high, is 33.19 ºC. Another factor is rain. The occurrence of heavy rain, even if the temperature drops. However, this affects the amount of energy that receives from the PV production system.

Therefore, when judged by these two techniques, it was found that finding the potential of the system with PRcorr is higher than PR. Because the correct temperature of the panel results in higher production efficiency as the effects of seasonal changes are reduced. Techniques for determining the average temperature of the panel are important for the analysis of PRcorr. PVsyst is useful to estimate the performance ratio of the PV system with the measurement data in the study area. In the case of the grid-connected PV system, if the main grid has some problems due to voltage drop or outage. The PV system is disconnected from the main grid, even though the PV system capable of producing electricity. Finally, both PRcorr and PR calculations are affected.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Sathiracheewin, K. Prommaneewat, "Short Term Analysis of PV System Performance in Different Areas of Thailand," in 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 270-273, 2019, doi: 10.1109/ECTI-CON47248.2019.8955369.

[2] Ministry of Energy Department of Alternative Energy Development and Efficiency. Thailand Alternative Energy Situation report, Bangkok, Thailand, 2017, doi: http://www.dede.go.th.

[3] C. Li, "Comparative Performance Analysis of Grid-Connected PV Power Systems with Different PV Technologies in the Hot Summer and Cold Winter Zone," International Journal of Photoenergy, 2018, doi: 10.1155/2018/8307563.

[4] M. Dhimish, S. Silvestre, "Estimating the impact of azimuth-angle variations on photovoltaic annual energy production," Clean Energy, **3**, 47-58, 2019, doi: 10.1093/ce/zky022.

[5] H. Singh, C. Sirisamphanwong, S. R. Santhi Murthy, "Effect of Tilt and Azimuth Angle on the Performance of PV Rooftop System," Applied Mechanics and Materials, **839**, 159-164, 2016, doi: 10.4028/www.scientific.net/AMM.839.159.

[6] T. Dierauf, A. Growitz, S. Kurtz, J.L.B. Cruz, E. Riley, C. Hansen, "Weather-Corrected Performance Ratio (No. NREL/TP-5200-57991)," National Renewable Energy Lab (NREL), Golden, CO, United States, 2013.

[7] H.A. Basson, J.C. Pretorius, "Risk Mitigation of Performance Ratio Guarantees in Commercial Photovoltaic Systems," International Conference on Renewable Energy and Power Quality Journal, **14**, 120-125, 2016. doi: https://doi.org/10.24084/repqj14.244.

[8] E. Syahputra, R.A. Fajarin, E.A. Setiawan, "Characteristic Analysis of Photovoltaic On-Grid System in Tropical Region for Weather-Corrected Performance Ratio Calculation Method Implementation," E3S Web of Conf, **67**, 1-7, 2018, doi: https://doi.org/10.1051/e3sconf/20186701025.

[9] R. Srivastava, V.K. Giri, "Design of Grid Connected PV System using PVsyst," Journal on Electrical Engineering, **10**, 92-96, 2016, doi: https://doi.org/10.26634/jee.10.1.8195.

[10] N.M. Kumar, M.R. Kumar, P.R. Rejoice, M. Mathew, "Performance Analysis of 100 kWp Grid Connected Si-Poly Photovoltaic System Using PVsyst Simulation Tool," Energy Procedia, **117**, 180-189, 2017, doi: https://doi.org/10.1016/j.egypro.2017.05.121.

[11] I. Jamil, J. Zhao, L. Zhang, R. Jamil, S.F. Rafique, "Evaluation of Energy Production and Energy Yield Assessment based on Feasibility, Design, and Execution of 3× 50 MW Grid-Connected Solar PV Pilot Project in Nooriabad," International Journal of Photoenergy, **11**, 2017, doi: https://doi.org/10.1155/2017/6429581.

[12] L. Boughamrane, M. Boulaid, A. Tihane, A. Sdaq, K. Bouabid, A. Ihlal, "Comparative Analysis of Measured and Simulated Performance of the Moroccan First MV Grid Connected Photovoltaic Power Plant of Assa, Southern Morocco," J. Mater. Environ. Sci, **7**, 4682-4691, 2016.

[13] Ministry of Energy Department of Alternative Energy Development and Efficiency, The project to improve the solar energy potential map from satellite images for Thailand in 2017 report, Bangkok, Thailand, 2017. doi:http://www.dede.go.th

[14] M.T. Boyd, "Comparative Performance and Model Agreement of Three Common Photovoltaic Array Configurations," Journal of solar energy engineering, **140**(1),014503, 2018, doi: https://doi.org/10.1115/1.4038314.

[15] J.S. Stein, PV Performance Modeling Methods and Practices, Sandia National Lab (SNL-NM), Albuquerque, NM, United States, 2017.

[16] IEC 61724, Photovoltaic system performance monitoring-guidelines for measurement, data exchange and analysis, International Electrotechnical Commission, 1998.

[17] A.Mermoud, B. Wittmer, PVSYST User's Manual, Switzerland, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 820-825 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Innovative Course Delivery using Analyze – Group – Design – Optimize (AGDO) Methodology: Case Study of Entity-Relationship Model

Aparna Sharma*, Rishabh Singh, Prathamesh Churi, Mahesh Mali

School of Technology Management and Engineering, NMIMS University, Mumbai, 400056, India

ARTICLE INFO

Article history:
Received: 12 August, 2020
Accepted: 25 September, 2020
Online: 12 October, 2020

Keywords:
AGDO
Design
Group
Analyze
ER Diagram

ABSTRACT

*Visualization plays an important role in teaching and learning. The ability of the learner to grasp the visual contents are better than that of textual contents. Traditional teaching methods often revolved around instructions and recitation techniques. However, most of these approaches were dormant and did not call for active learning. The proposed AGDO-Analyze, Group, Design & Optimize methodology was inspired by these obstacles and creates an engaging and wholesome experience for the students. A sample of 47 undergraduate students was used to analyses the feasibility and effectiveness of this teaching mechanism. Students first tackled and **analyzed** the problem statement with their preconceived knowledge. They were then segregated into **groups** wherein they **designed** solutions collectively. Finally, they tried to find an **optimized** solution considering all the suggested designs. Observations were validated based on feedback and how well the students were able to perform. The results revealed that approaches such as AGDO facilitate an immersive learning environment and ensure the quality of teaching. The implications and methodology were further discussed.*

1. Introduction

The current education system in our country has been blamed for all sorts of evil for hundreds of years. Rote learning is still a trouble for our education system, students mug up concepts/theory/methodologies to score marks. The main problem associated with the system is the lack of effective learning pedagogies [1], [2], knowledgeable instructors, and trusted online content. The current assessment process comprises testing the crammed information presented on paper, in such cases the knowledge and skill are given less importance. Teaching pedagogy must emphasize skill-based development which will help students in their professional life [3], [4]

The research work done in [5] reveals the fact that the use of digital video in the classroom increases student engagement. According to the case study in [5]

- It helps to increase the student's motivation to study.
- The learning experience is enhanced.
- Students may get good marks if they are made interested to learn the topic.
- It also increases the team building and communication skills.

Many researchers have observed, learners can gain positive indicators when they watch authentic and real-life clips [6] Information and communications technology is influencing the delivery of education in tertiary institutions. Researchers studied video usage and student beliefs about video instructions for 'beyond classroom' learning and concluded that videos are preferred by students as learning tools compared to other online learning media [7]. In a study conducted, one of the five reasons cited for effective teaching was easy to access to taped teaching sessions, hence rendering the use of video technology in the teaching-learning process as critical [8]. The implementation of this process is not without difficulties. Finding relevant and objective videos that are of optimal length is a challenging and time-consuming task [9]. The proposed AGDO methodology focuses on skill-based learning and makes use of video and audio tools to solve a case study. We agree that it does not apply to all the complex engineering problems.

We conducted a case study on 47 students, in which they had to follow the process of the proposed model. The students were given a problem statement that demanded the development of an entity-relationship model by them. In the initial stage, all students individually analyzed the given problem statement using some

*Corresponding Author: Aparna Sharma, NMIMS University, sharmaaparnasunil@gmail.com

media as a resource for the same. This media including text, images, and videos. After the analysis, the students were grouped and they had discussions about the problem statement and their respective analysis. By listening and understanding the analysis done by every student in the group they moved towards designing the model. The students adapted and improvised on the group's model based on their discussion, the key towards the development of an ideal model lied in the decision making and choices of the groups. After receiving the models from all the groups, their strengths and weaknesses are discussed. The models need to be compared and evaluated to obtain the most suitable model applying to the problem statement through optimization.

Through this research, we are proposing a model that helps the students and teachers to obtain better results by following four basic procedures of A-G-D-O that are- analyze, group, design, and optimize. Apart from the proposed model our results also showed an enhanced understanding level of concepts by the students while using visual data like images and videos rather than the traditional textual way. The study showed a gradual increase in effectiveness with the increase in visuals and animations i.e. from text to images and videos.

2. Motivation

2.1. Skill-based learning

Skill Based Learning (SBL) [10]-[12] is an adaptive method that provides students with external knowledge that is suitable for a particular skill. Psychological research says that by having students learn through the experience of solving problems in place of rote learning, they can learn both contents of study and analyzing strategies [13], [14]

In SBL, student learning focuses on a complex and innovative problem rather than a simple and traditional problem that does not have a single correct answer. Identifying appropriate complex and innovative tasks is the role of an instructor. A complex problem is a problem in which there are multiple ways to solve the problem OR may have to use multiple strategies or algorithms to solve a problem. A complex problem doesn't have a normal and direct answer, it requires thinking and analytical skills which always comes through experience.

According to recent research, Skill Based Learning helps in the following way:

- **Motivating creative skills** – SBL helps to think out of the box and motivates creative skills to solve a particular problem.
- **Develops critical thinking** [15]– SBL helps students to hone analytical and critical thinking skills in all the topics. In this way, rote learning can be avoided, and self-learning can be motivated.
- **Enhances collaborative problem solving** [16] – Constructive criticism is always necessary to solve a problem in the real world. This is only possible if students learn how to work collaboratively in a group in a selfless way. Skill-based learning helps in drawing enhancing strengths and important skills to achieve their goals. While organizing an event or a field-based activity, students work as a team to deliver the desired results.
- **Builds effective written and oral communication** [17] – SBL helps to improvise written and oral communication through reports, presentations, seminars, etc.
- **Hones leadership** – SBL in classrooms also helps to develop leadership skills in students and help them see beyond their self-interests.

However, it is difficult to indulge in the culture of analyzing the things in students by the teacher. A graphical representation/ pictorial view/video can be used to analyze the concepts better [18]. The advantages of Video and Graphics/Image are tabulated in Table 1.

Table 1: Advantages of the use of Text, Images, and Videos in SBL

Parameter	Text	Images /Graphics	Video
Facilitating thinking and problem solving	X	✔	✔
Inspiring and engaging students	X	✔	✔
Easier to be accepted by students	✔	✔	X
More than words can tell	X	✔	✔
Inspired Thinking	X	✔	✔

2.2. Effect of Textual Data, Images, and Videos on Human Brain

Humans are social creatures. They have metamorphosed to communicate in forms of stories, ideas, and visions. Digital storytelling endeavors to stimulate real-life ways of understanding and communication. This allows humans to acknowledge and conceptualize the idea, which is a stark contrast to the notion of the traditional methodology of imparting knowledge through textual scripts [19]. Similarly, a study conducted by Teleiquile also pointed to the fact that people acquire 83% knowledge from vision, 11% from hearing, and 5% from other sources [20]

With the digital boom all over the world, digital technology has become a part and parcel of one's life. These digital reformations have impacted the entire world dynamics, including the field of education [21]. Multimedia such as engaging images and videos arouses interest and enthusiasm. Images and videos accede students in going beyond the four walls of a classroom [22].

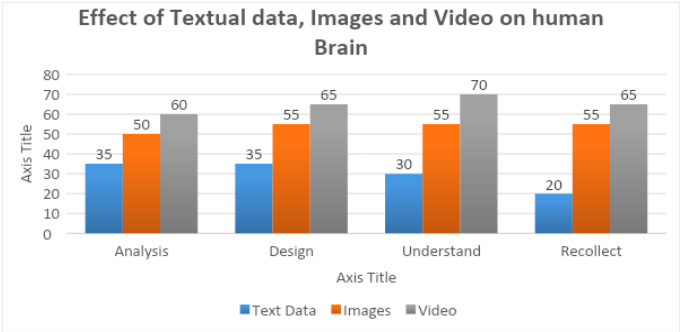


Figure1: Effect of Textual Data, Images, and Video on Human Brain

The problem statement was exercised with a group of 47 students. It was performed and segregated into three levels.

Students were entrusted with a problem definition in different forms and a survey was conducted thereafter, to analyze their understanding for the same. Level One consisted of the problem statement as a two-page textual script. In Level Two, the problem statement was depicted in the forms of images, similar to comics. In Level Three, a five-minute video was outlined in front of the students. All the three stages were recorded and the survey divulged the following results(See figure 1):

3. AGDO Methodology

In the proposed Analyze-Group-Design-Optimize (AGDO) method students work individually as well as in groups to identify basic requirements to solve a real-time complex problem. The AGDO methodology makes use of Graphical con-tents and Skill-Based Learning as described in previous sections. Students get engaged in self-learning and then apply their knowledge to the problem and reflect on what they learned. The instructor acts as a mentor who is passively involved in the learning process rather than actively providing knowledge.

The block diagram of AGDO methodology is drawn in figure 2 below. The diagram depicts the outcomes and strategies of each keyword with the description.

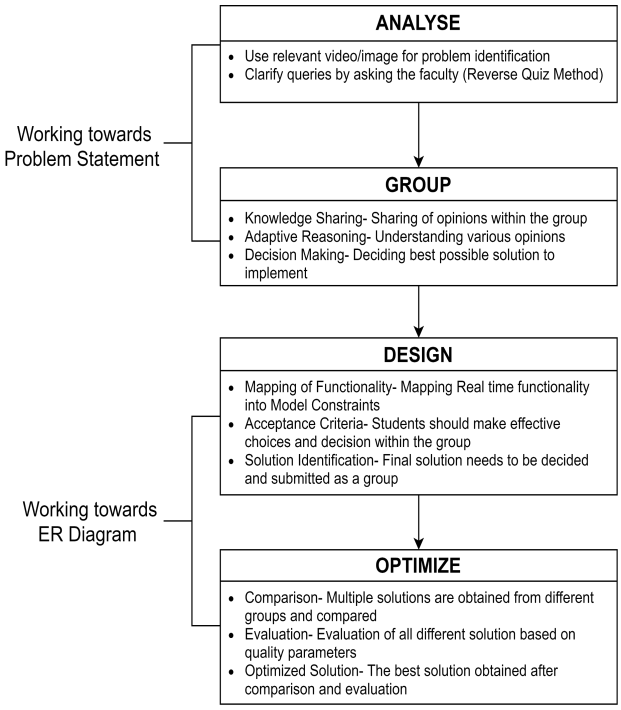


Figure 2: AGDO Methodology Block Diagram

We have experimented with our proposed AGDO methodology for conducting database management lectures to teach data modeling (Entity-Relationship Model). In the proposed AGDO methodology, students will be provided with visual artifacts to **analyze** the real-time problem and model an Entity-Relationship Model. By forming **groups** students will get an opportunity to collaborate and come up with different ideas. By using Adaptive reasoning and Decision-making ability of each group to map functionalities of the system in the **designed** ER diagram. Each group would present their proposed solution, these individual solutions would be evaluated and updated to arrive at an **optimized** solution.

- **Analyze**

  Individual students will analyze the real-time problem using relevant videos/images to felicitate thinking and problem defining abilities. Instructors will continuously monitor progress using the reverse quiz [23] method. The conceptual understanding and progress of the students are confirmed by asking them questions about the same in the reverse quiz method. With a proper understanding of the problems by the students and assurance of the same by the instructors, we move on to the next phase of our model.

- **Group**

  The students are then grouped and asked to have discussions about their analysis by the instructors. Group discussion will ensure active participation and the process of learning by listening to others. Adaptive learning will promote self-determination [24]-[27] among students to provide and accept design decisions proposed by peers. The formation of groups provides several different ideas as well as induction of innovation among peers. Listening skills are a key to this process, better understanding within the group results in better model development. After acceptance of the solution proposed by the peers, groups start with the designing phase of the model.

- **Design**

  After the discussion, adaptation, and improvisation of their respective analysis. Students will come up with mapping real-time functionality into model constraints. Groups will decide the acceptance criteria to make effective choices and decisions within the group. The final solution will be proposed by the project team [23].

- **Optimize**

  All the models developed by every group are collected, discussing the positives and negatives of each. Selecting one model for any problem statement in most cases would not result in the ideal solution. Hence, all proposed solutions are compared and evaluated to obtain the best possible solution using presentation by the groups.

4. Case Study using the AGDO Methodology – Entity-Relationship Model

A case study approach [28] allows us to explore real-life scenarios in a given sample size and identify solutions. Hence, this case study was established to give ideas for designing data models for students. It was effectively conducted with students of Semester III, Computer Engineering who had merely basic knowledge of design process and methodologies. The main objective of this case study was to make students aware of innovative and practical strategies to conceptualize problem statements and eventually, draw ER diagrams for the same. The course name was Relational Database Management System. The total time required for this case study was four hours.

The case study involved following the AGDO methodology and hence conducted in four buckets- Analyze, Group, Design & Optimize. See Figure 3 for the given problem statement.

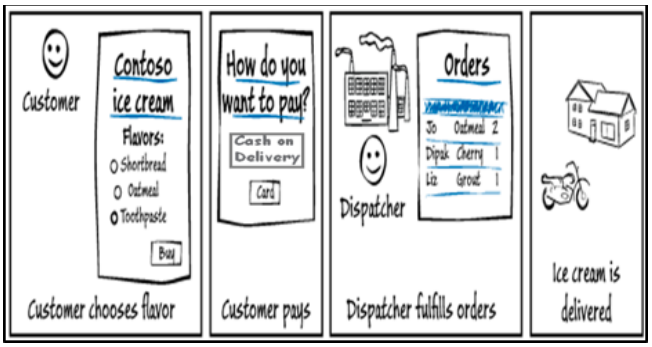


Figure 3: Problem statement

- **Analyze:** A group of 47 students was presented with the above picture. The problem statement involved forming a Relationship Entity Model based on the given picture. The students were expected to analyze the given problem statement, connect it with pre-existing knowledge, and form initial grounds of approaching this problem statement. The system environment consisted of a typical E-Commerce, involving customers (consumers), shopkeepers (vendor), and dispatchers (distributor).

The problem definition of Figure 3 is as follows:

The CoolPoint Ice Cream Delivery System has many flavors of ice cream. The system will allow customers can choose flavors of ice cream from the list, Customer places the order and pays the bill through card or chooses to pay on delivery, Staff gives the order dispatching list to the dispatcher and the dispatcher delivers the ice-cream orders to the respective customers. The system environment is for Customer, Shopkeeper, and Dispatcher.

- **Group:** After analyzing, the students were now grouped into clusters, containing 4 to 5 students. The team goal is to help one another understand the problem statement successfully and to identify the purpose. Each member of the group was accountable individually and collectively. Every member of the group had to work towards a common goal and perform his/her tasks that contribute towards it. In smaller groups, students could now freely discuss and debate their ideas. They also learn by effectively listening [29] to other peers to express their ideas. Grouping calls for an engaging and more confined platform which lets students feel much more confident and accepted within the smaller constraints of a group.

  A sample outcome of this step (Referring figure 3) is given below:

Identify Entities and attributes for each entity.

– To identify entities from the scenario we can use the 'Noun-phases' approach.

– This approach considers all nouns and phases in the problem as entities.

– Then we will take some steps to remove all vague or unnecessary Entities

– If an entity is depending on other entities or entities not having independent existence in the system such entities are enlisted as Weak entities.

– Some problem attributes are given in the problem statement itself. Sometimes attributes are not specified in the problem.

Example Above:

IceCream (Flavor, pricePerScoop)

Customer (Customer_Name, Customer_Mobile, Cust_Address)

Shopkeeper (Customer_Name, Customer_Mobile, Cust_Address)

Dispatcher (Order_Id, Shipping_Address, Ship_Status)

- **Design:** The third step involved them working collaboratively to build on the designing of the ER model. Completion of the ER diagram required each student's work to be combined with that of other students to produce a team effort. They started designing the model by forming and recognizing primitive stages of relations such as (see Figure 4 and 5):



Figure 4: Formation of relationships within entities

Approaches to Identify Relationships

– We can use conjunction used in problem for naming relationship.

– Example, Employee works for a department

– We can also use the name of two entities connected with an underscore as a conjunction for the naming relationship.

– Example, Employee works for a department

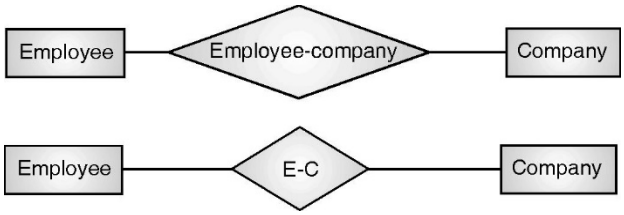


Figure 5: Formation of relationships within entities

- **Optimize:** This problem statement which closely tried to mimic real-world problems, had a wide range of parameters and constraints. The smaller clustered groups presented their contrived designs in front of the entire class. They gave and received feedback on their presented solution. The class then tried to find the optimum solution by comparing all the suggested diagrams by the various groups. The optimized solution was an amalgamation from diagrams received by various groups.

5. Results and Analysis

The usage of AGDO methodology accorded students the skills required to understand real-world problems using visual artifacts such as videos and images. The analyzing stage empowers them to use preconceived knowledge and relate it with the given problem statement. Reverse questioning can lead to a better conceptual

understanding [30] while ensuring continuous assessment by the teacher. Hence, making sure that the results of the analysis obtained are relevant to the topic. The grouping stage allows students to work together and share experiences which leads them to a better understanding of the problem at hand and integrates different perspectives. The designing data model in the group will improve the acceptability of solutions from other peers. Students experience active learning through problem-solving, brain-storming sessions, and group assignments for designing the solutions.

This case study was conducted with the help of student teams in a session regarding RDBMS during a class. After following the proposed AGDO methodology, the student teams came up with multiple solutions, and the one shown in figure 5 was found out to be the optimal one.

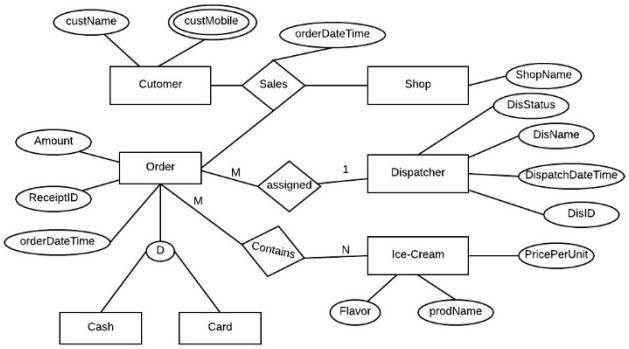


Figure 5: Optimized Solution

This led to the introduction of various metrics such as Problem Statement, Questions, Solutions, and Learning Outcomes that help affluently evaluate the proposed approach. Students were graded according to their performance on these parameters.

The following results can be drawn out of the assessment of the case study post implementing the AGDO methodology: Greater than half the strength of the course, 55.5% (26) of students scored over 90% in successfully understanding and conceptualizing the problem statement. This led us to the direction that students were able to harness the given visual artifacts into meaningful and coherent comprehension. In the category labeled 'Questions', it was observed that 44% of the class could surpass the 80% threshold. This paves the way for the reverse questioning method and group discussions which stimulate and encourage students to probe and ask questions, thereby fueling their curiosity. Similarly, 40.4% (19) scored more than 80% for the solutions that they provided for the given problem statement. 72.3% (34) of students secured over 80% when evaluating the learning outcomes. This, in turn, gave better insights into potential subject outcomes, thus accelerating learning and growth.

In totality, 42.3% (16) of the total students attained above 90% of the maximum marks allocated, while a staggering 93.6% (44) were able to achieve over 80% of the maximum marks. These numbers substantiate the proposed AGDO methodology.

The students gave feedback about the usage of this case study regarding their learning of viewing real-world scenarios for designing IT solutions. Students got the experience of active learning through problem-solving, brainstorming sessions, and group assignments for designing Real-World Systems. Feedbacks provide better insight and help verify the feasibility and efficacy of the proposed method.

Feedback was provided by the 47 participating students on various parameters. They were asked to give the rating on a scale of 0 to 5, zero being the least effective to 5 being most effective. For the first two parameters which were based on the relevancy and ease of understanding of the course, it was observed that images were the most effective medium. While the latter two which were based out of helpfulness and the overall impact indicated the video was the most preferred option. In all four parameters, the textual transmission was rated relatively poorly proving that it was not favored amongst the students. This table reaffirms the advantages of visual artifacts such as videos and images over textual data.

6. Limitations of the AGDO Methodology

AGDO methods motivate skill-based learning and encourage the group working culture among engineers. In industry. Working together in a group is the most important aspect. Engineering leadership is another emerging area of research [31]. The courses like software engineering, Technical Project, Mobile Application Development, Prototype Development, etc. wherein the problem definition and problem formulation is involved, the AGDO methodology is best suited. The theoretical subjects and other self-study-based subjects, the AGDO courses are not appliable.

Another limitation of this paper is the sample size, which is very small for any kind of analysis. Since AGDO methodology was new, we have tested its applicability on a smaller population. In near future, we would like to extend our work to a larger population and we would like to conduct the statistical measures to prove how AGDO methodology helped to build a perfect ER model.

7. Conclusion

There is evidence that suggests that the traditional education reforms, once considered significant, are now struggling to keep up with modernization. Historically, students often resort to ineffective methods such as rote learning or often feel disengaged from the learning process. In the AGDO modus operandi, it tries to eliminate the above by students working collaboratively in groups towards a common goal.

As a denouement of the case study, students were able to discover and learn the subject of the course on their own by questioning, interrogation, and logical ability. Grouping allowed them to weigh, judge, and preview others' ideas with their own. This sanctioned them to construct and assemble ideas which were also accepted by their peers. By using their primary knowledge and associating it with daily life, they were able to optimize the best possible outcome and come up with scalable and practical solutions. AGDO strives to foster an easy understanding of concepts and bring joy to learning. To conclude, this study showed that AGDO is an effective teaching method.

While applying this methodology, it was also observed that in most cases students using visual forms of data like images and videos had better efficiency than that of the students using traditional textual methods. This was a result of the increasing

proficiency of students in technology and a better understanding of concepts through visualization rather than textual descriptions.

References

[1] G. Domik, Fostering collaboration and self-motivated learning: Best practices in a one-semester visualization course, IEEE Computer Graphics and Applications, 2012.

[2] H. J. Holz, A. Applin, D. Joyce, H. Purchase, C. Reed, B. Haberman, Research Methods in Computing : What are they , and how should we teach them ?, Information Systems, 2006.

[3] F. David, R. Abreu, Information technology in education: Recent developments in higher education, in *Iberian Conference on Information Systems and Technologies, CISTI*, 2014, 2014.

[4] W. G. Spady, *Outcome-Based Education: Critical Issues and Answers*, 1994. 1994.

[5] P. Willmot, M. Bramhall, K. Radley, Using digital video reporting to inspire and engage students, The Higher Education Academy, 2012,.

[6] W. A. Almurashi, THE EFFECTIVE USE OF YOUTUBE VIDEOS FOR TEACHING ENGLISH LANGUAGE IN CLASSROOMS AS SUPPLEMENTARY MATERIAL AT TAIBAH UNIVERSITY IN ALULA, 2016.

[7] H. Forbes *et al.*, Use of videos to support teaching and learning of clinical skills in nursing education: A review, *Nurse Education Today*, 2016. 2016.

[8] G. C. Lee, C. C. Wu, Enhancing the teaching experience of pre-service teachers through the use of videos in web-based computer-mediated communication (CMC), *Innovations in Education and Teaching International*, 2006. 2006.

[9] S. Ivanova, Using explainer videos to teach web design concepts, in *The 13th International Scientific Conference eLearning and Software for Education*, 2017, 2017.

[10] K. Kraiger, J. K. Ford, E. Salas, Application of Cognitive, Skill-Based, and Affective Theories of Learning Outcomes to New Methods of Training Evaluation, Journal of Applied Psychology, 1993.

[11] P. Pastor, M. Kalakrishnan, S. Chitta, E. Theodorou, S. Schaal, Skill learning and task outcome prediction for manipulation, in *Proceedings - IEEE International Conference on Robotics and Automation*, 2011, 2011.

[12] C. Gomathi, V. Rajamani, Skill-based education through fuzzy knowledge modeling for e-learning, Computer Applications in Engineering Education, 2018.

[13] A. G. GREENWALD, Cognitive Learning, Cognitive Response to Persuasion, and Attitude Change, in *Psychological Foundations of Attitudes*, 1968, 1968.

[14] R. C. Schank, T. R. Berman, K. A. Macpherson, Learning by doing, in *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory*, 2013, 2013.

[15] M. J. Bezanilla, D. Fernández-Nogueira, M. Poblete, H. Galindo-Domínguez, Methodologies for teaching-learning critical thinking in higher education: The teacher's view, Thinking Skills and Creativity, 2019.

[16] J. Andrews-Todd, C. M. Forsyth, Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task, Computers in Human Behavior, 2020,.

[17] V. Garousi, G. Giray, E. Tuzun, C. Catal, M. Felderer, Closing the Gap between Software Engineering Education and Industrial Needs, IEEE Software, 2020.

[18] L. DE SOUSA, B. RICHTER, C. NEL, The effect of multimedia use on the teaching and learning of Social Sciences at tertiary level: a case study, Yesterday and Today, 2017,.

[19] B. R. Robin, The power of digital storytelling to support teaching and learning, *Digital Education Review*, 2016. 2016.

[20] M. Himanshoo, K. Sharma, Role of ICT in Improving the Excellence of Education, International Journal on Computer Science and Engineering (IJCSE, 2015.

[21] A.-M. Suduc, M. Bîzoi, G. Gorghiu, L.-M. Gorghiu, Digital Images, Video and Web Conferences in Education: A Case Study, Procedia - Social and Behavioral Sciences, 2012.

[22] L. Killian, New Textual Formats: Reading online is re-wiring the human brain and changing how we process information, Dalhousie Journal of Interdisciplinary Management, **5**(1), 2010. https://doi.org/10.5931/djim.v5i1.49

[23] P. Churi, K. Mistry, A. Dhruv, S. Wagh, Alchemizing education system by developing 5 layered outcome based engineering education (OBEE) model, in *Proceedings - 2016 IEEE 4th International Conference on MOOCs, Innovation and Technology in Education, MITE 2016*, 2017, 2017.

[24] J. C. Ortiz-Ordoñez, F. Stoller, B. Remmele, Promoting Self-confidence, Motivation and Sustainable Learning Skills in Basic Education, Procedia - Social and Behavioral Sciences, 2015.

[25] P. Ommundsen, Problem-based learning, in *Inspiring Students: Case Studies on Teaching Required Courses*, 2013, 2013.

[26] D. Boud, G. I. Feletti, *The challenge of problem-based learning*, 2013. 2013.

[27] J. R. Savery, T. M. Duffy, Problem based learning: An instructional model and its constructivist framework, Educational Technology, 1995.

[28] S. Crowe, K. Cresswell, A. Robertson, G. Huby, A. Avery, A. Sheikh, The case study approach, BMC Medical Research Methodology, 2011,.

[29] C. Edwards, Life-long learning, Communications of the ACM, 1993.

[30] J. P. Weaver, R. J. Chastain, D. A. DeCaro, M. S. DeCaro, Reverse the routine: Problem solving before instruction improves conceptual knowledge in undergraduate physics, Contemporary Educational Psychology, 2018.

[31] B. Ahn, M. F. Cox, J. London, O. Cekic, J. Zhu, Creating an instrument to measure leadership, change, and synthesis in engineering undergraduates, Journal of Engineering Education, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 824-840 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Fine Tuning the Performance of Parallel Codes

Sanaz Gheibi*, Tania Banerjee, Sanjay Ranka, Sartaj Sahni

Department of Computer and Information Science and Engineering, University of Florida, Gainesville, 32611, USA

ARTICLE INFO

Article history:
Received: 03 June, 2020
Accepted: 11 August, 2020
Online: 31 August, 2020

Keywords:
Parallel speed up
Large matrices
Performance fine tuning

ABSTRACT

We propose a multilevel method to speed highly optimized parallel codes whose runtime increases faster than their workload. This method requires the ability to solve large instances by decomposing them into smaller instances. Using a simple parallel computing model, we derive a mathematical model that predicts whether or not our method can improve performance and also predicts the amount of improvement attainable. Our method is tested and shown to be effective on three highly optimized BLAS (Basic Linear Algebra Subprograms) routines from Intel's Math Kernel Library (MKL). Those routines are cblas_dgemm, cblas_dtrmm and cblas_dsymm. On the Intel Knights Landing (KNL) platform our method speeds cblas_dgemm by 33%, cblas_dtrmm by 50% and cblas_dsymm by 49% on double-precision matrices of size $16K \times 16K$*, when the KNL's default memory-clustering configuration (cache-quadrant) is used.*

1 Introduction

This paper is an extension of our previous paper on fine tuning a group of linear algebra libraries using a multi-level approach [1].

Our problems of interest are a group of parallel codes whose runtime grows faster than the problem size. The following example provides a quick demonstration of how such parallel codes behave and what motivates our proposed method.

Table 1 contains the runtime values for multiplying square matrices of different sizes on the intel KNL multi-level memory multicore computer. The memory-clustering configuration is set to its default which is cache quadrant. The multiplications are done in double precision using the highly optimized cblas_dgemm function from Intel MKL. The runtimes are obtained using either 32 or 64 cores. For each set of cores and each matrix size of $2N \times 2N$, the table also provides the runtime ratio when the matrix size increases from $N \times N$ to $2N \times 2N$.

Since the matrix dimensions are increasing by a factor of 2 (thus increasing the workload by a factor of 8), we expect all the ratios to be 8. However, that is not the case and four of the ratios are greater than 8. That is where the runtime increases faster than the workload and that brings up possibilities of improvement. Here, for both 32 cores and 64 cores, we get better results if instead of performing a $16K \times 16K$ ($64K \times 64K$) multiplication, we perform 8 instances of $8K \times 8K$ ($32K \times 32K$) multiplications and combine the results (each input matrix is partitioned into 4 blocks of half the dimension; the result is obtained by doing 8 block multiplies and 4 block additions). The previous statement is valid as long as the overhead of performing 4 pairs of matrix additions is less than the runtime reduction obtained by multiplying 8 smaller instances.

We can further reduce the runtime by running the smaller instances in parallel. For example, instead of running one instance of $32K \times 32K$ multiplication using 64 cores, we can run 4 parallel pairs of $16K \times 16K$ multiplications using 32 cores in less total time. Again, we will have speed up if the overhead of adding 4 pairs of matrices is less than the obtained runtime reduction.

Table 1: Run times of double-precision matrix multiplications using 32 and 64 cores. Ratio is obtained by dividing the run time in the current row by the runtime in the previous row.

Matrix Size	32 Cores		64 Cores	
	Time(seconds)	Ratio	Time(seconds)	Ratio
$4K \times 4K$	0.37	-	0.32	-
$8K \times 8K$	2.32	6.27	1.51	4.71
$16K \times 16K$	20.56	8.86	12.32	8.16
$32K \times 32K$	159.01	7.73	97.52	7.92
$64K \times 64K$	1517.21	9.64	871.10	8.93

Our proposed method tries to improve the speed of the parallel codes whose runtime increases faster than the workload. This method uses multiple levels and works on applications that are of a decomposable nature. We start with a 2-level algorithm. At level 2, the problem is broken into a number of subproblems that are solved (serially or in parallel) using the 1-level algorithm and then

*Corresponding Author: Sanaz Gheibi, Dept. of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, USA, sgheibi@ufl.edu

the results are combined to form the final solution. (By 1-level algorithm we mean the parallel and highly efficient algorithm whose runtime grows faster than its workload). Now, suppose the 2-level algorithm itself demonstrates the property that its runtime grows faster than its workload. In this case, we can gain further speed up by using a 3-level algorithm in which the problem is broken into a proper number of subproblems which are solved using the 2-level algorithm. We can continue increasing the number of levels in a similar fashion.

While our proposed method has much in common with divide and conquer, block matrix multiplication and Cannon's method for matrix multiplication [2], there are subtle differences. In a typical divide and conquer algorithm [3]- [7] the division scheme is the same across the levels (an example being the Strassen's matrix multiplication algorithm [3]); whereas in our method the number of subproblems into which the problem is broken can be different in different levels. Block matrix multiplication has been used both in serial and parallel matrix multiplication algorithms [2, 8, 9]. It can be viewed as a 1-level divide-and-conquer algorithm; while our method uses multiple levels of decomposition and combination. Our method is different from Cannon's algorithm in both the number of levels and the block's movement scheme. Cannon's movement scheme is designed for mesh-connected parallel computers with wraparound and uses row-and-column circular shifts. Our method uses a serpentine movement scheme that will be described in section 3.2.

Regardless of the above-mentioned similarities, the value of our method is that it can speed up codes whose runtimes grow faster than their workload. Here, the codes to be optimized are viewed as black boxes that can be already parallel, highly optimized and have their own complexities. As a result, this method can be applied to a range of problems broader than just linear algebra libraries.

We use a simple parallel computing model to formulate the conditions when our method results in speed up and also predict the amount of speed up. In section 5 we will demonstrate the effectiveness of our algorithm and our prediction formula on three linear algebra libraries from Intel MKL.

Our experimental results are obtained using the Knights Landing (KNL) computer. Although this architecture has been discontinued by Intel, we believe our fine-tuning method remains relevant for the following reasons. First, Knights Landing clusters are still used in many national laboratories and second, our method can be used in future architectures that have multilevel memory.

The rest of the paper is organized as follows. Section 2 summarizes some of the related work. In Section 3 we describe a solution framework for general problems and then describe its specific application to a class of linear algebra functions. Section 4 gives a brief description of KNL architecture. Section 5 presents the experimental results; and we conclude in Section 6.

2 Related Work

This section summarizes some of the previous methods used to speed up applications.

Tiling has been used extensively to improve data locality and cache efficiency. This method works by breaking the problem into subproblems or tiles in such a way that the amount of memory reuse in faster levels of memory hierarchy increases. Tiling has been used in a large group of problems including sparse and dense matrix-matrix multiplication [10, 11], parallel tensor transpose [12] and LU factorization [11]. In [13] the authors use tiling along with thread batching and architecture specific optimizations to speed up Alternating Least Squares algorithm used in recommender systems. They use a fine grained tiling in order to mitigate the imbalance in threads' workload that results from the sparse nature of the problem at hand. In [14] the authors use temporal tiling alongside spatial tiling to improve cache efficiency of large scale stencil computations. In a particular class of stencil computations that is the focus of their work, each time step only depends on a limited number of previous time steps and hence temporal tiling can also be incorporated.

Data reordering has been done to change the order in which the input data are processed and increase the cache efficiency. This method performs by mapping the original data indices to the new indices and has been used to improve the efficiency of both regular applications such as dense matrix kernels [15] and irregular applications such as molecular dynamic simulation and hydrodynamic computations [16]. Other methods have applied data reordering to matrix multiplication [17] and embedded multimedia processing [18] with an extra assumption that tiling is already in place.

Reducing communication overhead has also been used to speed up applications. One way to reduce communication is through asynchrony. Asynchrony can be very effective in runtime reduction as synchronizing parallel processes results in much overhead. However, asynchrony can only be used for applications whose correctness is not hurt by lack of synchronization; an example being Stochastic Gradient Descent for sparse matrix factorization [19, 20]. Computation communication overlapping has also been used to hide the communication latency in a group of problems including three-dimensional Fast Fourier Transform [21], data-parallel training of Convolutional Neural Networks [22] and distributed Stochastic Gradient Descent [23]. Block reordering is another method used in [24] to reduce the amount of communication required in Strassen's algorithm. Reducing the communication traffic has also been used to speed up applications in large scale distributed systems such as mobile networks [25].

Load balancing has been used to improve the response time of distributed applications by evenly distributing the workload among the worker units and preventing one worker to be overloaded while the other workers are idle. Among the main areas of application are cloud computing[26] and edge computing[27]. In [28] the authors have used load balancing to speed up matrix-vector multiplication in large scale machine learning and data mining applications. The authors of [29] have developed a simulator to select the best balancing technique for scientific computing applications whose stochastic and irregular natures greatly contribute to load imbalance.

Automatic data placement can also be used to improve memory efficiency of memory bound applications. These methods work by analysing the memory access patterns of a program either in compile time [30] or during the runtime using a lightweight profiling method [31] and automatically placing data segments on optimal memory components.

All of the above-mentioned methods could be used in the 1-level algorithm that is our starting point; i.e, the algorithm for which the

runtime grows faster than the workload. What we do is essentially taking the highly optimized code as a black box and adding a layer of fine-tuning on top of that.

3 Proposed Approach

3.1 General framework

In this section, we provide a general framework for speeding up any parallel code that is decomposable and whose runtime increases faster than its workload. In the subsequent sections, we will demonstrate how to apply our method to speed up the highly optimized codes for general, triangular and symmetric matrix multiplication.

We have already described our proposed method in the introduction section. Here we give a brief summary and present the pseudo codes. As we have already mentioned, our method works in multiple levels. A 2-level implementation divides the problem into subproblems and solves them using a 1-level code either serially, in parallel or using a hybrid of the both. Then the solutions to the subproblems are combined to form the general solution. Similarly, an n-level code divides the problem into subproblems that are solved using an $(n-1)$-level algorithm and then combines the results.

Our n-level algorithm for $n > 2$ is described in Algorithm 1. Given a problem A of size s, we decompose it into s/s_b subproblems of size s_b each. In lines 2-4, the subproblems are solved serially using the $(n-1)$-level algorithm (GEN_{n-1}). Then in line 5 the solutions to the subproblems are combined to form the final solution C. Note that multithreading can be incorporated in all the functions in this section and in subsections 3.2, 3.3 and 3.4. All the threads can be used to copy the subproblems (as we will explain in the next paragraph) and to combine the results. If subproblems are solved serially, then all the threads will be used to solve each subproblem. However, if b subproblems are solved in parallel, then $1/b$ of the total threads will be assigned to each subproblem. Also note that combination of each partial solution with the final result can happen at the same time that the partial solution is produced. In that case, the same number of threads used to solve each subproblem are used to combine its result with the final solution.

GEN_{n-1} in its turn calls GEN_{n-2} and this process continues until GEN_2 is called. The last level of problem decomposition happens at GEN_2 and then the subproblems are solved using GEN_1 which is the highly optimized parallel code to be sped up. In order to use the fast levels of memory hierarchy, we copy the subproblems into a fast memory and that is done in level 2 which is the level at which the actual problem solving is done. Note that in many cases the fast memory is one of the levels of cache hierarchy which we can not directly access and copy blocks into it. In those cases we still make copies of the blocks in the main memory and that causes the blocks to be automatically brought into the cache. Due to the space constraints of the fast memory, the parallel solving of subproblems happens only at level 2. And that is also why we have two sets of pseudo codes one for level $n, n > 2$ and one for level 2. Algorithm 2 describes our two-level algorithm.

Using GEN_2, a problem $\overline{A}$ of size $\overline{s}$ is divided into $\overline{s}/\overline{s_b}$ subproblems of size $\overline{s_b}$ each. $\overline{C}$ is the output and b is the degree of concurrency. Meaning that, at each iteration, the input to b of the subproblems are copied into the fast memory (line 3), the b subproblems are solved concurrently using GEN_1 (lines 4-6) and then the results are combined to form part of the final solution (line 7).

3.2 Matrix multiplication

This section illustrates our method by demonstrating how it works to speed up matrix multiplication. Software packages such as AMD Core Math Library (ACML), OpenBLAS, ATLAS and Intel Math Kernel Library (MKL) provide efficient parallel implementations of matrix multiplication as a part of the Basic Linear Algebra Subroutines (BLAS). In level 1 of our implementation, we use cblas_dgemm that is the double precision matrix multiplication code from MKL.

Algorithm 1 n-level Algorithm

1: **function** $GEN_n(A)$ ▷ Problem A of size s is decomposed into s/s_b subproblems of size s_b each. C is the output. $C_p[I]$ is the solution to the subproblem $A[I]$.
2: **for** $I = 1$ to s/s_b **do**
3: $C_p[I] \leftarrow GEN_{n-1}(A[I])$
4: **end for**
5: Combine $C_p[I], I = 1, 2, ..., s/s_b$ with C
6: **return** C
7: **end function**

Algorithm 2 Two_level Algorithm

1: **function** $GEN_2(\overline{A})$ ▷ Problem $\overline{A}$ of size $\overline{s}$ is decomposed into $\overline{s}/\overline{s_b}$ subproblems of size $\overline{s_b}$. $\overline{C}$ is the output and $\overline{C_p}[I]$ is the solution to the subproblem $\overline{A}[I]$. b is the degree of concurrency.
2: **while** $I \leq \overline{s}/(\overline{s_b} \times b)$ **do**
3: Copy b subproblems $\overline{A}[I]$ to $\overline{A}[I + b - 1]$
4: **start** b **parallel function calls**
5: $\overline{C_p}[J] \leftarrow GEN_1(\overline{A}[J])$ for $J = I, ..., (I + b - 1)$
6: **wait for all functions to return**
7: Combine $C_p[J], J = I, ..., (I + b - 1)$ with C
8: $I \leftarrow I + b$
9: **end while**
10: return $\overline{C}$
11: **end function**

Algorithm 3 *L_level* Matrix Multiplication

1: **function** $MXM_L(A, B)$ ▷ Input matrices $A_{m\times p}$, $B_{p\times n}$ and output matrix $C_{m\times n}$ are divided into blocks of sizes $s_m \times s_p$, $s_p \times s_n$ and $s_m \times s_n$ respectively.
2: **for** $I = 1$ to m/s_m **do**
3: **for** $J = 1$ to n/s_n **do**
4: **for** $K = 1$ to p/s_p **do**
5: $C[I, J] \leftarrow C[I, J] + MXM_{L-1}(A[I, K], B[K, J])$
6: **end for**
7: **end for**
8: **end for**
9: **return** C
10: **end function**

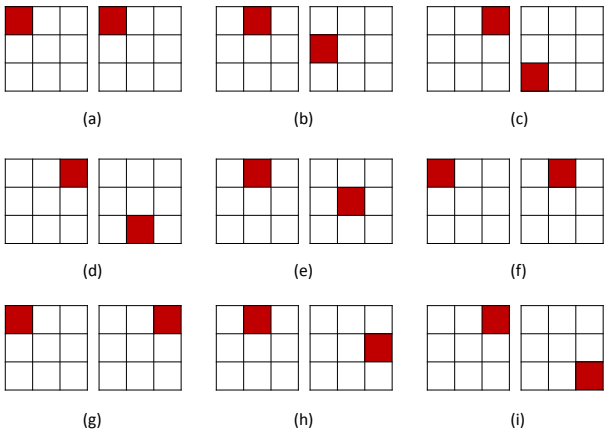


Figure 1: In parts (a)-(i) the chosen blocks in $M1$ (the matrix on the left) and $M2$ (the matrix on the right) are used to compute row 1 of $M3$. The resulting matrix $M3$ is not shown. $M3[1,1]$, $M3[1,2]$ and $M3[1,3]$ are computed in parts (a)-(c), (d)-(f) and (g)-(i) respectively. If we follow the path of the selected blocks in each matrix, a pattern similar to snake movement will result.

Our method for the matrix multiplication problem follows the general framework presented in the previous section. Here, the input is two matrices A and B of sizes $m \times p$ and $p \times n$ to be multiplied and the result is stored in matrix C of size $m \times n$. At level $L > 2$, the matrices are divided into blocks of sizes $s_m \times s_p$, $s_p \times s_n$ and $s_m \times s_n$ respectively. Product of matrix blocks $A[I,K]$ and $B[K,J]$ is computed using a lower level function and the result gets accumulated in the corresponding output block $C[I,J]$. Algorithm 3 describes our *L_level* algorithm (MXM_L).

MXM_2 is our two-level matrix-matrix multiplication algorithm. Input matrices $M1$ and $M2$ are multiplied and the result is stored in matrix $M3$. The matrix sizes are $\overline{m} \times \overline{p}$, $\overline{p} \times \overline{n}$ and $\overline{m} \times \overline{n}$ respectively.

As we previously mentioned, in level 2, the subproblems (matrix blocks in this case) are copied into the fast memory, so that the highly optimized level 1 code can benefit from the high bandwidth. To optimize the copying process, we use a serpentine pattern to iterate through the input sub-matrices. Figure 1 shows an example of this serpentine movement and how it reduces the amount of data transfer between the slow and the fast memories. Here, a 3×3 grid is used to divide matrices into smaller blocks. Degree of concurrency $b = 1$ that means one block from each matrix is selected at a time. Here, for computing $M3[1,1]$, blocks $[1,1]$, $[1,2]$ and $[1,3]$ are selected from $M1$ with their corresponding blocks from $M2$. Now, for computing $M3[1,2]$, instead of selecting blocks of $M1$ in the previous order, i.e. starting from $M1[1,1]$ and moving left-to-right, we start from $M1[1,3]$ and move right-to-left. As a result, we won't have to copy $M1[1,3]$ again as it is already in the fast memory. For $M3[1,3]$ we start from $M1[1,1]$ and move left-to-right again as then, $M1[1,1]$ is already present in the fast memory. If we consider the order in which we iterate the blocks of $M1$ and their corresponding blocks in $M2$, each of them forms a pattern like a snake movement. Using this iteration pattern we reduce the number of copied blocks by 2 for computing the first row of $M3$. The effect gets more significant for larger matrices with higher degrees of concurrency.

We should note that in our implementation, we also used serpentine pattern for MXM_L; however, as the blocks for $L-level$ matrix multiplication ($L > 2$) are usually too large to fit in the fast levels of memory hierarchy, there is not much benefit in using this pattern. Therefore, we eliminated the details from algorithm 3 to make it more concise and readable.

Our 2-level algorithm MXM_2 is described in algorithm 4. It follows the general structure of algorithm 2. Matrices $M1$, $M2$ and $M3$ of sizes $\overline{m} \times \overline{p}$, $\overline{p} \times \overline{n}$ and $\overline{m} \times \overline{n}$ are divided into blocks of sizes $\overline{s_b}_1 \times \overline{s_b}_3$, $\overline{s_b}_3 \times \overline{s_b}_2$ and $\overline{s_b}_1 \times \overline{s_b}_2$ respectively. In lines 8-15 b blocks from each of the two input matrices are copied, then in lines 16-19 b instances of our 1-level algorithm (cblas_dgemm) are called in parallel on each of the block pairs. Finally, the partial results are combined in lines 20, 21. The rest of the algorithm deals with the order in which the blocks are selected and whether or not they are already present in the fast memory (if a block is already present, then it should not be copied again). That is the serpentine pattern we previously talked about and is implemented by K_{dir} and J_{dir} that define the moving directions along the columns of $M1$ and rows of $M2$ respectively. Variables $copy_{M1}$ and $copy_{M2}$ define when the blocks of $M1$ and $M2$ should be copied.

Algorithm 4 Two_level Matrix Multiplication

1: **function** $MXM_2(M1, M2)$ ▷ Input matrices $M1_{\overline{m}\times\overline{p}}$ and $M2_{\overline{p}\times\overline{n}}$ and the output matrix $M3_{\overline{m}\times\overline{n}}$ are divided into blocks of sizes $\overline{s_b}_1 \times \overline{s_b}_3$, $\overline{s_b}_3 \times \overline{s_b}_2$ and $\overline{s_b}_1 \times \overline{s_b}_2$ respectively. b is the degree of concurrency. $C_p[U]$ stores the partial result of $M1[I,U] \times M2[U,J]$. MXM_1 is cblas_dgemm() in our case.
2: $copy_{M1} \leftarrow true$, $copy_{M2} \leftarrow true$
3: $K \leftarrow 1$, $J \leftarrow 1$, $K_{dir} \leftarrow 1$, $J_{dir} \leftarrow 1$
4: **for** $I = 1$ to $\overline{m}/\overline{s_b}_1$ **do**
5: **for** $\overline{n}/\overline{s_b}_2$ iterations involving J **do**
6: **for** $\overline{p}/\overline{s_b}_3$ iterations involving K **do**
7: **if** $copy_{M1}$ **then**
8: Copy b blocks from $M1$:
9: $(M1[I,K]$ to $M1[I, K + (b-1) \times K_{dir}])$
10: **end if**
11: **if** $copy_{M2}$ **then**
12: copy b blocks from $M2$:
13: $(M2[K,J]$ to $M[K + (b-1) \times K_{dir}, J])$
14: **end if**
15: **start** b **parallel function calls**
16: $C_p[U] \leftarrow MXM_1(M1[I,U], M2[U,J])$
17: for $U = K, ..., K + (b-1) \times K_{dir}$
18: **wait for all functions to return**
19: $M3[I,J] \leftarrow M3[I,J] + C_p[K] +$
20: $... + C_p[K + (b-1) \times K_{dir}]$
21: $K \leftarrow K + b \times K_{dir}$; $copy_{M1} \leftarrow true$
22: **end for**
23: $K_{dir} \leftarrow -K_{dir}$; $copy_{M1} \leftarrow false$
24: $J \leftarrow J + J_{dir}$; $copy_{M2} \leftarrow true$
25: **end for**
26: $J_{dir} \leftarrow -J_{dir}$; $copy_{M2} \leftarrow false$; $copy_{M1} \leftarrow true$
27: **end for**
28: **return** $M3$
29: **end function**

After all blocks corresponding to the K variable in the block

matrix multiplication formula

$$M3[I,J] = \sum_{K=1}^{\overline{p}/\overline{s_{b3}}} M1[I,K] \times M2[K,J] \quad (1)$$

have been used, we change the direction in which we select those blocks for computing $M3[I, J+1]$. That is when blocks from $M1$ are reused. Similarly, after $M3[I,1]...M3[I,\overline{n}/\overline{s_{b2}}]$ have all been computed; we compute $M3[I+1,J]$s in reverse direction and b blocks from $M2$ are reused.

3.3 Triangular Matrix Multiplication

This section illustrates how our methodology may be employed to multiply two matrices where one of the matrices is (upper/lower) triangular. The highly optimized function we use at the first level is cblas_dtrmm from Intel MKL.

Algorithm 5 *L_level* Triangular Matrix Multiplication

function $TrMXM_L(A,B)$ ▷ Input matrices $A_{m\times m}$, $B_{m\times n}$ and output matrix $C_{m\times n}$ are divided into blocks of sizes $s_{b1}\times s_{b1}$, $s_{b1}\times s_{b2}$ and $s_{b1}\times s_{b2}$ respectively. b is the number of block multiplications done in parallel.
  for $I = 1$ to m/s_{b1} **do**
    for $J = 1$ to n/s_{b2} **do**
      for $K = 1$ to I **do**
        if $I == K$ **then**
          $C[I,J] \leftarrow C[I,J] +$
          $TrMXM_{L-1}(A[I,K], B[K,J])$
        else
          $C[I,J] \leftarrow C[I,J] + MXM(A[I,K], B[K,J])$
        end if
      end for
    end for
  end for
  return C
end function

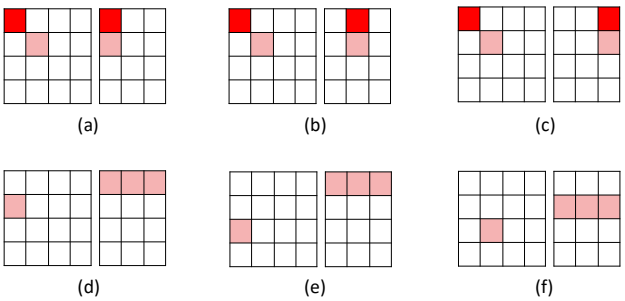


Figure 2: Phase 1 and 2 of $TrMXM_2$. Parts (a)-(c) show three steps from phase 1 where at each step two triangular blocks from the matrix $M1$ are multiplied in parallel to the corresponding blocks of $M2$. Each pair of blocks that are multiplied at each step are shown with the same shade. The resulting matrix $M3$ is not shown. Parts (d)-(f) show three steps from phase 2 where an ordinary block from $M1$ is multiplied concurrently to three blocks from $M2$. For all the parts (a)-(f), $M1$ is the matrix on the left, $M2$ is the matrix on the right and $M3$ is not shown.

Algorithm 6 Two_level Triangular Matrix Multiplication

function $TrMXM_2(M1,M2)$ ▷ Input matrices $M1_{\overline{m}\times\overline{m}}$ and $M2_{\overline{m}\times\overline{n}}$ and the output matrix $M3_{\overline{m}\times\overline{n}}$ are divided into blocks of sizes $\overline{s_{b1}}\times\overline{s_{b1}}$, $\overline{s_{b1}}\times\overline{s_{b2}}$ and $\overline{s_{b1}}\times\overline{s_{b2}}$ respectively. $b1$ is the number of triangular block multiplications done in parallel and $b2$ is the number of ordinary block multiplications done in parallel. $TrMXM_1$ is cblas_dtrmm() in our case.
  for $I = 1$ to $\overline{m}/\overline{s_{b1}}$ increment = $b1$ **do** ▷ phase 1
    for $J = 1$ to $\overline{n}/\overline{s_{b2}}$ **do**
      Copy $b1$ blocks of $M1$ along the main diagonal:
        i.e. $M1[I,I]$ to $M1[I+b1-1, I+b1-1]$
      Copy $b1$ blocks of $M2$ along column J:
        i.e. $M2[I,J]$ to $M2[I+b1-1, J]$.
      start $b1$ **parallel function calls**
        $C[U,J] \leftarrow C[U,J] +$
          $TrMXM_1(M1[U,U], M2[U,J])$
        for $U = I, ..., (I+b1-1)$
      wait for all functions to return
    end for
  end for
  $cnt \longleftarrow 0$
  for $I = 2$ to $\overline{m}/\overline{s_{b1}}$ **do** ▷ phase 2
    for $J = 1$ to $I-1$ **do**
      for $K = 1$ to $\overline{n}/\overline{s_{b2}}$ **do**
        Copy $M1[I,J]$; Copy $M2[J,K]$; $cnt \longleftarrow cnt+1$
        if $cnt == b2$ **then**
          start $b2$ **parallel function calls**
            Call $b2$ instances of $\overline{MXM}$ over
            blocks copied from $M1$ and $M2$
          wait for all functions to return
          Sum up partial results of $\overline{MXM}$
            to $b2$ corresponding blocks of $M3$
          $cnt \longleftarrow 0$
        end if
      end for
    end for
  end for
  return $M3$
end function

Without loss of generality, we assume that the triangular matrix appears at the left side of multiplication. Therefore, in our triangular matrix multiplication problem, matrices A and B of sizes $m\times m$ and $m\times n$ respectively, are multiplied resulting in a matrix C of size $m\times n$. Notice that A being triangular means that it should be a square matrix. Again, without loss of generality, we assume A is lower triangular.The general framework is also applicable here. However, the nature of the problem implies some modifications in our implementation. Here, we divide A into blocks of size $s_{b1}\times s_{b1}$. If $g = \frac{m}{s_{b1}}$, then we will have $g\times g$ blocks, g of which are triangular, $\frac{g\times g-g}{2}$ of which are ordinary (non _ triangular) blocks and $\frac{g\times g-g}{2}$ of which are zero blocks.The zero blocks are ignored while the ordinary blocks are involved in ordinary matrix multiplication and the triangular blocks get involved in triangular matrix multiplication.

Algorithm 5 describes the *L_level* algorithm ($TrMXM_L$) where $L > 2$. Here, we break A, B and C into blocks of sizes $s_{b1}\times s_{b1}$,

$s_{b1} \times s_{b2}$ and $s_{b1} \times s_{b2}$ respectively. In lines 6, 7 the triangular blocks from A are multiplied to the corresponding blocks from B whereas in line 9 the ordinary blocks from A are multiplied to the corresponding blocks from B. Here, MXM is an efficient presumably multilevel matrix multiplication algorithm.

For our two-level algorithm we make an observation: cblas_dtrmm that is used at the leaf level of triangular matrix multiplication is optimized to benefit from the lower amount of computations that result from the zero elements. Therefore, we can further improve the runtime by grouping the triangular blocks together and the ordinary blocks together. That way, the more efficient triangular matrix multiplications should not wait for the results of ordinary matrix multiplications. At each iteration, we operate either on $b1$ triangular block multiplications or on $b2$ ordinary block multiplications in parallel where $b1$ can be different from $b2$.

Our algorithm uses two phases in order to correctly select and group together the matrix blocks. In the first phase, $b1$ triangular blocks from the main diagonal of $M1$ are copied and at each step, they are multiplied concurrently to one other block from $M2$. In the second phase, ordinary blocks from $M1$ are picked in a top-to-bottom and left-to-right order. For each block x from $M1$, we iterate over all blocks of $M2$ to which x should be multiplied. When the total number of block pairs reaches $b2$, we stop and perform $b2$ multiplications concurrently.

The above-mentioned process is illustrated in figure 2. The top row shows three steps from phase 1 and the bottom row shows three steps from phase 2. In each part, $M1$ is the matrix on the left and $M2$ is the matrix on the right where a 4×4 grid is used to block $M1$ and a 4×3 grid is used to block $M2$. Here, $b1 = 2$ and $b2 = 3$. In part (a), two triangular blocks $M1[1, 1]$ and $M2[2, 2]$ are copied to the fast memory; and through steps (a) - (c), those blocks are concurrently multiplied to $M2[1, J]$ and $M2[2, J]$ respectively ($J = 1, 2, 3$). In part (d), the ordinary block $M1[2, 1]$ is copied three times in the fast memory and three parallel multiplications are done on $M1[2, 1]$ and $M2[1, J], J = 1, 2, 3$. In parts (e) and (f) the next ordinary blocks from $M1$ are chosen in the top-to-bottom and left-to-right order and multiplied concurrently with three blocks of $M2$. Note that if the degree of concurrency for the ordinary blocks ($b2$) was equal to 4, then in part (d), block $M1[3, 1]$ would also be copied along with a second copy of block $M2[1, 1]$. Then there would be 4 parallel multiplications, three of them involving $M1[2, 1]\times M2[1, J](J = 1, 2, 3)$ and one of them $M1[3, 1] \times M2[1, 1]$. The cases for parts (e) and (f) would be similar.

Algorithm 6 describes our 2-level triangular matrix multiplication algorithm, $TrMXM_2$. We divide the input matrices $M1_{\overline{m}\times\overline{m}}$ and $M2_{\overline{m}\times\overline{n}}$ into blocks of sizes $\overline{s_{b1}} \times \overline{s_{b1}}$ and $\overline{s_{b1}} \times \overline{s_{b2}}$ respectively. The output matrix $M3_{\overline{m}\times\overline{n}}$ will also be divided into blocks of size $\overline{s_{b1}} \times \overline{s_{b2}}$. Lines 2-14 implement the first phase of the algorithm where we run $b1$ parallel cblas_dtrmm instances on the triangle-rectangle block multiplications. Lines 15-33 implement the second phase of our algorithm where we run $b2$ parallel instances of $\overline{MXM}$ over the square-rectangle block multiplications where $\overline{MXM}$ is an efficient matrix multiplication algorithm. The auxiliary variable *cnt* is used to keep track of the square blocks being multiplied in parallel.

3.4 Symmetric Matrix Multiplication

In this section, we illustrate our methodology on symmetric matrix multiplication. This is when one of the matrices involved in the multiplication is symmetric. The highly optimized parallel function we use at the first level is cblas_dsymm from Intel MKL.

Without loss of generality, we assume that the symmetric matrix appears at the left side of multiplication. Therefore, the problem can be stated as the following: matrices A and B of sizes $m \times m$ and $m \times n$ respectively where A is a symmetric matrix, are multiplied and the result is stored in matrix C of size $m \times n$. Notice again that symmetry is only defined for square matrices; that is why A is $m \times m$. The same methodology is used as the general framework. Here again, similar to triangular multiplication, the nature of the problem implies some modifications to the implementation details. We divide A into blocks of size $s_{b1} \times s_{b1}$. If $g = \frac{m}{s_{b1}}$, then we will have $g \times g$ blocks, g of which are symmetric. The ordinary blocks are involved in ordinary matrix multiplication and the symmetric blocks get involved in symmetric matrix multiplication.

Algorithm 7 describes the *L_level* algorithm ($SymMXM_L$) where $L > 2$. Here, we break A, B and C into blocks of sizes $s_{b1} \times s_{b1}$, $s_{b1} \times s_{b2}$ and $s_{b1} \times s_{b2}$ respectively. In lines 6, 7 the symmetric blocks from A are multiplied to the corresponding blocks from B whereas in line 9 ordinary blocks from A are multiplied to the corresponding blocks from B. Here, similar to that of triangular matrix multiplication, MXM is an efficient matrix multiplication algorithm that could be multilevel.

Algorithm 7 *L_Level* Symmetric Matrix Multiplication

1: **function** $SymMXM_L(A, B)$ ▷ Input matrices $A_{m\times m}$, $B_{m\times n}$ and output matrix $C_{m\times n}$ are divided into blocks of sizes $s_{b1} \times s_{b1}$, $s_{b1} \times s_{b2}$ and $s_{b1} \times s_{b2}$ respectively.
2: **for** $I = 1$ to m/s_{b1} **do**
3: **for** $J = 1$ to n/s_{b2} **do**
4: **for** $K = 1$ to m/s_{b1} **do**
5: **if** $I == K$ **then**
6: $C[I, J] \leftarrow C[I, J] +$
7: $SymMXM_{L-1}(A[I, K], B[K, J])$
8: **else**
9: $C[I, J] \leftarrow C[I, J] + MXM(A[I, K], B[K, J])$
10: **end if**
11: **end for**
12: **end for**
13: **end for**
14: **return** C
15: **end function**

Algorithm 8 describes our 2-level symmetric matrix-matrix multiplication algorithm, $SymMXM_2$. We divide $M1_{\overline{m}\times\overline{m}}$ and $M2_{\overline{m}\times\overline{n}}$ into blocks of sizes $\overline{s_{b1}} \times \overline{s_{b1}}$ and $\overline{bs_1} \times \overline{bs_2}$ respectively. The resulting matrix $M3_{\overline{m}\times\overline{n}}$ will also be divided into blocks of size $\overline{s_{b1}} \times \overline{s_{b2}}$. Similar to triangular matrix multiplication, we group the symmetric blocks together and the ordinary blocks together. The algorithm works in two phases where at each step of the first phase, $b1$ symmetric blocks multiplications and at each step of the second phase $b2$ ordinary block multiplications are performed in parallel. $b1$ and $b2$ are not necessarily equal. Lines 2-14 implement the first phase

and lines 16-33 implement the second phase of our algorithm. The auxiliary variable *cnt* is used to keep track of the ordinary blocks being multiplied in parallel. $\overline{MXM}$ is an efficient matrix multiplication algorithm.

Algorithm 8 Two_level Symmetric Matrix Multiplication

1: **function** $SymMXM_2(M1, M2)$ ▷ Input matrices $M1_{\overline{m}\times\overline{m}}$ and $M2_{\overline{m}\times\overline{n}}$ and the output matrix $M3_{\overline{m}\times\overline{n}}$ are divided into blocks of sizes $\overline{s_{b1}} \times \overline{s_{b1}}$, $\overline{s_{b1}} \times \overline{s_{b2}}$ and $\overline{s_{b1}} \times \overline{s_{b2}}$ respectively. $b1$ is the number of symmetric block multiplications done in parallel and $b2$ is the number of ordinary block multiplications done in parallel. $SymMXM_1$ is cblas_dsymm() in our case.
2: **for** $I = 1$ to $\overline{m}/\overline{s_{b1}}$ increment $= b1$ **do** ▷ phase 1
3: **for** $J = 1$ to $\overline{n}/\overline{s_{b2}}$ **do**
4: Copy $b1$ blocks of $M1$ along the main diagonal:
5: i.e. $M1[I, I]$ to $M1[I + b1 - 1, I + b1 - 1]$
6: Copy $b1$ blocks $M2$ along column J:
7: i.e. $M2[I, J]$ to $M2[I + b1 - 1, J]$.
8: **start** $b1$ **parallel function calls**
9: $C[U, J] \leftarrow C[U, J] +$
10: $SymMXM_1(M1[U, U], M2[U, J])$
11: for $U = I, ..., (I + b1 - 1)$
12: **wait for all functions to return**
13: **end for**
14: **end for**
15: $cnt \longleftarrow 0$
16: **for** $I = 1$ to $\overline{m}/\overline{s_{b1}}$ **do** ▷ phase 2
17: **for** $J = 1$ to $\overline{m}/\overline{s_{b1}}, I \neq J$ **do**
18: **for** $K = 1$ to $\overline{n}/\overline{s_{b2}}$ **do**
19: Copy $M1[I, J]$
20: Copy $M2[J, K]$
21: $cnt \longleftarrow cnt + 1$
22: **if** $cnt == b2$ **then**
23: **start** $b2$ **parallel function calls**
24: Call $b2$ instances of $\overline{MXM}$ over
25: blocks copied from $M1$ and $M2$
26: **wait for all functions to return**
27: Sum up partial results of $\overline{MXM}$
28: to $b2$ corresponding blocks of $M3$
29: $cnt \longleftarrow 0$
30: **end if**
31: **end for**
32: **end for**
33: **end for**
34: **return** $M3$
35: **end function**

3.5 Performance Analysis

In this section, a simple shared-memory model in which processes communicate by reading/writing from/to a shared memory is used to describe the runtime of the parallel codes. Using this model, we formulate the conditions under which our method will be effective and also the speed up we can expect. Let $T^i_{x,y}$ be the time our algorithm takes at level i to solve a problem of size y using x processors. Table 2 illustrates our notation for a level 1 code. Here, b and c are integer multipliers.

Table 2: Notations used for the runtime of 1-level algorithm depend on the problem size and the number of processors.

	p processors	bp processors
problem size= n	$T^1_{p,n}$	$T^1_{bp,n}$
problem size= cn	$T^1_{p,cn}$	$T^1_{bp,cn}$

Let,

$$r = \frac{T^1_{bp,cn}}{T^1_{p,cn}} \tag{2}$$

$$k = \frac{T^1_{p,cn}}{f(c) \times T^1_{p,n}} \tag{3}$$

Where $f(c)$ is defined as follows: if we decompose a problem of size cn into subproblems of size n; then $f(c)$ denotes the number of those subproblems. In the case of matrix multiplication, for example, if the initial problem is multiplying two $2n \times 2n$ matrices and we break each matrix into four matrices of size $n \times n$; then we should multiply eight matrix pairs of size $n \times n$ each. Here, we have $c = 2$ and $f(c) = 8$.

Let $T_i = T^i_{bp,cn}$. We wish to find the ratio T_i/T_1. We consider two cases: $i = 2$ and $i > 2$. The reason we consider two cases is that we use parallelism only at the second level.

We define $g(c, n)$ to be the time it takes to combine $f(c)$ subproblems of size n each. For a 2-level algorithm, if the subproblems run serially, then all the bp processors can be used for each of them. We will have:

$$T^2_{bp,cn} = f(c)T^1_{bp,n} + g(c, n) \tag{4}$$

However, if we run b subproblems in parallel at a time, then only p processors can be used for each of the subproblems. In that case, the runtime equation will be:

$$T^2_{bp,cn} = \frac{f(c)}{b}T^1_{p,n} + g(c, n) \tag{5}$$

In the rest of this subsection, we use equation 5 to describe T_2. That is

$$T_2 = \frac{f(c)}{b}T^1_{p,n} + g(c, n)$$

Therefore,

$$\frac{T_2}{T_1} = \frac{1 + rbk\delta}{rbk} \tag{6}$$

where

$$\delta = \frac{g(c, n)}{T_1} \tag{7}$$

For level $i > 2$, we consider c_1 and c_2 to be divisors of c, i.e. $c = c_1c_2$. The time to solve a problem of size $cn = c_1c_2n$ using bp processors and an i-level algorithm will be:

$$T_i = T^i_{bp,cn} = f(c_1)T^{i-1}_{bp,c_2n} + g(c_1, c_2n) \tag{8}$$

where $g(c_1, c_2n)$ is the time to combine $f(c_1)$ subproblems of size c_2n.

3.6 Speed Up Analysis

In this section, we formulate the conditions for when increasing the number of levels results in speed up. Throughout this section, we consider the problem size to be cn and the total number of available processors to be bp. Also, as mentioned before, we use different sets of conditions for T_2 and $T_i, i > 2$ and the reason is that we only run the subproblems in parallel at level 2.

3.6.1 $T_2 < T_1$

The condition for a 2-level algorithm to be faster than a 1-level algorithm is as follows:

$$T_2 < T_1 \Rightarrow \frac{T_2}{T_1} < 1$$

$$\Rightarrow \frac{1 + rbk\delta}{rbk} < 1 \Rightarrow rbk(1-\delta) > 1 \tag{9}$$

In the formulation for δ, we assume that the combination time is negligible compared to T_1 i.e. $g(c,n) \ll T_1$. Then $\delta \to 0$ and the condition for $T_2 < T_1$ will be:

$$rbk > 1 \tag{10}$$

The expected speed up is ${}^{T_1}/_{T_2} = rbk$ (as $\delta \to 0$).

3.6.2 $T_i < T_1 (i > 2)$

Using equation (8) we get the following condition for an i-level algorithm ($i > 2$) to be faster than a 1-level algorithm:

$$T_i < T_1 \Rightarrow \frac{T_i}{T_1} < 1$$

$$\Rightarrow f(c_1)\,\frac{T^{i-1}_{bp,c_2n}}{T_1} + \frac{g(c_1,c_2n)}{T_1} < 1 \tag{11}$$

Again, assuming $g(c_1, c_2n) \ll T_1$, the inequality (11) will reduce to

$$f(c_1)\,\frac{T^{i-1}_{bp,c_2n}}{T_1} < 1 \tag{12}$$

Using equations 5 and 8 for $i = 3$ we get:

$$T^{i-1}_{bp,c_2n} = \frac{f(c_2)}{b} T^{i-2}_{p,n} + g(c_2,n) \tag{13}$$

$$T_{i-1} = T^{i-1}_{bp,cn} = \frac{f(c_1c_2)}{b} T^{i-2}_{p,n} + g(c_1c_2,n) \tag{14}$$

and for $i > 3$

$$T^{i-1}_{bp,c_2n} = f(c_2) T^{i-2}_{bp,n} + g(c_2,n) \tag{15}$$

$$T_{i-1} = T^{i-1}_{bp,cn} = f(c_1c_2) T^{i-2}_{bp,n} + g(c_1c_2,n) \tag{16}$$

We make the assumption that $f(.)$ is separable, meaning that $f(c_1c_2) = f(c_1)f(c_2)$. Assuming that $g(c_2,n)$ and $g(c_1c_2,n)$ are negligible compared to T_1; for all values of $i > 2$ we will have:

$$f(c_1)\,\frac{T^{i-1}_{bp,c_2n}}{T_1} = \frac{T_{i-1}}{T_1} \tag{17}$$

Combining equations (17) and (12), we get

$$\frac{T_{i-1}}{T_1} < 1 \tag{18}$$

that is (by induction) $rbk > 1$.

3.6.3 $T_i < T_{i-1} (i > 3)$

In the previous section, the condition for $T_i < T_1$ was established. Even if $T_i < T_1$, it may be better to stay at level $i-1$ and not to increase the level. To decide whether or not we can use another level, we can consider the $(i-1)$-level code as a new 1-level code. Then recalculate r, k and δ for this code and set $b = 1$ (that is because we don't use parallelism for $i > 2$). We call these newly computed values r', k', δ' and b'. Then the condition for $T_i < T_{i-1}$ will be $r'b'k' > 1$. As $b' = 1$ the condition will reduce to $r'k' > 1$.

3.7 Performance Analysis: Matrix Multiplication

The general model can be applied to matrix multiplication. Without loss of generality, we assume the matrices are square $cn \times cn$ matrices divided into blocks of size $n \times n$. The model details will be as follows:

- $f(c) = c^3$ therefore by plugging it in (3) we get

$$k = \frac{T^1_{p,cn}}{c^3 T^1_{p,n}} \tag{19}$$

- $g(c,n) = (c-1)c^2a(n)$. Here $a(n)$ is the time it takes to add two matrix blocks of size $n \times n$ and we should add $(c-1)c^2$ of them to get the final result for a $cn \times cn$ matrix multiplication. By plugging it in (7) we will have:

$$\delta = \frac{(c-1)c^2a(n)}{T_1} \tag{20}$$

Therefore, for matrix multiplication, equations (5) and (8) will be:

$$T^2_{bp,cn} = \frac{c^3}{b} T^1_{p,n} + (c-1)c^2a(n) \tag{21}$$

$$T_i = {c_1}^3 T^{i-1}_{bp,c_2n} + (c_1-1)c_1^2 a(c_2n) \tag{22}$$

As $f(c)$ is separable and the addition time is negligible compared to multiplication time for large enough matrices (and therefore we can safely assume that $\delta \to 0$), all the speed up analysis remains valid.

3.8 Performance Analysis: Triangular Matrix Multiplication

Our multilevel method for triangular matrix multiplication is a compound method. Meaning that for solving the subproblems, we should call both triangular matrix multiplication and ordinary matrix multiplication functions on the smaller instances. Consider the following equations for $T^2_{bp,cn}$ and $T^i_{bp,cn} (i > 2)$ respectively, where - as before - $c = c_1c_2$:

$$T^2_{bp,cn} = \frac{f_1(c)}{b1} T^1_{p_1,n} + \frac{f_2(c)}{b2} \tau_{p_2,n} + g(c,n) \tag{23}$$

$$\begin{aligned} T^i_{bp,cn} &= f_1(c_1) T^{i-1}_{bp,c_2n} \\ &+ f_2(c_1)\tau_{bp,c_2n} + g(c_1,c_2n) \end{aligned} \tag{24}$$

In both (23) and (24) $f_1(.)$ is the number of triangular matrix multiplication subproblems and $f_2(.)$ is the number of ordinary matrix multiplication subproblems. $b1$ is the degree of concurrency for triangular block multiplication and $b2$ is the degree of concurrency for ordinary block multiplication. $p_1 = {}^{(bp)}/_{b1}$ and $p_2 = {}^{(bp)}/_{b2}$. We use only one $g(.,.)$ function because the result of multiplying a triangular block with an ordinary block will not necessarily be triangular and therefore there is no need to distinguish between the combination times. Also, $\tau_{x,y}$ is the runtime of an ordinary matrix multiplication function on blocks of size y using x processors.
Let,

$$\theta = \frac{b2}{b1} \tag{25}$$

$$\lambda = \frac{\tau_{p_2,n}}{T^1_{p_1,n} \times \theta} \tag{26}$$

Now, if we define:

$$b = b1 \tag{27}$$

$$f(x) = f_1(x) + \lambda f_2(x) \tag{28}$$

equations (23) and (24) will turn to

$$T^2_{bp,cn} = \frac{f(c)}{b} T^1_{p,n} + g(c,n) \tag{29}$$

$$T^i_{bp,cn} = f(c_1) T^{i-1}_{bp,c_2 n} + g(c_1, c_2 n) \tag{30}$$

that are the same as equations 5 and 8. Again, without loss of generality, we assume that the matrices are square, their sizes are $cn \times cn$ and their block sizes are $n \times n$. As a result,

- $f_1(c) = c^2$ and $f_2(c) = \frac{c^3 - c^2}{2}$. By using (28) we get:

$$f(c) = \frac{\lambda c^3 + (2 - \lambda)c^2}{2} \tag{31}$$

- $g(c,n) = \frac{c^3 - c^2}{2} a_n$ where a_n is the time to add two $n \times n$ blocks.

We can safely assume that $g(c,n) \ll T_1$ for large enough matrices and therefore the condition for $T_2 < T_1$ still holds. However, $f(c)$ is not separable, therefore the conditions for $T_i < T_1$ for $i > 2$ will not necessarily work.

3.9 Performance Analysis: Symmetric Matrix Multiplication

The performance analysis for symmetric matrix multiplication has a lot in common with the performance analysis for triangular matrix multiplication.Equations 23 - 30 are valid for symmetric matrix multiplication. The definitions for $f_1(.)$, $f_2(.)$, $b1$, $b2$, θ and λ are the same as those in section 3.8 and the same equations $b = b1$ and $f(x) = f_1(x) + \lambda f_2(x)$ work for symmetric matrices. The differences are in details as described in the following (again, without loss of generality, we assume that the matrices to be multiplied are both square matrices of sizes $cn \times cn$ which are divided into blocks of sizes $n \times n$):

- $f_1(c) = c^2$ and $f_2(c) = c^3 - c^2$. By using (28) we get:

$$f(c) = \lambda c^3 + (1 - \lambda)c^2 \tag{32}$$

- $g(c,n) = c^2(c-1)a_n$ with a_n being the time to add two $n \times n$ blocks.

Once again, we can assume that $g(c,n) \ll T_1$ for large enough matrices and therefore the condition for $T_2 < T_1$ holds. However, the function $f(c)$ is only separable if $\lambda = 1$ and that is the only case where the conditions $T_i < T_1 (i > 2)$ will work.

4 KNL Architecture

4.1 Architecture Overview

Knights Landing (KNL) [32] is the codename for the second generation Intel Xeon Phi product family. It is a many-core architecture enabling highly parallel workloads. KNL CPU includes up to 36 active tiles. Each tile includes 2 cores, 2 vector processing units per core and one L_2 cache shared between the cores. A 2D mesh interconnect provides connection between the tiles, memory controllers and other on-board elements. The mesh supports the MESIF (modified, exclusive, shared, invalid, forward) cache-coherence protocol and uses a distributed tag directory to keep all the L_2 caches coherent in all the tiles.

KNL includes 2 types of memory: 16GB of multichannel DRAM (MCDRAM) and up to 384 GB of double data rate (DDR) memory. MCDRAM is a high bandwidth memory (HBM) that provides an aggregate bandwidth of more than 450 GBps. The aggregate bandwidth of DDR is more than 90 GBps.

4.2 Memory Modes

The two memory modules can be configured in three modes explained below:

4.2.1 Flat Mode

In this mode, MCDRAM is treated as an addressable memory alongside DDR. Flat mode gives the user the ability to allocate data either from DDR or MCDRAM. The downside is that it requires software modification.

4.2.2 Cache Mode

In cache mode, MCDRAM is configured as a memory-side cache for the whole DDR. Here, the user has no control over MCDRAM usage, but no software modification is required.

4.2.3 Hybrid Mode

This mode is a combination of flat mode and cache mode. In hybrid mode, a portion of MCDRAM (either 0.25 or 0.5) is used in cache mode and the remaining portion is used in flat mode.

4.3 KNL Clustering Modes

The mesh interconnect provides five different clustering modes. Each of these clustering modes defines the affinity properties of tiles, tag directories and memory controllers. These modes are explained in the following:

4.3.1 All-to-All

This is the most general mode that lacks any affinity between tiles, directory and memory.

4.3.2 Quadrant

This mode divides the KNL chip into four virtual quadrants. A request from any tile can go to any directory, but the directory can only access the memory in its own quadrant. Meaning that this mode provides affinity between directory and memory. However, there is no affinity between a tile and the directory or memory.

4.3.3 Hemisphere

This mode is similar to the quadrant mode, with the only difference that the chip is divided into 2 hemispheres instead of 4 quadrants.

4.3.4 SNC4

SNC stands for Sub-NUMA Clustering. SNC4 is a more restricted version of the quadrant mode where the quadrants are viewed by the OS as nonuniform memory access (NUMA) nodes (or clusters). Here, a request form a tile accesses the directory in the same cluster and the directory will also access a memory controller in that same cluster. Meaning that this mode provides affinity between tiles, memory directories and memory controllers.

4.3.5 SNC2

This mode is similar to SNC4 the only difference being the number of clusters. Here the number of clusters is two instead of four.

5 Experimental Results

We implemented our algorithms using C++ and OpenMP. The total number of threads was set to 64. Using a larger number of threads led to performance degradation. We used the Intel icpc compiler with the flags setting "-O3 -xMIC-AVX512 -mkl -lmemkind -qopenmp". In "flat" memory mode, we have to manually allocate memory from MCDRAM. This was done using "hbw_posix_memalign" command. To allocate memory from DDR, the command "posix_memalign" was used.

For all three of our algorithms, namely matrix-matrix multiplication, triangular matrix-matrix multiplication and symmetric matrix-matrix multiplication, we used double precision square matrices as test data for both the operands. For matrix pairs of sizes up to $16K \times 16K$, the reported times are the average over 10 run times. For matrix pairs of sizes $32K \times 32K$ and $64K \times 64K$, we used only 5 runs. That is because those runs were very time consuming and also because at those sizes, the runtime values were quite stable and 5 runs seemed to be sufficient. The error bound for our reported times is at most 4%.

With 3 memory modes and 5 clustering modes, there are 15 possible architectural configurations in the KNL. Of these, the *cache, all-to-all* mode is not supported. So, in reality, the KNL has 14 architectural configurations. In our experiments, we do not consider the hybrid memory mode. This leaves us with 9 configurations to consider. There are at least three different scenarios for configuration selection; the selection of scenario being limited by system and application constraints.

1. S1: Each algorithm can select the configuration to run on based on the size of the matrices to be multiplied.
2. S2: The configuration is determined by the application and cannot be changed by the algorithm.
3. S3: Each algorithm must use the same configuration for all instances; different algorithms can use different configurations.

In Sections 5.1-5.3, we give the measured average run times for the three versions of matrix multiplication considered in this paper for all 9 of the architectural configurations considered. A comparison of these run times for each of the three scenarios (S1-S3) of configuration selection is done in Section 5.4. The speed up values reported in Section 5.4 are the percentage of runtime reduction $(T_{old}-T_{new})/T_{old}$.

5.1 Matrix Multiplication Times

Table 3 shows the runtime values for the code we use as the 1-level matrix multiplication algorithm (cblas_dgemm) for different matrix sizes and different KNL memory/clustering configurations.

Table 4 contains the runtimes for our 2-level algorithm. For all the matrix sizes, the input and output matrices are divided into sub-matrices using a 4×4 grid. The degree of concurrency used in our 2-level implementation is ($b = 4$), meaning that at each step 4 pairs of matrix blocks are multiplied in parallel. We have also experimented with other block arrangements such as 2×2 and other degrees of concurrency such as 2 and 8; however, the current configuration resulted in the best runtimes.

Tables 5 and 6 show the runtimes for our 3 and 4-level algorithms respectively. For both the levels, we used 2×2 grids to divide the input and output matrices into sub-matrices. The block pairs were multiplied sequentially.

5.2 Triangular Matrix Multiplication Times

Table 7 shows the run times for the one-level code (cblas_dtrmm). This table together with Table 3 show the speed up potential using our method. Consider $64K \times 64K$ multiplications as an example. Instead of performing one $64K \times 64K$ triangular matrix multiplication; we can multiply six $32K \times 32K$ blocks and combine the results. Out of those block multiplications, four of them are triangular matrix multiplication and two of them are ordinary matrix multiplication. If we consider the combination time to be small enough, then on all the memory-clustering configurations, the time to sequentially perform two ordinary multiplications of $32K \times 32K$ matrices plus the time to sequentially perform four triangular multiplications of $32K \times 32K$ matrices is less than the time to perform one $64K \times 64K$ triangular matrix multiplication. That notifies us about the potential of speed up. Note that this is just an initial evaluation and in practice we can also use parallelism and different settings for block sizes.

Table 8 shows the run time values for our two-level algorithm. Here, matrix A is divided using a 2×2 grid and matrices B and C are divided using a 2×1 block arrangement. Degree of concurrency for

Table 3: Run time (in seconds) of the 1-level matrix multiplication using cblas_dgemm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$4K \times 4K$	0.65	0.39	0.30	0.33	0.30
$8K \times 8K$	3.41	1.99	1.64	1.49	**1.39**
$16K \times 16K$	23.21	15.42	12.74	11.92	**10.25**
$32K \times 32K$	119.81	121.47	102.53	101.58	102.63
$64K \times 64K$	1002.30	843.25	843.04	826.23	824.15
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$4K \times 4K$	0.48	0.42	—	0.32	**0.29**
$8K \times 8K$	1.94	1.71	—	1.51	**1.39**
$16K \times 16K$	10.89	11.32	—	12.32	10.36
$32K \times 32K$	88.90	87.39	—	97.52	**80.18**
$64K \times 64K$	**618.26**	781.83	—	871.10	652.60

Table 4: Run time (in seconds) of the 2-level matrix multiplication algorithm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$4K \times 4K$	0.36	0.32	0.30	0.30	**0.28**
$8K \times 8K$	2.42	1.52	1.27	1.21	1.25
$16K \times 16K$	113.51	14.82	10.12	8.87	8.85
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$4K \times 4K$	0.31	0.29	—	0.30	0.32
$8K \times 8K$	1.62	1.36	—	**1.17**	1.21
$16K \times 16K$	11.71	11.54	—	**8.15**	8.61

Table 5: Run time (in seconds) of the 3-level matrix multiplication algorithm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$32K \times 32K$	151.12	119.23	80.62	70.50	70.28
$64K \times 64K$	10540.52	884.25	768.97	720.37	653.97
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$32K \times 32K$	93.53	94.26	—	**64.98**	68.35
$64K \times 64K$	729.14	805.44	—	686.36	**651.94**

Table 6: Run time (in seconds) of the 4-level matrix multiplication algorithm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$64K \times 64K$	1292.55	945.01	640.35	560.46	557.92
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$64K \times 64K$	744.34	745.16	—	**517.99**	544.31

Table 7: Run time (in seconds) of the 1-level triangular matrix multiplication using cblas_dtrmm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$4K \times 4K$	0.46	0.40	**0.34**	0.36	0.41
$8K \times 8K$	2.43	1.50	1.31	1.41	**1.12**
$16K \times 16K$	10.97	10.38	7.83	9.12	6.21
$32K \times 32K$	73.78	79.32	62.85	71.24	62.82
$64K \times 64K$	556.95	588.33	503.83	556.77	501.58
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$4K \times 4K$	0.49	0.38	—	0.37	0.38
$8K \times 8K$	2.04	1.43	—	1.61	**1.12**
$16K \times 16K$	7.99	7.93	—	15.40	**6.20**
$32K \times 32K$	54.03	59.04	—	77.93	**50.15**
$64K \times 64K$	450.17	445.05	—	596.00	**389.26**

Table 8: Run time (in seconds) of the 2-level triangular matrix multiplication algorithm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$4K \times 4K$	0.44	0.40	0.35	**0.33**	0.35
$8K \times 8K$	1.65	1.32	1.16	1.33	**1.04**
$16K \times 16K$	10.02	9.52	6.91	7.72	6.02
$32K \times 32K$	69.81	59.07	52.05	51.58	51.64
$64K \times 64K$	868.32	566.61	382.54	380.42	380.55
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$4K \times 4K$	0.35	0.35	—	**0.33**	0.32
$8K \times 8K$	1.11	1.13	—	1.26	**1.04**
$16K \times 16K$	6.87	6.60	—	7.64	**5.97**
$32K \times 32K$	50.60	50.57	—	49.77	**42.98**
$64K \times 64K$	402.92	414.56	—	422.38	**356.40**

triangular blocks is $b1 = 2$ and for ordinary blocks is $b2 = 1$. Moreover, for square-rectangle block multiplication, we called MXM_2 instead of cblas_dgemm. In the MXM_2 algorithm we used, the grids imposed on the matrices were 4×4, 4×8 and 4×8 for A, B and C respectively.

5.3 Symmetric Matrix Multiplication Times

Table 9 shows the run time values for the one-level symmetric matrix-matrix multiplication code (cblas_dsymm). Similar to what we did for triangular matrix multiplication, here again we can use this table along with Table 3 to make sure that there is the potential for performance improvement. Consider the multiplication time for $64K \times 64K$ matrices as an example. We can break matrices into $32K \times 32K$ blocks and perform 8 block multiplications 4 of which are symmetric matrix multiplication and the remaining 4 are ordinary matrix multiplication. For all memory-clustering configurations (except for flat-snc2) the time to sequentially perform four ordinary $32K \times 32K$ multiplication plus the time to sequentially perform four symmetric $32K \times 32K$ multiplications is less than the time to perform one symmetric $64K \times 64K$ multiplication. As a result, the initial test has passed and the potential for speed up exists for a simple sequential model using 2×2 grids to block the matrices. We can also use parallelism and different settings for block dimensions.

Table 10 shows the run time values for our two-level algorithm. In this implementation, matrix A is divided using a 2×2 grid and matrices B and C are divided using a 2×1 grid each. The degree of concurrency for both symmetric and ordinary blocks is $b1 = b2 = 2$.

5.4 Run Time Comparison For Different Configuration Selection Scenarios

Throughout this section, we use the layout convention given in Table 11 when we refer to "our algorithm". For example, the 3-level algorithm for matrix multiplication is the same one for which we reported runtimes in Section 5.1 and has the same parameters.

Figures 3, 4 and 5 compare the algorithms under scenario S1 (each algorithm uses the best configuration for each matrix size), for matrix multiplication, triangular matrix multiplication and symmetric matrix multiplication respectively. For matrix multiplication, our algorithm results in 20.5% speed up for $16K \times 16K$ matrices, 19.0% for $32K \times 32K$ matrices and 16.2% for $64K \times 64K$ matrices. The corresponding speed up values are 3.7%, 14.3% and 8.4% for triangular matrix multiplication and -0.5%, 11.5% and 10.5% for symmetric matrix multiplication.

Tables 12, 13 and 14 show the comparison results using scenario S2 (all algorithms use the same configuration) for matrix multiplication, triangular matrix multiplication and symmetric matrix multiplication respectively. The comparisons are expressed as the percentage runtime reduction obtained by our algorithm over the 1-level code for each of the above-mentioned applications. In all the three tables, the negative percentages (where our algorithm has performed worse than the 1-level code) appear at the configurations involving snc2 or snc4 (Table 14 is exception, but the negative speed up values we get for other configurations have very small absolute values). We will briefly explain in Section 5.5 why our method performs poorly in the Sub-NUMA Clustering modes. If we ignore the modes involving snc2 and snc4, then for matrix multiplication on $16K \times 16K$ matrices, the speed up we get is in range [13.7%, 33.8%]; for $32K \times 32K$ matrices it is in range [14.8%, 33.4%] and for $64K \times 64K$ matrices it is in range [16.6%, 40.5%]. The corresponding ranges for triangular matrix multiplication are [3.1%, 50.4%], [14.3%, 36.1%] and [8.4%, 31.7%] and for symmetric matrix multiplication are [−0.9%, 50.9%], [10.3%, 28.2%] and [11.6%, 30.1%].

For the third scenario S3, we make the following observations ($conf_1$ is the best configuration for the 1-level algorithm and $conf_m$ is the best one for our multilevel algorithm):

- For matrix multiplication, $conf_1$: cache-hemisphere and $conf_m$: cache-quadrant (for $conf_m$, results of 2-level, 3-level and 4-level algorithms have been considered collectively).
- For triangular matrix multiplication, $conf_1$: cache-hemisphere and $conf_m$: cache-hemisphere.
- For symmetric matrix multiplication, $conf_1$: cache-hemisphere and $conf_m$: cache-hemisphere.

Figures 6, 7 and 8 compare our algorithms vs. the 1-level code for scenario S3. Using our speed up method on matrix multiplication, we get 21.3% speed up for $16K \times 16K$ matrices, 19.0% for $32K \times 32K$ matrices and 20.6% for $64K \times 64K$ matrices. The corresponding speed up values are 3.7%, 14.3% and 8.4% for triangular matrix multiplication and -0.9%, 10.3% and 11.6% for symmetric

Table 9: Run time (in seconds) of the 1-level symmetric matrix multiplication algorithm using cblas_dsymm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$4K \times 4K$	0.67	0.57	**0.47**	0.54	0.56
$8K \times 8K$	4.47	2.63	2.33	2.97	1.99
$16K \times 16K$	24.78	20.75	15.12	32.13	**11.96**
$32K \times 32K$	130.27	152.23	122.94	120.16	123.49
$64K \times 64K$	1006.21	957.23	994.72	1002.23	1005.31
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$4K \times 4K$	0.67	0.55	—	0.53	0.55
$8K \times 8K$	2.98	2.50	—	2.87	**1.97**
$16K \times 16K$	13.43	13.98	—	31.66	12.00
$32K \times 32K$	99.14	109.71	—	153.31	**98.17**
$64K \times 64K$	791.85	**778.54**	—	1346.62	797.71

Table 10: Run time (in seconds) of 2-level symmetric matrix multiplication algorithm

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$4K \times 4K$	0.75	0.51	0.41	0.39	0.39
$8K \times 8K$	3.59	2.59	2.15	2.27	1.80
$16K \times 16K$	24.09	19.95	14.78	15.78	**12.02**
$32K \times 32K$	93.36	164.51	89.49	89.71	90.72
$64K \times 64K$	1631.96	1312.09	711.02	711.50	709.60
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$4K \times 4K$	0.45	0.45	—	**0.36**	0.38
$8K \times 8K$	2.08	2.16	—	2.19	**1.77**
$16K \times 16K$	12.85	13.41	—	16.05	12.11
$32K \times 32K$	94.60	100.01	—	110.11	**88.03**
$64K \times 64K$	826.87	765.08	—	941.59	**704.79**

matrix multiplication. Note that for the latter two applications as $conf_1 = conf_m$, the speed up values are the same as the ones in the "cache-hemisphere" column of Tables 13 and 14 respectively.

Table 11: The algorithm we use (number of levels) depends on the target application and the size of the input matrices.

	$16K \times 16K$ matrices	$32K \times 32K$ matrices	$64K \times 64K$ matrices
Matrix Multiplication (MXM)	2-level algorithm	3-level algorithm	4-level algorithm
Triangular MXM	2-level algorithm		
Symmetric MXM	2-level algorithm		

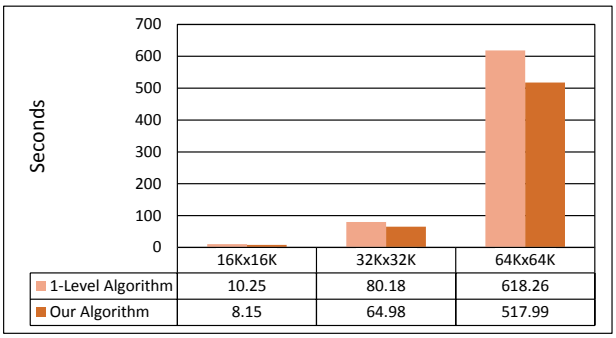


Figure 3: Comparison of our method against the 1-level matrix multiplication algorithm (cblas_dgemm) using scenario S1.

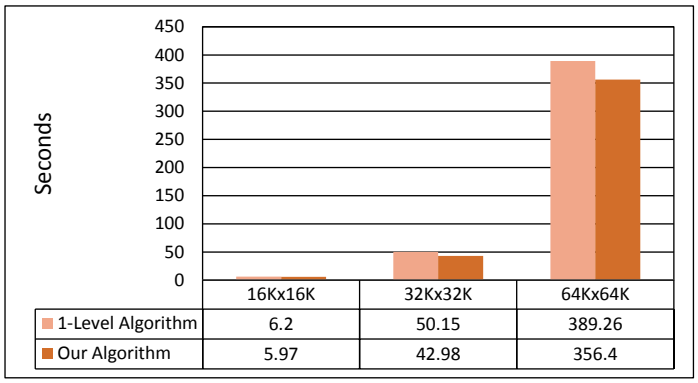


Figure 4: Comparison of our method against the 1-level triangular matrix multiplication algorithm (cblas_dtrmm) using scenario S1.

5.5 *Model Validation*

In this section, we validate the mathematical model we developed in Section 3.5 to predict performance improvement. The validation is done on matrix multiplication. It is similar for triangular and symmetric matrix multiplication problems. We show that the $rbk > 1$ condition correctly predicts the potential for speed up. Also, we compare the actual $\frac{T_i}{T_1}$ ratios with those predicted using our formulation. For predicting the ratios, we use equations (5) and (8) ignoring the $g(.,.)$ terms. We ignore the $g(.,.)$ terms by assuming

that $g(c,n) \ll T_1$ (the same assumption as the one we made several times in Section 3.5). Meaning that we use (33) and (34) to predict the runtime ratios.

$$\frac{T_2}{T_1} = \frac{f(c)}{b} \times \frac{T^1_{p,n}}{T_1} \tag{33}$$

$$\frac{T_i}{T_1} = \frac{f(c_1)T^{i-1}_{bp,c_2 n}}{T_1} \tag{34}$$

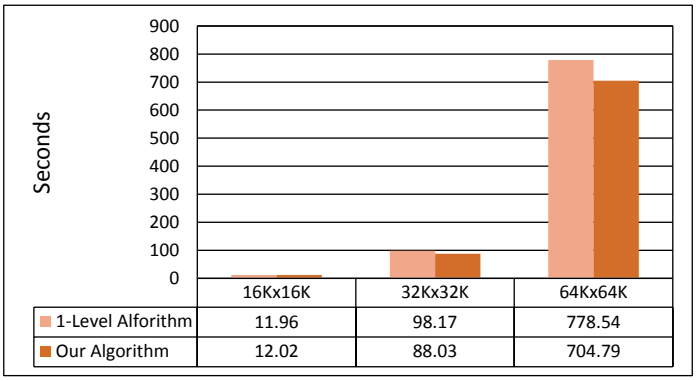


Figure 5: Comparison of our method against the 1-level symmetric matrix multiplication algorithm (cblas_dsymm) using scenario S1.

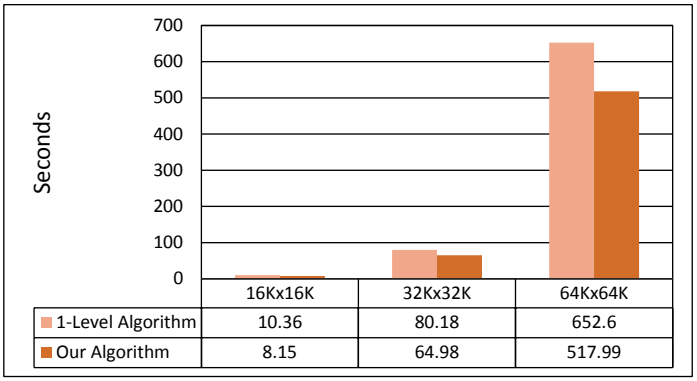


Figure 6: Comparison of our method against the 1-level matrix multiplication algorithm (cblas_dgemm) for scenario S3.

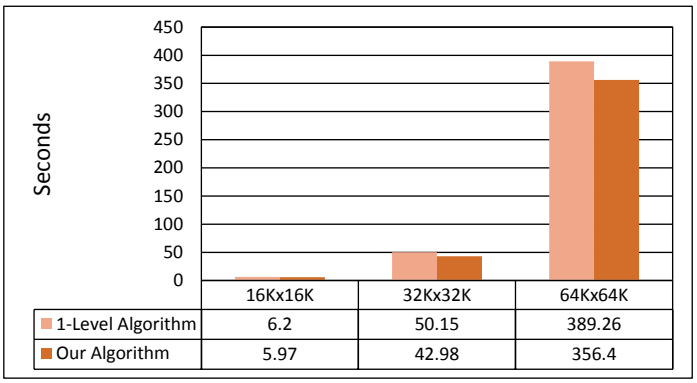


Figure 7: Comparison of our method against the 1-level triangular matrix multiplication algorithm (cblas_dtrmm) for scenario S3.

In our analysis we ignore matrix addition time as well as the time spent on copying the matrix blocks into the fast memory. The reason is that theoretically those are $O(n^2)$ operations while matrix multiplication is an $O(n^3)$ operation. In practice, for our two - level matrix multiplication code in the default KNL mode (cache - quadrant) for instance, the ratio (multiplication time)/(data transfer time) is 3 for matrices of size $4K \times 4K$ and 42 for matrices of size $16K \times 16K$. The ratio (multiplication time)/(addition time) is 21 and 199, for matrix sizes of $4K \times 4K$ and $16K \times 16K$ respectively. Similar ratios can be obtained for other matrix sizes and using other configurations.

Table 15 shows the validation results for our T_2 formula in Equation 5. Here, $cn = 16K$, number of blocks $c = 4$ and therefore $n = 4K$. The degree of concurrency b is 4. Here and also in the following validation experiments, the total number of processors $bp = 64$. The error between the computed value of T_2/T_1 and the measured value ranges from 1.27% (Flat-all2all) to 19.44% (Cache-snc4) in all the modes except for Flat-snc4 (86.09%).

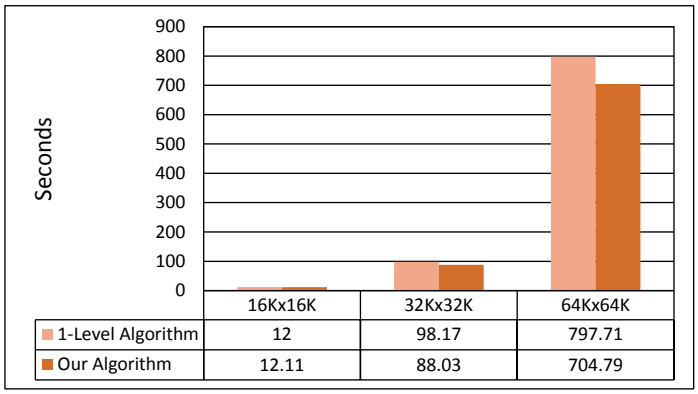


Figure 8: Comparison of our method against the 1-level symmetric matrix multiplication algorithm (cblas_dsymm) for scenario S3.

Table 16 shows the validation results for our T_3 formula. Here, matrix size $cn = 32K$, number of blocks in the first (outermost) level $c_1 = 2$ (arranged as a 2×2 grid) and therefore the outermost blocks are of size $16K \times 16K$. The error between computed T_3/T_1 and the observed value ranges from 0% (Flat-snc2, Flat-all2all) to 1.85% (Cache-snc2) for all the modes except for Flat-snc4 where it is 502%.

Table 17 shows the validation results for our T_4 formula. Here, matrix size $cn = 64K$, number of blocks in the first (outermost) level $c_1 = 2$ (arranged as a 2×2 grid) and therefore the outermost blocks are $32K \times 32K$. Note that in tables 15, 16 and 17 the computed and the actual runtime ratios have been rounded to the nearest hundredth and that the error values are computed based on those rounded ratios. Here, the error between the computed and the observed values ranges from 0% (Flat - quadrant) to 1.69% (Cache - quadrant) in all the modes except for the Flat-snc4 mode where it is 6.2%.

Figures 9, 10 and 11 visually show the actual values versus the computed values for T_3/T_1, T_2/T_1 and T_4/T_1 respectively.

Our validation results show that our model works fairly well for all the KNL configurations except for those involving snc2 or snc4. The reason is that in the Sub-NUMA Clustering modes, each half (or quadrant) of the chip is exposed as a separate domain to the OS, essentially making the chip analogous to a 2 or 4 socket Xeon. This restricts data sharing and memory access time will increase as a result.

The condition for speed up, namely $rbk > 1$ works correctly for 8 out of 9 KNL configurations (the only exception being Flat-snc4). Note that the *rbk* column is the same in all the three tables. That is because its value only depends on the parameters of levels 1 and 2.

Table 12: The speed up obtained by our algorithm over the 1-level matrix multiplication code (cblas_dgemm) for scenario S2,

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$16K \times 16K$	-389.0%	3.9%	20.6%	25.6%	13.7%
$32K \times 32K$	-26.1%	1.8%	21.4%	30.6%	31.5%
$64K \times 64K$	-29.0%	-12.1%	24.0%	32.2%	32.3%
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$16K \times 16K$	-7.5%	-1.9%	—	**33.8%**	16.9%
$32K \times 32K$	-5.2%	-7.9%	—	**33.4%**	14.8%
$64K \times 64K$	-20.4%	4.7%	—	**40.5%**	16.6%

Table 13: The speed up obtained by our algorithm over the 1-level triangular matrix multiplication code (cblas_dtrmm) for Scenario S2.

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$16K \times 16K$	8.7%	8.3%	11.7%	15.4%	3.1%
$32K \times 32K$	5.4%	25.5%	17.2%	27.6%	17.8%
$64K \times 64K$	-55.9%	3.7%	24.1%	**31.7%**	24.1%
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$16K \times 16K$	14.0%	16.8%	—	**50.4%**	3.7%
$32K \times 32K$	6.3%	14.3%	—	**36.1%**	14.3%
$64K \times 64K$	10.5%	6.9%	—	29.1%	8.4%

Table 14: The speed up obtained by our algorithm over the 1-level symmetric matrix multiplication code (cblas_dsymm) for Scenario S2.

Matrix Dimensions	Flat-SNC4	Flat-SNC2	Flat-All2All	Flat-Quadrant	Flat-Hemisphere
$16K \times 16K$	2.8%	3.9%	2.2%	**50.9%**	-0.5%
$32K \times 32K$	**28.3%**	-8.1%	27.2%	25.3%	26.5%
$64K \times 64K$	-62.2%	-37.1%	28.5%	29.0%	29.4%
Matrix Dimensions	Cache-SNC4	Cache-SNC2	Cache-All2All (not supported)	Cache-Quadrant	Cache-Hemisphere
$16K \times 16K$	4.3%	4.1%	—	49.3%	-0.9%
$32K \times 32K$	4.6%	8.8%	—	**28.2%**	10.3%
$64K \times 64K$	-4.4%	1.7%	—	**30.1%**	11.6%

Table 15: Evaluation of T_2 formula against different KNL configurations - All time values are in seconds

	T_1	T_2	$T^1_{p,n}$	$T^1_{bp,n}$	$T^1_{p,cn}$	Computed T_2/T_1	Actual T_2/T_1	Error	*rbk*
Flat - SNC4	23.21	113.51	0.98	0.65	46.04	0.68	4.89	86.09%	1.48
Flat - SNC2	15.42	14.82	0.84	0.39	51.67	0.87	0.96	9.37%	1.15
Flat - All2All	12.74	10.12	0.64	0.3	52.88	0.80	0.79	1.27%	1.24
Flat - Quadrant	11.92	8.87	0.57	0.33	35.62	0.77	0.74	4.05%	1.31
Flat - Hemisphere	10.25	8.85	0.59	0.3	38.02	0.92	0.86	6.98%	1.09
Cache - SNC4	10.89	11.71	0.88	0.48	38.03	1.29	1.07	19.44%	0.77
Cache - SNC2	11.32	11.54	0.81	0.42	42.56	1.14	1.02	11.76%	0.87
Cache - Quadrant	12.32	8.15	0.55	0.32	35.79	0.71	0.66	7.58%	1.40
Cache - Hemisphere	10.36	8.61	0.56	0.29	38.33	0.86	0.83	3.61%	1.16

Table 16: Evaluation of T_3 formula against different configurations - All time values are in seconds

	T_1	T^2_{bp,c_2n}	T_3	Computed T_3/T_1	Actual T_3/T_1	Error	*rbk*
Flat - SNC4	119.81	113.51	151.12	7.58	1.26	501.59%	1.48
Flat - SNC2	121.47	14.82	119.23	0.98	0.98	0.00%	1.15
Flat - All2All	102.53	10.12	80.62	0.79	0.79	0.00%	1.24
Flat - Quadrant	101.58	8.87	70.5	0.70	0.69	1.45%	1.31
Flat - Hemisphere	102.63	8.85	70.28	0.69	0.68	1.47%	1.09
Cache - SNC4	88.9	11.71	93.53	1.05	1.05	0.00%	0.77
Cache - SNC2	87.39	11.54	94.26	1.06	1.08	1.85%	0.87
Cache - Quadrant	97.52	8.15	64.98	0.67	0.67	0.00%	1.40
Cache - Hemisphere	80.18	8.61	68.35	0.86	0.85	1.18%	1.16

Table 17: Evaluation of T_4 formula against different KNL configurations - All time values are in seconds

	T_1	T^3_{bp,c_2n}	T_4	Computed T_4/T_1	Actual T_4/T_1	Error	*rbk*
Flat - SNC4	1002.3	151.12	1292.55	1.21	1.29	6.20%	1.48
Flat - SNC2	843.25	119.23	945.01	1.13	1.12	0.89%	1.15
Flat - All2All	843.04	80.62	640.35	0.77	0.76	1.32%	1.24
Flat - Quadrant	826.23	70.5	560.46	0.68	0.68	0.00%	1.31
Flat - Hemisphere	824.15	70.28	557.92	0.68	0.68	0.00%	1.09
Cache - SNC4	618.26	93.53	744.34	1.21	1.20	0.83%	0.77
Cache - SNC2	781.83	94.26	745.16	0.96	0.95	1.05%	0.87
Cache - Quadrant	871.1	64.98	517.99	0.60	0.59	1.69%	1.40
Cache - Hemisphere	652.6	68.35	544.31	0.84	0.83	1.20%	1.16

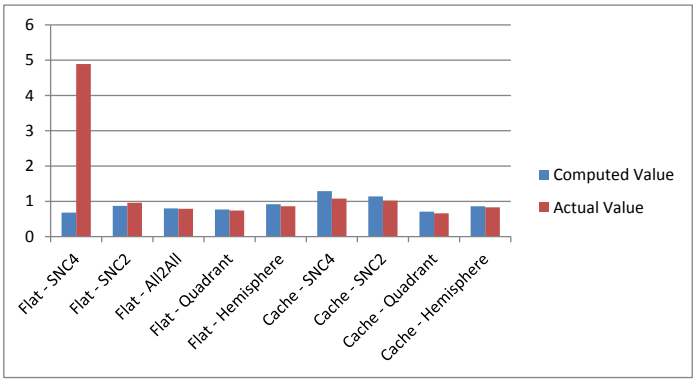


Figure 9: T_2/T_1: Computed Values vs. Actual Values. Demonstrated over KNL memory-clustering configurations.

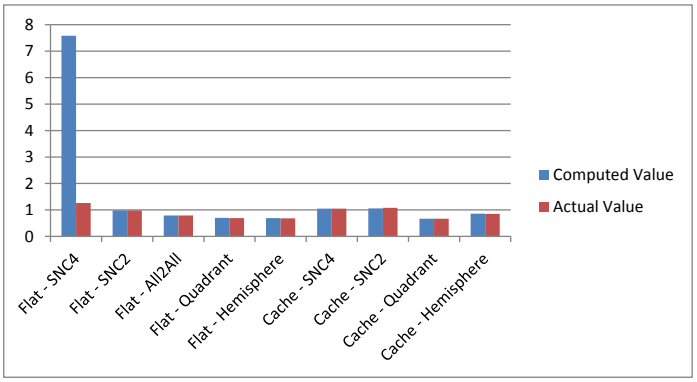


Figure 10: T_3/T_1: Computed Values vs. Actual Values. Demonstrated over KNL memory-clustering configurations.

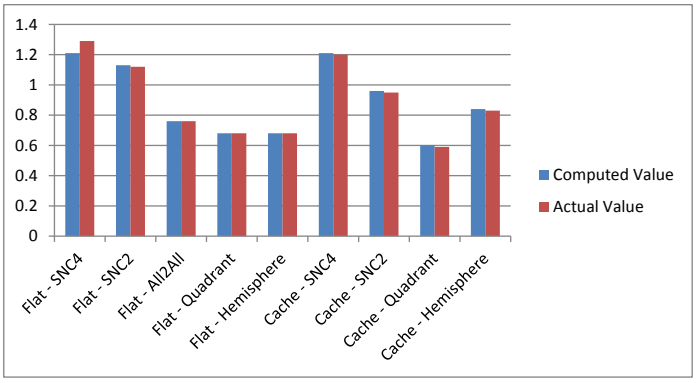


Figure 11: T_4/T_1: Computed Values vs. Actual Values. Demonstrated over KNL memory-clustering configurations.

If we leave out the Flat-snc4 configuration, then the speed up predictions of our model (either negative or positive) are within 2%, 2% and 20% of the actual values for matrix sizes $64K \times 64K$, $32K \times 32K$ and $16K \times 16K$, respectively.

6 Conclusions

We have proposed a multilevel method to speed parallel codes whose runtime grows faster than their workload. The target problems are the ones that can be decomposed into smaller subproblems. The original highly optimized parallel code is treated as a black box and no change is made to it. Using a simple parallel computing model, we derived formulas to predict whether or not any speed up is possible and also the amount of attainable speed up.

We demonstrated the effectiveness of our multilevel method on the highly optimized parallel BLAS routines cblas_dgemm, cblas_dtrmm and cblas_dsymm that are in the MKL library using the Intel KNL platform. Our proposed method is, however, general and can be potentially applied to other optimized parallel codes and on other platforms.

In our experiments, we considered 3 possible application scenarios. In the first of these (each algorithm is run using the best memory/clustering configuration for each problem size), we obtained speed up values of 16.2%, 8.4%, and 10.5%, respectively, for matrix multiplication, triangular matrix multiplication, and symmetric matrix multiplication for $64K \times 64K$ matrices. In the second scenario (all algorithms use the same configuration), these percentages are up to 40.5%, 31.7%, and 30.1%. The default mode in the KNL is cache - quadrant configuration. For this mode, the speed up percentages are 40.5%, 29.1%, and 30.1%. In the third scenario (each algorithm uses the same configuration regardless of problem size; different algorithms may use different configurations), the speed up values were 20.6%, 8.4%, and 11.6%.

For the matrix multiplication problem, the simple predictive mathematical model developed by us is able to correctly predict whether or not speed up is possible for 8 out of the 9 memory/clustering configurations tested. Also, for 8 out of the 9 configurations and for the matrix sizes of $64K \times 64K$, the amount of runtime increase or decrease our formula predicted for matrix multiplication was within 2% of the actual value. For both predicting the potential of speed up and predicting the amount of attainable speed up, Flat-snc4 was the configuration for which our model failed.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work was funded, in part, by the National Science Foundation, under Contract No. 1748652.

References

[1] S. Gheibi, T. Banerjee, S. Ranka, S. Sahni, "Multilevel Approaches to Fine Tune Performance of Linear Algebra Libraries," in 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 1–6, IEEE, 2019.

[2] L. E. Cannon, A cellular computer to implement the Kalman filter algorithm, Ph.D. thesis, Montana State University-Bozeman, College of Engineering, 1969.

[3] V. Strassen, "Gaussian elimination is not optimal," Numerische mathematik, **13**(4), 354–356, 1969.

[4] H. Prokop, Cache-oblivious algorithms, Ph.D. thesis, Massachusetts Institute of Technology, 1999.

[5] R. D. Blumofe, M. Frigo, C. F. Joerg, C. E. Leiserson, K. H. Randall, "An analysis of dag-consistent distributed shared-memory algorithms," in SPAA, volume 96, 297–308, 1996.

[6] J. Demmel, D. Eliahu, A. Fox, S. Kamil, B. Lipshitz, O. Schwartz, O. Spillinger, "Communication-optimal parallel recursive rectangular matrix multiplication," in 2013 IEEE 27th International Symposium on Parallel and Distributed Processing, 261–272, IEEE, 2013.

[7] B. Lipshitz, G. Ballard, J. Demmel, O. Schwartz, "Communication-avoiding parallel strassen: Implementation and performance," in SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, 1–11, IEEE, 2012.

[8] G. H. Golub, C. F. Van Loan, Matrix computations, volume 3, JHU press, 2012.

[9] R. A. Van De Geijn, J. Watts, "SUMMA: Scalable universal matrix multiplication algorithm," Concurrency: Practice and Experience, **9**(4), 255–274, 1997.

[10] M. M. A. Patwary, N. R. Satish, N. Sundaram, J. Park, M. J. Anderson, S. G. Vadlamudi, D. Das, S. G. Pudov, V. O. Pirogov, P. Dubey, "Parallel efficient sparse matrix-matrix multiplication on multicore platforms," in International Conference on High Performance Computing, 48–57, Springer, 2015.

[11] Q. Xiangzhen, "Cache performance and algorithm optimization," in High Performance Computing on the Information Superhighway, 1997. HPC Asia'97, 12–17, IEEE, 1997.

[12] D. I. Lyakh, "An efficient tensor transpose algorithm for multicore CPU, Intel Xeon Phi, and NVidia Tesla GPU," Computer Physics Communications, **189**, 84–91, 2015.

[13] J. Chen, J. Fang, W. Liu, T. Tang, C. Yang, "clmf: A fine-grained and portable alternating least squares algorithm for parallel matrix factorization," Future Generation Computer Systems, **108**, 1192–1205, 2020.

[14] C. Yount, A. Duran, "Effective use of large high-bandwidth memory caches in HPC stencil computation via temporal wave-front tiling," in Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS), International Workshop on, 65–75, IEEE, 2016.

[15] S. Chatterjee, V. V. Jain, A. R. Lebeck, S. Mundhra, M. Thottethodi, "Nonlinear array layouts for hierarchical memory systems," in Proceedings of the 13th international conference on Supercomputing, 444–453, ACM, 1999.

[16] J. Mellor-Crummey, D. Whalley, K. Kennedy, "Improving memory hierarchy performance for irregular applications using data and computation reorderings," International Journal of Parallel Programming, **29**(3), 217–247, 2001.

[17] E. Athanasaki, N. Koziris, "Fast indexing for blocked array layouts to improve multi-level cache locality," in Interaction between Compilers and Computer Architectures, 2004. INTERACT-8 2004. Eighth Workshop on, 107–119, IEEE, 2004.

[18] C. Kulkarni, C. Ghez, M. Miranda, F. Catthoor, H. De Man, "Cache conscious data layout organization for embedded multimedia applications," in Proceedings of the conference on Design, automation and test in Europe, 686–693, IEEE Press, 2001.

[19] B. Recht, C. Re, S. Wright, F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in Advances in neural information processing systems, 693–701, 2011.

[20] W.-S. Chin, Y. Zhuang, Y.-C. Juan, C.-J. Lin, "A fast parallel stochastic gradient method for matrix factorization in shared memory systems," ACM Transactions on Intelligent Systems and Technology (TIST), **6**(1), 2, 2015.

[21] S. Song, J. K. Hollingsworth, "Designing and auto-tuning parallel 3-D FFT for computation-communication overlap," in Proceedings of the 19th ACM SIGPLAN symposium on Principles and practice of parallel programming, 181–192, 2014.

[22] S. Lee, D. Jha, A. Agrawal, A. Choudhary, W.-k. Liao, "Parallel deep convolutional neural network training by exploiting the overlapping of computation and communication," in 2017 IEEE 24th International Conference on High Performance Computing (HiPC), 183–192, IEEE, 2017.

[23] H. Wang, S. Guo, R. Li, "Osp: Overlapping computation and communication in parameter server for fast machine learning," in Proceedings of the 48th International Conference on Parallel Processing, 1–10, 2019.

[24] J. Huang, T. M. Smith, G. M. Henry, R. A. van de Geijn, "Strassen's algorithm reloaded," in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 59, IEEE Press, 2016.

[25] R.-I. Ciobanu, C. Dobre, M. Bălănescu, G. Suciu, "Data and task offloading in collaborative mobile fog-based networks," IEEE Access, **7**, 104405–104422, 2019.

[26] V. Priya, C. S. Kumar, R. Kannan, "Resource scheduling algorithm with load balancing for cloud service provisioning," Applied Soft Computing, **76**, 416–424, 2019.

[27] D. Puthal, R. Ranjan, A. Nanda, P. Nanda, P. P. Jayaraman, A. Y. Zomaya, "Secure authentication and load balancing of distributed edge datacenters," Journal of Parallel and Distributed Computing, **124**, 60–69, 2019.

[28] A. Mallick, M. Chaudhari, U. Sheth, G. Palanikumar, G. Joshi, "Rateless codes for near-perfect load balancing in distributed matrix-vector multiplication," Proceedings of the ACM on Measurement and Analysis of Computing Systems, **3**(3), 1–40, 2019.

[29] A. Mohammed, A. Eleliemy, F. M. Ciorba, F. Kasielke, I. Banicescu, "An approach for realistically simulating the performance of scientific applications on high performance computing systems," Future Generation Computer Systems, **111**, 617–633, 2020.

[30] I.-J. Sung, J. A. Stratton, W.-M. W. Hwu, "Data layout transformation exploiting memory-level parallelism in structured grid many-core applications," in Proceedings of the 19th international conference on Parallel architectures and compilation techniques, 513–522, ACM, 2010.

[31] G. Chen, B. Wu, D. Li, X. Shen, "PORPLE: An extensible optimizer for portable data placement on GPU," in Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture, 88–100, IEEE Computer Society, 2014.

[32] A. Sodani, R. Gramunt, J. Corbal, H.-S. Kim, K. Vinod, S. Chinthamani, S. Hutsell, R. Agarwal, Y.-C. Liu, "Knights landing: Second-generation intel xeon phi product," Ieee micro, **36**(2), 34–46, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 826-833 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Deaf Chat: A Speech-to-Text Communication Aid for Hearing Deficiency

Mandlenkosi Shezi*, Abejide Ade-Ibijola

Formal Structures, Algorithms and Industrial Applications Research Cluster, Department of Applied Information Systems, University of Johannesburg, Bunting Road Campus, Johannesburg, 2006, South Africa

ARTICLE INFO

Article history:
Received:31 July, 2020
Accepted:15 September, 2020
Online:12 October, 2020

Keywords:
Speech-to-Text
Hearing Aids
Communication Aids
Multiple-Speaker Classification
AI for Social Good

ABSTRACT

Hearing impairments have a negative impact in the lives of individuals living with them and those around such individuals. Different applications and technological tools have been developed to help reduce this negative impact. Most mobile applications that have been developed that use Speech-to-Text technology have been inconsistent such that they are not inclusive of all types of hearing impaired individuals, only work under specifically predefined environments and do not support conversations with multiple participants. This makes the present tools less effective and makes hearing impaired participants feel like they are not completely part of the conversation. This paper presents a model that aims to address this by introducing the use of Multiple-Speaker Classification technology in the design of mobile applications for hearing impaired people. Furthermore we present a prototype of a mobile application called `Deaf Chat` *that uses the newly designed model. A survey was conducted in order to evaluate the potential that this application has to address the needs of hearing-impaired people. The results of the evaluation presented a good user acceptance and proved that a platform like* `Deaf Chat` *could be useful for the greater good of those who have hearing impairment.*

1 Introduction

Hearing is a very important ability that many people take for granted until it weakens, or they lose it [1]. Communication impairments have been studied and proven numerous times to have a negative impact on many aspects of an individual's life such that most communication impaired individuals experience mental and emotional challenges [2]–[5].

Hearing disorders have led into negative effects on the social and professional lives of older adults [6, 7]. Hearing impairment has a mass effect on the performance of individuals irrespective of their age group. Children who have any form of hearing impairment have been found to be experiencing continuous behavioral problems and issues with language development [8]. While examining hearing-impaired children, it was found that those who have hearing aids perform far way better compared to those who do not [9]. Using hearing aids reduces listening effort and the possibility of going through mental fatigue as less effort is made for processing speech [10]. This shows the greater potential that technology-based hearing aids could play on the lives of hearing-impaired individuals.

On many instances hearing-impaired individuals must depend on interpreters (mostly family members) to help them communicate with other people [11, 12]. Some also resort to using sign language for communication and others even go to the extent of lip reading [13, 14]. Despite the efforts made, the failure to accommodate the hearing-impaired and the use of insufficient and ineffective communication methods further contributes to depression, loneliness and social anxiety [15]. Alternatively, some people use SMS messaging to communicate with hearing impaired individuals [16], but that is only effective when both parties are at the same geographical location. It is highly imperative to note that SMS is not that much effective when individuals who are engaging in a conversation are not at the same place because one can never know if the other party viewed or received the message or not [17].

The understanding of the problems that hearing-impaired individuals of different ages face gave us motivation to create a platform that can help reduce these problems. With advances in Artificial Intelligence (AI), there has been a movement to create more AI tools for social good. Here, we introduce a model and a tool (named `Deaf Chat` to communicate with hearing impaired individuals based on AI. AI for Social Good can be indelicately described as AI that aims to improve people's lives in more meaningful ways [18]. Hence, creating an AI-powered software tool to support people with hearing impairment is considered as AI for Social Good [19].

*Corresponding Author: Mandlenkosi Shezi, University of Johannesburg, Bunting Road Campus, Johannesburg, South Africa, +27 74 875 1559, mandlashezifbi@gmail.com

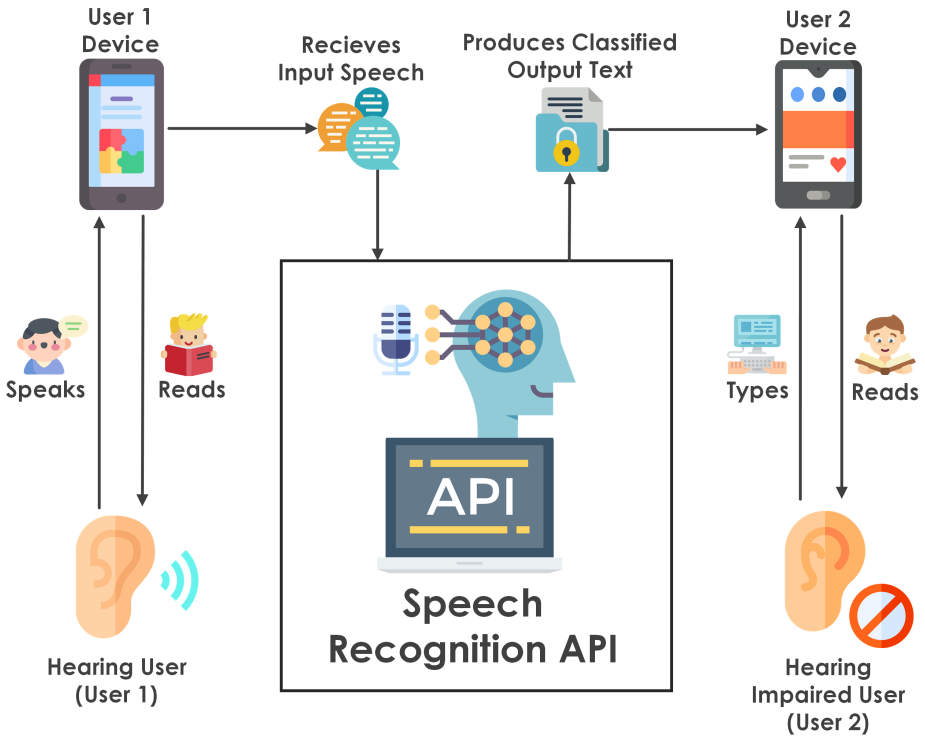


Figure 1: Graphical representation of messaging using the Deaf Chat application

Figure 1 gives a high level overview of how `Deaf Chat` facilitates communication between both hearing and hearing-impaired users. `Deaf Chat` supports both one-to-one chat and group chats, in both modes speech is analysed and classified in the interest of confirming if more than one voice is embedded in a voice note. The process is as follows. A hearing user, pair or group of hearing users will record their speech and the speaker recognition API will process the speech then convert it into classified text. The output text will then be sent to the hearing-impaired user in place of the voice note. Depending on the classification of the hearing-impairment and the preference of the hearing-impaired user, he or she will respond using either speech or text.

The following are the contributions of this paper. We have:

1. conducted an extensive literature review on previous tools used as hearing aids and tools,
2. developed a prototype of a mobile based messaging application called `Deaf Chat` that uses Multiple Speaker Recognition technology to serve as a communication aid for people with hearing deficiency, and
3. evaluated `Deaf Chat` using a survey in order to analyse its potential impact.

This paper is organized as follows. The following section will explore the Background. Section 3 will then follow with Related Work. Section 4 presents the design of the `Deaf Chat` application. Section 5 presents Implementation and Results. Section 6 presents the evaluation. Lastly, Section 7 will conclude the paper by presenting the Conclusion, Recommendations and directions for Future Work.

2 Background

This section provides a detailed view of all the tools that are similar to `Deaf Chat`. The first subsection discusses our problem statement which shows why the development of this solution is necessary.

2.1 Problem Statement

Mobile applications that have been developed to aid hearing-impaired individuals to communicate with each other and with people without disabilities have been focused on creating solutions that are not inclusive of all categories of hearing-impaired individuals, most of them are scenario based and none of them currently support multiple speakers. In order to explore and solve the following problem, the following questions will be addressed:

1. How has technology been used in the design of hearing aids and tools previously?
2. How can Multiple-Speaker Recognition software be incorporated in the design of a new tool to enhance communication with hearing deficient individuals?
3. Can we evaluate this new tool, and show that it can be helpful for hearing impaired individuals?

The following subsection gives a brief background on previously developed tools and how they affect the design of this new tool.

2.2 Classifying Hearing Impaired Individuals

Hearing impairments are one of the most well-known types of impairments around the world. According to the World Health Organization (WHO) [20], there are about 6.1% people who are suffering from hearing loss around the world. It is very important to note that there are different types of hearing impairments, these are distinguished according to the degree of hearing loss one is experiencing [21]. Firstly, some people experience partial hearing loss, an individual living with such a condition is labelled as one who is "hard of hearing" [22]. Secondly, other hearing-impaired individuals cannot hear a thing but can speak, an individual of such a group is called "deaf" [23]. Thirdly, some hearing-impaired individuals can not hear nor speak, and they are called "deaf mute" or "deaf and mute" [24]. Most people use the term "deaf" to describe all three groups but they are not the same.

It is important to understand this as it is a contributing factor that should be considered in the design of mobile applications for the hearing-impaired. The following subsection discusses the two types of hearing aids which are normally the primary aid that hearing-impaired individuals use.

2.3 Types of Hearing Aids

The most common hearing aids that exist are of two types, namely analog and digital hearing aids. Analog hearing aids are less expensive but what makes them unique is that when they amplify sounds, they do not filter noise. Digital hearing aids on the other hand are more advantageous because they can remove unwanted signals of which eliminates noise [25]. Digital hearing aids have been developed spontaneously such that newer versions of them have also gave birth to "smart hearing aid systems" [1]. Smart hearing aid systems consist of digital hearing aids and software for mobile applications that help with providing more flexibility and performance. Smart hearing aid systems enable users to set their environments using the mobile applications of which enables their digital hearing aids to filter noise more suitably [26, 27].

Technology is continuously evolving as time moves. It is important to note that, hearing aids are better suited for people who are living with or rather are experiencing partial hearing loss. This stresses the need for assistive technologies as they are more inclusive. The following section discusses these technologies.

2.4 Assistive Technologies

Assistive technologies can be used to aid people with different special needs and they have created new possibilities for people with living with limiting conditions [27]. These possibilities include help with day-to-day activities, rehabilitative services and others. As a result, these technologies make it possible or rather easier for people with disabilities to complete tasks that were previously difficult, or close to impossible. In the context of hearing-impaired individuals, these are better suited because they cater for all three groups of hearing-impaired individuals.

Some assistive technologies have been developed for educational purposes. GameOhm [28] is a serious game that was developed for electronic engineering students which is based on the Android platform. A serious game is a game which is developed with the purpose of contributing to teaching and learning. Serious games have been applied in numerous fields such as STEM (Science, Technology, Engineering and Medicine), education, business and more. We can easily categorize students as young people as the dominating age group of students is them. On the other hand, it is important to note that assistive technologies are not solely designed for young people.

Elderly people must also be included [26]. Bearing in mind that it is highly possible for elderly people to have more than one disability, the design of assistive mobile application technologies then becomes a major concern. Typically, most elderly people do not easily welcome technology. Making applications that are not user-friendly for them would further contribute negatively. Upon developing mobile applications for the hearing-impaired, developers must go for a user-centered approach and designs must be familiar to that of familiar applications [26, 29]. This enables the user training process to be less complex, which also increases motivation and user acceptance [30].

Upon collecting requirements for assistive technologies, it is important to consider both hardware and software capabilities in accordance to the needs of the targeted disabled users [31]. Different types of software could be used in accordance to the different classifications of hearing-impaired users such as Haptic Feedback software and Speech and Language Processing software. Haptic software technology uses touch to communicate with operators. The most common Haptic software implementations are that of vibration responses made by devices upon different user inputs. The following subsection discusses Speech and Language Processing.

2.5 Overview of Speech and Language Processing

Speech and Language Processing is basically concerned with the conversion of Speech-to-Text and Text-to-Speech. Text-to-Speech software synthesizes written natural language into an audio output, in simple terms it uses Natural Language Processing (NLP) technology to read written text. On the other hand, Speech-to-Text software does the same process but the other way around, it synthesizes spoken language and processes it into written natural language using NLP technology. Such software is often incorporated in the designs of voice assistants, automatic speech recognition (ASR) engines and speech analytic tools [31].

Applications that use Speech and Language Processing technology must be able to convert and transfer messages timely. Time, message transfer and real-time presentation of written text are some of the major challenges of converting speech-to-text in real time [32]. Instant speech-to-text conversion aims to synthesize spoken language into written text nearly concurrently. This is very important as it would enable people living with hearing-impairments to take part in conversations as they would have normally done if they were not experiencing any hearing problems. This would enable their conversations to be more engaging as they do not feel like they slow down participation because of taking too much time to process statements.

It is highly imperative to note that there are some challenges that might arise when converting speech to text in real-time [33]. Challenges are often caused by the nature of how Automatic Speech Recognition (ASR) based tools are built. Automatic speech recogni-

tion software typically needs one to train their speech recognition engine [34, 35]. It is advisable to have a minimum signal-to-noise ratio as noisy environments are inadequately fit. The amount of time spent training the speech recognition engine is directly proportional to the speech recognition engine's performance. However, this condition does not apply in the context of this research. Automatic speech recognition can be used for various purposes ranging from speaker identification [36], speaker classification [37] to speaker verification. This research is mainly concerned with speaker classification.

Speaker classification can separate audio input according to the voices of different speakers. It is not necessary to train a speech recognition engine for speaker classification because speaker classification is aimed at separating speech input from different voices not identifying and verifying predefined speakers. All speakers' voices are treated the same and are not compared against anything as they are viewed as new every time a voice input is synthesized during the use of a newly created speech client. *No literature currently exists on the application of speaker classification techniques for the design of mobile applications for hearing-impaired individuals* but the following subsection discusses some similar work.

3 Related Work

In this section we present an overview of current tools that are similar to `Deaf Chat` and how our tool differs.

3.1 Overview of Current Tools

The most common tools that exist in the market are mobile applications that use Automatic Speech Recognition software to help convert speech to text [17]. The requirements for these tools must be further studied in order to improve usability and usefulness [38]. Assistive technologies for the hearing-impaired must be designed to fit for all and more commercialised as many are only designed for research purposes [39, 27, 40]. Many solutions that are present are flawed as most are one-sided, they either work for a certain group of hearing deficient individuals (deaf, hard of hearing, or mute) or they work only under specifically predefined situations. This is evident in tools like Sign Support [41], which works as an aid in pharmacies only and LifeKey [42], which only works in cases of emergency.

3.2 Examples of Existing Tools

VoIPText is a mobile application that was developed to assist deaf and hard of hearing people. It attempted to apply VoIP (Voice Over IP), speech to text software and instant Automatic Speech Recognition (ASR) functionalities to include hearing deficient individuals in the use of VoIP technology for communication [17]. This tool was based online (dependent on internet connectivity) and its development realized the potential VoIP technology can assist hearing impaired people.

BridgeApp is a mobile application that was designed to assist deaf and mute Filipino Sign Language and American Sign Language speaking people from the Philippines, their user acceptance was positive, and the tool was deemed as useful [31]. This tool makes use of Text-to-Speech, Speech-to-Text, Text-to-Image and Haptic Feedback technology and it only works offline. The tool was only applicable to only predefined scenarios such as being at church, the office and others. Another tool called Ubider was developed for Cypriotic deaf people and it had the ability to convert a natural language into Cypriotic Sign Language [21]. This tool is dependent on the internet. The following subsection shows the gap that exists in current tools.

3.3 The Gap

Many tools have been developed to assist hearing-impaired people and the most common characteristics of these tools are that they are either:

1. online or offline,
2. dedicated only to a specific audience (deaf, mute, hard of hearing or any combination of two of the three categories),
3. they make use of speech to text, text to speech, and automatic speech recognition software.

A model that is inclusive of all three categories of hearing-impaired individuals (deaf, mute or hard of hearing) would make a greater contribution to their lives. This is what we present in this paper — `Deaf Chat`.

4 Design of Deaf Chat

The biggest distinguishing factors that represent how `Deaf Chat` is unique are the algorithm behind it and technologies that it uses. The following subsection gives a brief explanation of the algorithm that represent the app flow in `Deaf Chat`.

4.1 The Deaf Chat Algorithm

`Deaf Chat` has a unique design compared to other messaging applications that exist. It uses speaker diarisation techniques recognize and classify speakers before sending converted speech to text from one user to another. Speaker diarisation refers to the process of separating different voices on a voice input. This term is also spelled as "diarization" and this paper uses both spellings interchangeably. Algorithms 1-3 give a high-level overview of the algorithms behind how `Deaf Chat` was designed and implemented. Algorithm 1 shows how a message is sent from one user to another and how the dialogue process works on a Chat Activity.

```
Data: Text Message
Result: Delivery Notification
initialisation;
message = MakeInput();
SEND message;
if User_2 responds then
    MakeInput();
else
    do nothing;
end
```

Algorithm 1: Chat Activity

Algorithm 2 gives the possibility of two input types which are either text or voice input. This algorithm is a function that takes an argument and results in a text output that is used in Algorithm 1. In the case where a user decides to type text, the input will just be sent as is. In contrast to that, if the user decides to record a voice note the speech will be sent to the Process Speech function.

```
Function MakeInput():
begin
    Data: txt_input or voice_input
    Result: text_message
    if (input==txt_input) then
    begin
        text_message = txt_input;
    end
    else
    begin
        text_message = ProcessSpeech(voice_input)
    end
    end
    return text_message
end
```

Algorithm 2: Make Input Function

As called in Algorithm 2 (`MakeInput` Function), this algorithm shows how `Deaf Chat` interacts with the API which facilitates the classification of speakers. Algorithm 3 takes the voice input from the Make Input function and sends it to the API which then transcribes the speech and returns a string of words. This is continued until the user decides to either stop the recording using the stop button or by just simply sending the currently transcribed text.

```
Function ProcessSpeech():
begin
    Data: voice_input
    Result: text
    text = " ";
    while text = " " do
    begin
        SEND request to API;
        text = RECEIVED response from API;
        return text
    end
    end
end
```

Algorithm 3: Process Speech Function

The following subsection explains the graph structure of `Deaf Chat`.

4.2 Deaf Chat's Operation Flow

Figure 2 shows how users can interact with each other through all the functionalities that `Deaf Chat` has. Users can simply send a text message, media or make use of the diarisation functionality by recording a voice note. In order to save time in development, our prototype showcases the core functionality of our application of which is the separation of speakers from a voice input. The following section will give more thorough insights about the technologies that were used of which is a discussion about what they are and how they were implemented.

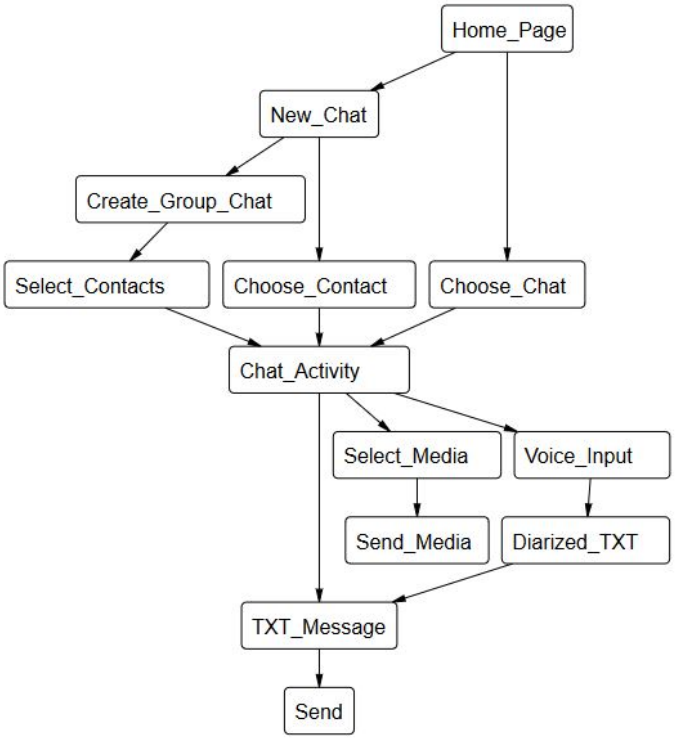


Figure 2: A Graph diagram, showing the operation flow of `Deaf Chat`

5 Implementation and Results

As indelicately mentioned on the previous section, our prototype showcases the core functionality of the `Deaf Chat` tool. By applying the core functionality we improved the quality of the results we were to obtain because incorporating an online chat service would have impacted the response time of the application. The following subsections discuss the two core technologies that we used in the implementation of `Deaf Chat`.

5.1 Android Studio

We used Android Studio IDE to develop our tool. We used the Java programming language to create all our back-end functions and XML (Extensible Markup Language) to create the front end in our activities. Figure 3 shows the chat activity whereby a voice input has been made and processed. The activity has an easy to use interface of which enhances the user experience. In order for a user or a group of users to make an input, one simply has to click the record button and start talking. There is no limit as to how long the user can speak. Furthermore, a user is allowed to edit the output text that has been produced before sending the generated text to the next user.

5.2 IBM Watson API

The IBM Watson API has a number of different features that can be used to build different applications ranging from Chat bots, Text-

to-Speech engines, Speech-to-Text engines and more. In the implementation of our tool, we used the IBM Watson Speech to Text API. It facilitated the conversion of speech from voice notes into classified text.



Figure 3: A Chat Activity showing processed voices of four distinct speakers.

The IBM Watson Speech to Text API manages to do this by applying Deep Neural Network (DNN) models. Within this context, Deep Neural Networks process speech by using powerful algorithms that can identify unique voices and further convert speech into text. The algorithms behind the API also filter noise by excluding unwanted sounds on the transcribed text. The IBM API only supports about 19 language models which include English (USA, Australia and UK), Mandarin, Dutch, French and more. We used the US language model. In order to identify different speakers we enabled diarization on our model and further created some methods that assisted us in formatting our output in a way that is easy to understand for different users.

6 Evaluation

This section presents the results we obtained from numerous individuals who tested the `Deaf Chat` prototype. We uploaded the application's installation file on the internet and then attached a link to an online survey which had 7 questions that respondents had to complete after testing the app. The results can be seen from Figure 4-10. Most of our respondents were between the ages 18-39, and they found the application useful and about 83.7% said that they would recommend it to a hearing impaired person. Only 14% of our respondents found it hard to navigate through the app of which shows that it is user friendly.

According to the results of the survey, `Deaf Chat` can convert Speech to Text and Identify Speakers successfully up to 86% and 88.5%, respectively. We accumulated those results by adding all participants who rated `Deaf Chat` 5 and above. Some respondents who experienced challenges with speech processing outlined that it might be because of their African accents. Some recommendations also included comments that said that `Deaf Chat` would be better if it could support local languages. A small percentage of our respondents outlined that the response rate is slow. Overall, we can conclude that `Deaf Chat` is very useful and performed even better than we had anticipated.

7 Conclusion and Future Work

7.1 Conclusion and Recommendations

In this paper, we have presented a new model and a software prototype for facilitating of communication with hearing-impaired individuals. Our tool can be used to communicate with all three types of hearing-impaired individuals and further makes a contribution into improving the lives of the targeted users. We evaluated our tool using a survey which included participants who are not hearing impaired. Most of our participants found the tool very useful and made recommendations as to how we can improve it so that it is more useful. Some participants had a bad experience because of their accents and we would like to recommend the use of a speech recognition engine that responds faster and supports more accents, and more especially those that are African.

There are many perceptions that the people of South Africa and Africa as a whole have about AI because it has not been used for much social good in South Africa. According to Stats SA [43], about 7.5% people in the population of South Africa are living with a disability. Furthermore, the South African National Deaf Association (SANDA) [44], certifies that it represents about four million deaf and hard of hearing people in South Africa which shows the greater need for more AI tools for Social Good. `Deaf Chat` is one of the first South African AI tools that were developed for Social Good and we perceive that it will open more opportunities for innovative research and development for this particular purpose.

7.2 Future Work

We are going to implement our model and design in an iOS and web based version of `Deaf Chat`. Moreover, we will add additional features such as quick access to emergency services for hearing impaired users.

Acknowledgment The authors would like to give special thanks to Nikita Patel for the rich illustration on this contribution and the whole Formal Structures Lab team for their support during the commencement of this study.

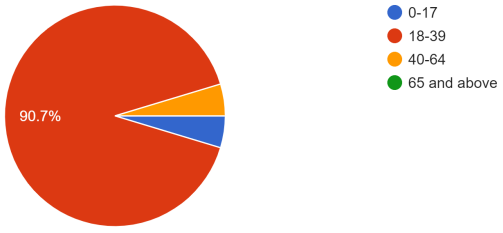


Figure 4: Age group of participants

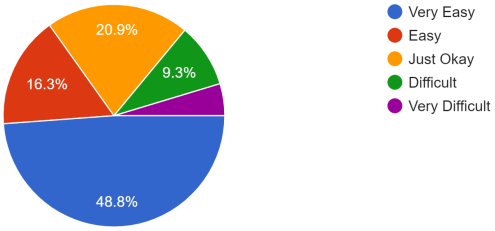


Figure 5: Ease of navigation

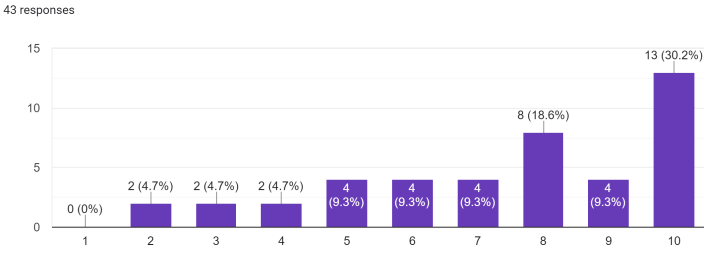


Figure 6: Rating of Speech-to-Text functionality

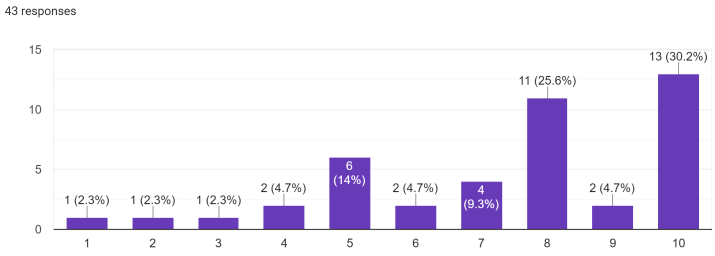


Figure 7: Rating of Speaker Diarization functionality

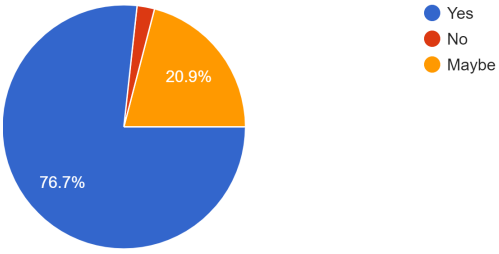


Figure 8: Usefulness of tool to hearing impaired people

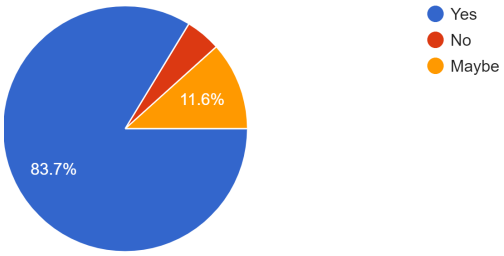


Figure 9: Likeliness to recommend app to hearing impaired person

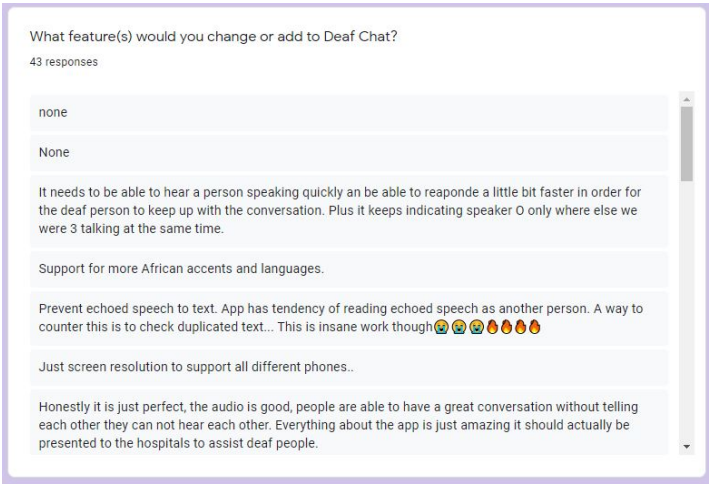


Figure 10: Recommendations

References

[1] W. Shehieb, M. O. Nasri, N. Mohammed, O. Debsi, K. Arshad, "Intelligent Hearing System using Assistive Technology for Hearing-Impaired Patients," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018, 2018, doi:10.1109/IEMCON.2018.8615021.

[2] D. G. Blazer, D. L. Tucci, "Hearing loss and psychiatric disorders: A review," Psychological Medicine, 2019, doi:10.1017/S0033291718003409.

[3] A. D. Palmer, P. C. Carder, D. L. White, G. Saunders, H. Woo, D. J. Graville, J. T. Newsom, "The impact of communication impairments on the social relationships of older adults: Pathways to psychological well-being," Journal of Speech, Language, and Hearing Research, 2019, doi: 10.1044/2018_JSLHR-S-17-0495.

[4] N. Shoham, G. Lewis, G. Favarato, C. Cooper, "Prevalence of anxiety disorders and symptoms in people with hearing impairment: a systematic review," Social Psychiatry and Psychiatric Epidemiology, 2019, doi: 10.1007/s00127-018-1638-3.

[5] M. I. Wallhagen, W. J. Strawbridge, G. A. Kaplan, "6-year impact of hearing impairment on psychosocial and physiologic functioning." 1996.

[6] A. D. Palmer, J. T. Newsom, K. S. Rook, "How does difficulty communicating affect the social relationships of older adults? An exploration using data from a national survey," Journal of Communication Disorders, 2016, doi:10.1016/j.jcomdis.2016.06.002.

[7] L. C. Gellerstedt, B. Danermark, "Hearing impairment, working life conditions, and gender," Scandinavian Journal of Disability Research, 2004, doi: 10.1080/15017410409512654.

[8] J. Stevenson, D. McCann, P. Watkin, S. Worsfold, C. Kennedy, "The relationship between language development and behaviour problems in children with hearing loss," Journal of Child Psychology and Psychiatry and Allied Disciplines, 2010, doi:10.1111/j.1469-7610.2009.02124.x.

[9] T. Most, "The effects of degree and type of hearing loss on children's performance in class," Deafness and Education International, 2004, doi:10.1179/146431504790560528.

[10] B. W. Hornsby, "The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands," Ear and Hearing, 2013, doi:10.1097/AUD.0b013e31828003d8.

[11] M. A. HARVEY, "Family Therapy with Deaf Persons: The Systemic Utilization of an Interpreter," Family Process, 1984, doi:10.1111/j.1545-5300.1984.00205.x.

[12] P. Kumar, S. Kaur, "Sign Language Generation System Based on Indian Sign Language Grammar," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), **19**(4), 1–26, 2020.

[13] C. N. Asoegwu, L. U. Ogban, C. C. Nwawolo, "A preliminary report of the audiological profile of hearing impaired pupils in inclusive schools in Lagos state, Nigeria," Disability, CBR and Inclusive Development, 2019, doi:10.5463/dcid.v30i2.821.

[14] R. Conrad, "Lip-reading by deaf and hearing children," British Journal of Educational Psychology, **47**(1), 60–65, 1977.

[15] J. F. Knutson, C. R. Lansing, "The relationship between communication problems and psychological difficulties in persons with profound acquired hearing loss," Journal of Speech and Hearing Disorders, 1990, doi:10.1044/jshd.5504.656.

[16] F. Bakken, "SMS Use Among Deaf Teens and Young Adults in Norway," in The Inside Text, 2005, doi:10.1007/1-4020-3060-6_9.

[17] B. Shirley, J. Thomas, P. Roche, "VoIPText: Voice chat for deaf and hard of hearing people," in IEEE International Conference on Consumer Electronics - Berlin, ICCE-Berlin, 2012, doi:10.1109/ICCE-Berlin.2012.6336486.

[18] Google, "AI for Social Good– Google AI," 2020.

[19] J. Cowls, T. King, M. Taddeo, L. Floridi, "Designing AI for Social Good: Seven Essential Factors," SSRN Electronic Journal, 2019, doi:10.2139/ssrn.3388669.

[20] WHO, "Estimates," 2018.

[21] K. Pieri, S. V. G. Cobb, "Mobile app communication aid for Cypriot deaf people," Journal of Enabling Technologies, 2019, doi:10.1108/JET-12-2018-0058.

[22] A. Stratiy, "The Real Meaning of "Hearing Impaired"," Deaf World: A Historical Reader and Primary Sourcebook, 203, 2001.

[23] T. K. Holcomb, "Deaf epistemology: The Deaf way of knowing," American Annals of the Deaf, 2010, doi:10.1353/aad.0.0116.

[24] A. Ballin, The deaf mute howls, volume 1, Gallaudet University Press, 1998.

[25] P. V. Praveen Sundar, D. Ranjith, T. Karthikeyan, V. Vinoth Kumar, B. Jeyakumar, "Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture," International Journal of Speech Technology, 2020, doi:10.1007/s10772-020-09686-y.

[26] A. Holzinger, G. Searle, A. N. Witzer, "On some aspects of improving mobile applications for the elderly," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2007, doi:10.1007/978-3-540-73279-2_103.

[27] G. Kbar, A. Bhatia, M. H. Abidi, I. Alsharawy, "Assistive technologies for hearing, and speaking impaired people: a survey," 2017, doi:10.3109/17483107.2015.1129456.

[28] A. Jaramillo-Alcazar, C. Guaita, J. L. Rosero, S. Lujan-Mora, "Towards an accessible mobile serious game for electronic engineering students with hearing impairments," in EDUNINE 2018 - 2nd IEEE World Engineering Education Conference: The Role of Professional Associations in Contemporaneous Engineer Careers, Proceedings, 2018, doi:10.1109/EDUNINE.2018.8450948.

[29] C. Siebra, T. B. Gouveia, J. Macedo, F. Q. Da Silva, A. L. Santos, W. Correia, M. Penha, F. Florentin, M. Anjos, "Toward accessibility with usability: Understanding the requirements of impaired uses in the mobile context," in Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication, IMCOM 2017, 2017, doi:10.1145/3022227.3022233.

[30] B. Adipat, D. Zhang, "Interface design for mobile applications," 2005.

[31] M. J. C. Samonte, R. A. Gazmin, J. D. S. Soriano, M. N. O. Valencia, "BridgeApp: An Assistive Mobile Communication Application for the Deaf and Mute," in ICTC 2019 - 10th International Conference on ICT Convergence: ICT Convergence Leading the Autonomous Future, 2019, doi:10.1109/ICTC46691.2019.8939866.

[32] S. Wagner, "Intralingual Speech-to-text conversion in real-time: Challenges and Opportunities," in Challenges of Multidimensional Translation Conference Proceedings, 2005.

[33] S. Rossignol, O. Pietquin, "Single-speaker/multi-speaker co-channel speech classification," in Eleventh Annual Conference of the International Speech Communication Association, 2010.

[34] J. P. Campbell, "Speaker recognition: A tutorial," Proceedings of the IEEE, **85**(9), 1437–1462, 1997.

[35] H. Beigi, "Speaker recognition: Advancements and challenges," New trends and developments in biometrics, 3–29, 2012.

[36] H. E. Secker-Walker, B. Liu, F. V. Weber, "Automatic speaker identification using speech recognition features," 2017, uS Patent 9,558,749.

[37] P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grezl, L. Burget, J. H. Cernocky, "Analysis of DNN approaches to speaker identification," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2016, doi:10.1109/ICASSP.2016.7472649.

[38] S. S. Nathan, A. Hussain, N. L. Hashim, "Studies on deaf mobile application: Need for functionalities and requirements," Journal of Telecommunication, Electronic and Computer Engineering, 2016.

[39] S. Hermawati, K. Pieri, "Assistive technologies for severe and profound hearing loss: Beyond hearing aids and implants," Assistive Technology, 2019, doi:10.1080/10400435.2018.1522524.

[40] M. C. Offiah, S. Rosenthal, M. Borschbach, "Assessing the utility of mobile applications with support for or as replacement of hearing aids," in Procedia Computer Science, 2014, doi:10.1016/j.procs.2014.07.079.

[41] M. B. Motlhabi, M. Glaser, M. Parker, W. D. Tucker, "SignSupport: A Limited Communication Domain Mobile Aid for a Deaf patient at the Pharmacy," in Southern African Telecommunication Networks & Applications Conference, 2013.

[42] L. Slyper, M. K. Kim, Y. Ko, I. Sobek, "LifeKey: Emergency communication tool for the deaf," in Conference on Human Factors in Computing Systems - Proceedings, 2016, doi:10.1145/2851581.2890629.

[43] S. SA, "Stats SA profiles persons with disabilities," 2020.

[44] SANDA, "Homepage," 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 838-844 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Effect of Heat Retention Time and Pouring Temperature on Graphite Shape and Mechanical Properties of Gray Cast Iron

Hong-Nga Thi Pham*

Faculty of Mechanical Engineering, HCMC University of Technology and Education, HCM City, 71307, Vietnam

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 11 November, 2020
Online: 27 November, 2020

Keywords:
Graphite
Gray cast iron
Microstructure
Holding time
Pouring temperature

ABSTRACT

In this study, the ASTM A48 Class 30 gray cast iron was cast in a medium frequency induction furnace. The ultimate tensile strength (UTS), ultimate flexural strength (UFS) and microstructure were determined in accordance with ASTM E8/E 8M-80 and ASTM A438-80 (1997). From the results of this study, the graphite size at overheating temperatures 1350°C was smaller than that at 1300°C. Moreover, if the holding time was sufficient, the nested sheet graphite would be completely destroyed. Particularly, in the case of holding time from 5 to 20 minutes, the UTS, UFS are increased and the graphite shape got fine. The smallest graphite size and the highest UTS, UFS values were achieved if 20 minutes holding time is carried out. However, the holding time increased to 30 minutes, the graphite size tended to grow larger, and the UTS, UFS values suffered a decline. In addition, decreasing the average pouring temperature from 1350°C, 1300°C to 1260°C caused the upward trend of the graphite average sizes which were fit with the downward trend in the UTS and UFS values. From the above results, it is possible to determine the casting conditions that can improve the quality of cast iron products in induction furnace are: the optimal average pouring temperature is 1300-1350°C, while the optimal holding time is 10-30 minutes.

1. Introduction

Gray cast iron is a widely-used material in the machinery industry. A typical product of gray cast iron is engine segments. Currently, Vietnamese engine segments are still not high-quality products due to the contamination of cupola furnace technology. In addition, the insoluble coke in the casting process will generate graphite in the form of nest sheet graphite, as shown in Figure 1. As a result, the graphite sheet in the engine segments is also usually in the form of a nest [1]. This is an undesirable form of graphite that negatively impact the mechanical properties of cast iron since the graphite nests break up the continuous of the cast-iron structure, as also the place that the stress concentrated. Therefore, the cast iron has a low flexural and tensile strength [2-4].

Graphite is extremely soft and is considered as a crack that separates the cast iron matrix. The size, shape, and distribution of graphite strongly affect the mechanical properties of cast iron graphite through the microstructure. In [5] the author presented that the main influences on flexural strengths are the shape and the graphite size. Author in [6] obtained that the impact of graphite shape on corrosion vulnerability of gray cast iron in an acidic environment is eclipsed; the vermicular shape of graphite, profitable from the perspective of mechanical and tribological properties might facilitate the destruction of cast iron under the corrosion erosion environment. A similar result was reported by research [7] that the tensile strength of the unidirectionally solidified samples is not as strong as that of spheroidal graphite cast iron because the fracture occurs at the graphite-terminated area in the aligned graphite samples. Study [8] investigates four flaky graphite cast irons of different graphite structures with a pearlitic matrix, results showed that the eutectic graphite structure wore down faster than the flaky graphite structure did, the type of graphite morphology influenced the specimen wear rate as strongly as the graphite volume fraction did in flaky graphite cast irons of this experimental range. The tensile strength of cast iron graphite is only 0.3-0.5 of the tensile strength of steel whose composition corresponds to the base of gray cast iron. Flexural strength and tensile strength of cast iron graphite is only 0.4-0.7 of flexural strength and tensile strength compared to the corresponding values of the steel that has the same composition as the iron matrix.

*Corresponding Author: Hong-Nga Thi Pham, HCMC University of Technology and Education, HCM City, Vietnam, +84-948691160, hongnga@hcmute.edu.vn

In general, the mechanical properties depend on the size, shape, and distribution of graphite which dependent components and pre-cast conditions treatment, additives, slag, furnace, temperature and holding time, the chemical composition of cast iron, cooling rate... [9,10]. Particularly, the graphite with nest sheet shape leads to the low value of the mechanical properties of cast iron [11,12]. However, the studies of the pouring temperature and holding time on the mechanical properties of gray iron are very limited. In [13,14], the authors described the effect of cooling rate on microstructure and mechanical properties of gray cast iron and the effect of pouring temperature on fluidity, porosity, and surface roughness of gray cast iron. According to [5], the pouring temperature plays a key role in constructing the tensile strength and elongation of the cast iron, besides the graphite factor, chemical composition and matrix microstructure also contribute to the tensile strength. Research [15] proves that the faster solidification in the microstructure, the lower the values of the mechanical properties. Results of [16] show that the mechanical properties and microstructure of compacted graphite iron depend on holding time, increasing the holding time from 10 min to 17 min results in lowering the Mg content from 0.031% to 0.021% and as a result lower nodularity was obtained, in addition, lowering the thickness from 20mm to 5mm increases the tendency of the dendritic structure as a result of increasing the cooling rate. In which, the studies on the influence of shape, size, and distribution of graphite on the mechanical properties of gray iron is still very limited. This report focuses on controlling the pouring temperature and holding time to destroy the nested sheet graphite and form a better graphite shape to improve the mechanical properties of cast iron.

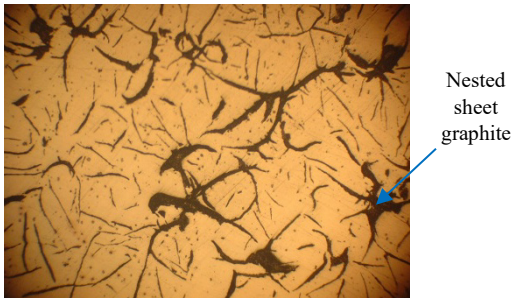


Figure 1: Microstructure of nest sheet graphite cast iron (50X magnification)

2. Experimental

2.1. Materials and Experimental Method

The material composition for the experiment is shown in Table 1. This composition is used to cast gray cast iron with grade 30 of ASTM International standard A48 for small electric motor covers whose capacity is less than 5kW. Cast iron experiments are made at Muoi Hoi Binh Duong Co., Ltd., as shown in Figure 2, without using the nodulizing elements. Initially, the mixed ingredients are fed into a medium frequency induction furnace, with a capacity of 1000kg/batch. The furnace runs for 10 minutes with low amperage to stabilize the frequency. The current then gradually increases to the maximum intensity for 1 hour 30 minutes. The crucible is covered by the acidic powder that allows desulfurization. Due to the rotating magnetic field, cast iron begins to melt and mix relatively well. The furnace is maintained at a constant current for half an hour to allow the cast iron to completely melt and the cast iron temperature is stable. Then, the cast iron temperature is increased to the pouring temperature as required. The pouring pot is pre-dried with a charcoal furnace to a temperature that is greater than 1000°C. The old sand mixture with good ventilation is used, mold is dried naturally. The mold is made by a pneumatic system to ensure uniform tightness of the mold, thus ensuring the heat transfer and exhaust mode.

Table 1: Material composition for experiment

No.	Composition
1	Cast iron scrap, recovery materials in casting production (filling system, shrinkage, broken casting)
2	Ferrosilicon 30-70, with 1-2%
3	Ferromanganese 50-50
4	$CaCO_3$



Figure 2: Medium frequency induction furnace at Muoi Hoi Binh Duong Co., Ltd

Table 2: Overheating temperatures, holding times, and average pouring temperatures

Sample	Average overheating temperatures (°C)	Holding times (minutes)	Average pouring temperatures (°C)
S1	1350	5	1350
S2		10	
S3		20	
S4		30	
S5	1300	5	1300
S6		10	
S7		20	
S8		30	
S9	1300	5	1260

Table 2 shows the cast iron samples are cast with different overheating temperatures, holding times, and average pouring temperatures. The layout of the experimental samples in the sand mold and the samples after casting are shown in Figure 3. The temperature is measured with the AS892 Infrared Thermometer. Some basic specifications are described as following: temperature range 200°C ~2200°C, accuracy ± 2°C, distance spot ratio 80:1, emissivity 0.10~1.00 adjustable, resolution 0.1°C, wavelength 900~1700 µm, response time 500ms, repeatability ±1°C.

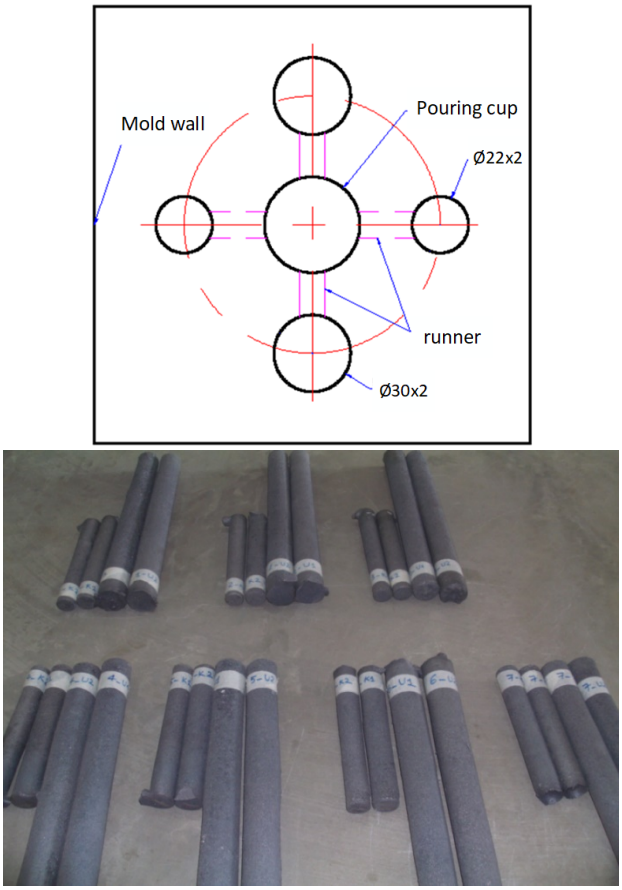


Figure 3: The layout of the experimental samples in the sand mold and the samples after casting

2.2. Methods of Observing Microstructure and Testing Mechanical Properties

Tensile test and Flexural test are conducted at Quatest 3 center, Vietnam, following the ASTM E 8/E 8M-80 (Figure 4) and ASTM A 438-80 (1997) (Figure 5) standards.

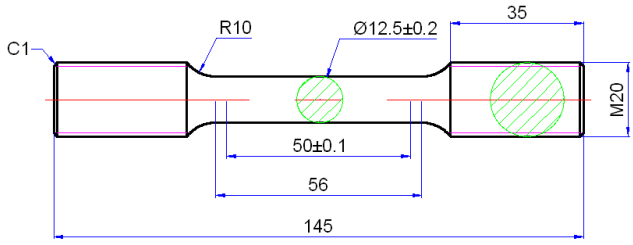


Figure 4: Tensile sample size standard ASTM E 8/E 8M-80

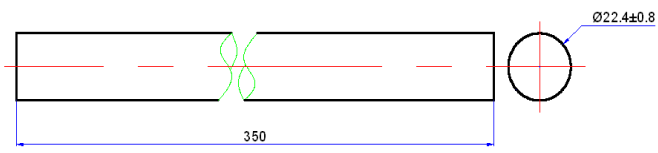


Figure 5: Flexural sample size standard ASTM A 438-80

The microstructure of the samples are observed at Materials Laboratory at Mechanical Faculty, University of Technology and Education, HCMC. The graphite size calculation is based on TCVN 3902-84.

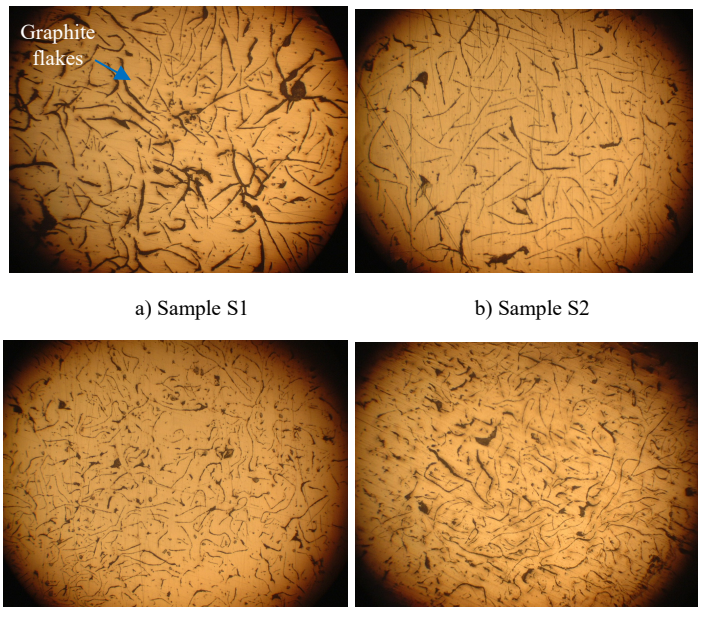


Figure 6: Universal testing machine – SANS

3. Results

3.1. Effect of Holding Time on the Graphite Shape

This section considers the effect of holding times and average pouring temperatures to identify the influence of these factors on the amount of graphite and the time needed to destroy the nested sheet graphite. Figure 7 illustrates the microstructure of the samples that overheated at 1350°C and holding at 5-30 minutes. These microstructures indicate that increasing the holding time leads to a decline in the graphite amount. Moreover, the graphite shape is also different with different holding times. Improving the holding time also reduces the number of nested sheet graphite. Interestingly, if the holding time is long enough, the nested sheet graphite will disappear, as shown in Figure 7(d).

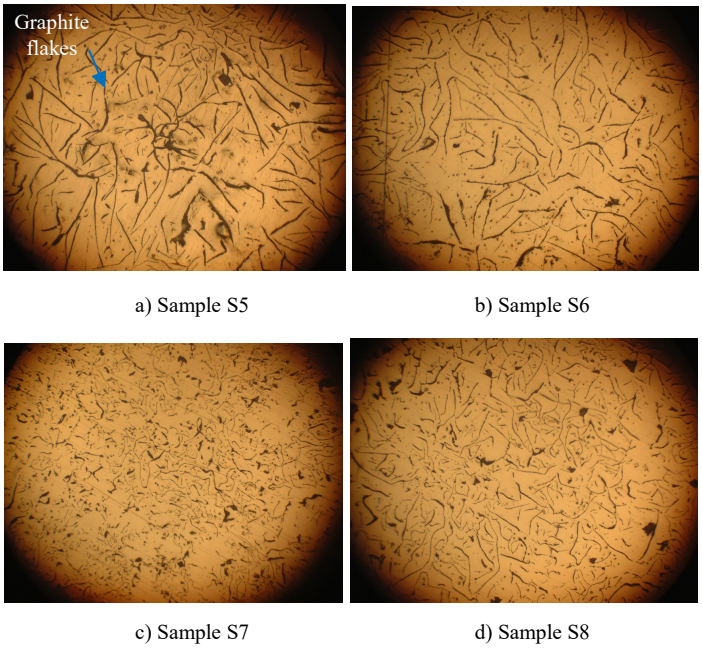


a) Sample S1 b) Sample S2

c) Sample S3 d) Sample S4

Figure 7: Microstructure of samples pouring at 1300oC at different holding time: (a) Sample S1 (holding time 5 minutes), (b) Sample S2 (holding time 10 minutes), (c) Sample S3 (holding time 20 minutes), and (d) Sample S4 (holding time 30 minutes) (50X magnification)

Figure 8 exhibits the microstructure of the samples that overheated at 1300°C and holding at 5-30 minutes. In general, improving the holding time results in a lower amount of graphite. In addition, the graphite shape is also different with different holding times. Similar to 1350°C cases, if the holding time is long enough, the nested sheet graphite will be destroyed, as shown in Figure 8(d).



a) Sample S5 b) Sample S6

c) Sample S7 d) Sample S8

Figure 8: Microstructure of samples pouring at 1350oC at different holding time: (a) S5 (holding time 5 minutes), (b) S6 (holding time 10 minutes), (c) S7(holding time 20 minutes), and (d) S8 (holding time 30 minutes) (50X magnification)

The average size of the graphite sheet at different overheating temperatures and different holding times is showed in Figure 9. Initially, increasing the holding time from 5 to 20 minutes results in finer graphite. However, if the holding time increases to 30 minutes, the graphite size will get bigger. The smallest graphite size is achieved by holding at 20 minutes. Moreover, the graphite size at 1350°C is smaller than at 1300°C.

The reason is that with a higher temperature and a longer holding time, the liquid cast iron is stirred stronger as the viscosity of the cast iron is lower. The stronger stirring results in finer graphite. The higher temperature also causes a more stable state, and a higher trend of overcooling of the casting liquid. However, holding too long could lead to an unstable graphitization process [3,4]. When overheated, the amount and the shape of graphite changes. As the overheating of cast iron increases, the graphite becomes smaller and changes its distribution characteristics. Firstly, the eutectic form of graphite changes, then the number of dendrites will increase. Overheating facilitates the formation of graphite between dendrites which also depends on the composition of the cast iron (carbon and silicon). When the overheating temperature increases, the graphite will strongly smooth out, proving that there is some solubility of these phases and increasing the dispersion of these phases. However, overheating does not facilitate the complete dissolution or agglomeration of these phases. Traditionally, the graphite in gray cast iron could be viewed as a void, and the cracks initiate at the tip of the graphite flakes due to the micro notch stress concentration. The deformed graphite flakes, an optimized microstructure, reduced the stress concentration and prevented the expansion of cracks at the tip of graphite [11]. In [9], the author explains that cast iron slug samples all had average flake lengths of 80-100 µm, and maximum flake lengths of 1000-1400 µm due to the drastically slower cooling rate in this mold configuration compared to the cooling rate in the commercial rotors. This suggests that either there is a limit to how much the diffusivity can be improved by increased flake length or that other factors limited the diffusivity of the gray iron from these large castings. According to [16], holding time displays a reasonable effect on the progression of residual magnesium that identifies with the fading process.

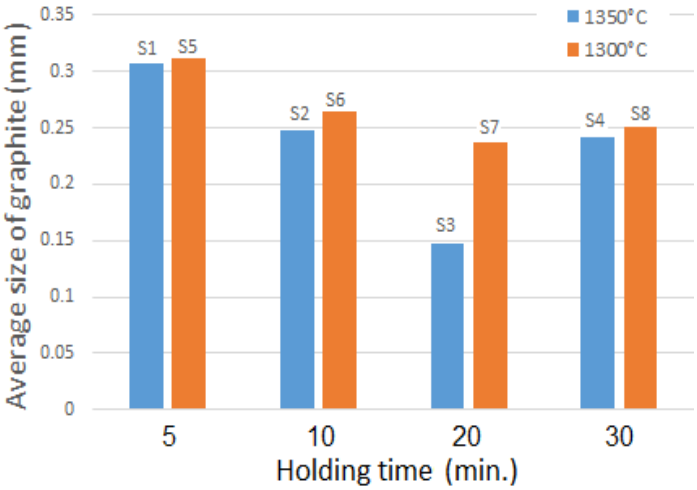


Figure 9: The average size of graphite sheet with different holding times and overheating temperatures

In conclusion, the optimal microstructure is obtained by holding at 10-30 minutes at 1350°C. Casting the gray iron in the medium frequency induction furnace can destroy the nested sheet graphite.

3.2. Effect of the Holding Time on the Mechanical Properties

Table 3 and Figure 10 present the ultimate tensile strength (UTS) of the samples at different holding times and overheating temperatures. These results indicate that the UTS rises as increasing the holding time from 5 minutes to 20 minutes and gaining the best UTS value at 20 minutes. Thereafter, the UTS suffers a decline as the holding time continues to increase from 20 minutes to 30 minutes.

Table 3: Tensile test results

Sample	Fracture force P_b (kN)	Tensile strength $\sigma_b=P_b/F_0$ (MPa)
S1	17.6	145
S2	18.1	150
S3	18.9	154
S4	18.0	148
S5	15.7	129
S6	16.0	132
S7	16.4	135
S8	15.8	130

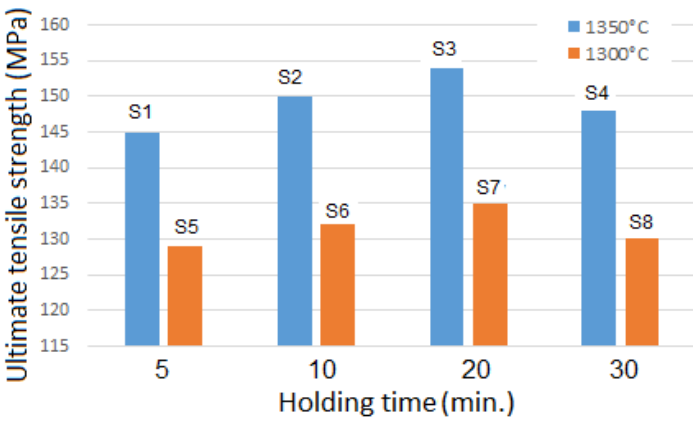


Figure 10: Ultimate tensile strength of the samples with different holding times and overheating temperatures

Table 4 and Figure 11 represent the ultimate flexural strength (UFS) of the samples at different holding times and overheating temperatures. Similar to the UTS value, the UFS increases as improving the holding time from 5 minutes to 20 minutes and reaching the highest value at 20 minutes. Then, this value decreases as the holding time increases from 20 minutes to 30 minutes. Interestingly, the results of both UTS and UFS values are consistent with the change of graphite size in section 3.1. In conclusion, the downward trend of the graphite size is fitted with an upward trend in the UTS and the UFS values. Therefore, decreasing the graphite size could lead to higher mechanical properties.

Table 4: Flexural test results

Sample	Resistance moment $W=\pi d^3/32$ (mm^3)	Ultimate flexural force P_u (kN)	Flexural moment $M_u=P_uL/4$ (kNmm)	Flexural strength $\sigma_{bu}=M_u/W$ (N/mm^2)
S1	1148.36	4.31	328.43	286
S2	1118.27	4.41	336.60	301
S3	1148.36	4.71	359.44	313
S4	1133.25	4.49	342.24	302
S5	1133.25	4.03	307.11	271
S6	1118.27	4.11	313.12	280
S7	1148.36	4.37	333.02	290
S8	1133.25	4.24	322.98	285

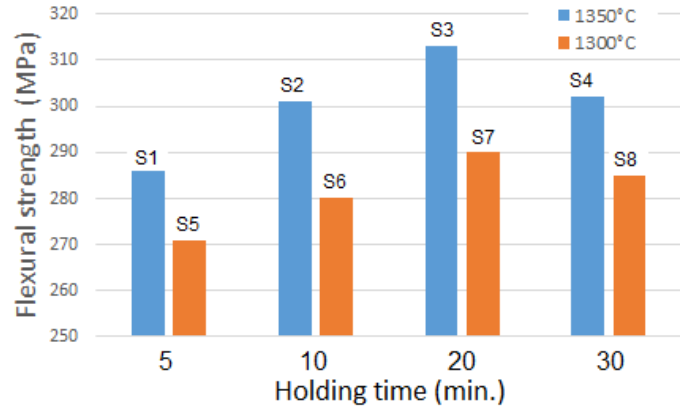


Figure 11: Flexural strength of the samples with different holding times and overheating temperatures

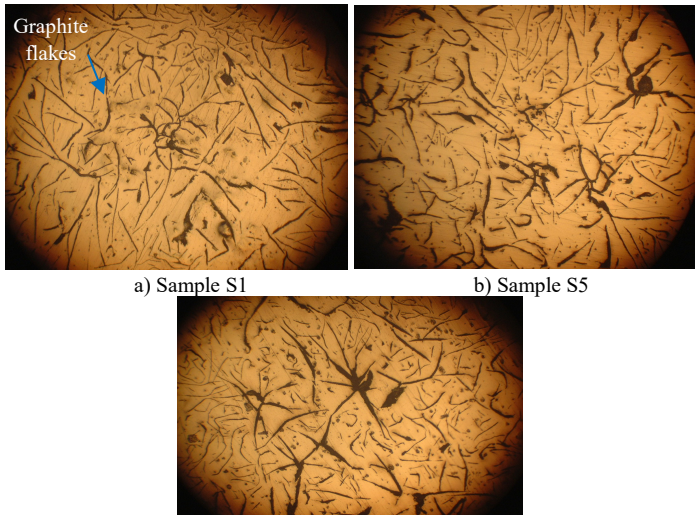
It can be seen that S4 and S8 samples with a holding time of 30 minutes have lower UTS and UFS values than the S3 and S7 samples with a holding time of 20 minutes. The reason is that, as the temperature and holding time increase, the viscosity will increase. The chemical composition and the amount of carbon are evenly distributed, therefore, the quality of the casting increases, the graphitization process happened more fully. However, if the temperature is too high and the holding time is too long, it will prevent the formation of liquid slag, increase the burn rate. In addition, the temperature difference between the liquid iron and the mold causes the graphitization process to occur unstably [4]. Therefore, choosing a holding time of 10-30 minutes is optimal.

The samples overheated at 1350°C and 1300°C gave different mechanical results. The reason is overheating changes the nature of the liquid phase and the differentiation conditions when solidified. On the one hand, overheating increases the stability of the liquid phase and increases its tendency to overcool. On the other hand, under the ideal solidification of a casting in a mold, overheating causes the preheating of the mold and changes the cold condition of the cast iron. Overheating requires a regular increase in the amount of bound carbon (reducing the amount of graphite) due to increased liquid phase stability and its tendency to over-cool. Overheating will limit impurities present in the liquid phase and make them more evenly distributed. In addition, the different overheating temperature will result in different morphology of graphite. The study [11] showed that the distribution and morphology of graphite has a direct influence on the properties of gray cast iron oriented the space orientation of the graphite flakes in gray cast iron and improved both strength and ductility.

3.3. Effect of the Average Pouring Temperature on the Graphite Size and the Mechanical Properties

This section surveys the effect of the average pouring temperature on the graphite size and the mechanical properties of the gray iron. The samples S1, S5, and S9 are chosen and analyzed.

Figure 12 and Table 5 show the microstructure of the cast iron and the average size of their graphite. These results indicate that decreasing the average pouring temperature will cause a rise in the graphite average size.



a) Sample S1 b) Sample S5

(c) Sample S9

Figure 12: Microstructure of samples pouring at different temperatures: (a) Sample S1, (b) Sample S5, and (c) Sample S9 (50X magnification)

Table 5: The effect of the average pouring temperature to the graphite size

Sample	Holding time (minutes)	Average pouring temperature (°C)	Average size of graphite (mm)
S1	5	1350	0.307
S5	5	1300	0.312
S9	5	1260	0.339

Table 6: UTS and UFS values at different average pouring temperatures

Sample	Average pouring temperature (°C)	UTS value (MPa)	UFS value (N/mm^2)
S1	1350	145	286
S5	1300	129	271
S9	1260	126	270

The UTS and UFS of the sample S1, sample S5, and sample S9 values at different average pouring temperatures is demonstrated in Table 6. The result points out that reducing the pouring temperature leads to a decline in the UTS and UFS values. The result is that the lower pouring temperature leads to the larger graphite size. Therefore, the larger graphite size will result in lower mechanical characteristics. Moreover, the higher temperature of the casting liquid will cause a lower viscosity and a better chemical uniformity. At a low pouring temperature of 1260°C (sample S9), the viscosity of the casting liquid is high, the chemical composition is not evenly distributed. To achieve a higher level of uniformity, the casting liquid can behold for a longer time. However, the longer time will cause a loss in the chemical composition as they are burned out or going to the slag. Generally, to improve the mechanical properties of the gray iron, increasing the temperature is better than increasing the holding time. Increasing to too high a temperature will result in a worse slag, a higher level of chemical burning, and the more unstable the graphitization [14].

In conclusion, the optimal pouring temperature to obtain the best mechanical properties is 1300-1350°C.

4. Conclusion

Increasing the holding time from 5 to 20 minutes at 1350°C and 1300°C results in finer graphite, a smaller amount of nested sheet graphite. The finest graphite size is achieved at 20 minutes. Holding up to 30 minutes could destroy the nested sheet graphite, in which at 1300°C the holding time will be longer than overheating and keep the heat at 1350°C to completely lose the nested sheet graphite. However, increasing the holding time to 30 minutes will cause an increase in the graphite size.

- At 1350°C and 1300°C, the UTS and UFS values increase as the holding time improve from 5 to 20 minutes, and reaching the highest values at 20 minutes. Thereafter, these mechanical properties suffer a decline as the holding time increase from 5 to 20 minutes. The finer graphite results in higher UTS and UFS values.

- Declining the average pouring temperature from 1350°C to 1300°C, and then 1260°C (with holding time in 5 minutes and same cooling condition) leads to a finer graphite size and the better UTS and UFS values.

From the above results, the casting conditions that can improve the product quality is the optimal average pouring temperature is 1300-1350°C, while the optimal holding time is 10-30 minutes which will completely lose the nest graphite.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We acknowledge HCMC University of Technology and Education, Dong Nhan Phat Co., Ltd and Material Testing Laboratory (HCMUTE). They gave me an opportunity to join their team, accessed the laboratory and research machines. Without their appreciated support, it would not be possible to conduct this research.

References

[1] T. H. N. Pham, N. T. Tran, P. J. Pritadewi, N. Y. P. Vo. "Research on Factors Influencing the Formation Graphite and Effect of Graphite on Mechanical Properties of Gray Cast Iron". 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi, Vietnam, 2019, doi:10.1109/ICSSE.2019.8823478.

[2] Y. Liu, J. Xing, Y. Li, S. Wang, D. Tao. "Tomographical Study of the Effect of Graphite on Properties of Cast Iron", Steel research international, 2018, doi:10.1002/srin.201800086.

[3] X. Huang, Y. Ye, X. Shen, X. Chang. "The mechanical properties of gray cast iron and metallographic structure effect on the chip shape", Advanced Materials Research, **339**, 200-203, 2011, doi:10.4028/www.scientific.net/AMR.339.200.

[4] I. Svensson, T. Sjögren. "On Modeling and Simulation of Mechanical Properties of Cast Irons with Different Morphologies of Graphite", International Journal of Metalcasting, **3**(4), 67-73, 2009, doi:10.1007/BF03355460.

[5] S. Kumar, P. Kumar, H. S. Shan. "Optimation of Tensile Properties of Evaporative Casting Process through Taguchi's Method". Journal of Materials Processing Technology, **204**, 59-69, 2008, doi:10.1016/j.jmatprotec.2007.10.075.

[6] E. Lunarska. "Effect of graphite shape on the corrosion of gray cast iron in phosphoric acid", Materials and Corrosion, **47**, 539-544, 1996, doi:10.1002/maco.19960471002.

[7] H. Nakae, H. Shin. "Effect of Graphite Morphology on Tensile Properties of Flake Graphite Cast Iron". Materials Transactions, **42**(7), 1428-1434, 2001, doi:10.2320/matertrans.42.1428.

[8] M. Tsujikawa, K. Nagamine, A. Ikenaga, M. Hino. "Influence of graphite morphology on dry sliding wear of flake graphite cast irons". International Journal of Cast Metals Research, **21**(1-4), 76-80, 2008, doi:10.1179/136404608X361701.

[9] R. L. Hecht, R. B. Dinwiddie, H. Wang. "The effect of graphite flake morphology on the thermal diffusivity of gray cast irons used for automotive brake discs". Journal of Materials Science, **34**(19), 4775-4781, 1999, doi:10.1023/a:1004643322951.

[10] M. C. Rukadikar, G. P. Reddy. "Influence of chemical composition and microstructure on thermal conductivity of alloyed pearlitic flake graphite cast irons". Journal of Materials Science, **21**, 4403-4410, 1986, doi:10.1007/bf01106563.

[11] W. Wang, T. Jing, Y. Gao, G. Qiao, X. Zhao. "Properties of a gray cast iron with oriented graphite flakes". Journal of Materials Processing Technology, **182**(1-3), 593-597, 2007, doi:10.1016/j.jmatprotec.2006.09.028.

[12] R. Hinrichs, M. A. Z. Vasconcellos, W. Österle, C. Prietzel. "Amorphization of Graphite Flakes in Gray Cast Iron Under Tribological Load". Materials Research, 21(4), 2018, doi:10.1590/1980-5373-MR-2017-1000.

[13] M.M. J. Behnam, P. Davami, N. Varahram. "Effect of cooling rate on microstructure and mechanical properties of gray cast iron", Materials Science and Engineering, **A528**, 583-588, 2010, doi:10.1016/j.msea.2010.09.087.

[14] Suyitno, Sutiyoko. "Effect of Pouring Temperature and Casting Thickness on Fluidity, Porosity and Surface Roughness in Lost Foam Casting of Gray Cast Iron", Procedia Engineering, **50**, 88-94, 2012, doi:10.1016/j.proeng.2012.10.011.

[15] T. Sjogren, P. Vomacka, I. L. Svensson. "Comparison of mechanical properties in flake graphite and compacted graphite cast irons for piston

rings". International Journal of Cast Metals Research, **17**(2), 65-71, 2004, doi:10.1179/136404604225017474.

[16] H. Megaheda, E. El - Kashifa, A. Y. Shashab, M. A. Essam. "Effect of holding time, thickness and heat treatment on microstructure and mechanical properties of compacted graphite cast iron", Journal of Materials Research and Technology, **8**(1), 1188-1196, 2019, doi:10.1016/j.jmrt.2018.07.021.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 853-864 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

An Enhanced Conceptual Security Model for Autonomous Vehicles

Abdulla Obaid Al Zaabi[*,1,2], Chan Yeob Yeun[1,2],Ernesto Damiani[1,2], Gaemyoung Lee[3]

[1]*Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi 127788, United Arab Emirates*

[2]*Center for Cyber-Physical Systems, Khalifa University, Abu Dhabi 127788, United Arab Emirates*

[3]*Department of Electrical Engineering, Jeju National University, Jeju, 63243, South Korea*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 01 November, 2020
Online: 27 November, 2020

Keywords:
Autonomous
Vehicles
Security

ABSTRACT

Connected and self-driving cars have emerged over the last decade as a leading example of cyber-physical systems, which seek to considerably enhance traffic safety, reduce emissions, decrease costs, and improve efficiency. Google, TESLA, Uber ATG are becoming pioneers in the autonomous vehicles industry. Autonomous vehicles can have a large codebase and with a large volume of messages exchanged. The concepts of connected driving, cooperative driving, and intelligent transportation systems increase the connectivity of vehicles to the internet or other cloud services. High connectivity, misconfiguration, and insecure coding widen the attack surface of autonomous vehicles. A conceptual model of autonomous vehicles is presented in this paper to better understand autonomous vehicles' attack surface. In addition, specific threat modeling techniques are discussed. Experiments were carried out to demonstrate the security risks to autonomous vehicles due to third party Electronic Control Units.

1 Introduction

The term autonomous vehicles is used to designate technology-oriented vehicles with robotic and self-driving features (self-parking, GPS sensing tools, and higher automation capabilities). Both connected and self-driving cars have emerged over the last decade as a leading example of cyber-physical systems, which seek to enhance traffic safety, reduce emissions, decrease costs and improve efficiency. At the same time, the automobile industry has experienced a major transformation from monolithic manufacturing processes to distributed procurement. The paradigm of distributed procurement allows automobile manufacturers to acquire software and sensor/hardware components from numerous third parties.

Distributed systems usually comprise of a complex connectivity layer with different nodes. The features of connectivity in autonomous vehicles provide real-time connections to the web, sensors, radar, GPS with user privacy, and other measurement units [1, 2]. In other words, autonomous vehicles are empowered to acquire all relevant information needed to perform the self-driving functions [3].

As the market penetration of autonomous vehicles has increased, there is a growing demand among organizations, consumers, and regulatory bodies for the underlying technologies to be tested to guarantee their overall safety, as well as evaluate their cloud and cyber security together with data privacy. The augmented diversity of software, sensors, and communication channels has driven this demand [4]. During the process of critical-information exchange between the vehicle and the external infrastructure, a new attack surface can emerge, as several stakeholders (telematics, hardware, and software components, etc.) and rapid modifications (i.e., software processes) are involved.

Sensors can provide an extensive amount of information on the environment surrounding the vehicle together with sensing the internal components. Electronic Control Units (ECUs) and sensors are the focal point of information exchange inside autonomous vehicles, as they provide critical information to the engine control unit of the vehicle. The data obtained from sensors can be shared and utilized for all types of assessments and calculations by other automotive systems.

Sensors and ECUs contain a large codebase that was enhanced to provide features required by autonomous vehicles. This "functional abundance" has widened the attack surface, adding to the existing vulnerabilities that have not been identified in the old code bases. Sensors and ECUs can talk to each other with different networks equipped in modern vehicles. The design of such networks has focused on efficiently wiring these components rather than on security.

There are 6 different levels of autonomous driving defined by

[*]Corresponding Author: Abdulla Al Zaabi, Abu Dhabi 127788, abdulla.alzaabi@ku.ac.ae

the Society of Automotive Engineers as shown in Figure 1. Modern vehicles nowadays can achieve levels between two and three.

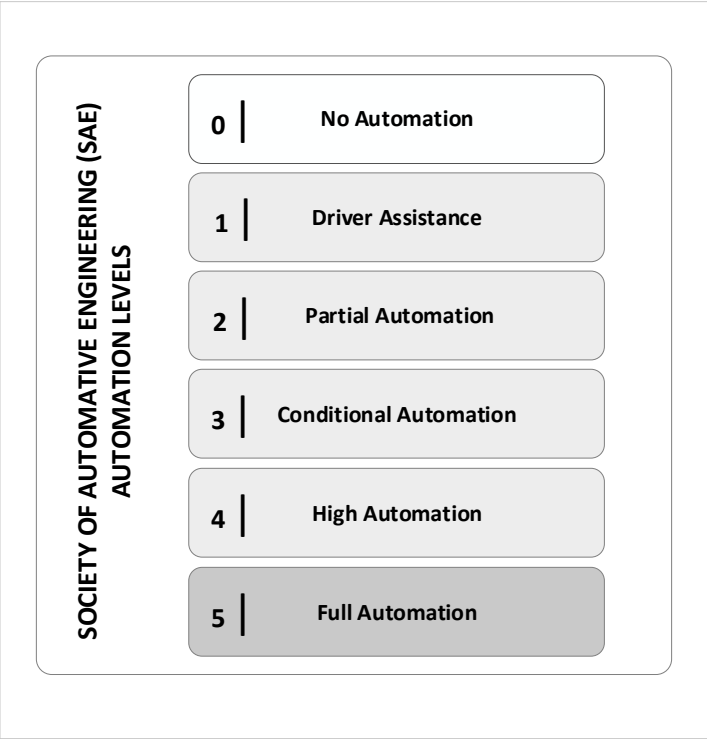


Figure 1: SAE automation levels

The motivation behind this work is to develop a conceptual model for autonomous vehicle to better understand the attack surface of autonomous vehicles and support their security testing. This will allow researchers and manufacturers to better understand the attack surface on autonomous vehicles and bridge the gap in the areas that have not been explored.

This paper develops on our previous work on defining a conceptual model for autonomous vehicles [5], where we modeled the autonomous vehicles' assets, their vulnerabilities and the corresponding threats. Here, we take a closer look to modeling the assets related to the Controller Area Network (CAN) of the vehicles. The rationale behind this choice is that many aftermarket gadgets and smart applications taking advantage of the CAN bus could contain vulnerabilities exploitable by malware. We conduct an experiment on a simulated CAN bus to illustrate different security attack scenarios such as traffic sniffing, replay attack and packet injection. Based on our experiment, we discuss and compare different threat modeling techniques that can be used to identify security threats based on the conceptual model presented.

The contribution of this article is, therefore: (i) enhanced conceptual model of autonomous vehicles to better understand their attack surface, (ii) analysis of Controller Area Network and cyber security attacks on modern vehicles, (iii) analysis of specific threat modeling techniques, (iv) demonstration of the cyber security risks to autonomous vehicles due to third party Electronic Control Units.

The paper is organized as follows. In Section 2 we present an autonomous vehicle conceptual model. Section 3 describes Controller Area Network protocols. Section 4 overviews cyber security attacks on modern vehicles. Section 5 focuses on a remote attack that had an impact on millions of vehicles. Section 6 provides an overview of different threat modeling techniques that could be used for threat modeling of autonomous vehicles. Section 7 illustrates the experimental results of sniffing and injecting traffic on CAN network using a simulated environment. In section 8 we summarize and discuss the future direction of research.

2 Autonomous Vehicle Conceptual Model

This section describes our enhanced conceptual model of autonomous vehicles [5]. The model consists of three main areas; vehicle, communications, and edge-cloud (Figure 2).

2.1 Vehicle

The vehicle component of the conceptual model focuses on the physical boundary of the vehicle. In the conceptual model, we focused on the important items that will support future work in security modeling and analysis of autonomous vehicles. The main categories of the vehicle layer are Electronic Control Units (ECUs), they are the brain of the vehicle, Sensors, and Inter-Vehicle communication. In addition, it is important to highlight the internal connectivity items such as the On-Board Diagnostic port (OBD-II). Different types of networks exist in autonomous vehicles with different speeds. The ECUs in the vehicle are usually connected to at least one or more internal vehicle network. The gateway is used to facilitate communication between different communication buses. The OBD-II connection provides access to the vehicle's internal network and interaction with various low-level sensors.

ECUs are essential components of modern vehicles. Each ECU is responsible for performing actions such as door locks, power steering, fuel injection, and many other critical functions. ECUs can also be categorized into subcategories based on their functionality [6]. The list below provides categories for ECUs found in modern vehicles:

- Powertrain: controls many factors in vehicles and responsible for transmission, emission controls, and onboard control modules.
- Safety: vehicle safety-related operations such as collision detection, active parking, and air-bag system.
- Body: less critical operations of the vehicle such as power windows and mirrors, cooling and heating, etc.

Examples of some of the key ECUs from the categories above are:

- ECM: engine control module
- VVS: vehicle vision system
- IRM: inflatable restrain module
- NCM: navigation control module

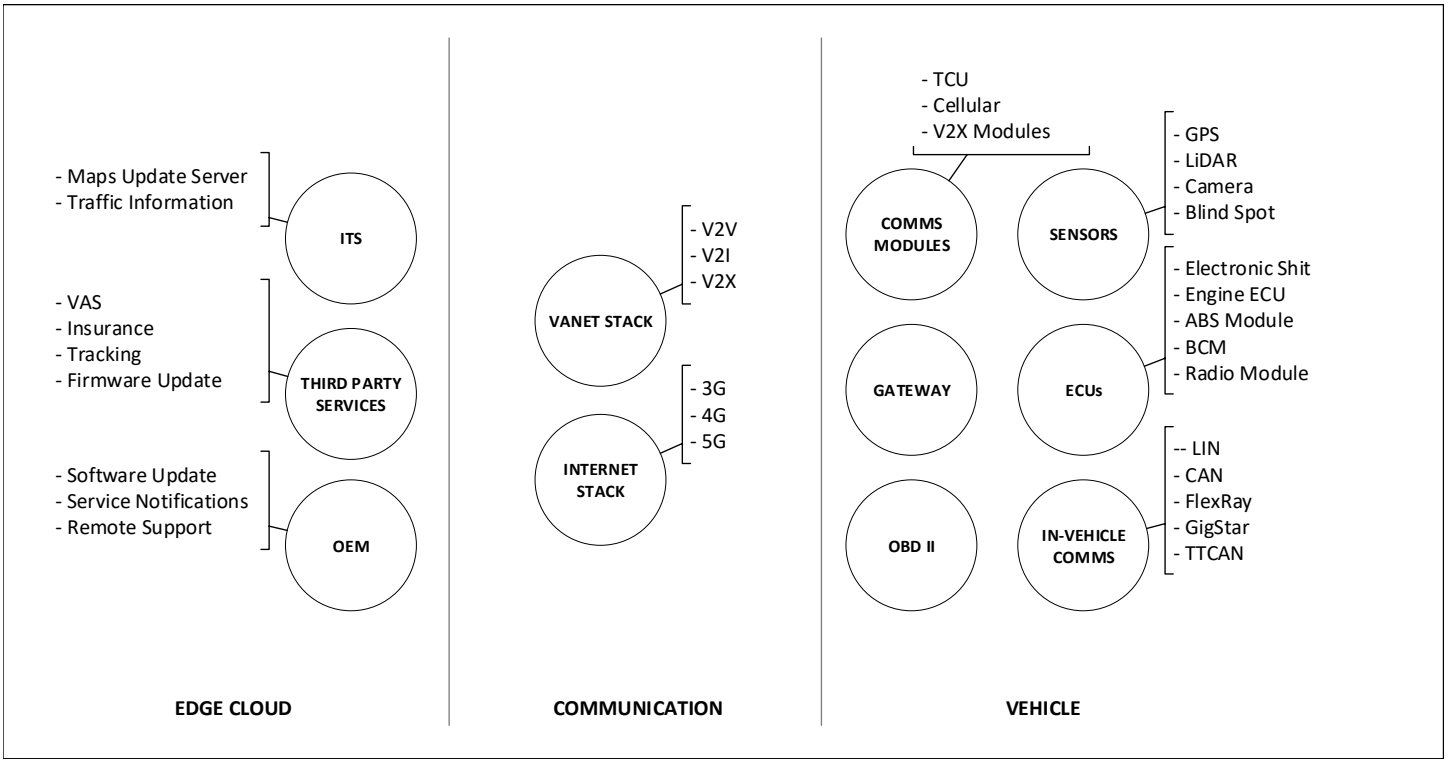


Figure 2: Autonomous Vehicle Conceptual Model

Sensors are used in autonomous vehicles to collect information about the vehicle's surroundings. Vehicles are equipped with various sensors, examples of sensors equipped are GPS, cameras, LiDAR. The vehicle uses the information received from the sensors to perform different actions such as warning messages, navigation capability, distance calculation, and others.

The Inter-vehicle communication component of the conceptual model focuses on the communication networks in the physical boundary of the vehicle. The main purpose of these networks is to allow efficient communication between electronic units in the vehicle. There are different types of networks in vehicles such as CAN, TTCAN, FlexRay, and Local Interconnect Network (LIN) [7, 8]. The CAN is widely used in automotive to support critical applications with a bandwidth of 125 kBits/s. The LIN bus runs at a lower bandwidth (10 kBit/s) to support less critical applications such as power windows and doors lock. FlexRay is a more expensive implementation of CAN, used commonly in BMW. FlexRay has 2 channels at a speed of 5 and 10 Mbit/s [9, 10].

2.2 Communication

The communication layer of the model focuses on the external communication medium for autonomous vehicles. Future transportation system is envisioned to support cooperative driving and information exchanging to enhance the driving experience, increase safety, and reduce collisions. There are two main categories in the communication layer; VANET stack and Internet Stack.

VANET Stack is the vehicular Ad Hoc Networks (VANETs). VANETs can use wireless networking technology to allow vehicles to communicate with each other and share information using Dedicated Short Range Communication (DSRC). There are different types of communication in VANETs. The first one is V2V, which stands for Vehicle to Vehicle communication. V2I: vehicle to station communication. V2R: vehicle to Road Side Unit (RSU) communication. V2X: vehicle to everything.

There are two types of messages in VANETs; beacon messages and special purpose messages [11, 12]. Vehicles broadcast messages that act as heartbeats using beacon messages. Special purpose messages carry information about collisions and warning messages to nearby vehicles [13]. VANETs using DSRC or even cellular network is anticipated as the future of intelligent transportation system [14].

Internet Stack component of the model considers the cellular network as the communication means. Internet stack includes 4G and 5G data link and physical link [15]. Initial research in autonomous driving focused on VANETs as the future of communication for vehicles using DSRC. Modern vehicles are equipped with cellular network connectivity that could enable VANET communication over cellular network.

2.3 Edge Cloud

Edge Cloud layer of the model focuses on the vehicle communication with external entities. Original Equipment Manufacturers (OEM) perform remote updates to vehicles remotely for applying patches, issues fix or enhance features. Many services are currently offered by third party providers such as media services, fleet management, vehicle tracking, and many other applications and services. Below are some examples:

- Maps update service.

- Traffic information.
- OEM firmware update.
- Third-party dongles firmware update.
- Value Added Services.

Some insurance companies nowadays mandate drivers to install OBD-II compatible dongles to monitor driving behavior and adjust insurance tariffs accordingly. Vehicles are expected to be connected to edge cloud services to support smart and green transportation [16]. As part of intelligent transportation system management [17, 18], vehicles are expected to be connected to transportation authorities and share various types of information.

3 Overview of Controller Area Network

The Controller Area Network (CAN) is a standardized network adopted by many car manufacturers for efficiency, safety, and manufacturing purposes. CAN ensures minimum latency, together with other features to handle functional networks of a vehicle. It has been adopted since early 1980s to primarily handle automotive electronics, and more specifically, different types of electronic control units [19]. The CAN protocol has many distinctive features such as being able to discover defects in the communication network and electronic control units (ECUs) [20]. ECUs play a crucial role in operating an autonomous vehicle. By allowing ECUs to exchange information via CAN, the autonomous vehicle can support new intelligent features. In comparison with other networks, however, CAN have some weaknesses such as limited electrical loading, and vulnerability to overflow of undesirable interactions or communications [21].

Known vulnerabilities of CANs have encouraged experts to develop a new security network for ECUs, especially for automated and self-driving vehicles [20]. One of the major motives for developing a new network is that CAN was not designed for communications from outside the vehicle, i.e. vehicle-to-Internet communications, vehicle-to-cloud, vehicle-to-pedestrian, and other types of connectivity. As a result, cyber security challenges, particularly denial-of-service attacks are a major hurdle when using CAN for autonomous vehicles. Moreover, third party dongles have been introduced by different service providers such as insurance or fleet management companies. These devices interface with the vehicle through the On-Board Diagnostics port (OBD-II). Physical OBD-II connection provides access to the vehicle's internal network and interaction with various low-level sensors.

3.1 CAN Protocols

CAN is a simple protocol that was deployed vehicles since 1996. It allows Electronic Control Units (ECUs) to communicate with each other. CAN run on two wires; CAN High (CANH) and CAN Low (CANL). At the physical layer, CANH uses differential signaling, which is used in environments where fault-tolerance to noise is a must. When a signal comes in, CANH raises the voltage on one line and drops the other line of an equal amount.

3.1.1 The OBD-II Connector

Most vehicles are equipped with On-Board Diagnostic Connector named OBD-II. It is usually located under the steering wheel of a vehicle. It is also known as a diagnostic link connector (DLC). CAN Connections can be found in supported vehicles from the OBD port in dual-wire pairs. CANH and CANL can be found on pin 6 and 14 or the OBD-II connector.

- CANH: High-Speed CAN Lines
- CANL: Low Speed or Mid Speed CAN Lines (LS-CAN and MS-CAN).

3.1.2 CAN Packets Layout

There are two types of CAN packets; standard and extended. Extended packets are like the standard ones but with larger space to hold IDs. The Complete frame of standard CAN is shown in Figure 3. Fields of standard packets are described below.

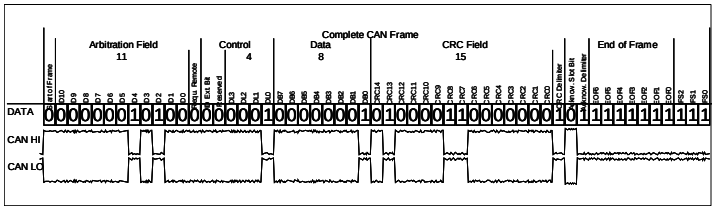


Figure 3: CAN Frame of standard packet [22]

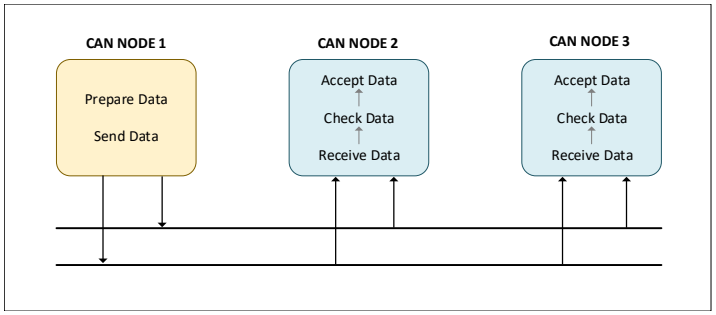


Figure 4: CAN communication based on broadcast messages

- Arbitration ID: a broadcast message that identifies the ID of the target device it tries to communicate with.
- Identifier Extension (IDE): always 0 for standard CAN
- Data length code (DLC): the size of the data, which ranges from 0 to 8 bytes
- Data: data to be transmitted.

Communication on CAN bus is broadcast messages, similar to UDP packets on ethernet networks as shown in Figure 4. All controllers see every packet transmitted on CAN. Extended packets are similar to standard packets, but the can be changed together to create longer IDs. They are designed to fit inside standard CAN packets with backward compatibility.

3.1.3 The ISO-TP Protocol

The ISO 15765-2 is a standard for sending packets over CAN. It extends the limit of sending 8 bytes to 4095 bytes. This protocol is commonly used for diagnostic messages (Unified Diagnostic Services) and KWP messages (alternative protocol to CAN). The first byte is used for extended addressing, leaving only 7 bytes for data per packet. Using ISO-TP can easily flood the CAN bus, thus it should be used carefully [23].

3.1.4 The Local Interconnect Network Protocol

The Local Interconnect Network (LIN) is a master-slave facility that was designed to complement CAN. It can have up to 16 slaves

4 Overview of Cyber Attacks on Modern Vehicles

This section focuses on recent cyber security attacks on modern vehicles. Firstly, the section provides an overview of the interesting vulnerabilities discovered by researchers in different modern vehicles. In the second part of this section discusses in detail one of the most interesting attacks on vehicles, the Jeep Cherokee 2015 attack.

4.1 Vulnerabilities Overview

4.1.1 Jeep Cherokee

Uconnect is an entertainment system that relies on cellular network to deliver entertainment services and applications in Jeep Cherokee entertainment system [24]. The entertainment system had a diagnostic bus port that runs on 6667. Since the port had no authentication requirements, some sensitive information was obtained by connecting to the port, such as GPS information, VIN, and other private information. At the time of the research, around 1.4 million vehicles using the Uconnect system were vulnerable. To identify vulnerable vehicles, a mass scan for the diagnostic bus port could be performed on the IP range of the service provider. The researchers used Sprint Cellular Network IP ranges to identify vulnerable vehicles on 21.0.0.0/8 and 25.0.0.0/8 address ranges. Additional details on the sequence of execution from Uconnect system to cyber-physical actions are explained in next section.

4.1.2 GM Chevy Impala

Another entertainment system related vulnerability was discovered in the GM Chevy Impala by researchers at the University of California at San Diego and the University of Washington [25]. This time, the entertainment system had a buffer overflow vulnerability. The vulnerability was exploited by playing a specially crafted mp3 file in the in-car call system. Successfully exploiting the vulnerability allows an attacker to access the vehicle's system and control most of its critical functions.

4.1.3 BMW

Security researchers identified a vulnerability in BMW ConnectedDrive service to unlock the vehicle remotely. Confidentiality requirement was missing in the communication between the vehicle and the BMW server for ConnectedDrive, the communication was not encrypted. Due to this, the attack performed using a fake cellular network and spoofed the BMW server. The vehicle is equipped with a communication capability using a pre-installed SIM card. An unlock message was sent to the vehicle using the spoofed server and it unlocked the vehicle successfully [26]. At the time, over 2.2 million vehicles were affected by the vulnerability. BMW released a security patch to fix the vulnerability by adding encryption to the communication between the vehicle and the BMW server.

4.1.4 Corvette

Corvette attack was mainly an attack on a third-party device, specifically on an insurance dongle connected to the vehicle via OBD port. The insurance dongle was installed to track driver behavior to determine insurance tariffs. The impact of the vulnerability depends on the commands allowed via OBD port. In the case of the Corvette, the researchers were able to trigger cyber-physical actions on the vehicle via the connected dongle. Through exploiting the vulnerability, the researchers were able to send an SMS to remotely apply breaks on the vehicle. Depending on the commands allowed from the OBD, other vehicles the attack could result in controlling more critical functions [27].

4.1.5 Tesla S Model

Different attackers were performed in Testa S Model, some required physical access to the vehicle and others exploited remotely. For the physical access, researchers were able to connect a laptop to the vehicle's in-car entertainment system and identified multiple vulnerabilities. Researchers were able to successfully implant a trojan in the vehicle's system to control it remotely. All the vulnerabilities identified in the research were patched by Tesla [28].

4.1.6 Nissan LEAF

Nissan application named Companion Application is used to allow users to have access to some information on the vehicle and perform specific actions. Researchers were able to identify vulnerabilities in the Companion Application and gain access to the historical GPS information of the vehicle [29].

5 Analysis of Cyber Security Attack on JEEP Cherokee

In this section, we analyze the attack on Jeep Cherokee that affected over 1.4 million vehicles. The researchers were able to perform the cyber attack remotely, with no physical access to the vehicle. Initial research required reverse engineering some of the components of the vehicle to plan the attack and chain different stages of the attack to reach from remote exploitation to perform physical actions on the vehicle such as applying breaks, vehicle's steering, and others. Although the experiment was conducted on Jeep Cherokee, it was not limited to this model. Vehicles with the same entertainment unit may have suffered similar vulnerabilities and impacts. In research, the attack surface was analyzed and the feasible attack chain was

chosen to perform the full attack chain, from initial access to physical action execution. The first part of this section discusses the CAN connectivity architecture seen in Jeep Cherokee, then discusses the attack phases. This is important to understand how to send messages between two networks in the vehicles (CAN buses).

5.1 CAN Connectivity

In the example of Jeep Cherokee, two CAN buses are used; CAN-HIS and CAN-C connected to different networks. ECUs talk to each other via CAN bus, each CAN bus is connected to a different set of ECUs. The CAN buses of Jeep Cherokee are explained below. The first CAN bus in the vehicle is CAN-IHS, stands for Controller Area Network for Interior High Speed. The CAN-IHS bus is low speed CAN bus that is used for non-critical systems such as climate controls, radios, and others. Examples of the ECUs connected to CAN-IHS bus are as follow:

- AMP Ampilifer Radio Module.
- Body Control Module (BCM).
- AHLM Module – Headlamp Leveling.
- PAM Module – Park Assist.
- DLC Data Link Connector

The second CAN bus is CAN-C bus, which is high-speed bus that is responsible to connect critical ECUs. Examples of the ECUs connected to CAN-C bus is as follow:

- Anti-Lock Brakes Module (ABS)
- Body Control Module (BCM)
- RADIO Module
- Electronic Parking Brake – EPB.
- Electronic Power Steering – EPS.

As we can see in the examples above, some ECUs are connected to both CAN buses, two examples are the radio module and the BCM Module. The direction the researchers took in the research was to target an ECU unit that could talk to both networks to widen the impact of the attack and reach critical ECUs.

5.2 Attack Phases

The attack chain developed by the researchers started from remote device exploitation to execute commands and sending control signals to critical ECUs. In section, the analysis is divided into four phases; target identification, remote code execution, sending CAN messages, and CAN payloads.

5.2.1 Identify Target

Initially, the researchers identified D-Bus service port 6667 by scanning the vehicle internally; connecting to the in vehicle's WiFi network. After examining the networking configurations of Uconnect, the researchers discovered an additional interface with a public IP address, which could result in remote exploitation of Uconnect system. To identify vulnerable Uconnect systems, performing a scan on port 6667 over the provider network. It was initially assumed that the provider blocks connections from other providers. The researchers took an additional step by exploiting a femtocell from the same provider to ensure that traffic is not blocked. It was later discovered that the port is open to all providers which increases the impact of the vulnerability. To avoid false positives, the service error message of services determines if the port belongs to Uconnect system or any other service such as IRC (Internet Relay Chat).

5.2.2 Gain Remote Code Execution

The researchers found more than one way to gain remote execution on the target system via the D-Bus service of Uconnect. Different vulnerabilities identified for remote code execution on the D-Bus service in addition to a feature that was implemented on the D-Bus service to allow code execution for an unauthenticated user. The first vulnerability idenfied was due to unsanitized user input on navTrailService function. The other vulnerabilities are either unauthenticated execution or command injection vulnerabilities.

Since arbitrary command execution achieved using the D-Bus service, researchers developed different payloads to interact with the CAN bus and obtain information using LUA scripts. Examples of the crafted commands sent via the D-Bus service.

- Uconnect system: From the Uconnect system, GPS coordinates can be queried.
- HVAC Module (A/C Heater): controlling heating and A/C of the vehicle.
- Radio Volume: controlling the volume of the radio.
- Knobs: disabling the D-Bus service make knobs unavailable and unresponsive.

5.2.3 Sending CAN Messages

The code execution achieved through the D-Bus service is performed on the OMAP chip. Although the Uconnect system is connected to both CAN buses of the vehicle; CAN-IHS and CAN-C, it cannot send messages to CAN directly. The communication from the OMAP to CAN buses are handled by Renesas V850ES/FJ3 chip. The researchers updated the V850 chip from the OMAP with a modified firmware to be able to send CAN messages. At this point, the execution from OMAP uses V850 chip as a proxy to communicate with CAN buses.

5.2.4 CAN Messages Payloads

The communication between OMAP chip and V850 chip uses a Serial Peripheral Interface (SPI). The V850 chip originally had a

predefined set of command IDs with hardcoded data to communicate with CAN bus. With flushed firmware, the V850 can send arbitrary CAN data from the OMAP chip.

At this point, arbitrary CAN commands can be sent remotely via the D-Bus service. After identifying the messages triggering cyber-physical actions, researchers were able to perform turn signals, locks, RPMS. These messages are considered normal CAN messages, the diagnostic CAN messages could cause greater cyber-physical impact but requires the vehicle to be traveling at a low speed.

6 Analysis of threat modeling frameworks

Threat modeling is in important phase of security engineering process. It involves identifying all security threats to the system regardless of the possibility to exploit them. It requires good understanding of the system. In order to develop meaningful security requirements, it is essential to have threats identified. Proper identification of threats in early stages can greatly reduce the ability of misusing the system [30]. There are different techniques of threat modeling [31]. In general, some of the main threat model techniques are:

- Attacker-centric
- Asset-centric
- Software-centric

These models are general and not specific to autonomous vehicles.

6.1 Attacker-centric approach

This threat modeling technique is built on attacker perspective by profiling an attacker's characteristics, goals and motivation to exploit vulnerabilities. The modeling technique aims to understand potential attackers who are likely to target the system and type of attacks that might be executed. Based on the understanding, mitigation strategies can be developed against studied attackers.

This approach uses tree diagrams for threat modeling. Examples of key elements of this approach include: the goals of the attacker, the techniques used by the attacker, the system components including software and assets and means of detection and mitigation.

A simple example is an attacker trying to obtain data from a database that is hosted in a backend system accessible from a website. Using this approach, the goal of the attacker would be to obtain the information stored in the database. Techniques include SQL injection (using unsensitized input commands) on the website, and use of automated tools by which exploitation could be carried out.

6.1.1 Intel's TARA (Threat Agent Risk Assessment)

This threat modeling technique focuses on prioritizing the most critical areas of the system and most vulnerable to attacks [32]. This ensures that the threat modeling efficiently address the areas that are exposed the most. Moreover, the framework focuses on threat agents that are most dangerous, what techniques and tradecrafts used to achieved their goals. The most exposed areas are then determined by cross-referencing the information obtained with known vulnerabilities. The process is shown in Figure 5. TARA's goal is to build security strategies with focus on areas with highest level of risk. This is achieved by determining the most likely attack vectors and applying efficient security strategy to address them. TARA relies on three main components:

- Threat Agents Library (TAL): unique library containing 22 threat agent and nine threat agent attributes.
- Common Exposure Library (CEL): list of known vulnerabilities and exposures.
- Methods and Objectives Library (MOL): contains threat agent objectives and methods used to achieve goals and objectives.

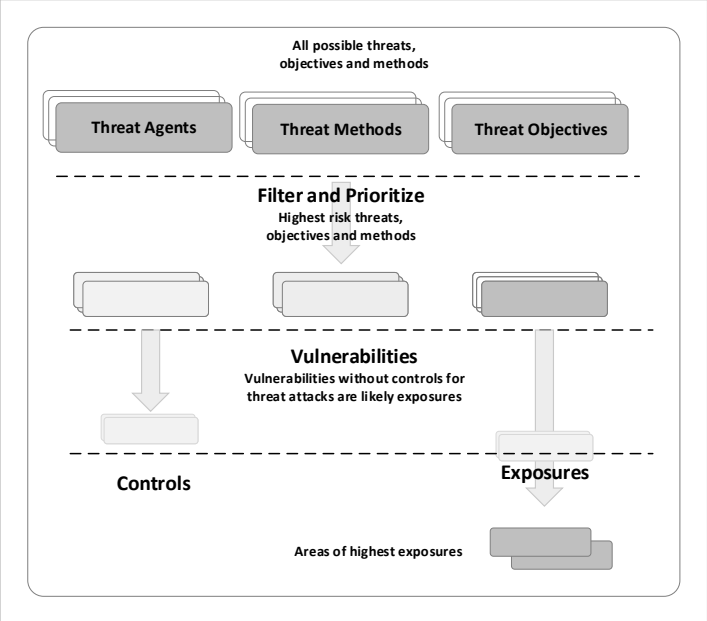


Figure 5: Narrowing down the fields of attacks.

6.1.2 Cyber Kill Chain

The Cyber-Kill threat modeling approach was developed by Lockheed Martin, an American company focusing on global aerospace, defense, security and advanced technologies [33]. The method focuses on the steps taken by an attacker to achieve desired goal. As name implies, the method is chained into multiple phases, each phase has its associated set of goals. Each phase outcomes contribute to the next phase. The cyber kill chain contains the following steps:

- Reconnaissance: research, identification, and selection of targets
- Weaponization: Combining remote access malware with exploit into deliverable payload
- Delivery: Delivering a payload to target (weaponized exploit).
- Exploitation: Exploit vulnerable application or system upon delivery

- Installation: Persistent access via means of installing backdoor in exploited system.
- Command and Control: Outside server communicates with the weapons providing "hands on keyboard access" inside the target's network

6.2 Asset-centric approach

This approach has two main components; assets and threats. The assets taxonomy involves identifying assets, asset groups, asset types, asset details [34]. The threat taxonomy utilizes attack trees, attack graphs and assets' own failure modes in order to prioritize associated risk levels and then mapped to assets. Methodologies based on the asset-centric approach include PASTA, OCTAVE and STRIDE

6.2.1 PASTA

PASTA (Process for Attack Simulation and Threat Analysis) is an asset-centric approach to threat modeling [34]. This methodology was developed by Minded Security and VerSprite. The obvious advantage of this methodology that it providers valuable information about risk reduction and security strategy that could target non-technical staff (example management or stakeholders). The methodology involves seven different stages illustrated below.

- Stage 1: Define Objectives
- Stage 2: Define Technical Scope
- Stage 3: Application Decomposition
- Stage 4: Threat Analysis
- Stage 5: Weaknesses and Vulnerability Analysis
- Stage 6: Attack Modeling and Simulation
- Stage 7: Risk Analysis and Management

The main goals of PASTA methodology are:

- Allows cost and time savings
- Integrates with SDLC
- Better understanding of most likely attack sources
- Better communication between security groups inside the company
- Increases the maturity of the organization in security

6.2.2 OCTAVE

OCTAVE Allegro (Operationally Critical Threat, Asset and Vulnerability Evaluation) a method developed by CERT Survivable Enterprise Management team. This method focuses information assets security in respect to the well-known security requirements, which are Confidentiality, Integrity and Availability (CIA). This method investigates the ways information assets are stored, transferred, used and processed. In addition, it extends to cover threats and vulnerabilities associated with information assets. The OCTAVE Allegro method consists of four different phases, with total of eight steps as shown in Figure 6.

- Phase 1: risk measurement criteria development
- Phase 2: identification and profiling of assets then the identification of containers to the identified assets
- Phase 3: threats to identified containers are identified with respect to assets
- Phase 4: mitigation strategies

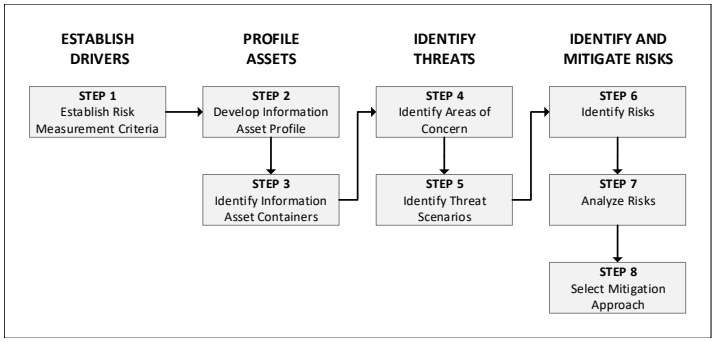


Figure 6: Phases and steps of OCTAVE Allegro.

6.2.3 STRIDE

The acronym STRIDE stands for six categories of security risks: Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privileges.

- Spoofing refers to the act of posing as someone else (i.e. spoofing a user) or claiming a false identity (i.e. spoofing a process). Spoofing threats affect the authenticity property.
- Tampering refers to malicious modification of data in transit, at rest, or within processes. Tampering threats affect the integrity property.
- Repudiation refers to the ability of denying that an action or an event has occurred. Repudiation threats affect the non-repudiation property.
- Information Disclosure refers to data leaks or data breaches. Disclosure threats affect the confidentiality property.
- Denial of Service (DOS) refers to causing a service or a network resource to be unavailable to its intended users. DOS threats affect the availability property
- Elevation of Privileges (EP) refers to gaining access that one should not have. EP threats affect the availability property

By considering these risks for each asset, STRIDE threat model allows to list the applicable risks (also called failure modes) to each asset, and to map them to threats and known attacks. This way measures can be identified to examine and address gaps in the security posture of applications.

7 Experimental Results

This section of the paper discusses and analyzes a proof of concept of two replay attack and packet injection on a simulated CAN bus. In this experiment, we setup a simulation environment to simulate random CAN traffic generation and identifying control signals to perform specific physical actions.

7.1 Simulation Environment Setup

The simulation environment is setup using a virtual CAN bus than communicated over TCP and other tools that interact with the virtual CAN bus. In this section, we described the tools used to set up the simulation environment.

CAN for vehicle communication have different drivers and software utilities with no unified CAN tools and interfaces. However, there is a set of tools with a common interface available in Linux. The SocketCAN is an interface that was created on the Open Source development site BerliOS in 2006. By extending the Berkeley socket API in Linux, SocketCAN can send and receive traffic from the CAN controller in the model of a network device.

SocketCAN is not limited to physical CAN devices and networks, it supports built-in CAN chips and card drivers, external communication via serial or USB, and virtual CAN devices. More importantly, it supports virtual CAN environments.

- Atmel AT91SAM SoCs
- Bosch CC770
- ESD CAN-PCI/331 cards
- Freescale FlexCAN
- Freescale MPC52xx SoCs (MSCAN)
- Intel AN82527
- Microchip MCP251x
- NXP (Philips) SJA1000

Figure 7 compares the implementation of traditional CAN software with the SocketCAN implementation, which is unified. In traditional CAN software, the application has its protocols stack that talks to a character device, whereas SocketCAN is implemented in the Linux kernel with its own CAN protocol family enabling applications to communicate with a CAN bus interface in a generic network interface fashion.

SocketCAN repository contains various utilities for Linux CAN subsystem. These utilities can work with CAN communications within the vehicle from Linux operating system. Examples of the functions supported by the suite:

- Basic tools to display, record, generate can traffic
- Tools to send, generate, sniff CAN frames.
- Serial line discipline (SLC) configuration
- Log files converters
- CAN access via IP sockets
- CAN bus management

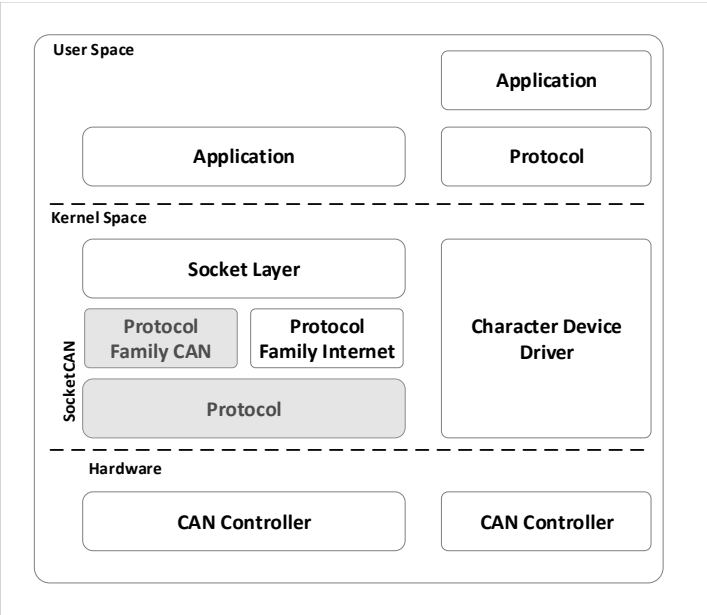


Figure 7: SocketCAN compared to traditional CAN software

Different tools are used in the experiment For CAN bus diagnosis, monitoring, and simulating target vehicle that interacts with the virtual CAN. For CAN diagnostic and monitoring, Kayak application is used. Kayak is a Java-based that can read CAN traffic log files and display information graphically on gadgets.

For target vehicle simulation and CAN traffic fuzzing, the instrument cluster simulator (ICSim) is used, which is an opensource tool developed by Open Garages. The ICSim software is Linux based and relies on virtual CAN devices. ICSim was designed to allow researchers, individuals, developers and others to safely familiarize themselves with CAN reversing. ICSim includes a graphical interface for some of the functions of CAN such as doors, engine RPM, and others.

7.2 Security Attacks on CAN Bus

The following experiment shows how the virtual environment setup can generate, display, and sniff CAN frames. Using the libraries and utilities described earlier in this section, a virtual CAN interface was created in Linux (Debian 4.19-28).

Using the virtual CAN interface is running as “vcan0”, CAN frames can be sent and received using can-utils suite. The current messages are random CAN messages, which are not currently analyzed as per their functions or character device they are trying to communicate with. The first step is generating random CAN messages with default bi trate. The messages generated can be viewed in Figure 8.

Since the messages are communicated over TCP/IP in the virtual environment, messages can be sniffed using different tools such as Wireshark and CANSniffer or Kayak.

```
vcan0  19B   [8]  1F 92 27 58 3A 92 CA 50
vcan0  280   [8]  A6 DD 51 25 FA 32 99 30
vcan0  00B   [8]  11 FD 6A 0B EC B4 3D 5A
vcan0  57A   [8]  40 24 FD 69 F7 30 A7 1A
vcan0  385   [8]  2F E0 CD 65 F5 B4 61 4D
vcan0  244   [8]  37 54 DE 4A 31 A7 9E 5F
vcan0  37D   [8]  5F 38 E1 22 38 AF 26 4B
vcan0  170   [8]  99 8B 43 61 18 FC 0A 5E
```

Figure 8: CAN frames dump using candump of can-utils

7.3 Injecting CAN Frames

This part explains injecting CAN frames into the virtual CAN bus and displays results in the simulator. To identify the CAN message responsible to perform a physical action, the ICSim was used to generate specific actions on the vehicle, and the generated CAN traffic was recorded. Then simply replaying the recorded messages could identify the messages related to the performed actions with other noise that can be discarded for this experiment. The sequence followed in the experiment is explained in Figure 9.

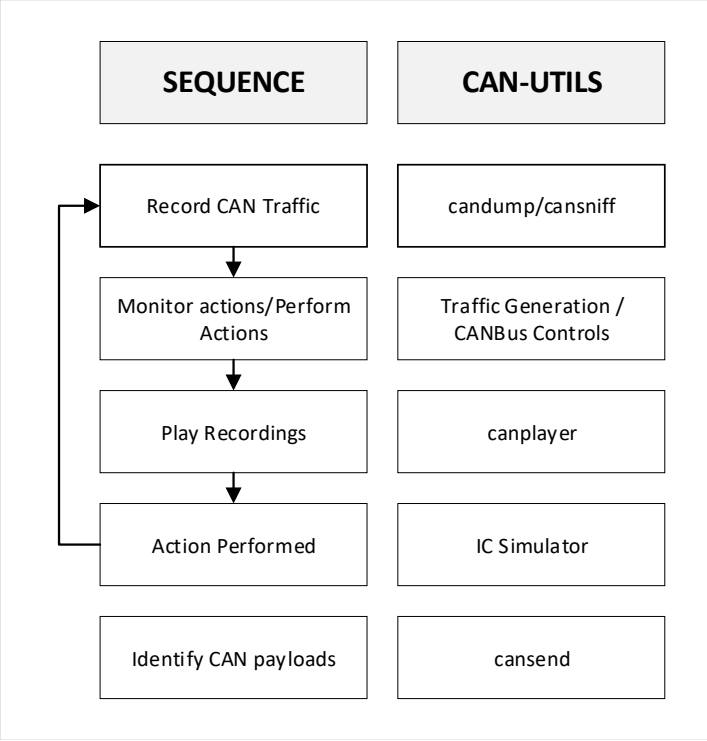


Figure 9: Sequence used to identify CAN messages

Using the generated traffic from the libraries explained earlier, the signals used to perform physical actions were identified. CAN traffic was generated on virtual CAN and recorded in log files. The signals responsible for door locks, RPM and signals identified as shown in Figure 10. At this point, using the CAN ID, we can send specific CAN messages to trigger doors, acceleration or light signals. For example, to accelerate, the CAN ID used is 244 (hex representation).

```
vcan0  517   [8]  09 B4 54 4A F6 4E D9 32
vcan0  188   [8]  B4 59 D4 54 36 DC D3 23
vcan0  0E6   [8]  A3 04 CD 47 E7 19 0C 7B
vcan0  0AF   [8]  8F 04 3A 50 F0 4E D2 18
vcan0  3D6   [8]  FD F6 28 0D 60 F3 28 6D
vcan0  19B   [8]  1F 92 27 58 3A 92 CA 50
vcan0  280   [8]  A6 DD 51 25 FA 32 99 30
vcan0  00B   [8]  11 FD 6A 0B EC B4 3D 5A
vcan0  57A   [8]  40 24 FD 69 F7 30 A7 1A
vcan0  385   [8]  2F E0 CD 65 F5 B4 61 4D
vcan0  244   [8]  37 54 DE 4A 31 A7 9E 5F
vcan0  37D   [8]  5F 38 E1 22 38 AF 26 4B
vcan0  170   [8]  99 8B 43 61 18 FC 0A 5E
```

Figure 10: Specific CAN messages for desired functions

After identifying the CAN messages to perform specific physical actions, two security attacks are performed. The first one is a replay attack. This was achieved by recording CAN traffic in a file. Since CAN communication (in the scenario described in the paper) does not use session IDs or encryption to protect against replay attacks, traffic was recorded were played on a different virtual CAN bus successfully. The second security attack performed is traffic injection. In a virtual CAN bus with random traffic being communicated, specific commands were sent to set RPM to a specific value, open doors, and start turn signals. The results of the signals sent are shown in Figure 11.

The tools used in the experiment do not simulate actual ECUs in terms of behavior. Some ECUs are wired directly and do not accept messages from the CAN bus directly. In such situations, additional effort is required to send signals from allowed ECUs. In diagnostic mode, messages might be accepted from the CAN bus. Additional research is required to fully understand what can be controlled directly from the CAN bus. Aftermarket ECUs expose CAN connectivity and expand the attack surface of modern vehicles. In the experiment, we focused on simple scenarios of packet sniffing and packet injection. The CAN bus is designed to allow real-time communication between ECUs but it is not secured by design. These types of attacks are not limited to aftermarket ECUs or gadgets, any successful exploitation or code execution on ECU connect to CAN bus could achieve similar or higher impact.

In the presence of aftermarket devices, new range of vulnerabilities will be introduced. The inter-vehicle communication was not designed to interact outside the perimeter of the vehicle. With the absence of defined threats, it is extremely difficult to provide assurance on security measures taken.

Starting from a conceptual model for autonomous vehicles is essential to capture different components of autonomous driving ecosystem. Threat modeling could be applied to identify potential threats pertaining autonomous vehicles and support the development of security requirements that cover scenarios discussed in this paper. This should support addressing the knowledge gap in the field of autonomous vehicle security.

Our future plans include the research of a test-based security verification framework for autonomous vehicles. Due to the lack of a centralized tests repository for security testing, we will focus on developing a security test-cases repository based on blockchain technology. In addition, the current research was limited to few

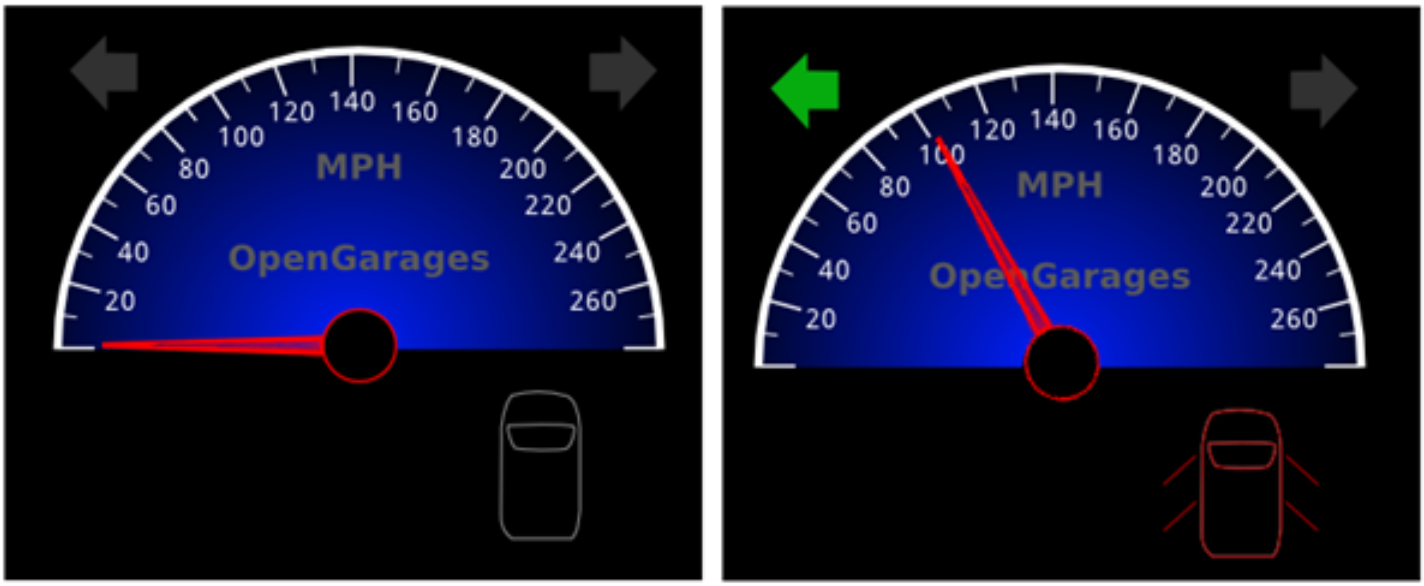


Figure 11: Results of the physical actions displayed on ICS Simulator

scenarios on attacks on CAN bus. We plan to design more attack scenarios against common vulnerabilities in aftermarket ECUs.

8 Conclusion

Several technologies in modern vehicles were designed and built without security in mind. As modern vehicles and future autonomous vehicles are becoming more and more computerized and connected, these technologies present new risks and widen the attack surface.

Threat modeling is an essential step for defining security requirements of autonomous vehicles, or any system in that regard. With the absence of defined threats, it is extremely difficult to provide assurance on security measures taken. The initial work conducted towards the research was focused on establishing a conceptual model for autonomous vehicles. This will support in defining the scope and basing the threat model against it.

The experiment showed briefly how easily CAN traffic can be captured or sniffed. In the presence of aftermarket devices, new range of vulnerabilities will be introduced. The inter-vehicle communication was not designed to interact outside the perimeter of the vehicle. However, these devices open new attack surfaces allowing some components of inter-vehicle communication to be exposed to the outside world.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work was supported by the Center for Cyber-Physical Systems, Khalifa University, under Grant 8474000137-RC1-C2PS-T3.

References

[1] A. S. Gajparia, C. J. Mitchell, C. Y. Yeun, "Supporting user privacy in location based services," IEICE transactions on communications, **88**(7), 2837–2847, 2005, doi:10.1093/ietcom/e88-b.7.2837.

[2] D. M. Konidala, C. Y. Yeun, K. Kim, "A secure and privacy enhanced protocol for location-based services in ubiquitous society," in IEEE Global Telecommunications Conference, 2004. GLOBECOM'04., volume 4, 2164–2168, IEEE, 2004, doi:10.1109/GLOCOM.2004.1378393.

[3] A. Talebpour, H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," Transportation Research Part C: Emerging Technologies, **71**, 143–163, 2016, doi:10.1016/j.trc.2016.07.007.

[4] B. Sheehan, F. Murphy, M. Mullins, C. Ryan, "Connected and autonomous vehicles: A cyber-risk classification framework," Transportation research part A: policy and practice, **124**, 523–536, 2019, doi:10.1016/j.tra.2018.06.033.

[5] A. O. Al Zaabi, C. Y. Yeun, E. Damiani, "Autonomous Vehicle Security: Conceptual Model," in 2019 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific), 1–5, IEEE, 2019, doi:10.1109/ITEC-AP.2019.8903691.

[6] A. M. Wyglinski, X. Huang, T. Padir, L. Lai, T. R. Eisenbarth, K. Venkatasubramanian, "Security of autonomous systems employing embedded computing and sensors," IEEE micro, **33**(1), 80–86, 2013, doi:10.1109/MM.2013.18.

[7] "L. S. P. Revision," 2016.

[8] C. P. Szydlowski, "CAN specification 2.0: Protocol and implementations," Technical report, SAE Technical Paper, 1992, doi:10.4271/921603.

[9] S. C. Talbot, S. Ren, "Comparision of fieldbus systems can, ttcan, flexray and lin in passenger vehicles," in 2009 29th IEEE International Conference on Distributed Computing Systems Workshops, 26–31, IEEE, 2009, doi:10.1109/ICDCSW.2009.15.

[10] M. Wolf, A. Weimerskirch, C. Paar, "Secure in-vehicle communication," in Embedded Security in Cars, 95–109, Springer, 2006, doi:10.1007/3-540-28428-1_6.

[11] L. Buttyán, T. Holczer, I. Vajda, "On the effectiveness of changing pseudonyms to provide location privacy in VANETs," in European Workshop on Security in Ad-hoc and Sensor Networks, 129–141, Springer, 2007, doi:10.1007/978-3-540-73275-4_10.

[12] M. Raya, J.-P. Hubaux, "The security of vehicular ad hoc networks," in Proceedings of the 3rd ACM workshop on Security of ad hoc and sensor networks, 11–21, 2005, doi:10.1145/1102219.1102223.

[13] P. Papadimitratos, L. Buttyan, T. Holczer, E. Schoch, J. Freudiger, M. Raya, Z. Ma, F. Kargl, A. Kung, J.-P. Hubaux, "Secure vehicular communication systems: design and architecture," IEEE Communications Magazine, **46**(11), 100–109, 2008, doi:10.1109/MCOM.2008.4689252.

[14] L. Bariah, D. Shehada, E. Salahat, C. Y. Yeun, "Recent advances in VANET security: a survey," in 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), 1–7, IEEE, 2015, doi:10.1109/VTCFall.2015.7391111.

[15] H. Seo, K.-D. Lee, S. Yasukawa, Y. Peng, P. Sartori, "LTE evolution for vehicle-to-everything services," IEEE communications magazine, **54**(6), 22–28, 2016, doi:10.1109/MCOM.2016.7497762.

[16] C. C. Forecast, "Global connected car market to grow threefold within five years," GSMA Connected Living Programme: mAutomotive, 2013.

[17] N. Lu, N. Cheng, N. Zhang, X. Shen, J. W. Mark, "Connected vehicles: Solutions and challenges," IEEE internet of things journal, **1**(4), 289–299, 2014, doi:10.1109/JIOT.2014.2327587.

[18] N. Liu, "Internet of Vehicles: Your next connection," Huawei WinWin, **11**, 23–28, 2011.

[19] A. K. Singh, A. Negi, S. Azad, S. Mudali, "Fuzzy based controller for lidar sensor of an autonomous vehicle," Energy Procedia, **117**, 1160–1164, 2017, doi:10.1016/j.egypro.2017.05.241.

[20] D. Fényes, B. Németh, P. Gáspár, "A predictive control for autonomous vehicles using big data analysis," IFAC-PapersOnLine, **52**(5), 191–196, 2019, doi:10.1016/j.ifacol.2019.09.031.

[21] D. J. Fagnant, K. Kockelman, "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," Transportation Research Part A: Policy and Practice, **77**, 167–181, 2015, doi:10.1016/j.tra.2015.04.003.

[22] C. Smith, The car hacker's handbook: a guide for the penetration tester, no starch press, 2016, doi:10.4271/1593277032.

[23] I. Foster, K. Koscher, "Exploring controller area networks," login. USENIX Association, **40**(6), 2015.

[24] C. Miller, C. Valasek, "Remote exploitation of an unaltered passenger vehicle," Black Hat USA, **2015**, 91, 2015.

[25] A. Greenberg, "GM took 5 Years to fix a full-takeover hack in millions of Onstar cars," Wired, September, 1059–1028, 2015.

[26] "J. Banter, "Open Thyself! – Security vulnerabilities in BMW's Connected-Drive," C'T Magazine for Computer Technique," 2020.

[27] "A. Charlton, "Hackers disable Corvette brakes by texting dongle meant to lower insurance risk," International Business Times," 2020.

[28] "D. Gilbert, "Tesla Model S hacked: Researchers discover six security flaws in popular electric car," International Business Times," 2020.

[29] "J. Billington, "Hacker takes control of Nissan electric vehicle from other side of the world through Leaf app," International Business Times," 2016.

[30] S. Myagmar, A. J. Lee, W. Yurcik, "Threat modeling as a basis for security requirements," in Symposium on requirements engineering for information security (SREIS), 2005, 1–8, Citeseer, 2005.

[31] S. F. Burns, "Threat modeling: A process to ensure application security," GIAC security essentials certification (GSEC) practical assignment, 2005.

[32] T. Casey, "Threat agent library helps identify information security risks," Intel White Paper, **2**, 2007, doi:10.13140/RG.2.2.30094.46406.

[33] D. Kiwia, A. Dehghantanha, K.-K. R. Choo, J. Slaughter, "A cyber kill chain based taxonomy of banking Trojans for evolutionary computational intelligence," Journal of computational science, **27**, 394–409, 2018, doi:10.1016/j.jocs.2017.10.020.

[34] T. UcedaVelez, M. M. Morana, Risk centric threat modeling, Wiley Online Library, 2015, doi:10.1002/9781118988374.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 865-870 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Low Power Fast Settling Switched Capacitor PTAT Current Reference Circuit for Low Frequency Applications

Muhammed Mansoor C. B.*, Hanumantha Rao G., Rekha S.

Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, 575025, India

ARTICLE INFO

Article history:
Received: 29 August, 2020
Accepted: 06 November, 2020
Online: 27 November, 2020

Keywords:
PTAT
Switched Capacitor
Log-Domain Filter

ABSTRACT

This paper presents a low voltage, low power, fast settling switched capacitor based Proportional to Absolute Temperature (PTAT) current reference circuit. Unlike in a conventional resistor based PTAT current source, the proposed circuit saves a significant amount of silicon area on chip and hence the circuit becomes less susceptible to process variations. It creates a reference current of 1 nA from a 0.5 V power supply at room temperature (27°C). It has PTAT characteristics in the temperatures from −10°C to 80°C. The circuit draws a very low power of 1.5 nW and exhibits a good supply voltage sensitivity of 3.2 %/V. A startup circuit connected to the PTAT source improves the transient response by reducing the settling time. To test the PTAT current reference circuit, a low power log-domain filter which can be used for biomedical applications is realised and biased with the proposed PTAT current source. Results show that the filter cutoff frequency is constant over temperature variations. The CMOS technology used for designing the circuits is UMC 65 nm and tool used for simulations is Cadence Virtuoso.

1 Introduction

Portable and wearable/implantable devices are gaining worldwide acceptance among consumer/biomedical electronics. Circuit designs in such devices have a bottleneck on power dissipation and chip area. Log-domain circuits are one of the best possibilities in exploring low power, high dynamic range analog designs. Log-domain filters and transconductance-capacitor ($G_m - C$) filters can be used in biological data acquisition system which involves low frequency signals like ECG (electrocardiogram) and EEG (electroencephalogram). Since the transistors in these log-domain filters are working in sub-threshold mode of operation (weak-inversion), a low cutoff frequency can be achieved with low values of bias currents (order of *nA*). This paper is an extension of work originally presented in [1].

The cutoff frequency (f_o) of log-domain and $G_m - C$ filters is given by (1) [2, 3].

$$f_o = \frac{G_m}{2\pi C} = \frac{I_o}{2\pi\eta V_T C} \tag{1}$$

where η, I_o and V_T are sub-threshold slope factor, bias current and thermal voltage respectively. Equation (2) gives the expression for thermal voltage.

$$V_T = \frac{kT}{q} \tag{2}$$

where k, T and q have their usual meanings.

Equation (1) shows the direct relation between cutoff frequency and bias current. I_o is usually generated from a current reference circuit. The current reference should be stable against variations in process and supply voltage. Supply voltage sensitivity is the parameter which shows the stability of the current reference against supply voltage variations. Lower supply voltage sensitivity indicates a better current reference. Depending on the mode of operation of transistors in the circuit or the topology of the filter architecture, the current reference can be a temperature independent or Proportional to Absolute Temperature (PTAT) current reference. From (1) and (2), in log-domain and $G_m - C$ filters, it is evident that f_o has an inverse relationship with temperature. To fix this problem, I_o is generated in such a way that it has also a linear relationship with temperature. Thus a PTAT current reference is used to bias log-domain/$G_m - C$ filters. As a cumulative effect, the filter cutoff frequency remains constant over temperature variations.

Many PTAT current reference designs are explored in the literature. Beta-multiplier based circuits discussed in [4]–[5] have adequate PTAT current characteristics. However, to generate low reference currents under low supply voltage for low frequency filters used in biological data acquisition system, larger values of resistors are needed which occupies large area on chip. Moreover the control over temperature coefficient of resistance also becomes

*Corresponding Author: Muhammed Mansoor C. B., Research scholar, Dept. of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, India, +91 9497368483, mansoorcb@gmail.com

difficult. A comprehensive research has been carried out on the conventional PTAT designs to tackle such shortcomings [6]–[7]. In [7], self-cascode transistors are used instead of resistor. This improves supply variation rejection, but linearity in current versus temperature characteristic has been compromised. In [8], composite transistors are used in the PTAT circuit to enhance the supply voltage sensitivity, but large resistor values are needed to generate smaller currents. In [2, 9], switched capacitors are used in place of resistor. But PTAT current is generated from PTAT voltage using opamp circuits like voltage to current converter which in turn increases the overall power dissipation of the circuit. [1] shows a good switched capacitor based PTAT reference. But the settling time of the reference current is compromised.

Settling time of current reference becomes a significant aspect especially in high speed applications. When the settling time of the reference current increases, the circuit which uses this reference current to generate its bias current also takes a significant amount of time to settle its bias current to steady state value. In case of a filter used in a wireless communication system, it takes a significant time to fix its cutoff frequency which may lead to attenuation of relevant signal information. For example, the circuit presented in [1] requires 8 ms to stabilise its current which means that it takes around 8 ms for the subsequent filter to fix its cutoff frequency which is biased by this reference current source. This work proposes a feasible solution to this problem.

In this paper, a fast settling PTAT current source is proposed. A startup circuit added to the PTAT source improves the settling time of reference current. The proposed circuit exhibits good PTAT characteristics and better stability against power supply fluctuations. This can be used to bias low frequency filters in biomedical applications. A low power log-domain filter is also designed and biased using the proposed PTAT current source. Results show that the cutoff frequency of the filter remains constant irrespective of temperature variations. This paper is organized as follows. Details of the proposed PTAT circuit along with the results are given in Section 2. Log-domain filter realisation and filter response are detailed in Section 3. Finally conclusions are drawn in Section 4.

2 Proposed Improved PTAT Reference

A conventional beta-multiplier based current source circuit [5] is shown in Figure 1. Depending on the mode of operation of transistors, it can be a constant-G_m biasing circuit [5] or a PTAT current source. It acts as former if all the transistors work in strong-inversion saturation region ($V_{GS} > V_{th}$ and $V_{DS} \geq V_{GS} - V_{th}$) and as latter if all the transistors work in weak-inversion saturation ($V_{GS} < V_{th}$ and $V_{DS} \geq 3V_T$).

When transistor operates in weak-inversion region, its drain current [10] can be expressed as following.

$$I_D = \frac{W}{L} I_{D0} e^{\frac{V_{GS}-V_{th}}{\eta V_T}} (1 - e^{\frac{-V_{DS}}{V_T}}) \tag{3}$$

where $I_{D0} = \mu_n C_{ox} (\eta - 1) {V_T}^2$; μ_n and C_{ox} are electron mobility and gate oxide capacitance per unit area respectively. Other terms have their usual meanings. For $V_{DS} \geq 3V_T$ (weak-inversion saturation region), (3) can be approximated to (4).

$$I_D = \frac{W}{L} I_{D0} e^{\frac{V_{GS}-V_{th}}{\eta V_T}} \tag{4}$$

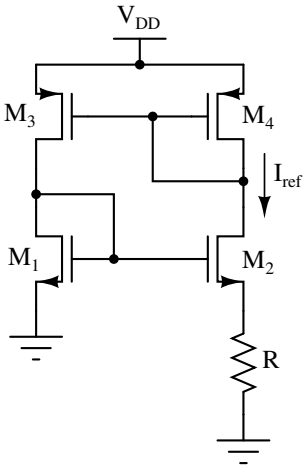


Figure 1: Beta-multiplier circuit

From Figure 1, the reference current I_{ref} which is also the drain current of M_2 (I_{D2}) can be written as (5).

$$I_{ref} = \frac{V_R}{R} \tag{5}$$

Voltage across resistor R (V_R) can be given as (6).

$$V_R = V_{GS1} - V_{GS2} \tag{6}$$

Using (4), gate-source voltages of M_1 and M_2 can be expressed as (7) and (8) respectively.

$$V_{GS1} = \eta V_T ln(\frac{I_{D1}}{I_{D0}(\frac{W}{L})_1}) + V_{th1} \tag{7}$$

$$V_{GS2} = \eta V_T ln(\frac{I_{D2}}{I_{D0}(\frac{W}{L})_2}) + {V_{th2}}' \tag{8}$$

${V_{th2}}'$ can be written as (9) [11] to include the body effect of M_2.

$${V_{th2}}' = V_{th2} + (\eta - 1) V_{SB} = V_{th2} + (\eta - 1)(V_R) \tag{9}$$

By substituting (7) and (8) in (6) and by assuming $V_{th1} = V_{th2}$, V_R can be derived as following.

$$V_R = I_{ref}.R = V_T ln(p) \tag{10}$$

where

$$p = \frac{(W/L)_2}{(W/L)_1}$$

Thus, I_{ref} can be expressed as (11).

$$I_{ref} = \frac{V_T ln(p)}{R} \tag{11}$$

Assuming R is thermally stable, I_{ref} is directly proportional to temperature thereby giving PTAT characteristics. For lower values of currents (order of nAs), large values of resistors are needed. For example, to generate a reference current of 1 nA using circuit shown in Figure 1, a 10 MΩ resistor is required if p is chosen as 1.5. To accommodate such large resistors on chip, huge area is needed

which makes the circuit more process dependent, hence reducing the accuracy of I_{ref}. In a typical SoC (System-on-Chip), where digital circuits and analog circuits co-exist together, a stable and precise clock will be available. In such systems, the resistor R in Figure 1 can be replaced by a switched capacitor driven by this clock which significantly enhances accuracy of I_{ref}.

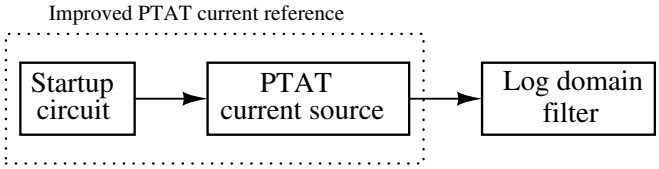


Figure 2: Block diagram

The entire block diagram of the circuit setup explained in this paper is shown in Figure 2. The startup circuit is added to the PTAT current generator to reduce the settling time of I_{ref} thereby improving the transient response.

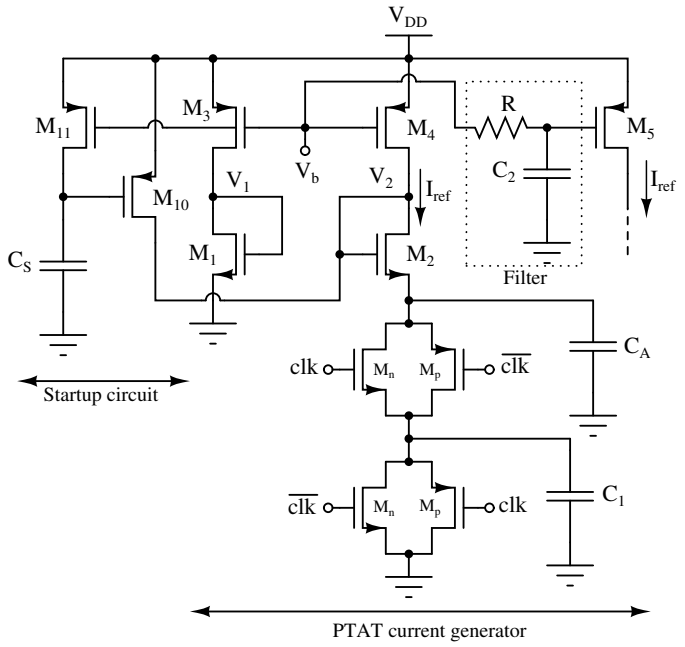


Figure 3: Proposed improved PTAT current reference

The transistor level implementation of PTAT current reference along with startup arrangement is shown in Figure 3. In the proposed circuit, every transistor is working in weak-inversion region. The value of p is chosen as 1.5 which means that the size of M_2 is 1.5 times the size of M_1. A switched capacitor comprising of capacitor C_1 and transmission gate switches (M_n and M_p) is used in place of resistor. M_p is selected 4 times larger than M_n in the transmission gate so that its on resistance does not depend on the voltage across it. Two clocks ('clk' and '$\overline{clk}$') which are complementary to each other and having frequency $\frac{f_{clk}}{}$ are used for the switching operation. When 'clk' becomes high ('$\overline{clk}$' becomes low), capacitor C_1 charges and when 'clk' becomes low ('$\overline{clk}$' becomes high), it discharges. The average resistance (R_{eq}) of the switched capacitor is given as in (12) [5].

$$R_{eq} = \frac{1}{f_{clk} C_1} \tag{12}$$

By substituting R_{eq} (12) in place of R in (11), we get (13).

$$I_{ref} = V_T f_{clk} C_1 ln(p) \tag{13}$$

From (13) it can be seen that, I_{ref} is directly proportional to C_1. If f_{clk} and p are chosen properly, smaller capacitor values can be obtained for the currents in the order of nA.

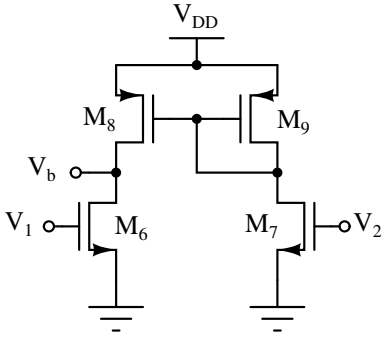


Figure 4: Error amplifier

The purpose of error amplifier which is shown in Figure 4 is to make the drain voltages of M_1 and M_2 equal thereby making currents through them equal. This kind of biasing minimises the dependency of I_{ref} on V_{DD} which improves supply voltage sensitivity.

It takes around *8 ms* for the reference current to settle to the steady state value of *1 nA* in the circuit presented in [1]. To reduce this delay, a startup arrangement comprising of M_{10}, M_{11} and C_S is added to the circuit. A similar kind of startup circuit is discussed in [12]. In Figure 3, when the circuit is switched on and before it reaches the steady state, initially the gates of $M_{1,2}$ and $M_{3,4}$ are at zero volt and V_{DD} respectively. Assuming zero initial voltage across the capacitor C_S, M_{10} turns on and pulls up the gates of $M_{1,2}$ to V_{DD}. $M_{1,2}$ start conducting and I_{ref} rises. Then, potential at the gates of $M_{3,4}$ drops and M_{11} turns on. C_S starts to charge towards V_{DD} and potential at the gate of M_{10} increases and once the circuit reaches steady state, M_{10} turns off. M_{10} and M_{11} do not have a static current path from V_{DD} to ground. Hence startup circuit dissipates zero static power.

Value of capacitor C_S and sizes of M_{10} and M_{11} determine how fast startup circuit can initialise the main circuitry and how fast it can be disconnected once main circuitry reaches the steady state. Transistors M_{10} and M_{11} operate in linear region and the linear on resistance of these transistors are given by (14).

$$R_{on} = \frac{1}{\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{th})} \tag{14}$$

How fast startup circuit is switched on depends on how fast transistor M_{10} turns on. The on time of M_{10} is limited by the input capacitance of M_{10} which comprises of gate to source and gate to drain capacitances. By properly choosing the size of M_{10}, the on time can be optimised.

When it comes to disconnection of startup circuitry, assuming M_{11} as a linear resistance given by (14), the RC time constant (τ) of the branch consisting of M_{11} and C_S is given by (15).

$$\tau = \frac{C_S}{\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{th})} \tag{15}$$

When capacitor C_S charges above the voltage $(V_{DD} - V_{th})$, M_{10} turns off. $(V_{DD} - V_{th})$ is approximately 150 mV in this design and time for C_S to charge to this voltage (30 % of final value) from 0.5 V supply through M_{11} is nearly 0.4τ.

A low pass filter which comprises of R and C_2 is added to maintain the ripple content in the output reference current within the accepted level. Capacitor C_2 is chosen as 0.1 pF and resistor R is tuned to obtain the ripple content in I_{ref} less than *1 %*. To remove high frequency noise during switching operation, capacitor C_A is used at the source of M_2. A 100 kHz clock is selected for this work. The capacitors, C_1, C_A, C_S are taken as 0.8 pF, 10 pF, 1 pF respectively. By substituting (13) in (1), f_o can be expressed as (16).

$$f_o = \frac{f_{clk} C_1}{2\pi\eta C} ln(p) \tag{16}$$

From (16), it can be seen that f_o is independent of temperature.

2.1 Simulation results

The proposed circuit is designed using UMC 65 nm CMOS process with a supply voltage of 0.5 V. Periodic Steady State (PSS) analysis is carried out to plot the reference current. Figure 5 shows the output current generated (1 nA) for one clock period (10 μs) at room temperature ($27°C$).

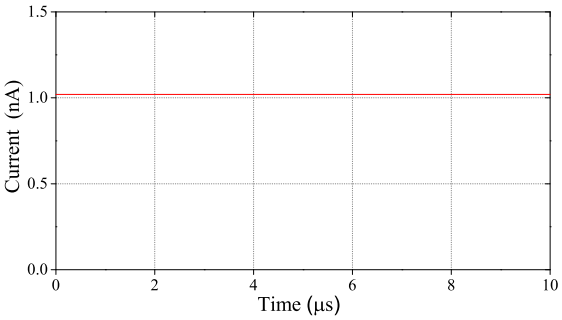


Figure 5: I_{ref} at V_{DD}= 0.5 V

Figure 6 shows I_{ref} variations when temperature is varied from $-10°C$ to $80°C$ for V_{DD} = 0.5 V. It can be observed that I_{ref} changes linearly with temperature and thus has a good PTAT characteristics over the range of temperature.

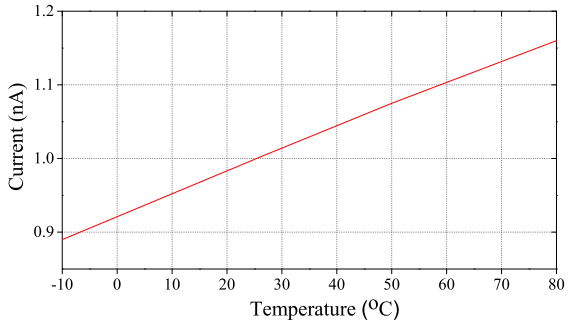


Figure 6: I_{ref} vs. Temperature plot

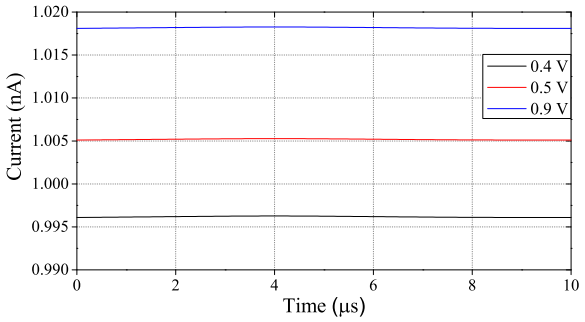


Figure 7: I_{ref} for different supply voltages

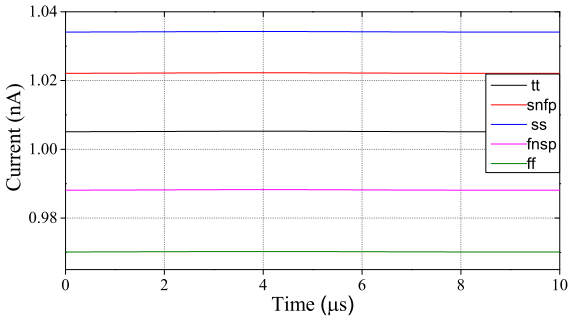


Figure 8: I_{ref} across the process corners

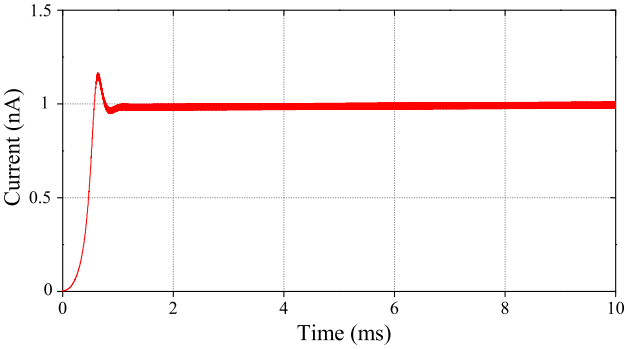


Figure 9: Transient response of the current flowing through M_2

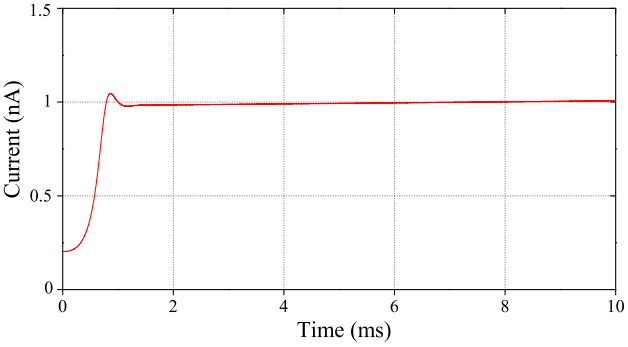


Figure 10: Transient response of the current flowing through M_5

I_{ref} is plotted for different values of V_{DD} at room temperature as shown in Figure 7. I_{ref} is *0.996 nA* and *1.018 nA* for V_{DD} of 0.4 V and 0.9 V respectively. The supply voltage sensitivity of I_{ref} is found to be *3.23 %/V* by considering these extreme values.

Figure 8 plots I_{ref} for various process corners at room temperature and V_{DD} = 0.5 V. I_{ref} deviates maximum in *ss* and *ff* corners (*1.034 nA* and *0.968 nA*). The maximum deviation is less than ±*3.6 %*.

The current flowing through the source of M_2 is shown in Figure 9. It has ripples having peak to peak magnitude of approximately *0.1 nA*. The ripple free current flowing through M_5 is shown in Figure 10. The settling time is found to be approximately *1 ms*.

Table 1: Comparison of the proposed work with other similar works

	[7]	[13]	[14]	[1]	This work
Technology (nm)	1500	180	40	65	65
Supply voltage (V)	1.2	1.2	1	0.5	0.5
Reference current (nA)	0.4	6	100	1	1
Settling time (ms)	-	-	-	8	1
Power (nW)	2	-	350	1.5	1.5
Supply voltage sensitivity (%/V)	6	6.47	2.9	2.7	3.2

The improved PTAT reference consumes a power of *1.5 nW* at room temperature from a supply voltage of 0.5 V. The proposed circuit along with existing PTAT current references in the literature are listed in Table 1 and a fair comparison is done. The proposed PTAT circuit functions under the lowest supply voltage and has a very low supply voltage sensitivity and settles faster than [1] (nearly 87 % improvement in settling time).

3 Log-Domain Filter Realization

The proposed PTAT current source shows excellent PTAT characteristics. To verify this, a low frequency log-domain filter is designed and biased with the proposed switched capacitor based PTAT current reference. This log-domain filter can be used in low power biomedical applications like processing of ECG, EEG signals. The expectation is that the cutoff frequency of the filter remains the same irrespective of the variations in temperature.
The transfer function of a first order low-pass filter is

$$H(s) = \frac{Y(s)}{X(s)} = \frac{\omega_o}{s + \omega_o} \tag{17}$$

where ω_o is the cutoff frequency of the filter. The corresponding time domain differential equation is given by

$$\frac{dy(t)}{dt} + \omega_o y(t) = \omega_o x(t) \tag{18}$$

where x(t) and y(t) represent the input and output currents of the filter respectively. By solving the differential equations and relating the current-voltage in transistor operating in subthreshold operation, we get

$$I_K I_{out} = I_{dc} I_{in} \tag{19}$$

where I_{in} contains both dc current I_{dc} and input current I_{signal}. I_K is the current flowing through M_3 which contains both capacitor current I_C and I_{dc}. I_{out} is the output current. Equation (19) can be realized using a translinear loop having four transistors. These equations can be implemented in the form of a translinear circuit as shown in Figure 11 [15].

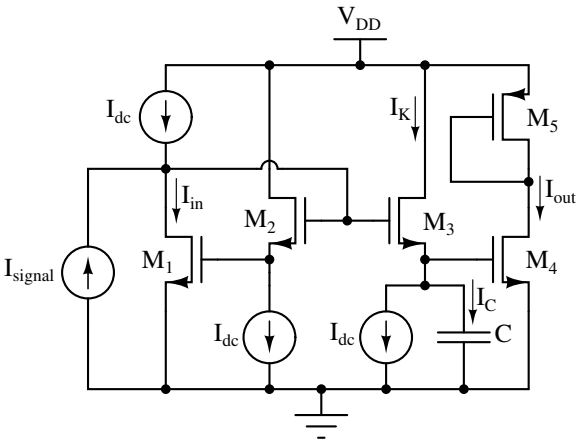


Figure 11: First order log-domain filter

In this circuit, the MOSFETs M_1–M_4 form the translinear loop. These transistors are working in sub threshold saturation region. The cutoff frequency of the filter can be tuned by varying the DC current and is given by following [16].

$$f_o = \frac{I_{dc}}{2\pi\eta V_T C} \tag{20}$$

The ideal DC current sources shown in Figure 11 can be replaced by current sources implemented using MOSFETs. In this work I_{dc} is taken as I_{ref} (1 nA). The cutoff frequency of 100 Hz is obtained by choosing C=37 pF.

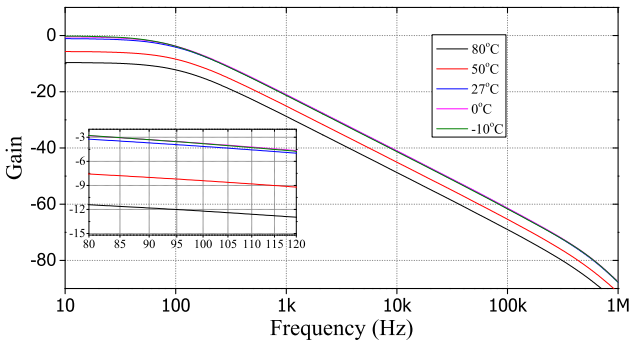


Figure 12: AC response of the filter biased with temperature independent current reference

3.1 Simulation results

The filter is designed using UMC 65 nm CMOS technology with a supply voltage of 0.5 V. For a comparison, I_{dc} is generated from a temperature independent current reference and ac response of the filter is plotted. Figure 12 shows the variation of cutoff frequency

in the temperatures ranging from $-10°C$ to $80°C$. The maximum deviation is found to be *16 %*. To verify the PTAT characteristics, I_{dc} is generated from the proposed PTAT source and ac response is plotted. Figure 13 shows that deviation of cutoff frequency is negligible and maximum deviation is found to be less than *2 %*. The band transitions are enlarged in corresponding inset graphs. Table 2 shows the deviation of filter cutoff frequency for both temperature independent and the proposed PTAT references.

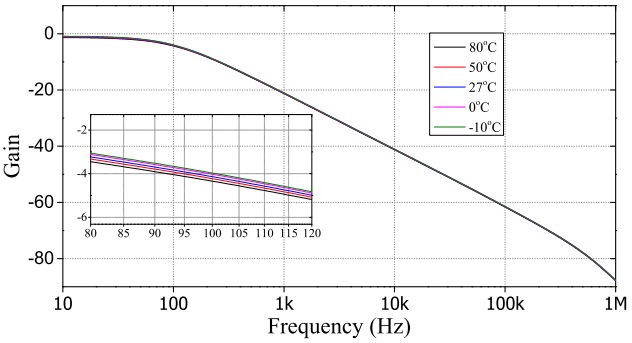


Figure 13: AC response of the filter biased with PTAT current reference

Table 2: Cutoff frequency deviation

Temperature (°C)	Cutoff frequency (Hz)	
	Temp. independent	PTAT
-10	116	101.9
0	107	101.2
27	100	100
50	93	99
80	88	98.4

Table 3 lists the parameters of the filter designed in this work.

Table 3: Parameters of the filter

Technology	65 nm CMOS
Supply voltage	0.5 V
Power	2.5 nW
Cutoff frequency	100 Hz
DC gain (at $27°C$)	0.94
I_{in_pp} at 25 Hz for 1% THD	0.82 nA
Input referred noise from 1 Hz to 100 Hz	3.26 pA_{rms}
Dynamic range	48 dB

4 Conclusion

This paper proposes a low voltage, low power, fast settling switched capacitor PTAT current reference circuit. Settling time is reduced with the aid of a startup circuit. The PTAT circuit exhibits good PTAT characteristics in the temperatures ranging from $-10°C$ to $80°C$ while operating under a very low power. The circuit is less susceptible to power supply fluctuations and can work at supply voltages as low as 0.4 V. To test the PTAT source, a low power log-domain low frequency filter is also designed and biased with PTAT reference. Simulation results imply that the filter cutoff frequency remains unchanged irrespective of temperature variations and this validates the theory explained in the paper.

References

[1] H. G. Rao, S. Rekha, "A 0.5 V, 1 nA Switched Capacitor PTAT Current Reference Circuit," in 2019 IEEE Conference on Modeling of Systems Circuits and Devices (MOS-AK India), 51–54, 2019, doi:http://idr.nitk.ac.in/jspui/handle/123456789/7046.

[2] A. Gerosa, A. Maniero, A. Neviani, "A fully integrated dual-channel log-domain programmable preamplifier and filter for an implantable cardiac pacemaker," IEEE Transactions on Circuits and Systems I: Regular Papers, **51**(10), 1916–1925, 2004, doi:https://doi.org/10.1109/TCSI.2004.835027.

[3] R. Schaumann, H. Xiao, V. V. Mac, Design of Analog Filters 2nd Edition, Oxford University Press, Inc., 2009.

[4] E. Vittoz, J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations," IEEE journal of solid-state circuits, **12**(3), 224–231, 1977, doi:https://doi.org/10.1109/JSSC.1977.1050882.

[5] B. Razavi, Design of analog CMOS integrated circuits, Tata McGraw-Hill, 2002.

[6] L. Magnelli, F. Crupi, P. Corsonello, C. Pace, G. Iannaccone, "A 2.6 nW, 0.45 V temperature-compensated subthreshold CMOS voltage reference," IEEE Journal of Solid-State Circuits, **46**(2), 465–474, 2011, doi: https://doi.org/10.1109/JSSC.2010.2092997.

[7] E. M. Camacho-Galeano, C. Galup-Montoro, M. C. Schneider, "A 2-nW 1.1-V self-biased current reference in CMOS technology," IEEE Transactions on Circuits and Systems II: Express Briefs, **52**(2), 61–65, 2005, doi: 10.1016/j.trc.2016.07.007.

[8] H. G. Rao, C. B. M. Mansoor, S. Rekha, "A 0.8 V, 5 nA PTAT current reference circuit with improved supply voltage sensitivity," in 2019 Global Conference for Advancement in Technology (GCAT), 1–4, 2019.

[9] S. Malik, M. Schlarmann, R. Geiger, "A low temperature sensitivity switched-capacitor current reference," in Proceedings of the European Conference on Circuit Theory and Design (ECCTD), 269–272, Citeseer, 2001.

[10] Y. Tsividis, C. McAndrew, Operation and Modeling of the MOS Transistor, Oxford Univ. Press, 2011.

[11] L. H. Ferreira, T. C. Pimenta, R. L. Moreno, "An ultra-low-voltage ultra-low-power weak inversion composite MOS transistor: Concept and applications," IEICE transactions on electronics, **91**(4), 662–665, 2008, doi: https://doi.org/10.1093/ietele/e91-c.4.662.

[12] S. Kumar, H. G. Rao, S. Rekha, "A 1-V, 8.6-nA Resistor-less PTAT Current Reference with Startup Circuit," in 2018 15th IEEE India Council International Conference (INDICON), 1–4, 2018.

[13] Y. Osaki, T. Hirose, N. Kuroki, M. Numa, "1.2-V supply, 100-nW, 1.09-V bandgap and 0.7-V supply, 52.5-nW, 0.55-V subbandgap reference circuits for nanowatt CMOS LSIs," IEEE Journal of Solid-State Circuits, **48**(6), 1530–1538, 2013, doi:https://doi.org/10.1109/JSSC.2013.2252523.

[14] V. Jorge, A. L. Aita, "A 1-V PTAT current reference circuit with 0.05%/V current sensitivity to VDD," in Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 502–505, 2016.

[15] A. Patil, C. B. M. Mansoor, H. G. Rao, S. Rekha, "Low power log-domain filter," in 2019 IEEE 16th India Council International Conference (INDICON), 1–4, 2019.

[16] J. Mulder, W. A. Serdijn, A. C. van der Woerd, A. H. van Roermund, Dynamic translinear and log-domain circuits: analysis and synthesis, volume 481, Springer Science & Business Media, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 871-882 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Blockchain Application in Higher Education Diploma Management and Results Analysis

Fernando Richter Vidal*, Feliz Gouveia, Christophe Soares

Faculty of Science and Technology, University Fernando Pessoa, Porto, Portugal

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 14 November, 2020
Online: 08 December, 2020

Keywords:
Blockchain
Blockchain Revocation
Higher Education
Digital Diplomas
Academic records
Blockcerts
Electronic document

ABSTRACT

Academic certifications are achievements desired by people, because they have a direct impact, positively, on their social lives. Such an important document, still widely issued in paper format, may be subject to forgery or impossibility of verification due to the unavailability of the issuing entity. This work consists of identifying, analyzing and testing some of the blockchain-based tools that are emerging, to offer more efficiency, reliability and independent degrees. A concept proof is presented, through the implementation of a prototype capable of issuing, verifying and sharing certificates. The results of this experiment are presented, analyzing the use of blockchain technology for this purpose. Finally, the work presents an overview of the current state of development and maturity in which these tools are found, reporting the advances and limitations, and exposing issues that still need to be resolved.

1 Introduction

This work is an extension of the work originally presented in 2019 at the international conference "Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)" [1], which aims to present an analysis of the use of blockchain technology in the education area, for issuing and verifying academic degrees in higher education.

Academic certificates attest to the certificate holder's abilities and skills and are accepted internationally [2]. These qualifications have a major impact on income and social position, both in emerging countries and in developed countries.

For example, in Brazil, data from *Pesquisa Nacional por Amostra de Domicilios Continuas* (PNADC) showed that the level of education is decisive for the Brazilian's income [3]. According to the results presented, those who complete higher education achieve almost triple the remuneration compared to those who only have a high school education.

In Europe, according to Eurostat [1], every year higher education graduates in Europe exceed 4.5 million, in which France and the United Kingdom are the leading countries with over 740,000 graduates per year. The fraction of people with university degrees between 30 and 34 years old almost doubled in fourteen years, going from 23 % in 2002, to 39 % in 2016 [4].

As professionals become qualified, knowledge is directly transformed into income, and consequently in a better quality of life. Furthermore, these numbers confirm a continuous increase and constitute a solid basis that justifies the creation of solutions to verify the authenticity of university degrees.

The weakness of the paper model, which is still widely used, was even more evident during the crisis caused by COVID-19, in which the use of digital resources became indispensable.

Although digital initiatives have emerged to address these weaknesses [5], these solutions still depend on the issuing entity, concerning authenticity verification. Fortunately, blockchain promises to offer, besides this innovation, other improvements not yet achieved by digital solutions. Using the resources that the technology offers, it is possible to make the check process disconnected from the issuer and still guarantee authenticity.

The second innovation is related to privacy. Certificates are documents that contain personal information and are sensitive to data leakage. If, on the one hand, student privacy needs to be preserved, on the other hand, its distribution benefits those interested [6].

*Fernando Richter Vidal, Sao Joao da Boa Vista-SP Brazil,37705@ufp.edu.pt
[1] `http://ec.europa.eu/eurostat`

Finally, blockchain solutions offer the third innovation compared to digital certificates, about timestamp. Due to the information being immutable and stored chronologically within the network, they accurately express the dates on which the events occurred.

This paper discusses the use of blockchain technology, aiming to benefit academic certificates with these possible innovations. The document is organized as follows: the next section provides a brief overview of blockchain technology; next, we describe the main differences between the blockchain solution and digital certificates; right after, we present the CertEdu prototype developed by *University Fernando Pessoa* (UFP), as well as a brief discussion of the results obtained and we conclude with some final remarks.

2 Background

2.1 Blockchain

Blockchain is a distributed system, formed by records that are organized in blocks and are linked to each other through cryptographic mechanisms. The technology became well known through the cryptocurrency Bitcoin [7], and soon expanded to several other applications.

There are three types of blockchain: public, private or consortium [8]. Public networks, also called permissionless blockchains, provides a free access environment to any participant who wants to join the network, however, the transaction validation rules are predefined and cannot be changed by any member. Typically, this type of blockchain implements the *Proof of Work* (PoW) consensus mechanism. In private networks, also called permissioned blockchain, the rules are defined according to the business interest. The organization that controls the network can define, for example, which users will decide on consensus or how to manage the network to accept new members. Finally, the consortium network is a category that merges properties from the public and private networks in the same environment. For example, in some scenarios, it would be interesting to keep the access public to the network, but also allow to make some data encrypted to preserve privacy and the anonymity of a participant.

Due to the consensus mechanisms and their *Peer-to-Peer* (P2P) topology, the network has gained strong protection against tampering. The success of an attempt of violation is conditioned to changes in all other previous blocks. Besides, attacks that would be able to take control of the network, such as a 51 % attack, would be extremely costly and could cost USD 500K per hour of processing [9]. In this aspect, the bigger the network, the more secure it becomes, which explains why Bitcoin is more secure when compared to other smaller networks [10]. This blockchain features results in an important property, which is immutability.

A block is always connected to a previous block, except for the genesis block, forming, then, a sequential chronological chain. The maximum size that a block can reach, depends on how each blockchain platform implements it. On Bitcoin, this value cannot exceed 1 Mb [11], while, on Ethereum there is no fixed limit, and the size change according to the number of gas units that can be spent per block (block gas limit).

As shown in figure 1, the block is organized in two areas: header and content. The header stores eighty bytes of control information. We find information like the hash header and the previous block; a software version, the target of the difficulty, a nonce, the root of Merkle [12], and a timestamp. Regarding the content area, they are stored as records. On average, a content area can store over five hundred transactions [5].

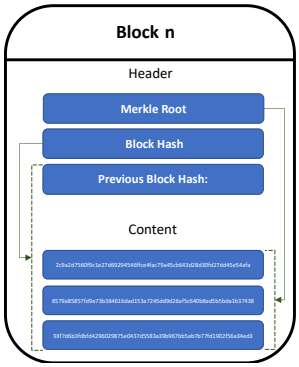


Figure 1: Overview of a Block Structure

Transactions are used to transfer an asset from one account to another account. The owner of an asset can move it inside the blockchain, through an account identified by a public address, which is controlled by its corresponding private key [13].

There is a conceptual difference between the public key and public address. The public key is used to verify the signatures generated by the private key's owner, while the address serves to identify the account. Using the example of Bitcoin, the process of generating the address is as follows: the public key is obtained by applying the function *Elliptic Curve Digital Signature Algorithm* (ECDSA) over the private key; the public key is subjected to two functions, *Secure Hash Algorithm* (SHA)-256, and *Race Integrity Primitives Evaluation Message Digest* (RIPEMD)-160; the resulting string size is reduced by passing the function *Base58Chech*, resulting in an address like *mgAzKQZZi*47*g*4*UMvmGJCsicbJ*4*P*3*B*8*SHRr* [14].

In the next sections, will be discussed how academic certificates can be used as assets in the blockchain. Through the CertEdu application implemented by UFP, questions such as authenticity will be analyzed, as well as universities and students can have their identities verified. Finally, the results obtained by systematic tests are analyzed.

2.2 Certification

Certification involves three processes: issuing, verification, and sharing [15]. This paper evaluates the application of blockchain in each one of the processes, mainly in the verification and sharing. The digital certificates innovate compared to the paper model but do not guarantee verification and sharing because they depend on central points. As Schär notes, academic certificates are useful, only if they can be verified [6].

The disruptive technology of the blockchain allows creating a structure capable of making the verification process independent [16]. The figure 2 presents a typical scenario, in which the university, the student, and the employer are involved. Note that Jane's certificates can be verified directly on the blockchain, without contacting the university.

Table 1: Comparative between digital certificates vs. blockchain solutions

Properties	Digital Certificate	Blockchain Certificate
Reliability	Relies only in the digital signature	Several cryptography mechanisms are used
Privacy	All information are available in the certificate	Only the hash is public
Autonomy	Depend on central regulatory bodies	Technology eliminates the need for third parties
Data Loss	Depend on backup mechanisms	Standard normal distribution
Proof of existence	Dates relies on the suitability of the subscriber	Timestamp represent the date of the facts

Table 2: Comparison between the tools analyzed

Criteria	BTCert	Hyperledger	EduCTX	Blockcerts
Blockchain agnostic	no	no	no	yes
Self-sovereign identity	yes	yes [2]	yes	yes
Community with active user	no	yes	no	yes
Public network	yes	no	no	yes
Privacy concerns	yes	yes	yes	yes

The public information available on the blockchain says nothing about the student. This occurs thanks to the use of the hash, which records the certificate in a unidirectional way on the network, not allowing to retrieve any personal information from Jane. As a result, Jane has the autonomy to share the digital certificate file, only with whomever she wants, and everyone who receives the file will also be able to check it independently on the blockchain.

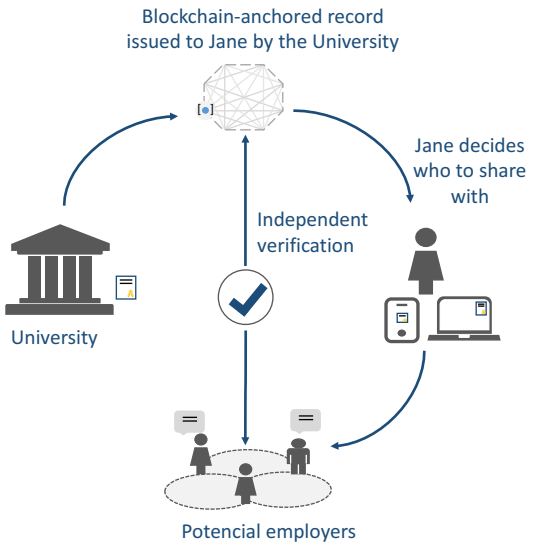


Figure 2: Diploma verification [1]

The table 1 presents a comparison between the properties of the digital certificate solution and the blockchain certificate solution. The blockchain innovation is realized in several aspects. The technology is more secure because unlike the digital certificate, in which all security depends relies only on the digital signature, the use of different cryptographic mechanisms, combined with the adoption of a distributed ledger, offers certificates a higher level of security.

Regarding privacy [17], digital certificates are much more sensitive to data leakage, considering that all personal information is contained in the certificate. This fact can restrict its replication. Analyzing this property in the blockchain solution, the distributed information says nothing about the student, and the certificate can be published without worries.

Digital signatures depend on central regulatory bodies. In some countries, there is not even an authority capable of certifying a signature [15]. In this aspect, blockchain offers complete autonomy, eliminating the need for third parties.

The blockchain, mainly in public architectures and consortium [8], offers a genuine backup mechanism since all information is replicated in pairs. Digital certificates, on the other hand, are easy to destroy electronically and depend on sophisticated backup mechanisms to avoid disappearing [15].

Finally, about the proof of existence, the reliability of the dates generated by digital certificates relies on the reliability of the subscriber. This can be a problem when a university has its private key stolen, and an attacker uses it to sign valid certificates with dates retroactive to the reported date of the theft. In this regard, the blockchain allows identifying the timestamp. As Ronning says, ”every credential issued with a stolen key must fail” [18].

3 CertEdu

3.1 Frameworks analyzed

The CertEdu project begins with the choice of the tool which will be used in the development. In 2018, when the project started, there were not many tools available for this purpose. The use of blockchain in the issue of academic certificate management is re-

[2] Shout combine Hyperledger Fabric + Hyperledger Indy

cent and the first students to receive their diplomas anchored in the blockchain, were from *Massachusetts Institute of Technology* (MIT), in 2017 [19].

We identified four potential platforms. The table 2 shows the comparison between them, and the used evaluated criteria.

Between the evaluated solutions, the Blockcerts tool was chosen to implement the prototype of this paper. The reason was that it better serves the requirements placed, especially about working on any blockchain. In reality, Blockcerts was the only certificate issuing solution that was born with the requirement to work for any blockchain. This requirement greatly increases the complexity of the solution, but it is important because it keeps the application life cycle long. Another point that drew attention is its active community of developers, which makes the project receive constant updates.

Interesting points were also noted on other platforms.

BTCerts, a project inspired by the MIT solution, address to the problem of centralizing the revocation process of Blockcerts. The model proposed by BTCerts solves the issue and can be easily adapted to any type of blockchain, but the costs are concernedly, mainly because it does not explain how the complementary revocation information would be registered, since the OP_RETURN_DATA field has a limited size of 83 bytes [20].

EduCTX, although having a more focused approach to digitizing credits *European Credit Transfer and Accumulation System* (ECTS), the solution brings an approach about how to use the multiple signatures, involving the student and institution to validate the transaction. The solution also raises an interesting question regarding the consensus mechanism used. Depending on the type of application, one consensus mechanism is more suitable than another. Considering education, it makes no sense to mine blocks to record transactions that contain academic certificates or ECTS credits. Universities are reliable nodes and responsible for the data they provide. Operating in a permissive way to create transactions, universities maintain autonomy in the issuing process, but they are unrelated about verification and sharing.

On Hyperledger's evaluation, its flexibility was identified as a strong point, and as a weak point, the lack of models capable of speeding up development. It would be interesting if pre-assembled open-source libraries existed for that purpose. Another issue involving Hyperledger is that to identify the university or the student, it would be necessary to combine the use of two greenhouse technologies (Fabric + Indy), which can make the development of the application even more complex.

Recently, the use of smart contracts has been evaluated for academic record systems[13, 21, 22, 6, 23, 24]. Prototypes are implementing this type of mechanism in the Blockcerts [3], but they are new and still being discussed. This type of approach is interesting but needs to be observed, because not every blockchain implements this type of contracts (example Bitcoin). Besides, the languages implemented by different blockchains can be different, which can make it difficult to create a standard. In the examples shown, the vast majority are based on Ethereum. We believe that because blockchain technology is recent, with many networks likely to disappear and others appearing in the future, the proposed solution should be compatible with any blockchain technology [5] [1]. Also, the blockchain scalability issue [25], may encourage applications to migrate to smaller networks.

3.2 Prototype

UFP has built a prototype and has been testing the application of academic certificates with blockchain technology. The application called CertEdu was built based on the Blockcerts platform and has its architecture designed according to the figure 5. As you can see, CertEdu issues electronic documents on Bitcoin and Ethereum networks. The objective of implementing two networks is precisely to assess the prototype's ability to achieve the desired blockchain compatibility property.

The implementation described in this work shows that even existing different ways to operate the blockchains (permissioned, permissionless, consortium), the implemented solution is easily adapted to operate on any type of network because of these reasons:

- The solution uses its own Merkle Proofs mechanism
- The technique of anchoring the diploma hash on the blockchain, allows verification and overcomes the space limitation
- The solution doesn't need the use of smart contracts

The technical standard of Blockcerts was designed to work with any blockchain, thus preventing the success of the project from being conditioned to the evolution of another product. In 2017 when the project was started, integration was only possible with Bitcoin, but it soon extended to Ethereum. In 2018, Universidad del Rosario, in Colombia, [26] built the integration with Hyperledger.

Blockcerts uses different layers that work together to create the hashes for each batch of certificates, issuing them on the blockchain and subsequently allowing web platforms to print the certificates by using *JavaScript Object Notation* (JSON) objects and verifying them on the blockchain [4].

The figure 3 shows a process of creating a digital certificate on the blockchain by using the components of Blockcerts. As we can see, the asset stored is the hash of the JSON generated file, which in practice means linking the digital file with the blockchain.

The next section describes CertEdu implementation.

3.2.1 Certificate templates

Cert-tools is responsible for creating the certificate templates that will later be signed and anchored on the blockchain. For each model, it is possible to customize information such as the title, logo, description, history. There are also customizable fields, so that peculiar information can be treated. These fields can be created globally (they will appear for all the certificates generated from the model in question) or by the recipient (they will only appear for a specific group of recipients). Next, part of the code of interaction with the component is presented.

[3] https://community.blockcerts.org/t/introducing-smart-contracts-to-blockcerts/2362

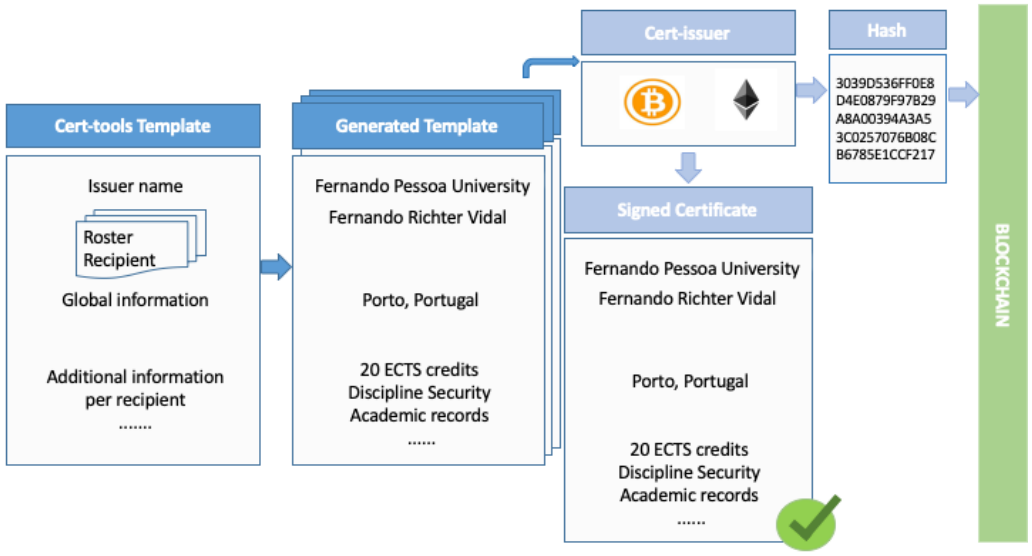


Figure 3: Diploma configuration [1]

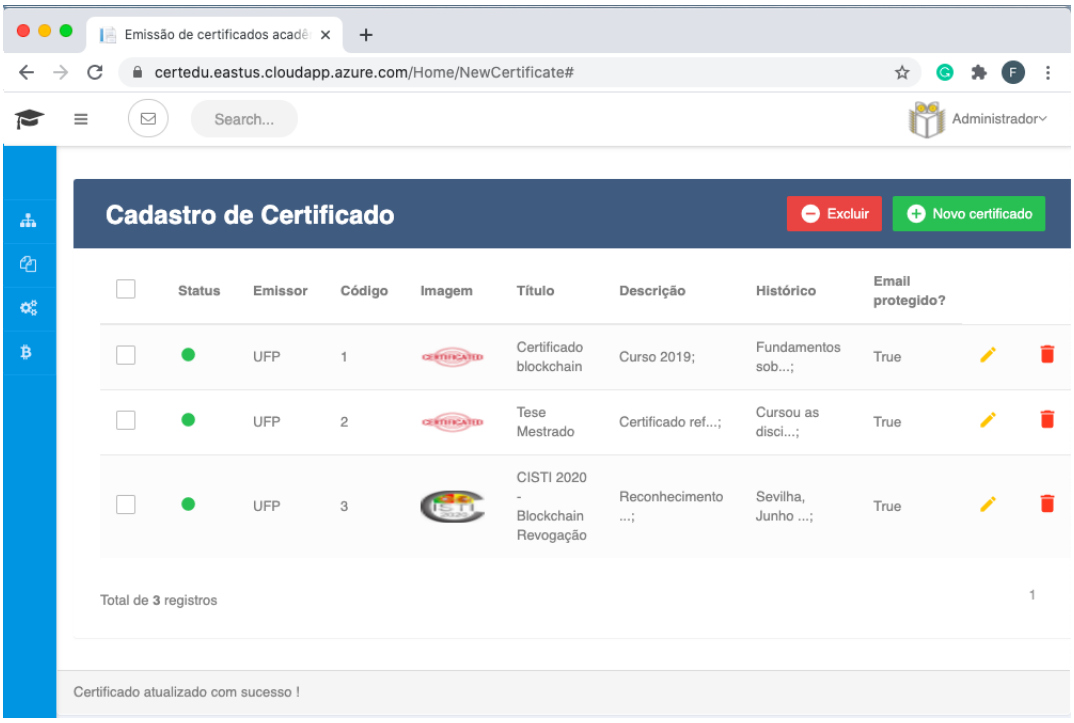


Figure 4: CertEdu's Certificate Template

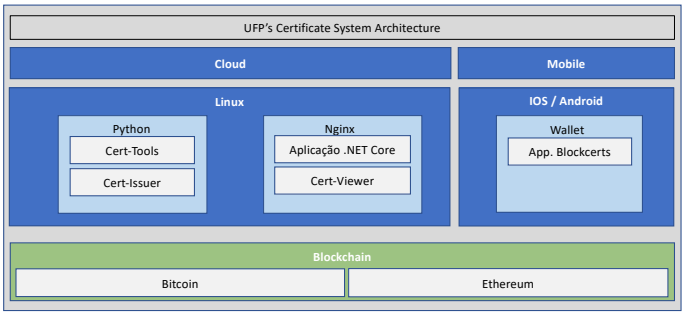


Figure 5: UFP Architecture [1]

```
create-certificate-template -c /ufp/model/ConfigJson_UFP.ini
--data_dir /ufp/ --template_dir /ufp/model
--template_file_name /ufp/model/UFP_3.json
--issuer_logo_file /ufp/image/UFP.png
--cert_image_file /ufp/image/UFP_3.png
--issuer_url https://certedu.eastus.cloudapp.azure.com/
--issuer_email ufp@ufp.edu.pt
--issuer_name 'Universidade Fernando Pessoa'
--issuer_id https://certedu.eastus.cloudapp.azure.com/UFP_1.json
--revocation_list
https://certedu.eastus.cloudapp.azure.com/rev_UFP_1.json
--issuer_certs_url https://certedu.eastus.cloudapp.azure.com
--certificate_description 'Reconhecimento de apresentação no congresso.'
--certificate_title 'CISTI 2020  - Blockchain Revogação'
--criteria_narrative 'Sevilha, Junho de 2020'
--issuer_public_key ecdsa-koblitz-pubkey:
```

```
mgAzKQZZi47g4UMvmGJCsicbJ4P3B8SHRr
--badge_id 370820c5-59ff-4a0b-8c9c-55faa0b9f431
--unsigned_certificates_dir /ufp/unsigned
```

Cert-tools receives as an input the information that will form the certificate and as output, returns a file JSON, ready to be signed by the next component. The file in question is generated by a Blockcerts module called cert-schema [4], which is based on the *Verifiable Credentials* (VC) [5]. The data types mapped by the standard follow the norm *Internationalized Resource Identifiers* (IRIs), the same used by XMLSchema [6]. All this concern in following these standards is to offer entities and interested parties a standardized format for certifications.

The figure 4 shows the implementation of cert-tools in CertEdu. The model stored in the database contains the necessary parameters for the issuance of a certificate. In another system interface, all it takes is to link the model code with a list of students in order to publish the certificates. Note that there is a status indicator, which checks when saving the record if all settings have been applied successfully. For example, there is a parameter that allows you to indicate the public address of the certificate. When it is activated, the application tries to publish on the indicated website with the parameterized access credentials. If it fails, the model remains pending and cannot be used until the error is fixed.

3.2.2 *Publishing area*

Cert-issuer is the component responsible for generating the transaction on the blockchain. The input receives the diploma file generated by cert-tools and returns as output the diploma hash published on the blockchain. Its role, in addition to the signing, is to allow blockchain compatibility, by providing a structure that allows connectors from other networks to be implemented. In the component area called blockchain_handlers, all it takes it to create three functions to connect a new network: connection (connectors.py), transaction (transaction_handlers.py), and subscriber (signer.py). On top of that, it is necessary to change the block below the main function of the component.

```
issue_certificate.py

def main(app_config):
    chain = app_config.chain
    if chain == Chain.ethereum:
        from cert_issuer.blockchain_handlers import ethereum
        certificate_batch_handler, transaction_handler,
        connector = ethereum.instantiate_blockchain_handlers(app_config)
    elif chain == Chain.bitcoin:
        from cert_issuer.blockchain_handlers import bitcoin
        certificate_batch_handler, transaction_handler,
        connector = bitcoin.instantiate_blockchain_handlers(app_config)
    else:
      new blockchain
      ...

    return issue(app_config, certificate_batch_handler,
    transaction_handler)
```

The standard is maintained by the open-source community, to support networks the Bitcoin and Ethererum. The integration to other networks is emerging as initiatives do, such as the University of Rosario, in Colombia [26], which built the connector for Hyperledger, and is testing it experimentally.

Another notable point of the figure 3 is the possibility of signing a group of models at once. Technically, cert-tools generates several certificate files and calculates the group's Merkle root, recording this value on the blockchain.

The figure 6 shows the structure of a certificate file that makes up a batch. All files in the batch have the same value as the merkleRoot field, and additionally store, in addition to the hash itself, the hashes of the nodes needed to verify the root of Merkle (proof 0, 1 and 2 of 6). In practice, when the verifier receives a file, it calculates the hash and checks on the blockchain whether this file belongs to the generated batch. With that, you only need to spend a single time, to be possible to check *n* certificates. Also, this check is very useful for revocation, because canceling a single blockchain registration automatically cancels the entire batch.



Figure 6: Certificate batch structure

Blockcerts implements an independent version of the Merkle root, the 2017 Merkle Proof Signature Suite [7]. This means that the field calculated in the JSON file, always follows the same pattern, contributing at this point for the tool to work on several networks. Each blockchain can implement the field differently, as is in the case Ethereum, which unlike Bitcoin, uses Merkle Patricia [27].

3.2.3 *Embedded authenticity checker*

The verifier has two roles: to inform the authenticity of a certificate and to represent it graphically to the user. The first versions of this component in Blockcerts were called cert-viewer, but later changed to blockcerts-verifier [8].

The technology is based on JavaScript, which makes it easier for applications, such as CertEdu (Figure 8), to embark on a universal certificate verifier within its structure.

[4] https://github.com/blockchain-certificates/cert-schema
[5] https://www.w3.org/TR/vc-data-model/
[6] http://www.w3.org/TR/xmlschema-2/
[7] https://w3c-dvcg.github.io/lds-merkleproof2017/
[8] https://github.com/blockchain-certificates/blockcerts-verifier

How the issuance process is implemented, directly impacts on the complexity of implementing this component. Notably, Blockcerts seeks to use standardized components of the blockchain, such as hashes, transaction recording, Merkle root. With this, the inclusion of new blockchain in the operation of the blockcerts-verifier becomes a simpler process. Considering a different scenario, in which the mechanisms use smart contracts, depending on the way it is implemented, this process of incorporating new networks can become complex.

There is still no defined standard on the display format of the digital certificate (figure 7 and 8). Although this is not a technical security problem, it can cause some distrust for an appraiser who receives the same diploma issued by the same broadcaster but in different formats. Fortunately, the verification function can guarantee the data veracity to the interested party, and can even validate the file in different universal verifiers, such as the one offered by Blockcerts [9]

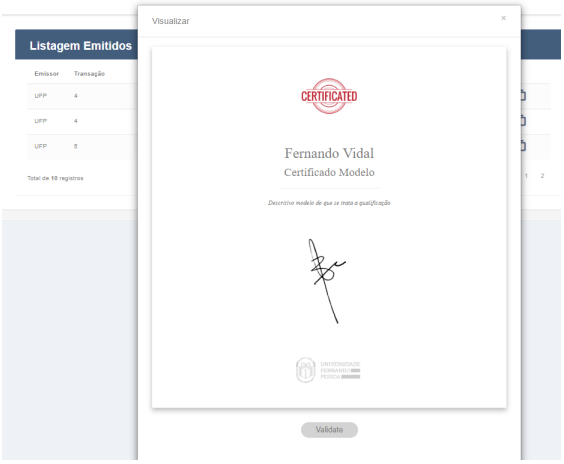


Figure 7: Model 1 for viewing a digital certificate issued by Blockcerts

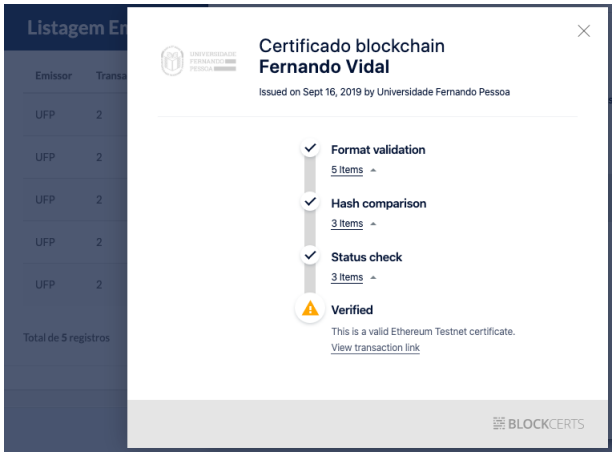


Figure 8: Model 2 for viewing a digital certificate issued by Blockcerts

[9]https://www.Blockcerts.org/

3.2.4 *Mobile application*

Cert-wallet is a Blockcerts application, also available in an open-source format, which aims to offer users autonomy over their records. Once the user downloads the application on their device, their identity is linked to a phrase created automatically by Blockcerts, which becomes a kind of private key.

The student is sovereign to register an issuer in his account and redeem the certificates that were issued by him. Autonomy is also evidenced by the fact that the student has the freedom to send his diploma to whomever he wishes, directly, without the intervention of any third parties. The difference between the digital process and the paper is that the registration received by the third party can be verified directly on a public blockchain network (in the case of this paper, Bitcoin or Ethereum), without any consultation with the issuing university.

The application has three main functions:

1. Automatically generate and send the student's public address to the sending systems
2. Store the digital certificates
3. Share the digital certificates

The figure 9 shows the operation of the first function of the cert-wallet. Note that the user informs an address of a sender profile file, and also a unique and disposable code, in English, called the nonce. This code serves as a kind of student credential to access the available service. Right after this procedure, cert-wallet accesses another address that is embedded in the profile file, and sends it to that address, the public key generated by the application, and also the code entered.

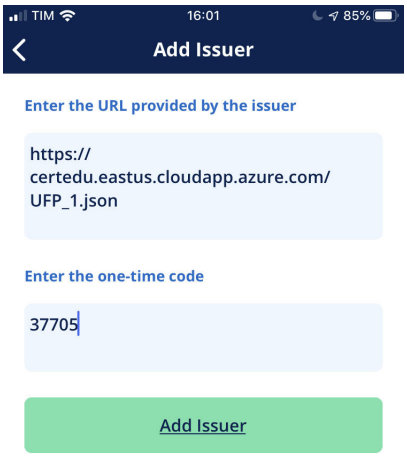


Figure 9: Cert-wallet application interface responsible for generating the student's public address

On the other side of the operation, in the second address mentioned software that is prepared to receive the public key and compare the received code is operating, validating, or not the message. If so, the cert-wallet receives a signal to register the issuer in its

database, and the student's public address is stored in the issuing application so that it can be used in future issues.

The second function of the application refers to the storage of certificates, which, as noted in the figure 10, can be done in two ways: informing an address *Uniform Resource Locator* (URL) in which the certificate is published, for example, `https://certedu.eastus.cloudapp.azure.com/certdata/2137705.json`, or by directly importing a JSON file.

Immediately after importing a credential, the application starts a process of validating and verifying the authenticity of the file, informing at the end of the process whether the certificate is valid or not.

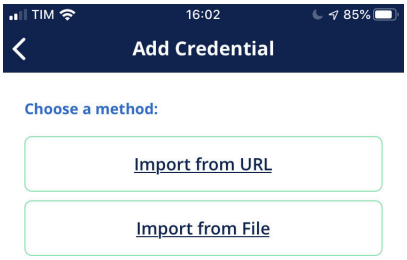


Figure 10: Ways to store a certificate in the cert-wallet

The third and final function of the cert-wallet, refers to the possibility given to the users to share their certificates. When selecting a certificate, it is possible to physically send the file to another user or simply provide the address that references it.

Table 3: Processes and challenges related to the blockchain solution

Process(s)	Challenge	Satisfies
Issuing	A feasible financial solution to large issuing	Yes
Issuing	Identify of universities	Yes
Issuing	Allow revocation	Partial
Verifying	Verifying em local database (decentralizing)	Partial
Verifying	Verifying in universal verifier	Yes
Verifying	High availability	Yes
Sharing	Identify of students	Yes
Sharing	Sovereign to share the certificates	Yes
Sharing	Graphical display of recorded information	Yes
All	Operating compatibility on any type of blockchain	Yes

[10] `https://faucet.ropsten.be/`
[11] `https://faucets.blockxlabs.com/ethereum`
[12] `https://coinfaucet.eu/en/btc-testnet/`
[13] `https://bitcoinfaucet.uo1.net/send.php`

3.2.5 Goals

The purpose of the application is to simulate an environment close to a real situation, for this reason, it was decided to configure the TestNet (Bitcoin) and Rospten (Ethereum) networks. Faucets [10], [11], [12], [13] were used to transfer credits to the application's accounts, and thus emulate a scenario close to reality, in which there are limited financial resources to generate certificates.

The table 3, presents an analysis by the process, about the main challenges to be addressed. In this way, the work assesses such issues through CertEdu, and highlights which issues are addressed and which still need to be addressed. Concerning the points not yet solved, some possible solutions to these problems are presented.

Below, the topics partially covered are discussed.

3.2.6 Disconnected checking and revocation

One of the innovations proposed by blockchain certificate solutions is disconnected verification. For example, it would be possible to verify the authenticity of the diplomas, using an off-line local copy. All certificates that are already there are perfectly verifiable.

However, the prototype found that this feature was not met, due to the way Blockcerts implements the verification process. There is a dependency on two external files (hosted on the issuer's server), one for university identification, and another for checking certificate revocation.

Regarding the first dependency, Learning Machine and NextID in late 2019, published an article [18] with a proposal to replace the issuer profile. This functionality must be present in versions 3.0 of blockcerts.

Regarding revocation, there is still no definitive solution to the problem. The paper [5], presents different approaches, such as smart contracts [28], control data [29] or even the combined use of *Interplanetary File System* (IPFS) and blockchain [30] , [31], [32].

3.3 Systemic tests

The following tests simulate some of the situations that can occur when issuing a blockchain certificate. Firstly, attempts are made to tamper the certificate, either through data editing or through the usurpation of property. Right after this test, it is evaluated how the validation system behaves in a disconnected way.

3.3.1 Tamper

In this simulation, a diploma issued by CertEdu is changed. Using a simple text editor, the original information of the diploma is modified, including a letter R in the student's name. It is notable that the digital diploma file is not encrypted, and can be easily edited by users who have a minimum of knowledge in text editing tools.

The result of the modified document validation operation can be seen on the figure 11, and the rules processed by the verifier can be observed through the table 4.

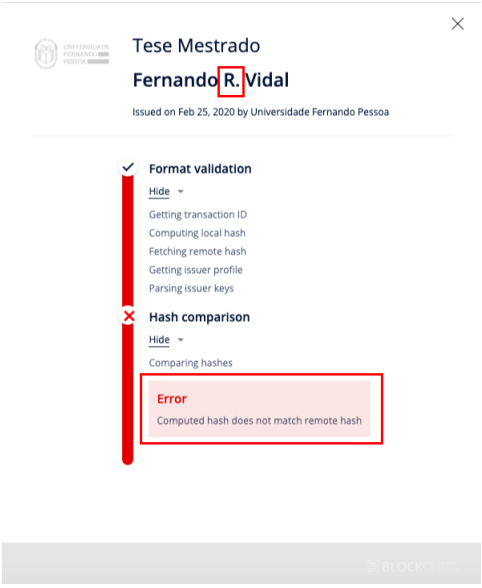


Figure 11: Tampered certificate validation

Table 4: Result of validations performed by the universal verifier

Nº	Rule	Satisfies ?
1	The transaction number contained in the certificate (sourceId) exists in the referenced blockchain ?	Yes
2	The computed hash over the information, match with the targetHash field?	Yes
3	The *targetHash* match with the blockchain stored content ?	No
4	The issuer's public address (server) is available ?	Yes
5	The embedded issuer's public key in the certificate, match with the key hosted in the issuer's server	Yes

We realized in this experience that the verifier can identify any adulterations of the file's content. Analyzing the verifier's rules, even if an attacker tampered a certificate and recalculated the hash data, the tamper would be detected, because the content of the *targetHash* field would not match the value stored in the blockchain.

3.3.2 *Unauthorized appropriation*

This other fraud simulation discusses the possibility for an attacker to change the issuers and /or recipients of a genuine certificate. For example, it would be interesting for criminals to replicate real university certificates with lesser visibility, changing them as if they were from prestigious universities. Likewise, a profitable fraud would be to offer genuine degrees to malicious recipients, who only intend to accumulate "achievements".

Blockcerts embed the student's public key in the certificate data, in the *recipientProfile* section, in the *publicKey* field. Likewise, in the *verification* section, the sender's information is embedded. This means that such fields are part of the hash calculation, which guarantees protection against tampering.

However, an attacker could alter the certificate data, informing other public keys of issuers and recipients, and recalculate the certificate hash. Additionally, to prevent rule 3 (table 4) from identifying the fraud, the attacker could change the *sourceId* field for a transaction that he created, which contains exactly the value of the hash of the defrauded certificate.

In this case, rule 3 will pass as true, because both the local calculation of the hash and the comparison with the information stored in the blockchain will be considered valid. However, Blockcerts can identify this fraud, because in addition to validating the content of the transaction, it checks whether the address of the transaction is owned by the issuing university. This is possible because the system compares the public address of the transaction with the public address hosted on the sender's server.

The figure 12 presents excerpts from the university's profile file, highlighting the public key information. The adulterated certificate, and the forged transaction, point to the *mgAzKQZZi47g4UMvmGJCsicbJ4P3B8SHRr* hash. However, the university's real public key points to the value *JKmg4UDWqBcnhklMEsDgdDHsRrF*.

As shown in figure 13, the validator detects the inconsistency of this information, preventing the fraud from being carried out.

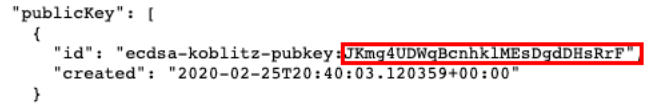


Figure 12: Issuer identification file

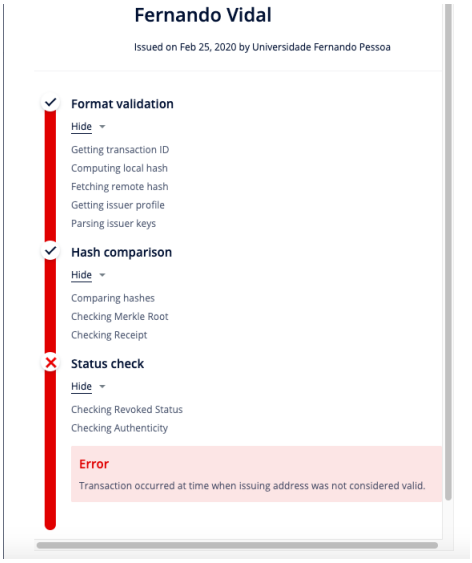


Figure 13: Certificate check with the issuer tampered with

3.3.3 *Stolen key and revocation*

This simulation evaluates Blockcerts' ability to identify frauds that may occur in situations where the university has had its private key stolen.

The digital diplomas issued by non-blockchain solutions, such as those that sign using a *Rivest-Shamir-Adleman* (RSA) key, do not guarantee in their signature proof that the creation date associated with their signature is consistent with reality. In most cases, the

issuer who signs with the keys to their property, must sign with the correct date and time [18].

In situations where the private key is stolen, the attacker may want to issue retroactive diplomas, thus managing to guarantee illegal conquests for any recipient.

When the university realizes that its key was stolen, it revokes the validity from the date on which the theft occurred. However, in non-blockchain solutions, since certificate dates cannot be trusted, it is very difficult to determine which diplomas are still valid.

Therefore, by using a blockchain's reliable timestamp, you can calculate the actual date of issuance of a diploma, and thus, determine the revocation/expiration. With that being said, it becomes more reliable to separate valid and invalid credentials. For example, all credentials that are referenced on a blockchain, before the reported date of the theft, remain securely valid and are not affected by any revocation process.

The figures 14 and 15, demonstrate how the date of the issuance of a certificate can be easily compared with its registration on the blockchain. By consulting the date of the corresponding transaction (indicated in the *sourceId* field), you can easily perform the comparison.



Figure 14: Certificate issue date

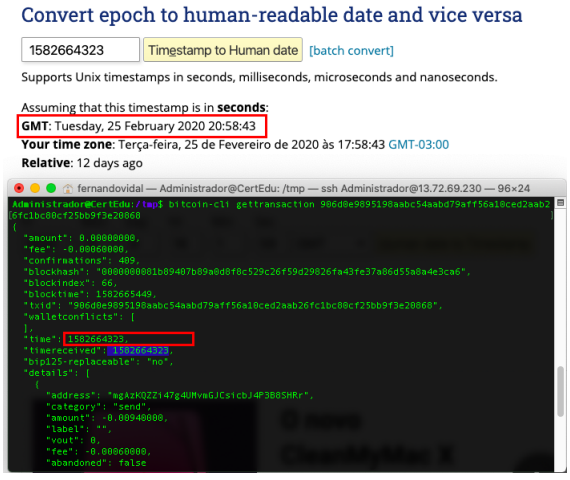


Figure 15: Registration date of certificate issuance on the blockchain

In this matter, it is concluded, that the blockchain solution offers a differential about the solutions without blockchain, since it guarantees students, that their old diplomas will not be affected, even if unexpected situations like a private key theft happens.

3.3.4 Validation without server access

This simulation is divided into two parts. First, the possibility of the issuer's server failing is evaluated and then the functionality of continuing to validate in a non-synchronized way, records on a local blockchain are tested.

In this first test, the sender's server is inaccessible. This simulation intends to evaluate the system's behavior when faced with a failure of the university server. For this, the server on which the profile and revocation files are hosted was turned off and was performed a check for a genuine certificate.

It can be seen through figure 16, that the verifier was unable to tell whether the certificate was valid or not.

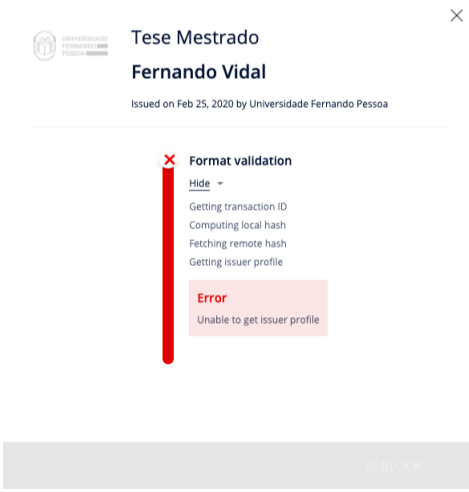


Figure 16: Local check, error when connecting transmitter

The test concludes that the solution is not yet capable of offering independence from the issuer. Besides, if an attacker manages to invade the server, even if it is temporarily, it would be possible to carry out the attacks described in the 3.3.1 section, due to the possibility of changing the information stored in the sender's profile.

In the second test considers a scenario in which the local blockchain is not synchronized. One of the benefits of using the blockchain would be the possibility to continue to validate certificates, on a local blockchain, even if it no longer exists.

Analyzing the libraries of the implemented prototype, it is noticed that the verifier uses interfaces *Application Programming Interface* (API), made available by block query tools, to verify them. This means that for a local validation it would be necessary to modify the verification address for the local base. As this is a relatively easy operation, it can be said that the tool meets this requirement.

Another situation that can be simulated is the verifier validating the issuance of a certificate that has not yet been authorized by the consensus mechanisms. As mentioned in the section, the validation of the blocks is one of the pillars that brings all the security of the technology. Validating a diploma from a registration that has not yet been confirmed could be considered a very serious failure.

Figure 17 shows the result of this operation, and as the result is shown, it can be said that the system also passed this security test.



Figure 17: Block not yet confirmed

4 Conclusion

Diploma fraud is far from being over. The combat mechanisms proved to be extremely inefficient, mainly in the paper models. Digital degrees have evolved with digitalization, but they do not offer privacy to the students, so they can have confidence in freely distributing their certificates.

Blockchain's disruptive technology offers a breakthrough in distribution and ensures protection from tampering. It was also shown that the blockchain offers the best resources to act in case of loss or theft of the university's private key, protecting the entity against undue retroactive emissions. Besides, the level of privacy offered by the technology, by recording only the hash certificate on the blockchain, makes the solution less prone to data leakage than the digital certificate solutions without blockchain.

The materialization of the objective of using the blockchain for the management of diplomas was achieved during the construction of the CertEdu prototype, in which it was possible to operate issuances, revocations, shares, and verifications of academic certificates.

However, tests also pointed out that disconnected operation is still an issue that needs to be worked on. Although some blockchains already offer resources, such as the smart contract, which would allow to easily resolve the centralization points placed, the premise of the solution operating on any type of blockchain has not been met, so those issues that still prevent decentralization are still raised by this work is pending.

There is also concern about the unpredictability of the issuance costs. Thinking of public networks like Bitcoin and Ethereum, you cannot predict the rate of transactions in the long run. This issue can inhibit the adhesion of universities.

Finally, it is concluded that the application of blockchain in the management of academic certificates is perfectly possible and that the technology already offers benefits in terms of privacy, distribution and revocation, compared to digital solutions. However, to be able to take advantage of the full potential of the technology, the centralization points addressed by this work, such as the validation of the issuer's profile and revocation, need to be migrated to features that operate within the blockchain itself.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors acknowledge the support of DuoTi Solucoes.

References

[1] F. Vidal, F. Gouveia, C. Soares, "Analysis of Blockchain Technology for Higher Education," in 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 28–33, IEEE, 2019, doi:10.1109/CyberC.2019.00015.

[2] N. Smolenski, D. Hughes, Academic Credentials in an Era of Digital Decentralization, Learning Machine Research, Anaheim, CA, 2016.

[3] IBGE, "Rendimento médio real mensal do trabalho principal, por níveis de instrução," 2018.

[4] M. Oliver, J. Moreno, G. Prieto, D. Ben\'\itez, "Using blockchain as a tool for tracking and verification of official degrees: business model," in Using blockchain as a tool for tracking and verification of official degrees: business model, 29th European Regional Conference of the International Telecommunications Society (ITS): "Towards a digital future: Turning technology into markets?", Trento, Italy, 1st - 4th August 2018, International Telecommunications Society (ITS), Trento, 2018.

[5] F. R. Vidal, F. Gouveia, C. Soares, "Revocation Mechanisms for Academic Certificates Stored on a Blockchain," in 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 1–6, IEEE, Sevilha, 2020, doi:10.23919/CISTI49556.2020.9141088.

[6] F. Schär, M. Fabian, "Blockchain diplomas: Using smart contracts to secure academic credentials," Beiträge zur Hochschulforschung, **48**, 2019.

[7] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Www.Bitcoin.Org, 2008, doi:10.1007/s10838-008-9062-0.

[8] M. Turkanović, M. Hölbl, K. Košič, M. Heričko, A. Kamišalić, "EduCTX: A blockchain-based higher education credit platform," IEEE Access, 2018, doi:10.1109/ACCESS.2018.2789929.

[9] M. Saad, J. Spaulding, L. Njilla, C. A. Kamhoua, S. Shetty, D. Nyang, A. Mohaisen, "Exploring the Attack Surface of Blockchain: {A} Systematic Overview," CoRR, **abs/1904.0**, 2019.

[10] J. a. Kroll, I. C. Davey, E. W. Felten, "The Economics of Bitcoin Mining, or Bitcoin in the Presence of Adversaries," The Twelfth Workshop on the Economics of Information Security (WEIS 2013), 2013, doi:June11-12,2013.

[11] A. M. Antonopoulos, Mastering Bitcoin: Unlocking Digital Crypto-Currencies, O'Reilly Media, Inc., 1st edition, 2014.

[12] R. C. Merkle, "A digital signature based on a conventional encryption function," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1988, doi:10.1007/3-540-48184-2_32.

[13] L. M. Palma, M. A. G. Vigil, F. L. Pereira, J. E. Martina, "Blockchain and smart contracts for higher education registry in Brazil," International Journal of Network Management, **29**(3), e2061, 2019, doi:10.1002/nem.2061.

[14] D. Yaga, P. Mell, N. Roby, K. Scarfone, "Blockchain technology overview," Technical report, National Institute of Standards and Technology, Gaithersburg, MD, 2018, doi:10.6028/NIST.IR.8202.

[15] A. Grech, A. F. Camilleri, Blockchain in Education, Publications Office of the European Union, 2017, doi:10.2760/60649.

[16] H. Haugsbakken, I. Langseth, "The Blockchain Challenge for Higher Education Institutions," European Journal of Education, **2**(3), 2019.

[17] N. Bore, S. Karumba, J. Mutahi, S. S. Darnell, C. Wayua, K. Weldemariam, "Towards Blockchain-enabled school information hub," in ACM International Conference Proceeding Series, 2017, doi:10.1145/3136560.3136584.

[18] A. Ronning, W. W. Chung, "Blockcerts V3 Proposal," 2019.

[19] M. L. MIT Media Lab, "Blockcerts-An Open Infrastructure for Academic Credentials on the Blockchain," 2016.

[20] M. Bartoletti, L. Pompianu, "An analysis of Bitcoin OP_RETURN metadata," CoRR, **abs/1702.0**, 2017.

[21] J. C. Cheng, N. Y. Lee, C. Chi, Y. H. Chen, "Blockchain and smart contract for digital certificate," in Proceedings of 4th IEEE International Conference on Applied System Innovation 2018, ICASI 2018, 2018, doi: 10.1109/ICASI.2018.8394455.

[22] T. Kanan, A. T. Obaidat, M. Al-Lahham, "SmartCert BlockChain Imperative for Educational Certificates," in 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 629–633, IEEE, 2019, doi:10.1109/JEEIT.2019.8717505.

[23] K. Mori, H. Miwa, "Digital university admission application system with study documents using smart contracts on blockchain," in Advances in Intelligent Systems and Computing, 2020, doi:10.1007/978-3-030-29035-1_17.

[24] K. Patel, M. L. Das, "Transcript Management Using Blockchain Enabled Smart Contracts," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, doi:10.1007/978-3-030-36987-3_26.

[25] M. Jirgensons, J. Kapenieks, "Blockchain and the Future of Digital Learning Credential Assessment and Management," Journal of Teacher Education for Sustainability, 2018, doi:10.2478/jtes-2018-0009.

[26] C. Iragorri, "Academic certificates on Hyperledger," 2018.

[27] G. Wood, "Ethereum: A Secure Decentralised Generalised Transaction Ledger. EIP-150 REVISION," 2017, 2017.

[28] J. Santos, K. H. Duffy, "A Decentralized Approach to Blockcerts Credential Revocation," 2019.

[29] L. Rujia, D. Galind, "BTCert," 2017.

[30] R. Kumar, R. Tripathi, "Implementation of Distributed File Storage and Access Framework using IPFS and Blockchain," in 2019 Fifth International Conference on Image Information Processing (ICIIP), 246–251, IEEE, 2019, doi:10.1109/ICIIP47207.2019.8985677.

[31] S. Wang, Y. Zhang, Y. Zhang, "A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems," IEEE Access, 2018, doi:10.1109/ACCESS.2018.2851611.

[32] A. Rajalakshmi, K. Lakshmy, M. Sindhu, P. Amritha, "A Blockchain and IPFS based framework for secure Research record keeping," International Journal of Pure and Applied Mathematics, **119**(15), 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 894-903 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Method of Analysis and Classification of Acoustic Emission Signals to Identify Pre-Seismic Anomalies

Marapulets Yury, Senkevich Yury, Lukovenkova Olga, Solodchuk Alexandra*

Laboratory of Acoustic Research, Institute of Cosmophysical Research and Radio Wave Propagation, Far Eastern Branch of the Russian Academy of Sciences (IKIR FEB RAS), 684034, Russia

ARTICLE INFO

Article history:
Received: 19 August, 2020
Accepted: 25 November, 2020
Online: 08 December, 2020

Keywords:
Acoustic Emission
Symbol Description
Signal Alphabet
Pre-Seismic Anomaly

ABSTRACT

A new method of analysis and classification of rock acoustic emission signals is proposed. It is based on symbol description of signals and involves the following processing. First, signal segments containing pulses are detected. Second, noise of the detected pulses is reduced by the wavelet filtration method. Fourth-order symlets and adaptive threshold scheme based on empirical Bayes method are used. Application of wavelet filtration allows us to increase the signal-to-noise ratio by 9 dB, on the condition that initial SNR is 8.9 dB. Each pulse is described by a descriptive matrix which represents a square binary matrix reflecting pulse amplitude-phase structure and characterizing the positions of pulse extrema relatively each other. Based on the similarity degree of descriptive matrices, classes are formed. The proposed algorithm classifies correctly 78% of pulses. Each class is considered as a symbol and a set of obtained symbols forms an alphabet. Then, changes of alphabet parameters on subsequent fragments of the signal are studied. The developed method was realized in the form of an application program with the help of which acoustic emission signals were analyzed. The signals were recorded at "Karymshina" site (IKIR FEB RAS), located in a seismically active region, Kamchatka peninsula, in 2016–2019. It was discovered during the analysis that in case of an acoustic emission anomaly, alphabet content redistributed in favor of the symbols described by larger-size descriptive matrices (up to 350 extrema). At the same time, the number of such symbols increased. Analysis of alphabet averaged dynamics discovered a well exhibiting tendency of alphabet cardinality decrease relatively an average 10–18 days before earthquakes and recovery of the value over the period from 2 to 8 days after them. The obtained estimates of the pre-seismic anomaly occurrence periods are consistent with the results of early studies. The proposed method updates the existing methodology of short-term earthquake precursor formation. Thus, applying the developed method for analysis, it is possible to identify pre-seismic anomalies of acoustic emission that is topical in the creation of strong earthquake preparation warning system operating in automatic mode.

1. Introduction

Acoustic emission (AE) in solid bodies is elastic oscillations occurring in the result of dislocation changes in a medium. In such a case, the characteristics of generated pulse emission are directly associated with plastic process features. That determines the interest in the emission investigation with the aim to develop methods for medium acoustic diagnostics.

AE is applied for different problems of diagnosis in seismology, geophysics as well as in non-destructive testing of objects and materials in industry. At the beginning of the XXI century it was shown that acoustic emission anomalies are observed before strong earthquakes and can be consider as their operative precursors [1-3]. In further investigations [4, 5] the nature of generation of pre-seismic acoustic anomalies and the features of their manifestations were considered. In the course of the research carried out at IKIR FEB RAS, it was shown that these anomalies are generated by the change of geomechanics stress field in earthquake preparation zone and appear as the result of the transformation, stress – near-surface rock deformation [6, 7]. The peculiarity of anomaly manifestation is geoacoustic emission intensity increase in the frequency range from hundreds of hertz to

*Corresponding Author: Solodchuk Alexandra, Email: aleksandra@ikir.ru

the first tens of kilohertz. It usually lasts for to several hours and occurs 1–3 days before an earthquake.

Statistics of pre-seismic acoustic anomaly observations in a seismically active region of Kamchatka peninsula, which is a part of the Pacific Ring of Fire, shows high efficiency and prospectivity of AE method to forecast catastrophic earthquakes [6]. Detailed analysis of AE signals shows that they a represented by a sequence of relaxation oscillation pulses with steep edge and smooth drop of different frequency filling [6]. In these circumstances, both separate pulses and their groups with significant overlapping are observed. In case of intensive emission, pulses merge into a continuous signal.

During a background (seismically calm) period, acoustic pulses with the repetition frequency within 0.1–0.5 pulses per a second are usually observed. During the increase of stress and of rock deformation velocity before an earthquake, both pulse amplitude and their number per a time unit grow. Pulse repetition frequency reaches the value of tens and even hundreds per a second [6].

At the present time one of the main and still unsolved problems is automation of search and operative detection of AE pre-seismic anomalies. Application of popular (widely spread in the world) methods for data analysis becomes complicated by several reasons, they are:

- wide dynamic range of recorded data is associated with great volumes of recorded information requiring special methods for storage, transfer, processing and analysis;
- strong irregularity of natural mediums and bad propagation of elastic oscillation causes signal significant distortion and attenuation which limits the capabilities of investigation methods;
- strong noisiness both by natural and industrial noises makes the analysis of inner (morphological) structure of AE pulses more complicated;
- wide variety of signal waveforms complicates the problem of classification and requires analysis methods adapting for a given signal.

The paper proposes a new approach to analyze and classify rock acoustic emission signals in order to identify AE pre-seismic anomalies. The approach is based on the description of acoustic pulses from which a signal alphabet is formed. Then the changes of alphabet parameters of signal subsequent fragments are studied.

2. AE signal model

In a general case, acoustic emission signals can be represented by a classical adaptive model, i.e. as a sum of noise $\varepsilon(t)$ and some function $s(t)$, the analytical expression of which is unknown,

$$x(t) = s(t) + \varepsilon(t), \tag{1}$$

where $\varepsilon(t)$ describes nonstationary background noise containing noises of natural and artificial origin and $s(t)$ is a "useful" AE signal. It is a flux of pulses of different waveform, amplitude and frequency. Moreover, we accept the following constraint $||s(t)|| > ||\varepsilon(t)||$. Figure 1 illustrates an example of AE signal.

Within the framework of the proposed approach, pulses of a "useful" signal $s(t)$ can be described by the means of their singular points. As it is known from the functional analysis theory, when there are no fractures of the first and of the second kind, these points coincide with local extrema and local inflections. It is assumed that signal function intermediate values lying between the singular points are less informative and may be neglected in the course of the analysis.

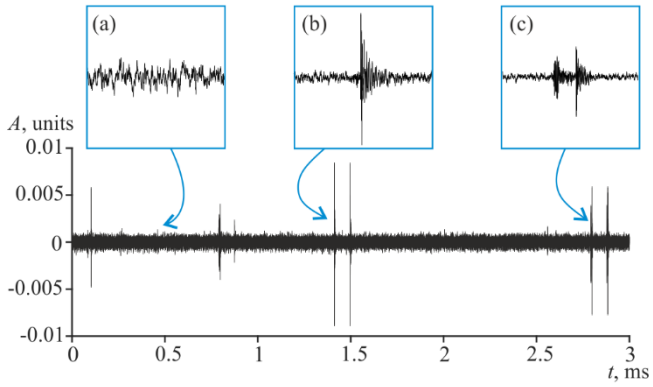


Figure 1: Acoustic emission signal: (a) – noise, (b), (c) – pulses

In order to understand the essence of further transformations, we consider an acoustic emission pulse signal (its graph is illustrated in Figure 2). We fix the values pulse local extreme amplitudes $\{a_i\} = a_0, a_1, \ldots, a_N$ and the time intervals between neighboring extrema $\{\tau_i\} = \tau_0, \tau_1, \ldots, \tau_{N-1}$. On the basis of these values we create a matrix of mutual relation of the fixed values (2). Matrices of **A** and **T** relation reflect pulse amplitude-phase structure. They characterize the positions of local extrema relatively each other and can be considered as a pulse model.

$$\mathbf{A} = \begin{pmatrix} a_0/a_0 & a_0/a_1 & \cdots & a_0/a_N \\ a_1/a_0 & a_1/a_1 & \cdots & a_1/a_N \\ \vdots & \vdots & \ddots & \vdots \\ a_N/a_0 & a_N/a_1 & \cdots & a_N/a_N \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} \tau_0/\tau_0 & \tau_0/\tau_1 & \cdots & \tau_0/\tau_{N-1} \\ \tau_1/\tau_0 & \tau_1/\tau_1 & \cdots & \tau_1/\tau_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{N-1}/\tau_0 & \tau_{N-1}/\tau_1 & \cdots & \tau_{N-1}/\tau_{N-1} \end{pmatrix} \tag{2}$$

We should note that if one knows any one value a_i, any one value τ_i and the relational matrix (2), it is possible to recover the initial waveform of a pulse more or less without loss.

To compose signal symbol description, the obtained pulse diversity is classified by comparing the relational matrices (2), i.e. based on the pulse waveform similarity criterion. However, such selection should take into account two key points: firstly, amplitude attenuation and phase distortions of acoustic pulses occurring during the propagation from a source to a receiver point; secondly, the distortions caused by the interference of additive noise on pulse waveforms at a receiver point. These problems inevitably result in significant increase of the number of pulse classes (redundant classification). The following transformations and computing operations allow us to solve, to a large extent, the problems arising during the classification.

When analyzing a pulse waveform, signal function numerical values at singular points often do not play the major role. In a

greater degree, to detect a pulse waveform, mutual position of these points is important [8]. In the framework of the applied approach, search for the differences in the observed pulse waveforms is of the main interest. Thus, relational matrices can be reduced to matrices of rank relation. In order to do that,it would be enough to transform matrices (2) into binary ones according to the following rule:

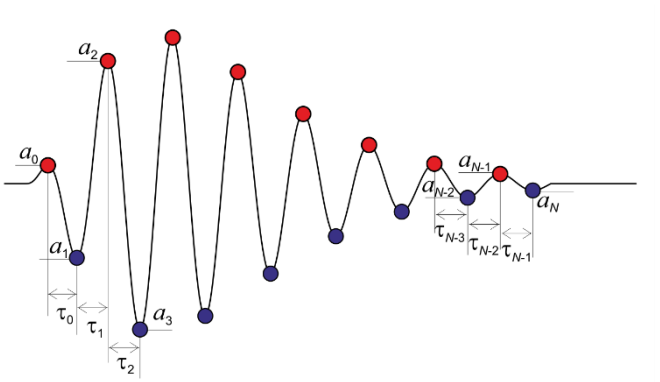


Figure 2: Local extrema of AE pulse

$$r_{i,j} = \begin{cases} 1, & a_i > a_j \\ 0, & a_i \le a_j \end{cases}, \qquad \omega_{i,j} = \begin{cases} 1, & \tau_i > \tau_j \\ 0, & \tau_i \le \tau_j \end{cases}, \tag{3}$$

where $r_{i,j}$ is the result of comparison of the i-th and j-th vales of extreme amplitudes; $\omega_{i,j}$ is the result of comparison of the i-th and j-th values of the intervals between the extrema. We create binary matrices of the comparison (3) results for amplitudes **R** and time intervals **W** separately.

$$\mathbf{R} = \begin{pmatrix} 0 & r_{0,1} & \cdots & r_{0,N} \\ r_{1,0} & 0 & \cdots & r_{0,N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N,0} & r_{N,1} & \cdots & 0 \end{pmatrix}$$
$$\mathbf{W} = \begin{pmatrix} 0 & \omega_{0,1} & \cdots & \omega_{0,N-1} \\ \omega_{1,0} & 0 & \cdots & \omega_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{N-1,0} & \omega_{N-1,1} & \cdots & 0 \end{pmatrix} \tag{4}$$

The obtained binary square matrices **R** and **W** (4) of the sizes N+1 and N, respectively (intervals between local extrema are less by 1) have redundant character since, owing to inequality asymmetry (if $a > b$, then $b \le a$), we have $r_{i,j} = \neg r_{j,i}$ and $\omega_{i,j} = \neg \omega_{j,i}$, and all the elements lying on the main diagonal are equal to 0. In order to represent the pulse waveform more compactly, we join significant elements of both matrices. Transform the matrices **R** and **W** into upper triangular ones zeroing the redundant elements according to (5)

$$r_{i,j} = \begin{cases} r_{i,j}, & i < j \\ 0, & i \ge j \end{cases}, \qquad \omega_{i,j} = \begin{cases} \omega_{i,j}, & i < j \\ 0, & i \ge j \end{cases}. \tag{5}$$

As the result we obtain triangular binary matrices $\mathbf{R}_\Delta$ and $\mathbf{W}_\Delta$ (6).

$$\mathbf{R}_\Delta = \begin{pmatrix} 0 & r_{0,1} & r_{0,2} & \cdots & r_{0,N} \\ 0 & 0 & r_{1,2} & \cdots & r_{1,N} \\ 0 & 0 & 0 & \cdots & r_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_{N-1,N} \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$
$$\mathbf{W}_\Delta = \begin{pmatrix} 0 & \omega_{0,1} & \omega_{0,2} & \cdots & \omega_{0,N-1} \\ 0 & 0 & \omega_{1,2} & \cdots & \omega_{1,N-1} \\ 0 & 0 & 0 & \cdots & \omega_{2,N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \omega_{N-2,N-1} \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{6}$$

We reduce the matrix $\mathbf{R}_\Delta$ size to the value N by removing zero elements of the lower line and the left column. We make transposition operation of the matrix $\mathbf{W}_\Delta$ and as the result obtain the matrix $\mathbf{W}_\Delta^{\mathrm{T}}$. After summing up the matrices $\mathbf{R}_\Delta$ and $\mathbf{W}_\Delta^{\mathrm{T}}$, we obtain a square binary matrix **D** of size N (7). It can be used as a model of pulse amplitude-phase structure.

$$\mathbf{D} = \begin{pmatrix} r_{0,1} & r_{0,2} & \cdots & r_{0,N-1} & r_{0,N} \\ \omega_{0,1} & r_{1,2} & \cdots & r_{1,N-1} & r_{1,N} \\ \omega_{0,2} & \omega_{1,2} & \cdots & r_{2,N-1} & r_{2,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_{0,N-2} & \omega_{1,N-2} & \cdots & r_{N-2,N-1} & r_{N-2,N} \\ \omega_{0,N-1} & \omega_{1,N-1} & \cdots & \omega_{N-2,N-1} & r_{N-1,N} \end{pmatrix} \tag{7}$$

We call (7) a pulse descriptive matrix. An example of a pulse and its descriptive matrix are illustrated in Figure 3.

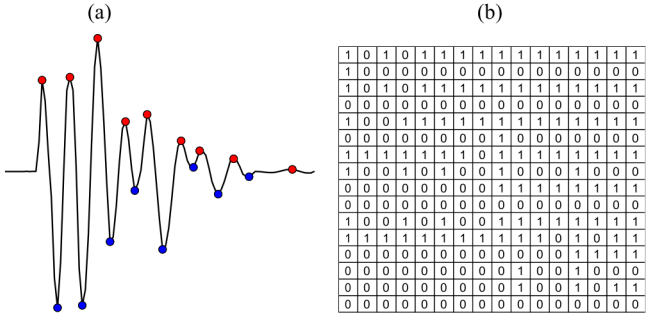


Figure 3: Representation of a pulse by descriptive matrix: (a) – pulse with detected extrema; (b) – pulse descriptive matrix

Descriptive matrices have invariant character under time and amplitude transformations that is proved by the following properties:

- invariance under time compression or expansion (if $k > 0$ and $\tau_i > \tau_j$, then $\tau_i \cdot k > \tau_j \cdot k$);
- invariance under amplitude compression or expansion (if $k > 0$ and $a_i > a_j$, then $a_i \cdot k > a_j \cdot k$).

The invariance is retained, if the coefficient k is substituted by any monotone positive function $f(x) > 0$. Attenuation processes

and nonlinear amplitude-phase distortions, changing AE pulse signal waveforms during its propagation in heterogeneous anisotropic mediums from a source to a receiver, have the character of monotone functions in the majority of practical cases. Thus, descriptive matrix invariance is of great importance when creating algorithms for identification and classification of acoustic emission pulses.

3. Model adequacy

To verify the proposed model adequacy, we generated two artificial AE signals.

It was shown in the paper [9] that AE typical pulses are described by modulated Berlage functions with a reasonable degree of accuracy

$$u(t) = A \cdot t^{n(p_{\max})\cdot\Delta} \cdot \exp\left(-\frac{n(p_{\max})\cdot\Delta}{p_{\max}\cdot t_{end}}\cdot t\right)\cdot\sin(2\pi ft+\varphi_0), \quad (8)$$

where A is the pulse amplitude; t is the time, $0 \le t \le t_{end}$; p_{max} is the pulse envelope maximum position defined relating to t_{end}; $n(p_{max})$ is the minimum value of the parameter affecting pulse envelope slope which is chosen so that $u(p_{max} \cdot t_{end}) \le 0.05 \cdot u(t_{end})$; Δ is the coefficient of pulse envelope slope variation, the more Δ value is, the steeper the envelope is, $\Delta > 1$; f is the pulse modulation frequency; φ_0 is the initial phase. To obtain different pulse waveforms, the parameter values p_{max}, Δ, t_{end}, f, φ_0 vary.

The first artificial signal (Figure 4a) contains five pulsesof similar waveform and the second one contains five pulsesof different waveform (Figure 4b). As long as the model should be invariant under amplitude and time transformations of the signal, the length and the amplitude of each pulse were drawn at random. Instead of the frequency f, the number of periods k on the time interval from 0 to t_{end} was defined. Parameters of the generated pulses are listed in Table 1.

Table 1: Parameters of test pulses

Wave-form	Number of periods, k	Maximum position, p_{max}	Steepness, Δ	Initial phase, φ_0
I	10	0.2	1.5	0
II	5	0.1	1.2	0
III	15	0.25	2	0
IV	11	0.01	1	$\pi/3$
V	20	0.4	1.5	0

Signal graphs and modeling results are illustrated in Figure 4a–b. Pulses having similar descriptive matrices are shown by the same color. Thus, application of the proposed model allowed us to determine correctly the similarity/difference of pulse waveforms. We should note that the pulses of the first signal have different frequency-time structures (Figure 4c), and would be attributed to different classes when applying frequency-time methods. Pulses of the second signal have the same frequency (Figure 4d), but different waveforms and would be attributed to the same class when applying frequency-time methods.

Figure 5 shows the real AE signal containing five pulses. Three pulse waveforms are distinguished visually. The suggested model describes real signals adequately. Pulses of the same waveform have similar descriptive matrices.

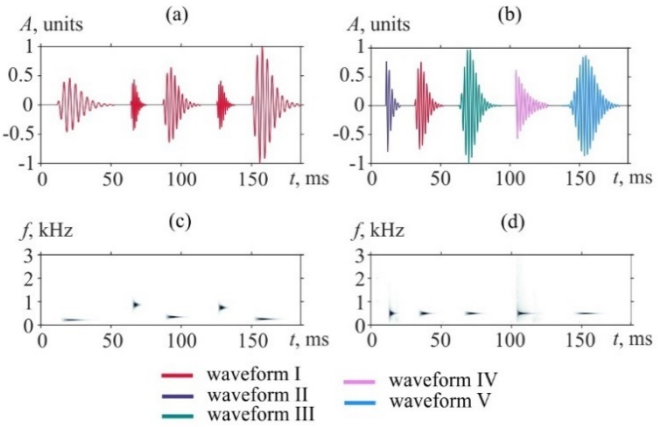


Figure 4: Modeling results for two signals. Pulses described by the same descriptive matrix are of the same color: (a) – signal containing five pulses of the same waveform; (b) – signal containing five pulses of different waveform; (c) – Wigner-Ville transformations of the first signal pulses; (d) – Wigner-Ville transformations of the second signal pulses

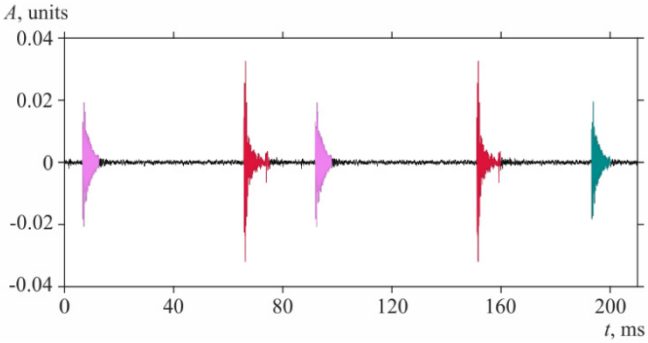


Figure 5: Modeling results for real AE signal. Pulses described by similar descriptive matrices are of the same color

In the considered examples we analyze the signals in which noises are either absent or insignificant. Real AE signals often contain noises of artificial and natural origin, thus perturbing a signal. Distortion may result in a gap, false detection or amplitude change of some extrema. Then the size and some elements of pulse descriptive matrices differ (Figure 6).

4. Classification of pulse diversity

We define initial conditions for classification. We put bounds on the size of descriptive matrices detected during the analysis by the values N_{min} and N_{max} from the bottom and from the top, respectively. Analysis epoch is the length of the time window in which an AE signal is observed. In practice, this index should provide formation of stable classes. N_{min} is determined by the minimum number of oscillations during pulse duration and cannot be set less than 3. N_{max} is defined according to the capabilities of a computing device and the maximum run-time.

Pulses are classified accorded to the results of calculation of descriptive matrix similarity coefficient. First, we restrict ourselves to descriptive matrices of the same size. In order to do that we divide descriptive matrices into groups $\mathbf{D}^{(Z)}$ of matrices of the same size Z, $N_{min} < Z < N_{max}$.

For the two descriptive matrices of the same size $\mathbf{D}_1^{(Z)}$ and $\mathbf{D}_2^{(Z)}$ the similarity coefficient $g(\mathbf{D}_1^{(Z)}, \mathbf{D}_2^{(Z)})$ is determined as a ratio of elements, coinciding in value and taking the same positions, to the

total number of elements in the matrix. In case of exact identity, the similarity coefficient is 1.

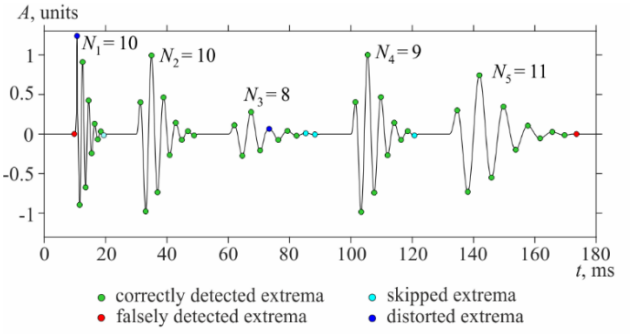


Figure 6: Examples of errors occurring during extreme detection

$$g(\mathbf{D}_1^{(Z)}, \mathbf{D}_2^{(Z)}) = \frac{\#(d_{1ij} = d_{2ij})}{Z^2}, \quad (9)$$

where #(*condition*) is the operator for calculation of the number of elements satisfying the condition, d_{1ij} are the elements of the matrix $\mathbf{D}_1^{(Z)}$, d_{2ij} are the elements of the matrix $\mathbf{D}_2^{(Z)}$.

The calculated similarity coefficient $g(\mathbf{D}_1^{(Z)}, \mathbf{D}_2^{(Z)})$ is compared with the selected threshold value G_0 (it is chosen empirically). If

$$g(\mathbf{D}_1^{(Z)}, \mathbf{D}_2^{(Z)}) > G_0,$$

then the pulses described by $\mathbf{D}_1^{(Z)}$ and $\mathbf{D}_2^{(Z)}$ belong to the same class.

Classification is the comparison of descriptive matrices of the same size and selection of the empirical threshold G_0 ($0 < G_0 \leq 1$) with which the similarity coefficient (9) is compared.

The possibility that two pulses of the same source will have exactly the same number of local extrema is very low. So, the obtained classification is too restrictive. Thus, we developed a procedure to compare descriptive matrices of different sizes. It takes into account false extrema or gaps of some extrema and allows us to reduce the number of detected classes. However, when descriptive matrix sizes are widely scattered, they are compared in erratic manner that may cause random merging of classes described by descriptive matrices of smaller sizes. We put bounds on the possibility for merging of descriptive matrices of smaller sizes by the threshold coefficient S_0:

$$N_L / N_M \geq S_0, \quad 0 < S_0 \leq 1,$$

where N_L and N_M are the sizes of smaller and larger descriptive matrices, respectively, S_0 is the empirically defined threshold.

Assume that there are two descriptive matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ of the sizes N_1 and N_2, respectively ($N_1 \geq N_2$). If $N_2/N_1 \geq S_0$ ($N = S_0 \cdot N_1$), then the matrices are comparable. For the meaningful comparison, the matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ are matched so that the latest N elements of the main diagonal $\mathbf{D}_2$ are overlapped on the first N elements of the main diagonal $\mathbf{D}_1$. Then we calculate the similarity coefficient (9) of matrices of size N in the domain bounded by the intersection. Elements are further compared after $\mathbf{D}_2$ shift along the main diagonal $\mathbf{D}_1$. The similarity coefficient (9) is calculated for each subsequent shift so far as the matrices in the domain bounded by the intersection have the size of not less than N (Figure 7). The comparison result is the maximum value of the similarity coefficient.

If the maximum similarity coefficient (9) exceeds the empirically defined threshold G_0 or is equal to it, the decision on sufficient closeness of descriptive matrices of compared pulses is made. In this case, the pulse with descriptive matrix of the smaller size is included into the class described by larger-size descriptive matrix (larger object absorbs a smaller one).

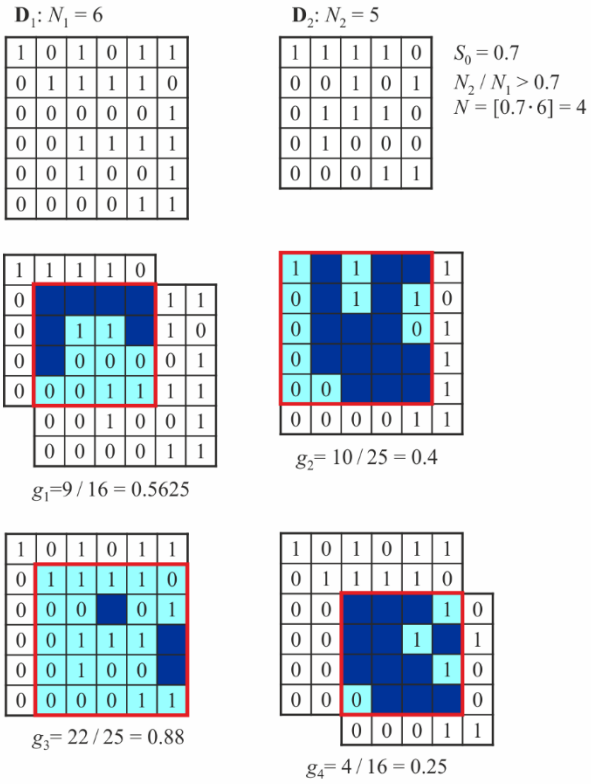


Figure 7: Comparison of descriptive matrices of sizes 6 and 5

In the result of classification many classes are formed. Each of them unites the pulses close in amplitude-phase structure with the accuracy up to the values determined by the coefficients G_0 and S_0. Each class can be considered as some symbol and the whole set of obtained symbols composes a signal alphabet.

In order to estimate the classification algorithm noise immunity for different values of the thresholds G_0 and S_0, we realized a simulation experiment. We processed the data containing 100 pulses of similar waveform (the alphabet cardinality is 1 symbol) with signal/noise ratio (SNR) from ∞ to 2.7 dB.

At the first stage of the experiment we compared the descriptive matrices with sizes differing by less than 30% ($S_0 \geq 0.7$). Similarity coefficient threshold was chosen to be $G_0 = 1$ (case when a matrix of smaller size is the sub-matrix of a larger-size matrix).

Figure 8 shows the dependences of alphabet cardinality on SNR for different values of the coefficient S_0 (0.7–1). Insignificant difference of the graphs proves that descriptive matrix element values significantly depend on local influence of noise on pulse waveform and are weakly sensitive to changes in the threshold S_0 at a high value G_0. When SNR values are less than 30 dB, the

classification does not work since matrices do not match with 100% agreement ($G_0 = 1$).

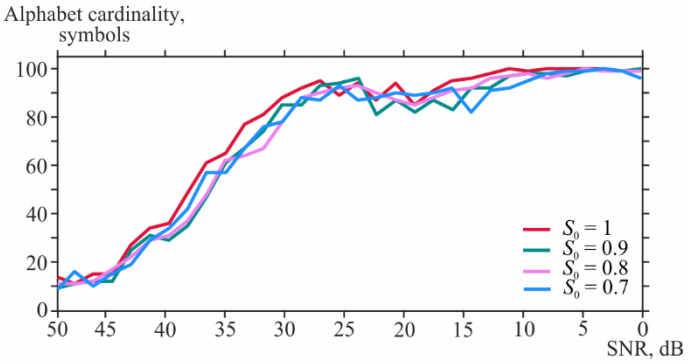


Figure 8: Dependence of alphabet cardinality on SNR for different values of the threshold S_0 (0.7–1) and for a fixed valued of the threshold G_0 =1

At the next stage of the computational experiment, we investigated class formation during threshold G_0 changes (from 0.6 to 0.9). In this case, matrices of the same size ($S_0 = 1$) were compared. The results of the second stage of the computational experiment are illustrated in Figure 9. The graphs show that decrease of the threshold G_0 causes significant reduction of alphabet cardinality. Moreover, the number of detected classes reduces with G_0 decrease.

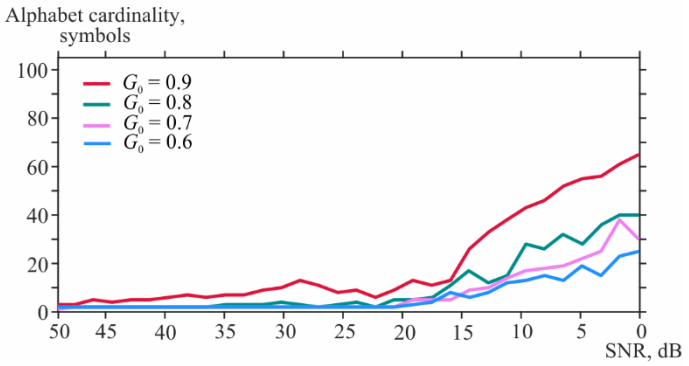


Figure 9: Dependences of alphabet cardinality on SNR for different values of the threshold G_0 (0.6–0.9) and for a fixed value of the threshold $S_0 = 1$

At the third stage of the computational experiment we investigated the dependence of alphabet cardinality on SNR for $S_0 = 0.7$ and different values of G_0. The results are shown in Figure 10.

The experiment shows that the classification method is nonlinearly sensitive to the choice of threshold functions of G_0 and S_0: decrease of S_0 causes pulse redistribution into classes with larger-size descriptive matrices and decrease of G_0 makes it possible to compact classes. Based on the experiment results, the recommended values of the thresholds S_0 (from 0.6 to 0.9) and G_0 (from 0.7 to 0.9) were selected.

As long as the developed method of classification is expectedly sensitive to the performance quality of the algorithms for pulse detection and noise reduction, one of the important tasks in AE signal investigation is the improvement of methods for signal pre-processing.

Classical digital filtration, empirical mode decomposition (EMD), and wavelet filtration were considered as methods of noise reduction and detection of signal useful component. Digital filtration was not used because the AE pulses have short duration (100–400 samples per pulse). Problem of mode functions selection complicated the usage of EMD for signal denoising (for example, in [10]), since rejected modes are different for different pulses. So, the authors chose wavelet filtration technology with adaptive threshold scheme. Wavelets are widely used in various signal processing tasks [11], in particular in noise reduction tasks [12, 13]. Denoising has three stages: direct wavelet transforms, application of a selected threshold scheme to every detailing component, inverse wavelet transforms. In order to choose a wavelet family, a threshold scheme and an algorithm for calculating threshold value, a computational experiment was carried out on model and real AE signals. The best results were obtained applying fourth-order symlets (maximum possible level of decomposition), median threshold scheme and empirical Bayes rule to calculate the threshold value [14, 15]. The selected wavelet filtration method allows us to increase the signal-to-noise ratio by about 9 dB, on the condition that initial SNR is 8.9 dB.

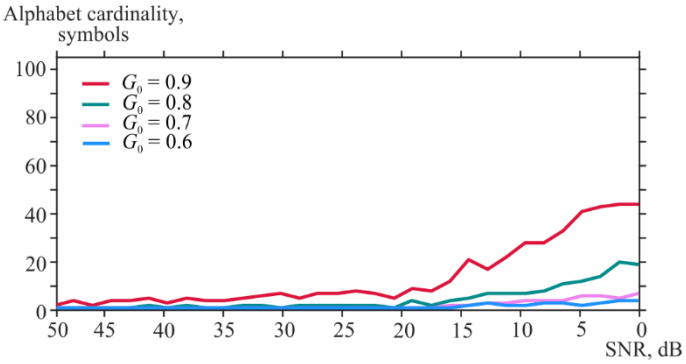


Figure 10: Dependences of alphabet cardinality on SNR for different values of the threshold G_0 (0.6–0.9) and for a fixed value of $S_0 = 0.7$

Figure 11a–b illustrates the process of detection of local extrema in a noised pulse and in a pulse with additive noise. Figure 11c shows how the wavelet filtering algorithm recovers the waveform of a perturbed signal.

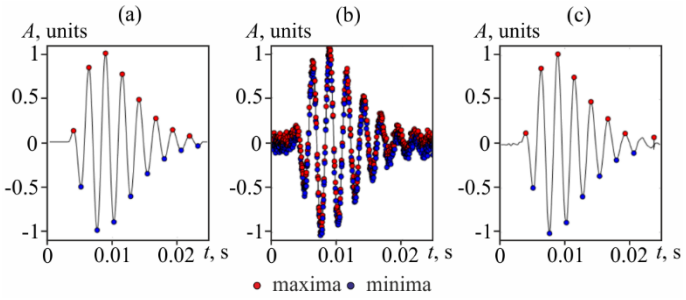


Figure 11: Noise effect on detection of local extrema: (a) – initial pulse; (b) – pulse with overlapped white noise, SNR = 12.4 dB; (c) – pulse denoised by wavelet filtration, SNR = 22.7 dB

Application of wavelet filtering with adaptive threshold schemes allows us to apply the proposed methods of analysis to perturbed and noisy signals.

Figure 12 illustrates the results of classification of model signal pulses with the signal/noise ratio of 8.9 dB. The signal contains 50 Berlage pulses described by three waveforms the parameters of which are listed in Table 3. Before being classified, each detected pulse was denoised by the selected method of wavelet filtering (after wavelet filtration, the SNR of a single pulse is about 20 dB).

When constructing descriptive matrices, we did not take into account the extrema the amplitudes of which were less than 15% from the maximum one. Classification algorithm thresholds were selected as $S_0 = 0.6$ and $G_0 = 0.7$. For a noisy signal (Figure 12a) we succeeded to define correctly the number of classes and to structure the pulses into classes (Figure 12b). In the course of multiple repetition of the experiment it was discovered that the obtained structuring into the classes is stable to the changes of initial conditions of white noise generator.

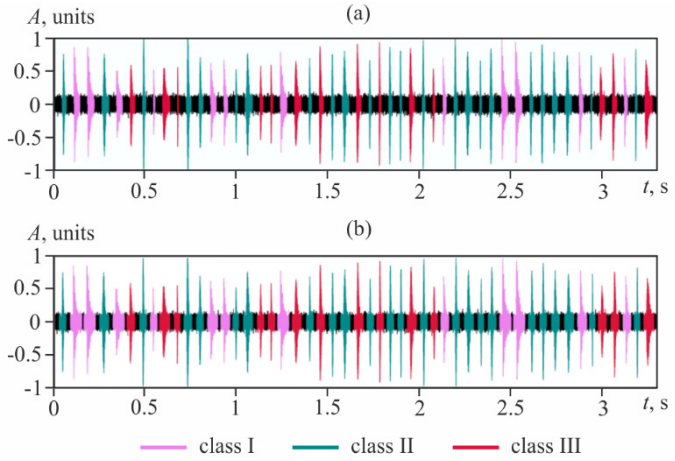


Figure 12: Classification results: (a) – signal with overlapped white noise and initial structuring into classes; (b) – classification of $S_0 = 0.6$, $G_0 = 0.7$; 3 classes were defined

Table 2: Waveform parameters

Class	Number of periods, k	Maximum position, p_{max}	Steepness, Δ	Initial phase, φ_0
I	7	0.05	1.1	0
II	20	0.35	3	$\pi/3$
III	40	0.2	2	$\pi/2$

The proposed algorithm classifies correctly 78% of pulses, on the conditions that thresholds are selected as $S_0 = 0.6$, $G_0 = 0.7$ and SNR is 8.9 dB before wavelet denoising. This estimate is based on the results of 100 experiments.

Efficiency of the proposed classification method can be estimated by a contraction coefficient

$$b = \frac{|A_{input}|}{|A_{output}|}$$

where $|A_{input}|$ is the alphabet cardinality (number of pulses) before classification, $|A_{output}|$ is the cardinality of the alphabet obtained as the result of classification.

Table 3 illustrates the classification results for 12 files containing 15-minute records of AE signals from IKIR FEB RAS archive data for January 1, 2018. The classification was carried out with the coefficients $G_0 = 0.8$, $S_0 = 0.8$.

As the result of further processing of AE data for the period 2016-2019, it was determined that contraction coefficient (10) varies from 50 to 200.

The developed methods of pulse description and classification were realized in the form of software. Preprocessing procedures such as centering, normalization, and noise reduction were performed by MATLAB system.

Table 3: Classification results

File number	Alphabet cardinality, $\|A_{input}\|$	Alphabet cardinality, $\|A_{output}\|$	Contraction coefficient, b
1	2864	37	77.4
2	3233	36	89.8
3	3581	46	77.8
4	3986	43	92.6
5	4593	45	102.0
6	3132	46	68.0
7	3778	43	87.9
8	4424	38	116.4
9	4836	36	134.3
10	3968	39	107.7
11	4734	36	131.5
12	3475	44	78.9

5. Examples of identification of pre-seismic AE anomalies

The developed method of classification was realized in the form of an application program. It was used to process and to analyze acoustic emission signals recorded in 2016–2019 at "Karymshina" site of IKIR FEB RAS (52.83º N, 158.13º E). AE signals are stored in the database in the form of 15-minute sound files (WAVE format). In the course of processing of each file, signal alphabet was constructed. Then alphabet symbols were ranked according to the size of their descriptive matrices and frequency (number of pulses described by the symbols).

Based on the conception of the relation of AE signal parameters with the state of a system generating a signal, we should analyze quantitative and qualitative parameters of symbol description of a signal for different time periods (from a day to a year) and to compare the defined anomalies with occurred earthquakes according to the data of regional seismic catalogue (Kamchatka Branch of the FRC GS RAS, Earthquake Catalogue for Kamchatka and Kuril Islands (1962 – present time), http://sdis.emsd.ru/info/earthquakes/catalogue.php).

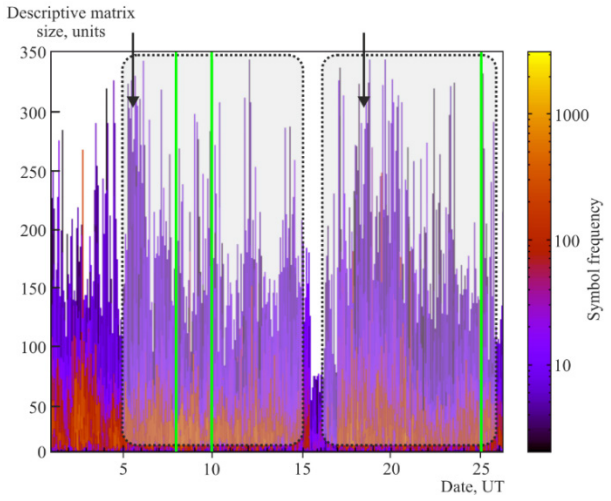


Figure 13: Anomalous dynamics of AE signal alphabet in October 2017

In order make the analysis, alphabets of each 15-minute AE signal were formed. Then three-dimensional images of a special type were drawn. Descriptive matrix size was measured on the vertical axis. Time (with 15-minute step) was measured on the horizontal axis. Color denotes symbol frequency.

Examples of anomalies observed 1–12 days before earthquakes and 1–6 days after earthquakes are illustrated in Figures 13–15.

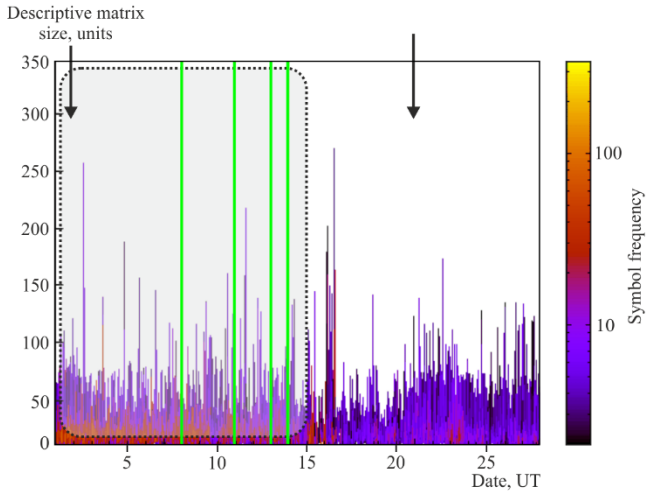


Figure 14: Anomalous dynamics of AE signal alphabet in February 2019

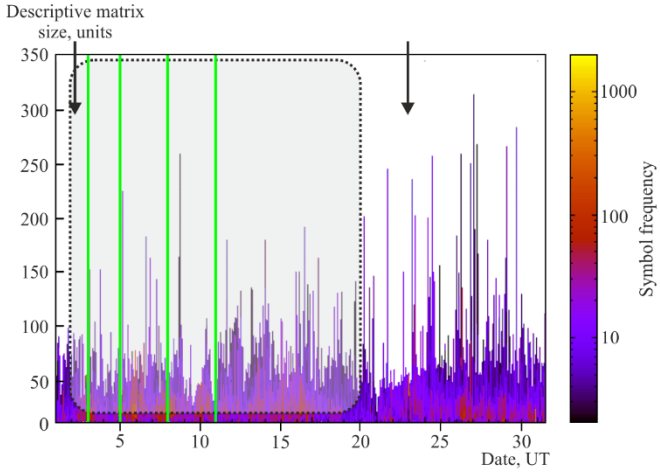


Figure 15: Anomalous dynamics of AE signal alphabet in October 2019

Anomaly peaks in the Figures are marked by black arrows. Green lines indicate earthquake times. These anomalies are alphabet content redistribution in favor of the symbols described by larger-size descriptive matrices (up to 350 extrema). Simultaneously with that the number of such symbols increases. Alphabet cardinality also grows from 20 to 150. Assumed areas of anomalies associated with earthquakes are marked by rounded rectangles. Short-term changes of alphabet content (less than 305 peaks) were not taken into consideration.

We should note that not all the analyzed AE anomalies in the form of redistribution of alphabet symbol frequency are associated with earthquake preparation processes. In the experiment the acoustic emission is recorded in sedimentary rocks and is a sensitive indicator of their stress-deformed state. Both global tectonic processes (for example, earthquake preparation) and local deformations, associated with the impact of different natural factors, may influence that state. Thus, not all the recorded AE anomalies are pre-seismic.

Figures 13–15 illustrate examples of AE anomalies observed during the periods when earthquakes occurred quite rarely, from one to four strong events per a month. Such a choice allows us to demonstrate the relation of detected anomalies with earthquakes. When earthquakes repeat quite frequently, the anomalous behavior is not so clear that may be explained by overlapping of fragments of evoked anomalous effects. Example of such behavior of alphabet content dynamics is shown in Figure 16.

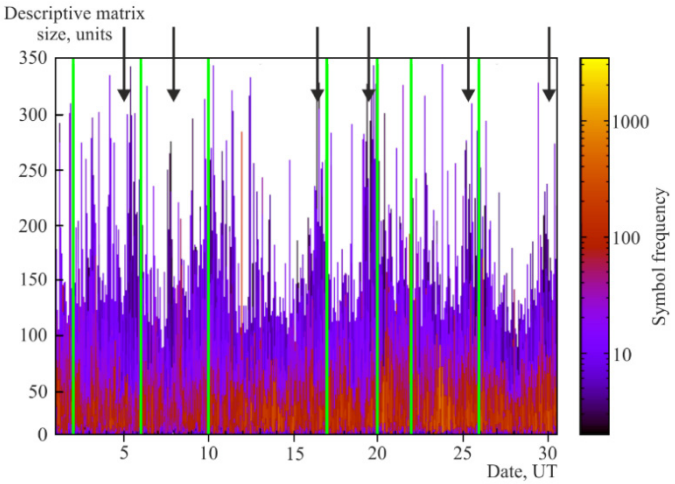


Figure 16: Anomalous dynamics of AE signal alphabet in April 2017

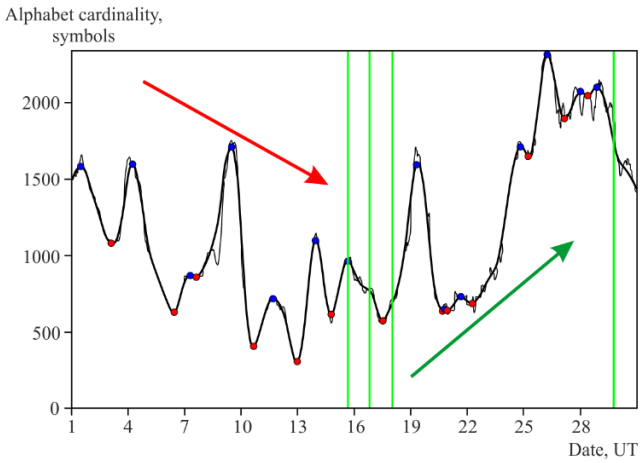


Figure 17: Graphs of AE signal alphabet cardinality dynamics in September 2017

Further analysis of the dynamics of AE signal alphabet parameters allowed us to find some regularities associated with the anomalies of the envelope of day-averaged alphabet cardinality. For the averaging we applied a sliding window of the width of 96 values of alphabet current cardinality. As an example, Figures 17–19 illustrate alphabet cardinality dynamics for one month, September 2017, December 2018 and September 2019. Thin gray line shows the result of alphabet averaging and the thick black line shows the result of approximation of the least square method. The graphs clearly show an oscillating process with small uncertain deviations of phase and amplitude the average period of which is from 2 to 3 days. The nature of this oscillating process was not determined and is under the study now. Blue dots on the graph are local maxima of detected oscillations. Red dots are their

minimums. Vertical green lines show the earthquake times. Red arrows indicate alphabet transformation tendencies before a series of earthquakes. Green arrows indicate transformation tendencies after a series of earthquakes. We can clearly distinguish the tendency of decrease of alphabet cardinality relatively the average 10–18 days before earthquakes and recovery of the value over the period from 2 to 8 days after them. This decrease is evidently determined by the fact that acoustic process before an earthquake becomes significantly more nonstationary. More pulses of similar waveforms appear and, consequently, alphabet set gets narrow.

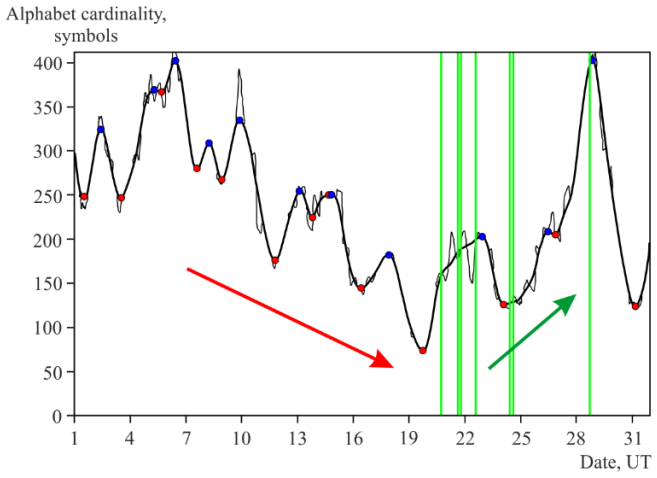


Figure 18: Graphs of AE signal alphabet cardinality dynamics in December 2018

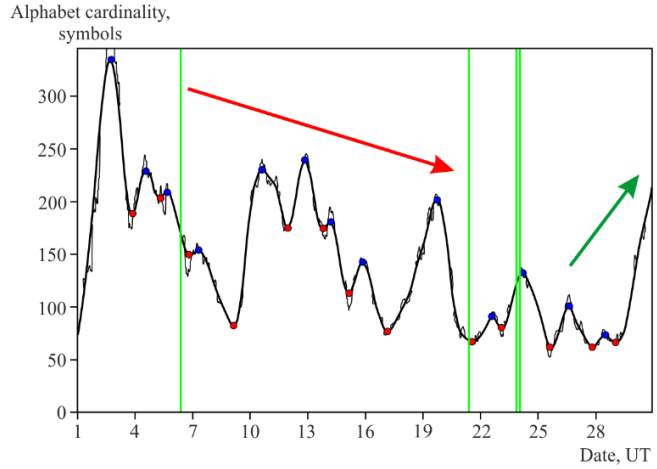


Figure 19: Graphs of AE signal alphabet cardinality dynamics in September 2019

The obtained estimates of the pre-seismic anomaly occurrence periods are consistent with the results of early studies. It was shown in [1, 2] that AE intensity increases at the frequency of 25 kHz 2–3 days before strong earthquakes. According to [16, 17], anomalies in the form of signal level increase appear in the frequency range of more than 6.5 kHz in the AE signals accumulated by 1 s. The results of time-frequency analysis of single AE pulses registered 1–7 days before earthquakes are presented in [18, 19]. According to these results, the number of pulses of a certain frequency increases before strong earthquakes. The above methods analyze signal frequency spectrum and, unlike the one described in the article, do not take into account waveform features. The proposed method of AE signals analysis updates the existing methodology and allows us to form more accurate short-term precursors of strong earthquakes.

Note that at all stages of the investigation, we had to face known nonlinear manifestations of the natural system under the study. These manifestations reflected in the chaotic character of AE signal dynamic characteristics. As it had been expected, the attempt of application of traditional methods for signal statistical processing, such as calculation of probability density function, ensemble averaging, estimation of statistic parameters of different orders, to nonstationary AE signals did not provide stable results. Thus, the attention was focused on the search for invariant transformations which would be applied to AE chaotic character properly. On the whole, we think that we are at the beginning of a perspective survey which, owing to the chosen approach and the developed computer tools for analysis, will allow us to investigate deeper and to extend general and personal knowledge on seismic process behavior and, consequently, to predict catastrophic effects of earthquakes at different stages and with more confidence.

6. Conclusions

A new method for analysis and classification of rock acoustic emission signals has been proposed. It is based on their symbol description. First of all, signal sections containing pulses are determined. Each of the sections is described by a descriptive matrix. Based on the similarity degree of the descriptive matrices, pulses are structured into classes. Each class is considered as a symbol and a set of obtained symbols composes an alphabet. Then we study the changes of alphabet parameters on signal subsequent fragments.

The developed method allows us to:

- distinguish informative features in AE signals based on their qualitative peculiarities;
- solve the problem of processing of a large volume of data by unifying pulse description and systematization;
- identify anomalies in AE signals distinguishing pre-seismic ones.

The method was implemented as an application program by the means of which we analyzed acoustic emission signals which were recorded at “Karymshina” site of IKIR FEB RAS in the seismically active region, Kamchatka peninsula, in 2016-2019. It was discovered during the analysis that in case of acoustic emission anomaly occurrence, alphabet content is redistributed in favor of the symbols described by larger-size descriptive matrices (up to 350 extrema). Simultaneously with that the number of symbols increases. Analysis of alphabet averaged dynamics detected a clear tendency of alphabet cardinality decrease relatively the average 10–18 days before earthquakes and recovery of the value over the period from 2 to 8 days after them.

Thus, a new approach to the analysis of acoustic emission signals has been developed. The approach updates the existing methodology of short-term earthquake precursor formation. On the basis of it, a system for automatic identification of pre-seismic AE anomalies can be created. It is topical for prediction of strong earthquakes which are catastrophic natural phenomena.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The research was supported by the Russian Science Foundation, Project No. 18-11-00087.

References

[1] G. Paparo, G. P. Gregori, U. Coppa, R. De Ritis, A. Taloni, "Acoustic Emission (AE) as a diagnostic tool in geophysics" Ann. Geophys-Italy, **45**(2), 401–416, 2002. https://doi.org/10.4401/ag-3511

[2] G. P. Gregori, G. Paparo, M. Poscolieri, A. Zanini, "Acoustic emission and released seismic energy" Nat. Hazard Earth Sys., **5**, 777–782, 2005. https://doi.org/10.5194/nhess-5-777-2005

[3] A. V. Kuptsov, "Variations in the geoacoustic emission pattern related to earthquakes on Kamchatka" Izv-Phys. Solid Eart+, **41**(10), 825–831, 2005.

[4] G. I. Dolgikh, V. A. Shvets, V. A. Chupin, S. V. Yakovenko, A. V. Kuptsov, I. A. Larionov, Y. V. Marapulets, B. M. Shevtsov, O. P. Shirokov, "Deformation and acoustic precursors of earthquakes" Dokl. Earth Sci., **413**(1), 281–285, 2007. https://doi.org/10.1134/S1028334X07020341

[5] G. P. Gregori, M. Poscolieri, G. Paparo, S. De Simone, C. Rafanelli, G. Ventrice, "Storms of crustal stress" and AE earthquake precursors" Nat. Hazard Earth Sys., 10, 319–337, 2010. https://doi.org/10.5194/nhess-10-319-2010

[6] Yu. V. Marapulets, B. M. Shevtsov, I. A. Larionov, M. A. Mishchenko, A. O. Shcherbina, A. A. Solodchuk, "Geoacoustic emission response to deformation processes activation during earthquake preparation" Russ. J. Pac. Geol., **6**(6), 457–464, 2012. https://doi.org/10.1134/S1819714012060048

[7] I. A. Larionov, Y. V. Marapulets, B. M. Shevtsov "Features of the Earth surface deformations in the Kamchatka peninsula and their relation to geoacoustic emission" Solid Earth, 5, 1293–1300, 2014. https://doi.org/10.5194/se-5-1293-2014

[8] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, 1982.

[9] A. B. Tristanov, Yu. V. Marapulets, O. O. Lukovenkova, A. A. Kim, "A new approach to study of geoacoustic emission signals" Pattern Recognit. Image Anal., 26, 34–44, 2016. https://doi.org/10.1134/S1054661816010259

[10] Yu. I. Senkevich, "The use of the empirical mode decomposition method to clean and restavration acoustic emission signal" E3S Web Conf., **62**, 03008, 2018. https://doi.org/10.1051/e3sconf/20186203008

[11] L. Alperovich, L. Eppelbaum, V. Zheludev, J. Dumoulin, F. Soldovieri, M. Proto, M. Bavusi, A. Loperte, "A new combined wavelet methodology applied to GPR and ERT data in the Montagnole experiment" J. Geophys. Eng., **10**(2), 1–17, 2013. https://doi.org/10.1088/1742-2132/10/2/025017

[12] O. Mandrikova, Yu. Polozov, N. Fetisova, T. Zalyaev, "Analysis of the dynamics of ionospheric parameters during periods of increased solar activity and magnetic storms" J. Atmos. Sol. Terr. Phy., **181**, 116–126, 2018. https://doi.org/10.1016/j.jastp.2018.10.019.

[13] O. Mandrikova, N. Fetisova, "Modeling and analysis of ionospheric parameters based on multicomponent mode" J. Atmos. Sol. Terr. Phy., **208**, 105399, 2020. https://doi.org/10.1016/j.jastp.2020.105399

[14] I. M. Johnstone, B. W. Silverman, "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences" Ann. Statist., **32**(4), 1594–1649, 2004. https://doi.org/10.1214/009053604000000030

[15] I. M. Johnstone, B. W. Silverman, "Empirical Bayes selection of wavelet thresholds" Ann. Statist., **33**(4), 1700–1752, 2005. https://doi.org/10.1214/009053605000000345

[16] Y. Marapulets, O. Rulenko, "Joint Anomalies of High-Frequency Geoacoustic Emission and Atmospheric Electric Field by the Ground–Atmosphere Boundary in a Seismically Active Region (Kamchatka)" Atmosphere, **10**(5), 267, 2019. https://doi.org/10.3390/atmos10050267

[17] O. P. Rulenko, Yu. V. Marapulets, Yu. D. Kuz'min, A. A. Solodchuk, "Joint Perturbation in Geoacoustic Emission, Radon, Thoron, and Atmospheric Electric Field Based on Observations in Kamchatka" Izv-Phys. Solid Eart+, **55**(5), 766–776, 2019. https://doi.org/10.1134/S1069351319050094

[18] O. Lukovenkova, A. Solodchuk, "Analysis of geoacoustic emission and electromagnetic radiation signals accompanying earthquake with magnitude $M_w = 7.5$" E3S Web Conf., **196**, 03001, 2020. https://doi.org/10.1051/e3sconf/202019603001

[19] O. Lukovenkova, A. Solodchuk, A. Tristanov, E. Malkin, "Complex analysis of pre-seismic geoacoustic and electromagnetic emission signals" E3S Web Conf., 127, 03001, 2019. https://doi.org/10.1051/e3sconf/201912703001

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 902-912 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Need of E-Recruitment System for Universities: Case of Pulchowk Campus, Nepal

Vijay Yadav*, Ujjwal Gewali, Suman Khatri, Shree Ram Rauniyar, Aman Shakya

Pulchowk Campus, Department of Electronics and Computer Engineering, Lalitpur, 44700, Nepal

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 28 September, 2020
Online: 12 October, 2020

Keywords:
e-recruitment
job recommendation
candidate filtering
HR management

ABSTRACT

This paper highlights the importance of on-campus online job recruitment system and its role in helping students grab the available job opportunities. It highlights the problems associated with the traditional way of hiring, especially for college students. It also presents some findings and results obtained through various surveys conducted within the campus before and after the deployment of this system. The work presented in this paper is based on an e-recruitment system built for one of the leading engineering institute in Nepal, Pulchowk campus. With some features like job recommendation based on various levels of skill, smart multi-criteria search, graduate tracking, this system proves to be useful for all i.e. companies, students and the campus as well.

1. Introduction

With an increase in production of manpower in various sectors, finding a proper job according to our skill is really a tough task to accomplish. Traditional method of hiring a candidate is proving to be inefficient. It is quite tedious and time consuming. Specially, those candidates who are freshers, don't have any idea regarding the whole process from advertisement to hiring face difficult times in landing to their first job. It's not that they don't have the right skills required for the job, rather most of the time its lack of right information at right time. Some sectors of job are really competitive and is difficult to make through it. But many other sectors have plenty of job opportunities, it's the lack of proper communication channel between employer and job seeker which makes it difficult. The advancement in electronics devices, IT sectors made people realize the need of e-recruitment system. In the beginning phase of development of e-recruitment system it was limited to only few countries and sectors. But today, it has widespread. Even the under developed and developing countries are working to build a proper electronic medium of hiring candidates to manage and mitigate the challenges posed by increasing unemployment in the country, to some extent.

In many developed countries, the universities and campuses have built their own job portal to make it easy for their students to get internships and jobs. This helps students know about the available opportunities at the right time and also establish proper communication channel with the employer. The employer can get their targeted candidate with little effort and less spending. The campuses can also be happy with the improvement in placement rate of their students.

Nepal has a range of web portals that work in the field of hiring of candidates for the jobs that they get through various companies from across the country. JobsNepal, KantipurJob, kumariJob, MeroJob are few of them. Though these commercial jobs are available in large numbers, the concept of job portal for universities and campuses is still a new idea to this country. It's hard to find such a portal in the institute of this country. Through the study and understanding of the importance of dedicated job portal for campuses we came up with the idea to develop an e-recruitment for our own campus. Before developing this system, we conducted surveys among the bachelor's students from the faculties of electronics, computer, electrical and civil engineering. The purpose of these surveys was to know their difficulties faced with traditional system, experiences of using available commercial job portals and perception towards having on-campus job portal.

Here, we are going to present a graphical representation of the process of hiring students that existed before this system and that exists after this system.

*Corresponding Author: Vijay Yadav, vizayyadav123@gmail.com

1.1. Before this system

The Figure 1 below depicts the traditional process of hiring of students that existed in our campus.

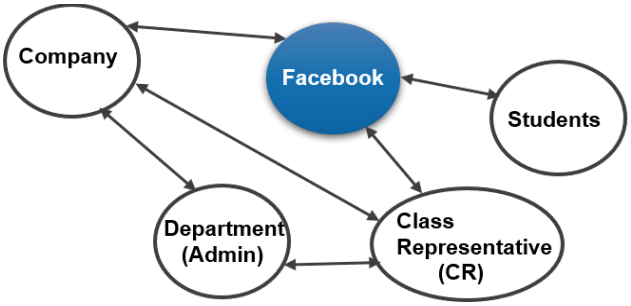


Figure 1: Traditional hiring process

In traditional system, in order to hire students, the company either contacted to various departments of our campus or Class Representative (CR) to circulate the information among students regarding any job vacancies. The department through CR and CR through Facebook page circulated that information. And then when student opened that page, they came to know about any job vacancies. The Facebook page was used for various purposes due to which the job-related information could not be properly communicated.

1.2. After this system

The Figure 2 shows the improvement we brought through the development of this system in the process of hiring students of the campus.



Figure 2: Present day hiring process

As shown in the figure above, a dedicated system is available to manage all the recruitment related activities. Now, company can directly post jobs in the web portal and students can see all the job postings in that portal without anyone's intervention. Companies will contact the department only once in their whole lifetime i.e. at the sign-up process just to verify their identity so that they can start posting jobs and communicate with their potential candidate directly.

1.3. Why this system?

Even though many commercial job portals were available in the market, why we felt the importance of building this system, what edge this system provides over those commercial job portals and how it will benefit the students, we will answer these questions in this section.

Let's have a look at a Figure 3.



Figure 3: Purpose of the system

According to the trend we have seen in last few years, when it comes to hiring a college student for the job, companies prefer students of one institute over others. Also, they post the same vacancies across variety of commercial job portals which is quite costly and time-consuming process. Now, they can simply post their job vacancies in our e-recruitment system which is not only free of cost but also reduces their time in the process of hiring the right candidate. Because it's a simple system with limited user access, the campus also can manage to run this system in free of cost because in the end, it's going to benefit their own students and themselves as they can keep track of the employment situation of their students.

2. Literature review

This paper is an extension of work originally presented in a conference, Artificial Intelligence for Transforming Business and Society (AITB), 2019 [1]. It contains an extensive explanation of the system using various diagrams. Some features which were improved in the system in the last one year have been included in this paper. Those improvements are job recommendation not only based on skills but also on the level of skill like beginner, skilled, advanced and expert, skill verification of students by the department, viewing the location of companies on map, knowing more about the location from where the job has been posted using Foursquare API. It also includes the analysis of data collected before and after the development of this system, through various surveys conducted among the students of the campus. Some data visualization tools have also been used to visualize the skill requirements of the country and places with more job opportunities across the country, using the data collected over the last one year.

Over the last few years technologies have developed a lot. But before that various traditional methods like advertisement through radio, television, newspapers were used to propagate information regarding vacancy opening [2]. Nowadays, people spend more time on internet finding a job rather than reading a newspaper or listening a radio every day. Many breakthroughs are being achieved regarding the speed of the internet which in turn have decreased the price of web surfing. At the same time, we cannot

deny the existence of traditional way of hiring a candidate. Still, the same process of publishing an advertisement on newspaper and hiring a candidate through an interview is considered as the best way to find the right candidate [3]. The major role of online recruitment system is to communicate information regarding the vacancies among the targeted candidates within the right time and in an organized manner in order to fight the increasing unemployment because in some places where even jobs were available, the candidates were unaware of those jobs due to lack of proper communication channel [4]. So, it becomes very important to study the existing situation around the place for which the system is being built. A proper study is needed to know the perception of both employer and job seekers especially the students regarding online recruitment system and traditional system of hiring [5]. Not all the employers looking to find candidate for the job through online process gets satisfactory result after spending considerable amount of money during the process [6]. So, once a system is built it is necessary to analyze the various parameters of the system like cost-effectiveness, security, performance, usability and reliability with the help of group of respondents who are actually going to use the system [7]. Charging a huge amount of money to the employers/ companies is not going to help, especially in the case of on-campus job portal because this will ultimately reduce the number of job postings through this system.

Knowing the purpose of building a job portal is very important because in the past as well many job portals emerged without analyzing the environmental situation required at that point of time and thus, later on were compelled to merge in order to provide better service [8]. Online job recruitment system cannot be useful in every scenario. There are some problems associated with e-recruitment system as well [9]. In some cases, it may prove to be more costly and inefficient. So, while developing a system all factors need to be considered. If done so it can save large amount of money which is spent on advertising and those saved money can be used for training and skill enhancement of the hired candidate [10]. Campuses where mostly non-income generating candidates are available, cost effective system is a must to reduce the economic burden which could come upon the student.

In a fast-paced world of today, it's very difficult to carry on the traditional way of hiring a candidate they are tedious, time consuming and stressful [11]. We live in the world of advertisement. No matter how good the product is, it needs an attractive presentation through advertisement to increase its sale and get desired price for that product. Same thing applies for job. It is directly linked with the life and dreams of an individual. So, he/ she needs to find something inspiring to apply for a job and an e-recruitment provides that platform to the employer to put a proper job posting in order to attract the targeted candidates [12]. Online medium and internet have brought drastic change in the recruitment process impacting both employer and job seeker in a positive way. Internet has become a major tool for both employer and job seeker as it has brought both of them on same platform. It has revolutionized the traditional method of hiring a candidate for the job [13]. It provides an edge to the company to get the best candidate suited for the job. Also, an edge to the job seeker to select from many options available. E-Recruitment system has virtually vanished the boundaries between the countries [14]. It has promoted the concept of globalization. It has reduced the work and time spent by Human Resource Department of companies on hiring process, significantly. It has made the whole process organized and paperless [15].

An individual has to go through various steps and screenings before they finally get a job and for that one should first be able to apply for a job of matched skills [16]. Companies perform those screenings to make sure they have found the right candidate who could sustain in the company for longer period of time because the human resource is what one company can have advantage over others [17]. Skilled and talented human resource makes wise use of resources available within the company and thus makes a company look different in comparison to its competitors. The human resources available within the company should be motivated and possess high morale and companies should help them with these because the conduct, work and efficiency of workforce of companies determine the survival of the company in the business they are involved in [18].

Lots of job are posted on an e-recruitment website on a daily basis. All the job seekers on those platforms don't have same set of skills and it takes lots of time for an individual to find the right job that match up with his/ her skills, if some mechanisms are not employed in the system to recommend jobs to the job seekers based on the skills they possess [19]. So, recommendation system is a must for any e-recruitment system because the main purpose of any e-recruitment system is to reduce the time in the process of hiring.

Nepal is still a developing economy. Many skilled manpower is forced to leave the country in search of job. So, a better and organized system is required to employ those workforces within the country to contribute to its economic development [20]. And, what could be a better than to start identifying the skill of individuals during the college times and helping them find the job that suits their skill. This will not only stop those skilled workforces from going abroad but also help them in contributing to the development of their own country. The college students are more in dilemma when they complete their study and look to find a job for themselves because they don't have any idea regarding where to apply, how the process takes place. They spend much time finding answer to those questions before they land to their first job. So, the e-recruitment system we have built will try to solve these issues. With this system, students will probably be able to get their first job before they complete their studies. Also, they would be familiar with the whole process in advance.

3. System overview

This application is a campus based online recruitment system. The need to build this system was realized when we faced with the traditional method of hiring students by the companies. The

traditional method was time consuming, tedious and unfair to most of the students because the job vacancies came through various individuals in the campus. There was not a dedicated platform for handling only job-related matters due to which the information related to vacancies could not reach to every students of the campus within the time. In fact, we faced this issue multiple times. Until we could get information related to vacancies, some other individuals were already hired. So, we thought of building a system which would treat each student fairly in terms of getting information timely and show casing their skills to the employer. We also defined the role faculty department (Admin) during the hiring process, in this system. There are three actors involved in this system. They are students, companies and faculty departments. Though the role of faculty department during hiring process is minimal, it's vital.

3.1. Basic features of the system

The basic and most important features of an e-recruitment system is shown in Figure 4 The system will help students to build their profile once they are logged in and this profile can be used to generate CV which they can download in the form of PDF. The companies can post job through their own profile. Based on the jobs posted by the companies, the system will recommend the jobs to the students that match with their skills. A secured login is provided to enhance the authenticity of the users.

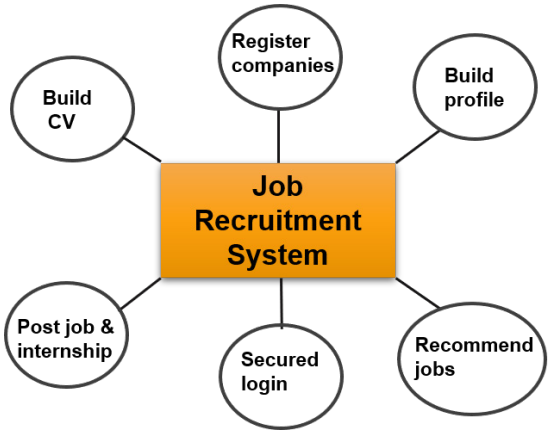


Figure 4: Overview of system

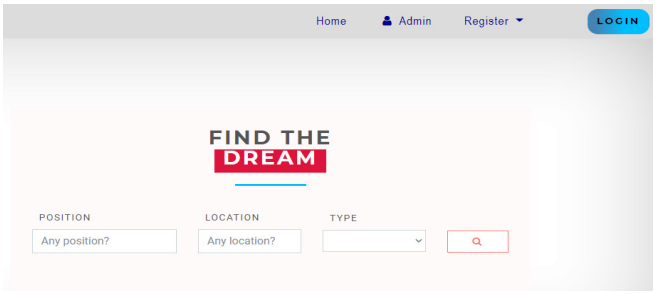


Figure 5: Home page of the system

3.2. Workflow of the system

The state diagram shown in Figure 6 represents various states and the workflow of the system. During the sign-up process, users need to identify themselves as either employer or student. Department does not have to create a separate account as it is already provided with the system. Students having an email account of the campus will only be allowed to create an account. Similarly, the companies can also create an account which they can use to post job vacancies. Students, Companies and Admin are the three actors involved in this system. The whole process of hiring is completed with interactions among these three actors.

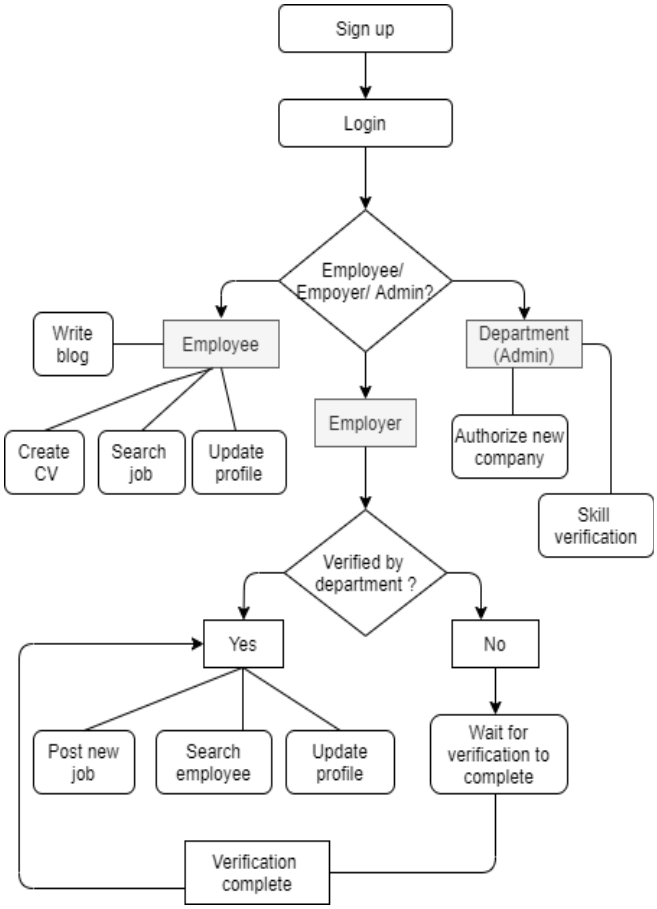


Figure 6: State diagram of the system

The role of all three actors are elaborated below:

Students

Students are the job seekers in this case. All they have to do is to first register themselves into the system. During the registration process, students need to verify their email address. This process does two things, first it verifies whether the email address used is of the campus or not (for e.g.: abcd@pcampus.edu.np), and the next thing is whether the provided email address is associated with particular user or not. The link sent to email for verification will be valid for only few hours. Once the email is verified, users will be asked to update their profile. After updating their profile, they can apply for the job. Also, students will now be visible to the companies in their search results. Students will be notified periodically through email to update their profile to increase the chances to find better job.

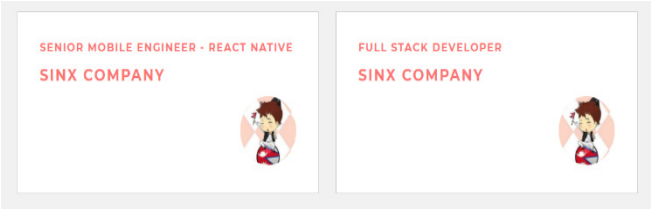


Figure 7: List of jobs the student has applied for

Companies

Companies are the job providers or employers in this case. All they have to do is to create an account to start posting jobs. They have to pass through two verification processes. First, they have to verify their email address and the second is to verify their identity which will be done by the Admin. Once an account is created, a notification will be sent to the Admin to verify whether the company is genuine or not. The admin can match the information provided by the companies during the creation of account or can directly contact with the companies to verify its existence and credibility. As soon as the two-verification processes are completed, companies can start posting jobs. There is no limit on how many jobs a company can post. Also, the posting of jobs will be free of cost to attract sufficient job opportunities towards the campus.

Department (Admin)

It is the center of the system. It controls all the activities of the system. The major task of this actor is to verify the identity of the companies signing up into the system. It can also keep track of the employment condition of the students even after their graduation.

The role of Admin has been extended from the previous version of the system. A separate account will be setup for project supervisors which will have very limited access control over the system. The one who has logged in using this account can only view an interface to verify the skills of the students. Once in a semester the project supervisor can use this account to verify the skill demonstrated by the student during the development of project. Student's profile will have green tick mark displayed along with their skills. This will be an added advantage to the student.

3.3. Software Development model

This application has been developed using prototype model. Since this system was itself a first of its kind in the country, it was very important to incorporate the feedbacks of various parties involved in this system directly of indirectly. Especially, it was very important to get the advices of the users who will actually be using this system. After refining so many prototypes, we finalized a prototype and started building this system. During prototyping, various faults within the design were identified. Also, it became easy for us to train them as most of them were familiar with the design. It also increased the acceptance among the users. The development and testing of the system were carried out simultaneously.

4. System architecture

The various components of this system have been elaborated below:

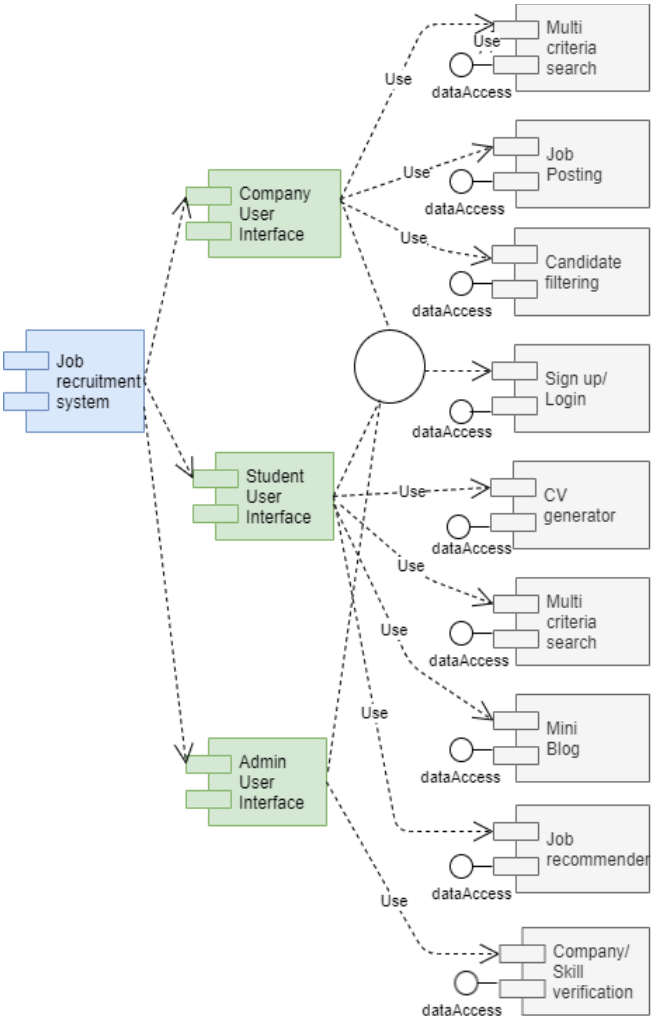


Figure 8: Component diagram of the system

Multi-criteria search

Multiple fields have been provided to both students and companies to make search. Students have been provided with fields like position, location and type to find the jobs more suited to them. Student can search for different positions like programmer, graphics designer, project manager, content writer and so on. Also, they can make search if jobs are available in the location of their choice. Type field represents the type of job like part time, full time, freelance or internship. Students can either use these fields individually or in combinations to make search.

Similarly, companies are provided multiple fields to make search for candidates best suited for the vacancies they have announced. The fields provided in search module of companies are position, qualification, skills, and years of experience. Position has same meaning as that of students. Qualification could be the degree that a particular candidate holds such as Bachelor in Computer Engineering, Masters in Electronics and communication

Engineering and so on. Skills could be technical skills like python, java and non-technical skills like communication, time management and so on. The companies can also use these fields individually or in various combinations.

Not every time the keywords used for searching match exactly to the keywords stored in database. So, Levenshtein distance has been used to approximate the search keyword. It helps to find the result that best represents the keyword entered by users during search process. It calculates the minimum edit required to transform one string to another [21]. The edit mentioned here could be insertion, deletion or substitution.

For example, let us consider someone is trying to make search using skill field and he/she enters PYTON instead of PYTHON.

P	Y	T	H	O	N
P	Y	T	O	N	

The Levenshtein distance between those two words is 3 as three changes are required. Also, we have fixed the maximum permissible to 3. So, in this case the user will successfully get all the results have skill value “PYTHON”.

Job posting

This allows a company to post jobs. There is no limit on how many jobs a single company can post. Also, the posting of job is free of cost. The company needs to add their address very carefully because we have used folium library to point to the address defined by the company, on map of Nepal. This could help student in finding the place easily in case if company calls that individual for the interview. Geopy library has been used to find the latitude and longitude of the place.

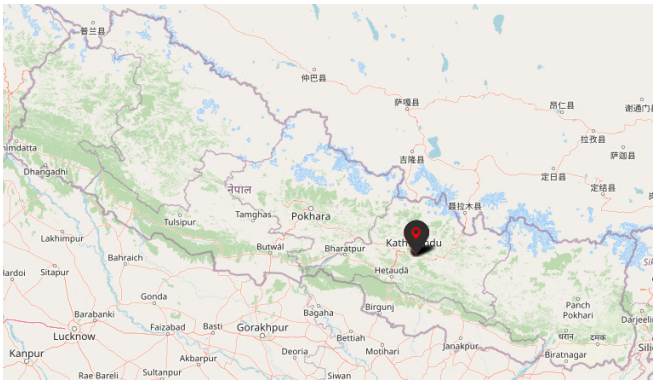


Figure 9: Example of map pointing to the address of company in New Baneshwor, Kathmandu, Nepal

There is also one feature provided called “Explore this place” which will help students get an overview of the place for which the job vacancy has been announced. It uses Foursquare API to find the top ten categories of venues within 500m of that place (if available within its dataset). Since it is still in an initial phase, the free version of Foursquare API has been used. Here also, Geopy library has been used to find the latitude and longitude of the place.

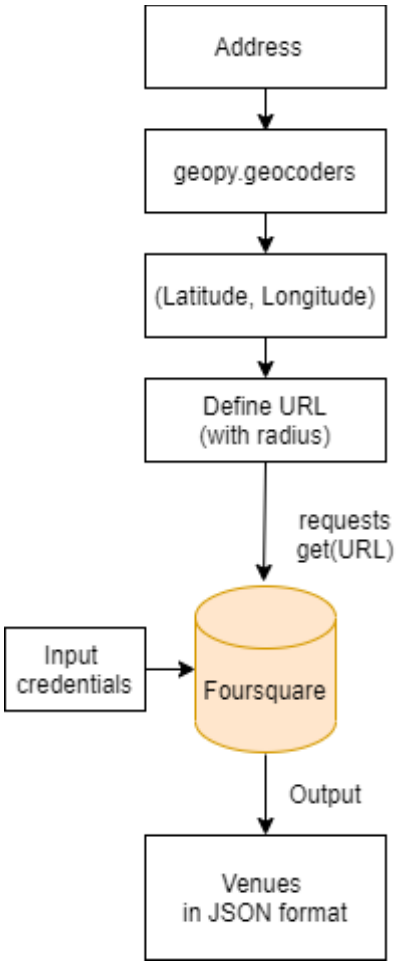


Figure 10: Fetching venues using Foursquare API

The venues fetched in JSON format is properly formatted and displayed in tabular form. Then, according to the availability of each category of venues, they are sorted in descending order to get the list of top ten categories of venues.

	Venues	Total Count
0	Asian Restaurant	5
1	Restaurant	5
2	Café	4
3	Hotel	4
4	Coffee Shop	3
5	Bar	2
6	Food	1
7	Garden	1
8	Himalayan Restaurant	1
9	Hostel	1
10	Bakery	1

Figure 11: Example of top 10 categories of venues within 500m radius of Kathmandu fetched by Foursquare

Candidate filtering

The company after receiving applications from many students regarding any job vacancy can shortlist the students based on their skill, qualification and experience required for the job in order to reduce the time for recruitment. They can also download the CVs of shortlisted students.

Registration/ login

This component is responsible for allowing the users to safely register and log in into the system. During the process of registration or login, it is very important to make sure authentic users are using the system. The number of attempts allowed to make to login into the system have been set to four. Also, for new users the time interval to verify the email has been set to a maximum one hour. Failing to do so will result in suspension of the account in both the cases which could only be activated by the Admin. Also, the password strength is checked during the registration process and during the password change.

CV generator

This component is built for students. It is used to generate CV using the information provided by students during profile creation. Five attractive templates have been provided using which CV can be generated. The generated CV can be downloaded in PDF format as well to use it outside this system. The information displayed on CV would be basic information, education, experience, projects, hobbies, certifications and trainings, skills and languages. Company, if finds the profile of any student interesting can also download his/her CV in PDF format for future reference. This feature saves time and money of students which they spent on various platform for building an attractive CV.



Figure 12: Example of a generated CV

Mini blog

This feature is provided to students which allow them to add short description of what projects or research they are working on. The companies can view these pieces of information when a student apply for the job to their company or when the company search for candidates using search component. Very limited functionalities have been provided with this component to keep it simple and yet useful.

Job recommender

Many jobs are posted in an e-recruitment system on a daily basis by different companies requiring different skills for different jobs. Every student/candidate doesn't possess same set of skills. It would be difficult for them to find the right kind of jobs suited to their skill. So, a recommendation system is a must for any e-recruitment system. Student does not need make search every time they login into the system. In fact, the system itself recommend some jobs to the students based on the skills they have mentioned on their profile. The system first matches the skills of students with the skill required for the job posted and if matched, it will then look for level of skills i.e. beginner, skilled, advanced or expert, required for the job and also possessed by students. The jobs whose both requirements matched with the profile of students will be displayed on top in the recommendation section. The recommendation section is displayed on the home page of students.

Company/ Skill verification

This component is developed for department (Admin). It allows admin to verify the identity of the companies and allow them to post jobs into the system. Also, a separate account is created for project supervisor within admin with limited access control, to verify the skills demonstrated by students during the project or research work. The student having verified skill will have an added advantage. As it is verified by the authorized body, companies can trust that student with those verified skills. It will help companies in taking decision during the process of hiring.

5. Results and discussion

An e-recruitment system we tried to build to solve the existing problems in our own campus, has been implemented successfully. The main motive of building this system was to overcome the shortcomings of traditional method of hiring and provide fair chance to every student to showcase their skills and grab the job opportunities that comes their way. It also helped to bring both the employers and the job seekers on common platform and establish direct communication channel between them.

5.1. Surveys

We will here present some very useful outcomes we got through the various surveys and questionnaires we conducted.

Survey 1

In this survey, we gave two sets of questionnaires were given to the participant. The purpose of this survey was to find the experiences of participants regarding the traditional method of hiring and also commercial job portals. we conducted this survey among 300 students of our campus (mainly comprising the third- and fourth-year students from the faculty of bachelor in computer, electronics and communication, electrical and civil engineering) and here are the results we got.

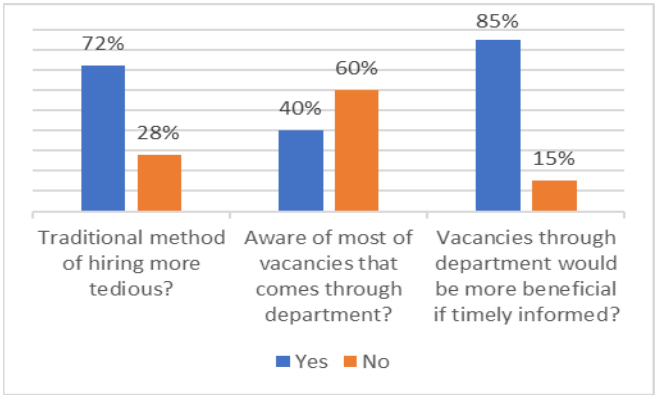


Figure 13: Responses of survey 1.1

The first questionnaire (survey 1.1) comprised 3 questions as shown in Figure 13 We found 72% participants felt traditional method of hiring is more tedious. Only 40% were aware of most of the vacancies that came through department and the other 60% either never came to know about some of the vacancies or knew when they were already expired. Despite that, 85% felt vacancies through the department would be more beneficial if proper communication channel is established between the companies and the students.

The second questionnaire (survey 1.2) comprised 3 questions. The responses are shown in Figure 14.

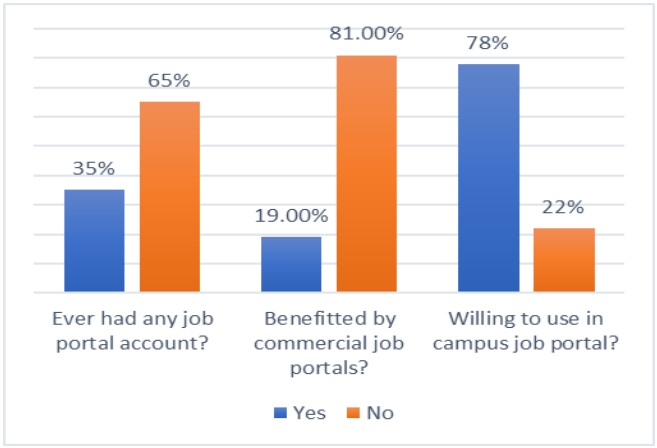


Figure 14: Responses of survey 1.2

In this survey, we found only 35% students had used any job portal website till date. Among them only 19% found worth using it. And 78% were willing to use the job portal site if build only for campus students. They felt it would be more beneficiary for students than those commercial sites that already exists in the market, as this on-campus job portal sites would provide even playing field to all the candidates.

Survey 2

This survey had one questionnaire comprising two questions. The number of participants were 200. Only the fourth-year bachelor's students from the faculty of computer, electronics and communication, electrical and civil engineering were involved.

Result for the question 1: "**What level of job do you look for in the 4th year of your study?**" is shown in Figure 15.

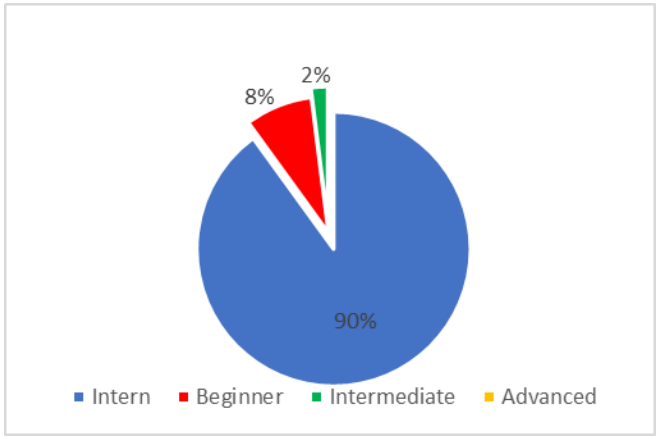


Figure 15: Responses to first question of survey 2

We can see 90% students look for intern position when they are in the fourth year of bachelor's study, as internship is the first step towards the journey of employment. There are some students who start working since the first or second year of bachelor. It is the result of this that in the fourth year, they start looking for beginner or intermediate level of jobs. According to our survey, 8% look for beginner level, 2% for intermediate level and no one for advanced level.

Result for the question 2: "**From which year do you start looking for jobs?**" is shown in Figure 16.

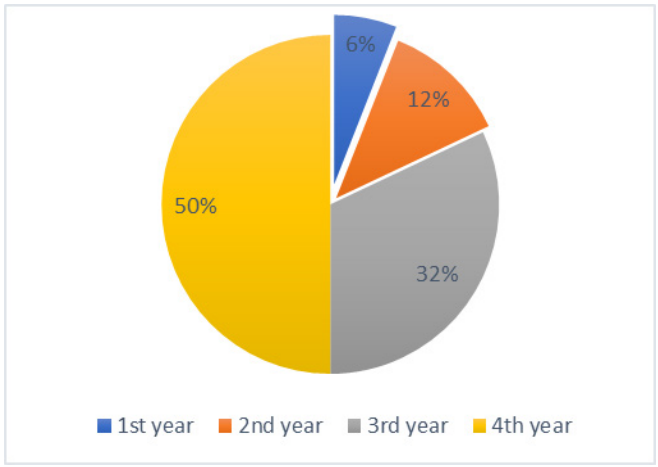


Figure 16: Responses to second question of survey 2

From the Figure 16, we can see 50% of students start looking for jobs from 4th year onwards. Similarly, 32% from 3rd year, 12% from 2nd year and only 6% from 1st year. This result shows as students' progress towards the end of their study, they more actively start looking for jobs. Especially, the numbers increase rapidly towards third and fourth year.

Survey 3

This survey was conducted after the system was brought in use to know if it has served its purpose for which it was developed. 280

students participated in this survey comprising the third- and fourth-year students from the faculty of bachelor in computer, electronics and communication, electrical and civil engineering

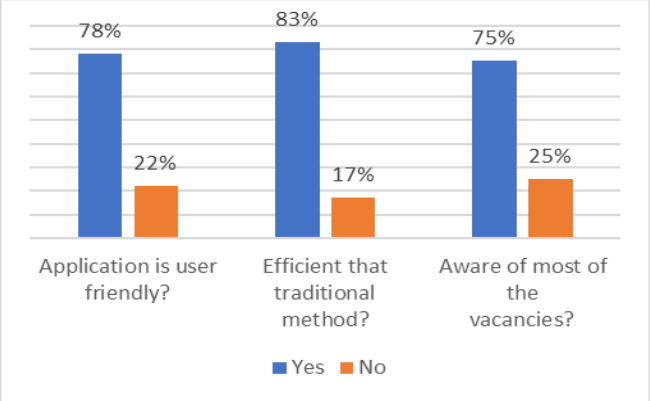


Figure 17: Responses to survey 3

From Figure 17, we can see 78% students found this application user friendly and 83% found it more efficient than the traditional method of hiring. Similarly, 75% who used this system were aware of most of the jobs posted, of their concern. This is actually a good number as it was only 40% in case of traditional method. Looking at these numbers, we feel the system has at least served its purpose to some extent.

Analysis and visualization of number of vacancies received over the period of one-year (i.e. from 2018/08/01 to 2019/07/31) using choropleth map:

5.2. Province wise analysis

From the Table 1 and Figure 18, we can see that Bagmati Province is way ahead in terms of generating employment opportunities. Then comes Province 5, Province 2 and Province 1 respectively. The remaining two provinces i.e. Karnali and Sudurpashchim provinces have not contributed in job openings through this system over the given period of time. This is because many industries and services sectors are situated in the Bagmati province. Also, its contribution to the GDP of the country is much higher than the other provinces.

Table 1: Number of vacancies across provinces of Nepal (Tabular representation)

Province	Number
Bagmati	77
5	15
2	12
1	11
Gandaki	4
Karnali	0
Sudurpashchim	0

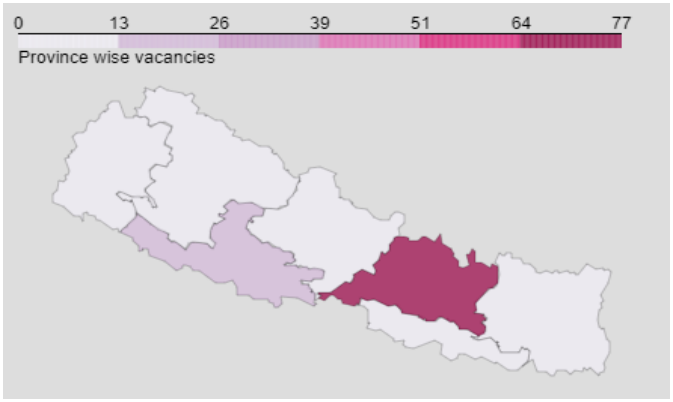


Figure 18: Number of vacancies across provinces of Nepal (Graphical representation)

5.3. District wise analysis

From the Table 2 and Figure 19, we can see Kathmandu district has the highest job openings over the period of one year. Also, we received job vacancies from only 16 out of 77 districts of Nepal. We received zero vacancy from the districts not mentioned in the Table 2. If we exclude Kathmandu and Lalitpur from the list we barely received vacancies from the other districts. This shows the concentration of economy within only few districts of Nepal.

Table 2: Number of vacancies across districts of Nepal (Tabular representation)

District	Number
Kathmandu	42
Lalitpur	24
Rupandehi	8
Morang	7
Dhanusa	5
Chitwan	5
Dang	5
Bhaktapur	4
Kaski	4
Sunsari	3
Parsa	3
Mahottari	2
Bara	2
Makwanpur	2
Kapilvastu	2
Bhojpur	1

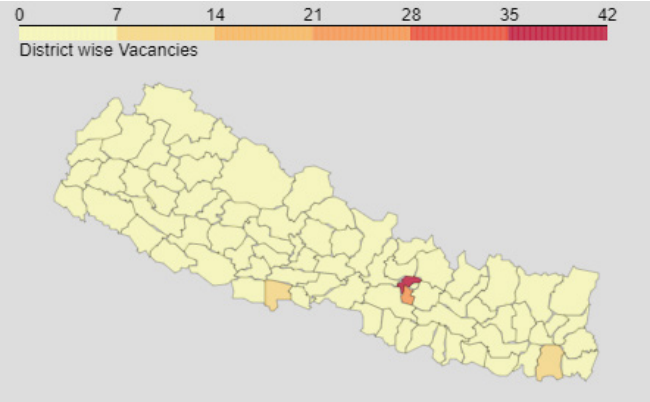


Figure 19: Number of vacancies across districts of Nepal (Graphical representation)

Visualization (using word cloud) of various types of skills that were required for the jobs we received over the period of one year (i.e. from 2018/08/01 to 2019/07/31):



Figure 20: Word cloud of skills that were mentioned for job vacancies

We can visualize in Figure 20 that skills like PHP, SQL, HTML, Python, Web design, Graphics design and so on were more in demand at least over that one-year period.

6. Conclusions and recommendations

We decided to build this e-recruitment system seeing the problems that were faced by the students of our campus. The main idea was to help students overcome the communication gap with the companies and to provide fair chance to every student in the campus to showcase their talent and skills in order to grab any job opportunity that comes their way. And to achieve this, we thought of building a system which could bring both the students and the companies on a single platform so that they can communicate directly and in an organized manner.

This system fulfills many of its purpose. But as it is said nothing is perfect, this system has also some shortcomings which we hope to improve in the upcoming future. We can make a provision to read the information of CV in PDF form, when uploaded to the system instead of enforcing the new users to fill the information manually. We can add blogging feature with advanced functionalities to develop it as a social hub where one can help others in finding a job. We can add online test features to verify the skills mentioned by students in their profile. We can improve the job recommendation using some advanced machine learning algorithms. The notification features informing the students about the job match can be improved. Some inbuilt data analysis tools can be added to provide insights on the current job scenario. Also, feature regarding the sharing of CV on different platforms through this platform can be added.

References

[1] V. Yadav, U. Gewali, S. Khatri, S.R. Rauniyar, A. Shakya, "Smart Job Recruitment Automation: Bridging Industry and University," in International Conference on Artificial Intelligence for Transforming Business and Society, AITB **2019**, 1-6, 2019, doi:10.1109/AITB48515.2019.8947445.

[2] E. Galanaki, "The decision to recruit online: A descriptive study," Career Development International, **07**, 243-251, 2002, doi:10.1108/13620430210431325.

[3] O. Shafique, "Recruitment in the 21 st century," Interdisciplinary Journal of Contemporary Research in Business, **04**(02), 887-901, 2012.

[4] J. Dorn, T. Naz, Integration of Job Portals by Meta-search, 401-412, 2007, doi:10.1007/978-1-84628-858-6_44.

[5] L. Khanam, M.A. Uddin, M.A. Mahfuz, "Students' Behavioral Intention and Acceptance of E-recruitment System: A Bangladesh Perspective," Proceedings of the 12Th International Conference on Innovation and Management, 1297-1303, 2015.

[6] E. Parry, S. Tyson, "An analysis of the use and success of online recruitment methods in the UK," Human Resource Management Journal, **18**(03) 257-274, 2008, doi:10.1111/j.1748-8583.2008.00070.x.

[7] M.G.G. Ventura, R.P. Bringula, "Effectiveness of online job recruitment system : evidence from the university of the east," IJCSI International Journal of Computer Science Issues, **10**(04), 152-159, 2013.

[8] A. Matas, "Internet becomes an on-line opportunity for employers and job hunters," The Washington Post, H2, Nov. 7, 1993.

[9] R. Ramaabaanu, & M. Saranya, "Importance and Problems of E-Recruitment," International Journal of Research (IJR), **01**(09), 2014.

[10] A.J. du Plessis, H. Frederick, "Effectiveness of e-recruiting: empirical evidence from the Rosebank business cluster in Auckland, New Zealand," Science Journal of Business Management, **2012**, 1-19, 2012, doi:10.7237/sisa/126.

[11] S. Mauno, U. Kinnunen, M. Ruokolainen, "Job demands and resources as antecedents of work engagement: A longitudinal study," Journal of Vocational Behavior, **70**(01), 179-171, 2007, doi:10.1016/j.jvb.2006.09.002.

[12] M.O. Fred, U.M. Kinange, "Effectiveness of E-Recruitment in Organization Development," Management and Economic Journal, **02**(04) 2018, 294-301.

[13] B. Raghavendra, C. Priyanka, G. Chaithanya, "A study on E recruitment,"

International Journal of Advance Research in Science and Engineering, **07**(03), 2018.
[14] R. Rani, "E-Recruitment and its impact upon job seekers: A contemporary approach," International Journal of Advanced Research and Innovative Ideas in Education, **02**(04), 2016.
[15] Y. Gupta, "Literature review on E-Recruitment: A step towards paperless HR," International Journal of Advance Research in Computer Science and Management Studies, **04**(01), 2016.
[16] F. Ramadhani, M. Zarlis, "Analysis of e-Recruitment System Design," International Journal of E-Education, e-Business, e-Management and e-Learning, **09**, 38–45.
[17] M.R. Karim, M.S. Miah, A. Khatun, "E-Recruitment in Practice: A Study on Jobseekers' Perception in Bangladesh," Global Disclosure of Economics and Business, **04,** 2015, doi:10.18034/gdeb.v4i1.150.
[18] R. et al Patel, "E-Recruitment Challenges and Methods used in Modern Organization," International Journal of Advance Research in Computer Science and Management Studies, **04**(12), 2016.
[19] P. Manjare, R. Munde, "An Effective Job Recruitment System Using Content-based Filtering," International Research Journal of Engineering and Technology(IRJET), **04**(03) 2017.
[20] B.R. Poudel, "Online Recruitment: A Cognitive Perspective of Job Seekers in Nepal," Journal of Business and Social Sciences, **02**(01) 2018, doi:10.3126/jbss.v2i1.22823.
[21] K. Kukich, "Techniques for Automatically Correcting Words in Text," ACM Computing Surveys (CSUR), 1992, doi:10.1145/146370.146380.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 904-911 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Fast Adaptive Time-delay-estimation Sliding Mode Controller for Robot Manipulators

Dang Xuan Ba*

Department of Automatic Control, Ho Chi Minh City University of Technology and Education (HCMUTE). Ho Chi Minh City, 71307, Vietnam

ARTICLE INFO

Article history:
Received: 14 August, 2020
Accepted: 25 November, 2020
Online: 08 December, 2020

Keywords:
Adaptive time-delay estimation
Gain adaptation
Robotic manipulators
Sliding mode control

ABSTRACT

A nonlinear adaptive robust controller is proposed in this paper for trajectory-tracking control problems of robot manipulators. On one hand, to effectively approximate the systematic dynamics, a simple time-delay estimator is first adopted. On the other hand, to minimize the control error, the controller is designed based on a sliding mode structure using the obtained estimation results. A fast learning mechanism is then proposed for automatically tuning control gains. Another proper adaptation law is furthermore developed to support the nominal inertia-matrix selection of the time-delay estimation. Effectiveness of the closed-loop system is intensively discussed using Lyapunov-based constraints and extended simulation results.

1. Introduction

Nowadays, robots play a crucial role in modern industry thanks to excellent performances in conducting complicated, repetitive, and dangerous tasks [1–4]. Day by day, high-precision intelligent robots are required to fulfill the given demands [2], [5]. To design such controllers, two important steps of the controller design are taken into account: nonlinearities and uncertainties in system models as well as unexpected effects of complex working environments have to be tackled to push the closed-loop system in a stationary state; in the second step, a driving robust control signal is generated to force the control objective stabilizing around origin [6–8].

Robot dynamics could be derived using Newton-Euler analyses, Lagrange formulations or virtual decomposition methods [6,9]. Although general systematic models could be obtained, but they contain lots of uncertainties due to particular structures of robots and disturbances [10,11]. To cope with the mathematical problems, soft-computing approaches are reasonable solutions [7,8]. Nonlinear uncertain dynamics could be well estimated by fuzzy-logic approximators and neural networks but heavy-computation burdens along with lots of tuning parameters are new obstacles in real-time applications. The time-delay estimation (TDE) is increasingly adopted as one of the most effective tools to reckon unknown dynamics in modern control, owing to the simplicity in design and easiness in real-time implementation [12,13]. This technique uses measured information of the acceleration and control signals to result in offset terms of the dynamics based on predefined nominal values of inertia matrices [5,14]. Effectiveness of the TDE methods has been gradually confirmed in both theoretical proofs and experimental validations [15,16].

To effectively complete control objectives in industrial applications, a proper control strategy with impressive adaptation and robustness ability is next required [17–19]. The sliding mode control (SMC) technique has a special attraction to designers thanks to the simplicity, widespread applications, and robustness to unknown system dynamics [20–23]. However, excellent control performances of the SMC controllers may be hard to maintain in different working cases. As a solution, gain-adaptive SMC schemes have been recently noted for servo systems [18,24–26]. Driving gains or robust gains were actuated by various learning mechanisms to effectively force the control objective to zero as fast as possible [25,26]. In these learning schemes, only one degree-of-freedom learning law is proposed, and the control error is stabilized to a bound around zero [17,24]. In another direction, intelligent approaches based on Type-2 fuzzy logic theory could be found as possible solutions of the gain adaptation [27,28]. Solving optimal problems of the closed-loop system under robustness constraints in a horizontal time yields the expected gains along with excellent control performances [29]. However, intensive behaviors of the gain variation that could judge the real-time control quality were ignored.

This paper is an extension of work originally presented in

*Corresponding Author: Dang Xuan Ba, HCMC University of Technology and Education (HCMUTE), Ho Chi Minh city, Vietnam, badx@hcmute.edu.vn

2019 International Conference on System Science and Engineering (ICSSE) [1]. In the previous work, both driving and robust gains of a conventional TDE sliding mode controller were actively self-tuned subjecting to a minimal control error. Its effectiveness was confirmed by intensive simulation results. The controller was operated with a fixed nominal value of the inertia matrix of the system dynamics. Improperly selecting this term could lead to instability of the closed-loop system [15]. An excellent control performance is obviously associated with a perfect value of the matrix [16]. Nevertheless, the best nominal matrix is not same for different running states. As a sequence, a dynamical learning mechanism for such the term is the main motivation of this research for maintaining high control precision in various working conditions.

In this study, an adaptive TDE sliding mode controller is developed for trajectory-tracking control of serial robot manipulators with the following contributions:

- The system behaviors are first approximated by a basic time-delay estimator (TDE). To support selection of a proper inertia matrix in the estimation process, an adaptation rule is designed to search the optimal nominal one.
- The control objective is then realized by a robust sliding mode control method. To provide the flexibility in tuning the control gains in divergent operation cases, a fast gain-learning law is proposed. The adaptation mechanism is worked as a damping low-pass filter that could yield smooth responses of the closed-loop system. Different from the previous work [1], nonlinear leakage functions are employed to speed up the learning process.
- Control performances of the overall system are analyzed by a Lyapunov-based approach and comparative simulation results.

The rest of this paper is structured as follows. Section 2 presents system modeling and problem statement. Section 3 describes design of the proposed controller incorporated with the adaptation laws. Section 4 discusses validation results, and the paper is then concluded in Section 5

2. System Dynamics and Problem Statement

Dynamics of a *n*-joint robotic manipulator are generally expressed as follows [6,9,17,24]:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) + \mathbf{f}(\dot{\mathbf{q}}) + \boldsymbol{\tau}_\mathbf{d} = \boldsymbol{\tau} \quad (1)$$

where $\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}} \in \Re^n$ are the joint angles, velocities, and accelerations, respectively, $\boldsymbol{\tau} \in \Re^n$ is the joint torques or control inputs, $\mathbf{M}(\mathbf{q}) \in \Re^{n\times n}$ is the inertia matrix, $\mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} \in \Re^n$ is the Centripetal/Coriolis vector, $\mathbf{g}(\mathbf{q}) \in \Re^n$ is the gravitational term, $\mathbf{f}(\dot{\mathbf{q}}), \boldsymbol{\tau}_\mathbf{d} \in \Re^n$ is the frictional torque, and $\boldsymbol{\tau}_\mathbf{d} \in \Re^n$ denotes disturbances.

The dynamics (1) could be simply rewritten in a following form:

$$\bar{\mathbf{M}}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{v}(\mathbf{q}) = \boldsymbol{\tau} \quad (2)$$

where $\bar{\mathbf{M}} \in \Re^{n\times n}$ is an proper nominal diagonal inertia matrix, $\mathbf{v} = (\mathbf{C}\dot{\mathbf{q}} + \mathbf{g} + \mathbf{f} + \boldsymbol{\tau}_\mathbf{d} + (\mathbf{M} - \bar{\mathbf{M}})\ddot{\mathbf{q}}) \in \Re^n$ is a lumped term that is defined as a combination of other dynamic terms such as the Centripetal/Coriolis, gravity, inertia variation, frictions, disturbances, modeling errors and unmodeled terms.

Remark 1: Assume that the system outputs $(\mathbf{q},\dot{\mathbf{q}})$ are measurable and a desired trajectory $(\mathbf{q_d})$ is bounded and differentiable up to the second order. The main objective of this paper is to find out an appropriate control signal to drive a tracking error of the system output $(\mathbf{q})$ and the given profile $(\mathbf{q_d})$ to zero or as small as possible. However, maintaining high control quality of the closed-loop system is a not-easy yet interesting work due to existence of the uncertain nonlinear dynamics and complicated working environments. Additional characteristics of the controller are required to be model-free, adaptive, and robust with high control performances.

3. An Adaptive Time-delay Nonlinear Controller

To accomplish the aforementioned features, the motion controller is built up based on a simple sliding mode framework incorporated with an adaptive time-delay estimator. Fast learning laws are then integrated to search proper control parameters. The proposed controller is sketched in **Figure 1.**

3.1. Simple Sliding Mode Control

The main control error is mathematically defined as

$$\mathbf{e} = \mathbf{q} - \mathbf{q_d} \quad (3)$$

A sliding manifold is employed as an indirect control objective of the closed-loop system:

$$\mathbf{s} = \dot{\mathbf{e}} + \boldsymbol{\Lambda}\mathbf{e} \quad (4)$$

where $\boldsymbol{\Lambda}$ is a selected positive-definite diagonal gain matrix.

For obtaining high-precision control results, the controller is designed with a model-free compensator - that could cope with the systematic uncertain nonlinearities and disturbances - and a robust control signal, as follows:

$$\boldsymbol{\tau} = \boldsymbol{\tau}_\mathbf{MOD} + \boldsymbol{\tau}_\mathbf{SMC} \quad (5)$$

Here, $\boldsymbol{\tau}_\mathbf{MOD}$ is a model-dynamical eliminator that is used to compensate for unexpected effect of the offset vector ($\mathbf{v}$) on the control behaviors and other terms of the surface dynamics (4). The signal is thus selected as follows:

$$\boldsymbol{\tau}_\mathbf{MOD} = \hat{\mathbf{v}} + \hat{\mathbf{M}}(\ddot{\mathbf{q}}_\mathbf{d} - \boldsymbol{\Lambda}\dot{\mathbf{e}}) \quad (6)$$

Note that the estimate $(\hat{\mathbf{v}})$ of the offset dynamics ($\mathbf{v}$) is computed using the basic time-delay estimator [15,22]:

$$\hat{\mathbf{v}} = \boldsymbol{\tau}_{t-T_s} - \hat{\mathbf{M}}\ddot{\mathbf{q}}_{t-T_s} \quad (7)$$

where the subscript $(t-T_s)$ is a time-delay value, T_s is the sampling time, and $\hat{\mathbf{M}} = \text{diag}(\hat{\mathbf{m}}) \in \Re^{n\times n}$ is estimate of the nominal matrix $\bar{\mathbf{M}}$.

The remaining term $\boldsymbol{\tau}_{\mathbf{SMC}}$ is designed to drive the sliding manifold from the stationary state back to around zero and stabilize thereafter by dealing with a model perturbance $\boldsymbol{\delta}$:

$$\boldsymbol{\tau}_{\mathbf{SMC}} = -\mathbf{K}_1\mathbf{s} - \mathbf{K}_2\,\text{sgn}(\mathbf{s}) \tag{8}$$

where $\mathbf{K}_{i|i=1,2} = \text{diag}(\mathbf{k}_i) = \text{diag}([k_{i1};...;k_{in}])$ is positive-definite diagonal gain matrices.

Remark 2: With the robust sliding mode design, excellent control outcome can be resulted in by a proper selection of the control gains $(\mathbf{K}_1, \mathbf{K}_2)$ and nominal inertia matrix $(\hat{\mathbf{M}})$ regardless of the existence of the disturbances. It is however sometimes a challenge to perfectly tune the gains to achieve good transient performance and especially ensure high control accuracies for unpredictable execution conditions in the real-time control.

3.2. Gain-adaptation Laws

To treat difficulty of the gain selection, the strategy for gain tuning is proposed in that the control gains $(\mathbf{K}_i|_{i=1,2})$ are separated into two terms: nominal gains $(\bar{\mathbf{K}}_i|_{i=1,2})$ and variation gains $(\breve{\mathbf{K}}_i|_{i=1,2})$. The nominal ones are chosen for stability of the closed-loop system. The variation ones are automatically tuned for the desired transient performance.

A fast adaptation rule for the automatic control-gain tuning is designed as follows:

$$\begin{cases} \dot{\breve{\mathbf{k}}}_1 = \boldsymbol{\Xi}_k\,\text{diag}(\mathbf{s})\mathbf{s} - \text{diag}(\boldsymbol{\eta}_1)\left(\mathrm{I}+\text{diag}\left(\mathbf{s}^2\right)\right)^{-1}\text{sat}\left(\breve{\mathbf{k}}_1\right) \\ \dot{\breve{\mathbf{k}}}_2 = \boldsymbol{\Xi}_k\,\text{diag}\left(\text{sgn}\left(\mathbf{s}\right)\right)\mathbf{s} - \text{diag}\left(\boldsymbol{\eta}_2\right)\left(\mathrm{I}+\text{diag}\left(\mathbf{s}^2\right)\right)^{-1}\text{sat}\left(\breve{\mathbf{k}}_2\right) \end{cases} \tag{9}$$

where $\boldsymbol{\Xi}, \text{diag}\left(\boldsymbol{\eta}_1\right)$ and $\text{diag}\left(\boldsymbol{\eta}_2\right)$ are positive-definite diagonal constant matrices.

On the other hand, properly choosing the nominal inertia matrix could reduce bound of the system perturbance ($\boldsymbol{\delta}$). To efficiently support this task, a new learning law is proposed as follows:

$$\dot{\hat{\mathbf{m}}} = \boldsymbol{\Xi}_m\,\text{diag}(\mathbf{s})\left(\ddot{\mathbf{q}}_{t-T_s} - \ddot{\mathbf{q}}_{\mathbf{d}} + \boldsymbol{\Lambda}\dot{\mathbf{e}}\right) - \text{diag}\left(\boldsymbol{\eta}_m\right)\text{diag}\left(\mathbf{s}^2\right)\hat{\mathbf{m}} \tag{10}$$

where $\boldsymbol{\Xi}_m$ and $\text{diag}\left(\boldsymbol{\eta}_m\right)$ are positive-definite diagonal constant matrices.

3.3. Stability Analysis

The theoretical effectiveness of the closed-loop system is investigated using the following theorem:

Theorem 1: Employing the model-free sliding mode control signals (3)-(8) to the robot dynamics (2), in which the driving-robust gains and nominal inertia matrix are automatically updated using the laws (9)-(10), asymptotic convergence of the closed-loop system is resulted in if the following condition holds:

$$\begin{cases} \min\left(\text{eig}\left(\bar{\mathbf{K}}_1\right)\right) > \max\left(\text{eig}\left(0.25\boldsymbol{\Xi}_m^{-1}\text{diag}(\boldsymbol{\eta}_m)\bar{\mathbf{M}}\bar{\mathbf{M}}\right)\right) \\ \bar{k}_{2i|i=1..n} \geq \max\left(|\delta_i|\right) \end{cases} \tag{11}$$

Proof:

The time derivative of the sliding manifold (4) along the system (2) is expressed

$$\dot{\mathbf{s}} = \bar{\mathbf{M}}^{-1}\left(\boldsymbol{\tau} - \mathbf{v}\right) - \ddot{\mathbf{q}}_{\mathbf{d}} + \boldsymbol{\Lambda}\dot{\mathbf{e}} \tag{12}$$

By substituting the control signal (5)-(8) into the dynamics (12), we have

$$\dot{\mathbf{s}} = \bar{\mathbf{M}}^{-1}\left(\boldsymbol{\delta} + \tilde{\mathbf{M}}\left(\ddot{\mathbf{q}}_{\mathbf{d}} - \ddot{\mathbf{q}}_{t-T_s} - \boldsymbol{\Lambda}\dot{\mathbf{e}}\right) - \mathbf{K}_1\mathbf{s} - \mathbf{K}_2\,\text{sgn}(\mathbf{s})\right) \tag{13}$$

where $\boldsymbol{\delta} = [\delta_1;\delta_2;...;\delta_n] = \bar{\mathbf{v}} - \mathbf{v} = \left(\boldsymbol{\tau}_{t-T_s} - \bar{\mathbf{M}}(\mathbf{q})\ddot{\mathbf{q}}_{t-T_s} - \mathbf{v}\right) \in \Re^n$ is the system perturbance mentioned above, and $\tilde{\mathbf{M}} = \text{diag}(\tilde{\mathbf{m}}) = \hat{\mathbf{M}} - \bar{\mathbf{M}} \in \Re^{n\times n}$ is the estimation error of the nominal matrix $\bar{\mathbf{M}}$.

Now, we consider a Lyapunov function:

$$V = 0.5\mathbf{s}^T\bar{\mathbf{M}}\boldsymbol{\Xi}_k\mathbf{s} + 0.5\sum_{i=1}^{2}\left(\breve{\mathbf{k}}_i^T\breve{\mathbf{k}}_i\right) + 0.5\tilde{\mathbf{m}}^T\boldsymbol{\Xi}_k\boldsymbol{\Xi}_m^{-1}\tilde{\mathbf{m}} \tag{14}$$

Substituting (9), (10), and (13) into the time derivative of the Lyapunov function leads to

$$\begin{aligned} \dot{V} &= \mathbf{s}^T\boldsymbol{\Xi}_K\left(\boldsymbol{\delta} - \bar{\mathbf{K}}_1\mathbf{s} - \bar{\mathbf{K}}_2\,\text{sgn}(\mathbf{s})\right) \\ &\quad - \sum_{i=1}^{2}\left(\breve{\mathbf{k}}_i^T\,\text{diag}(\boldsymbol{\eta}_i)\left(\mathrm{I}+\text{diag}\left(\mathbf{s}^2\right)\right)^{-1}\breve{\mathbf{k}}_i\right) \\ &\quad - \tilde{\mathbf{m}}^T\boldsymbol{\Xi}_k\boldsymbol{\Xi}_m^{-1}\text{diag}(\boldsymbol{\eta}_m)\text{diag}\left(\mathbf{s}^2\right)\tilde{\mathbf{m}} \\ &\quad + \tilde{\mathbf{m}}^T\boldsymbol{\Xi}_k\boldsymbol{\Xi}_m^{-1}\text{diag}(\boldsymbol{\eta}_m)\text{diag}\left(\mathbf{s}^2\right)\bar{\mathbf{m}} \end{aligned} \tag{15}$$

From the condition (11), there always exist two positive constants α_k and α_m such that

$$\begin{aligned} \dot{V} &= -\alpha_k\mathbf{s}^T\boldsymbol{\Xi}\bar{\mathbf{K}}_1\mathbf{s} - \sum_{i=1}^{2}\left(\breve{\mathbf{k}}_i^T\,\text{diag}(\boldsymbol{\eta}_i)\left(\mathrm{I}+\text{diag}\left(\mathbf{s}^2\right)\right)^{-1}\breve{\mathbf{k}}_i\right) \\ &\quad - \alpha_m\tilde{\mathbf{m}}^T\boldsymbol{\Xi}_k\boldsymbol{\Xi}_m^{-1}\text{diag}(\boldsymbol{\eta}_m)\text{diag}\left(\mathbf{s}^2\right)\tilde{\mathbf{m}} \le 0 \end{aligned} \tag{16}$$

Theorem 1 has been proven:

Remark 3: As seen in Figure 1 as well as stated in Theorem 1**,** the controller is just designed under a simple TDE sliding mode control framework to ensure the stability of the overall system**.** However, control performance of the system is significantly improved by injecting excellent adaptation laws for the control gains and nominal inertia matrix. The proposed controller hence possesses strong robustness, model-free, and self-learnable ability.

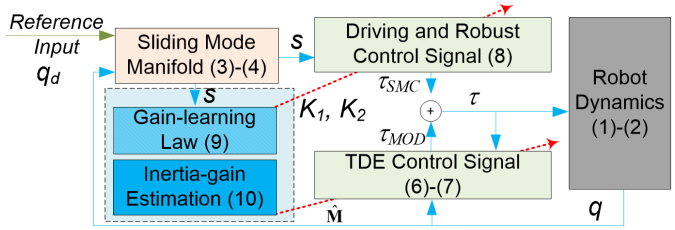


Figure 1: An overview of the designed controller.

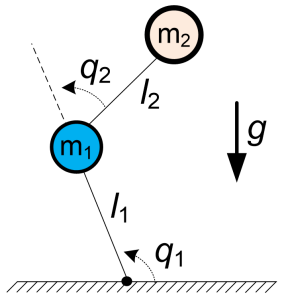


Figure 2: Configuration of a 2DOF manipulator.

Table 1: Selected Parameters for the simulation model

Description	Parameters	Values	Unit
Link length	l_1, l_2	0.2, 0.1	m
Gravitational Accel	g	9.81	m/s2
Friction coefficient	a_1, a_2	1	
Mass of links	m_1, m_2	10, 5	kg

4. Simulation Validation

4.1. Setup

To assess tracking performance of the proposed controller, extended simulations were carried out on a 2DOF robot. Configuration of the robot is depicted in Figure 2. The robot was selected to work in the vertical plane in which the gravity force was critically affected on the system performance. Dynamics of the robot are presented as follows [6,24]:

$$\mathbf{M} = \begin{bmatrix} m_2(l_2^2 + l_1^2) + 2l_1 l_2 m_2 \cos(q_2) + m_1 l_1^2 & m_2 l_2^2 + l_1 l_2 m_2 \cos(q_2) \\ m_2 l_2^2 + l_1 l_2 m_2 \cos(q_2) & m_2 l_2^2 \end{bmatrix}$$

$$\mathbf{C}\dot{\mathbf{q}} = \begin{bmatrix} -m_2 l_1 l_2 \sin(q_2)\dot{q}_2 (\dot{q}_2 + 2\dot{q}_1) \\ m_2 l_1 l_2 \sin(q_2)\dot{q}_1^2 \end{bmatrix}$$

$$\mathbf{g} = \begin{bmatrix} m_2 l_2 g \cos(q_1 + q_2) + l_1 g \cos(q_1)(m_1 + m_2) \\ m_2 l_2 g \cos(q_1 + q_2) \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} a_1 \dot{q}_1 \\ a_2 \dot{q}_2 \end{bmatrix}$$

where $q_{1,2}$ are the robot joint angles. The detailed parameters of the robot model are given in **Table 1**.

Another robust-gain-learning TDE controller (RLTDE) was also applied to the same system as a benchmark for evaluating the effectiveness of the proposed controller.

The design of the RLTDE controller was referred as in the previous work [24], as follows:

$$\begin{cases} \mathbf{s} = \dot{\mathbf{e}} + \mathbf{\Lambda}_c \mathbf{e} \\ \boldsymbol{\tau} = \bar{\mathbf{M}}\left(-\ddot{\mathbf{q}}_{t\text{-}T_s} + \bar{\mathbf{M}}^{-1}\boldsymbol{\tau}_{t\text{-}T_s} + \ddot{\mathbf{q}}_\mathbf{d} - \mathbf{\Lambda}_c \dot{\mathbf{e}} - \mathbf{K}_{c1}\mathbf{s} - \mathbf{K}_{c2}\,\mathrm{sgn}(\mathbf{s})\right) \\ \dot{\hat{k}}_{c2i} = \begin{cases} \varphi_i \left(\alpha_i^{-1}|s_i(t)|\right)^{\theta(t)} \theta(t), \text{if } \hat{k}_{c2i} > 0 \\ \varphi_i \alpha_i^{-1}|s_i(t)| \qquad \text{otherwise} \end{cases} \\ \theta(t) = \mathrm{sgn}\left(\|\mathbf{s}\|_\infty - \varepsilon\right) \end{cases}$$

The basic control gains of the two controllers were inherently selected as [24]:

$$\begin{cases} T_s = 0.001, \mathbf{\Lambda}_c = \mathrm{diag}([10;10]); \\ \boldsymbol{\alpha} = [0.6;1.57], \boldsymbol{\varphi} = [4000;3000]; \varepsilon = 0.015 \end{cases}$$

For a pair comparison, the RLTDE controller was operated with four distinct parameter sets as given in Table 2, denoted as RLTDE1, RLTDE2, RLTDE3 and RLTDE4, respectively.

Table 2: Different Control Parameters Selected for the RLTDE Controller Implementation

Name	Nominal Inertia Matrix ($\bar{\mathbf{M}}$)	Control Gain ($\mathbf{K}_{c1}$)
RLTDE1	diag([0.01;0.01])	diag([1;1])
RLTDE2	diag([0.04;0.04])	diag([1;1])
RLTDE3	diag([0.01;0.01])	diag([60;60])
RLTDE4	diag([0.04;0.04])	diag([60;60])

Besides, the additional parameters of the proposed controller were manually tuned as follows:

$$\begin{cases} T_s = 0.001; \mathbf{\Lambda} = \mathrm{diag}([10;10]), \\ \bar{\mathbf{K}}_2 = \mathrm{diag}([0.5;0.5]), \bar{\mathbf{K}}_1 = \mathrm{diag}([1;1]); \\ \mathbf{\Xi}_k = \mathrm{diag}([300;300]); \boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 = [1;1] \\ \mathbf{\Xi}_m = \mathrm{diag}([0.003;0.003]); \boldsymbol{\eta}_m = [3\times10^{-4};3\times10^{-4}]. \end{cases}$$

4.2. Simulation Results

In the first scenario, reference inputs of the robot joints were selected to be sinusoidal signals plotted in Figure 3. By applying the controllers to the system, the simulation results obtained are presented in Figures 4 and 5.

As seen in Figure 4, the RLTDE1 provided very good control performances, 0.53 (RMS deg) and 0.44 (RMS deg) of the control errors at joints 1 and 2, thanks to employment of the strong robust gain ($\mathbf{K}_2$) generated by a nonlinear learning algorithm [24]. Behaviors of the robust gain adaptation are demonstrated in Figure 5. The key purpose of using such the gain in the RLTDE control method was to attenuate effect of the incomplete TDE estimation. Figure 4 also reveals that the control performances were further improved if the more proper control gains were found out. When increasing only the nominal inertia matrix ($\bar{\mathbf{M}}$), or only the driving gain ($\mathbf{K}_{c1}$), or both of them ($\bar{\mathbf{M}}$ and $\mathbf{K}_{c1}$), higher

control accuracies were exhibited such as: the RLTDE2 controller obtained 0.17 (RMS deg) and 0.12 (RMS deg) of the control errors at joints 1 and 2, respectively; the RLTDE3 controller got 0.2 (RMS deg) and 0.11 (RMS deg) ones; and the RLTDE4 controller provided 0.09 (RMS deg) and 0.06 (RMS deg) ones. It seems larger the control gains would create better the control performance. In fact, structures of the RLTDE and proposed controllers are almost same.

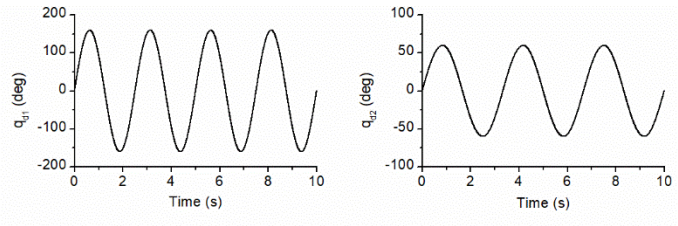


Figure 3: Reference inputs of the robot joints in the first simulation.

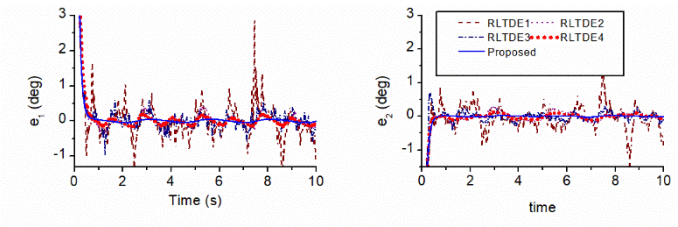


Figure 4: Control errors obtained in the first simulation

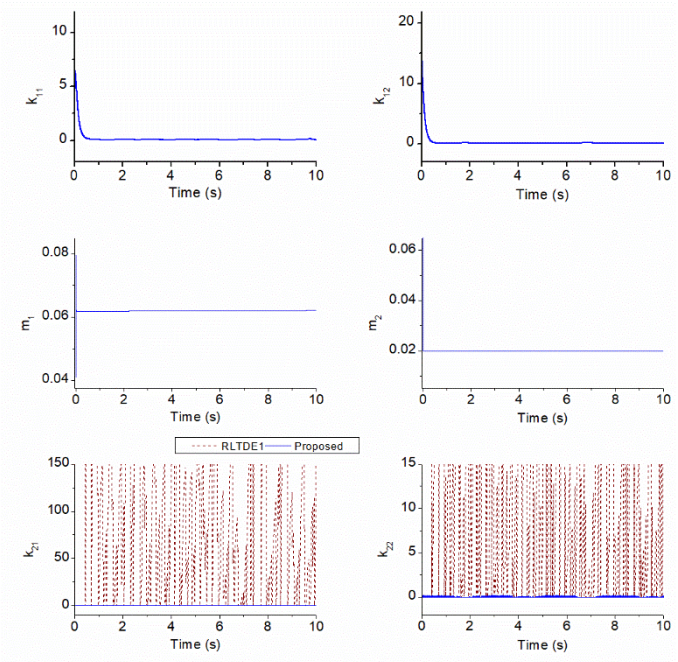


Figure 5: Comparative gains learning in the first simulation.

The core difference is that the proposed controller adopts the adaptation rules for the driving-robust gains and optimal nominal inertia matrix. Properly searching an optimal set of the driving gain and nominal inertia matrix could effectively reduce bound of the gain ($\mathbf{K}_2$) that could be clearly observed in Figure 5. Hence, as shown in Figure 4, the highest control precision, 0.03 (RMS deg) and 0.014 (RMS deg) of the control errors at joints 1 and 2, was resulted in by the proposed controller. Also noted in Figure 5, the optimal values did not need to be so large for the best control performance: the variation of the driving gain converged back to zero and the stationary nominal inertia matrix was found to be *diag(0.062; 0.021)*. This implies the flexibility of the proposed controller over the previous one.

In the second scenario, the robot joints were controlled to track smooth step signals. Applying the same controllers to the system, the simulation results are presented in Figure 6, 7 and 8.

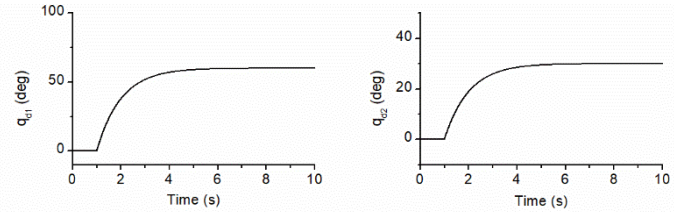


Figure 6: Reference signals of the robot joints in the second simulation.

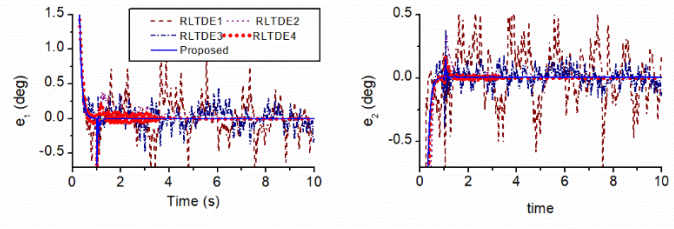


Figure 7: Control errors obtained in the second simulation.

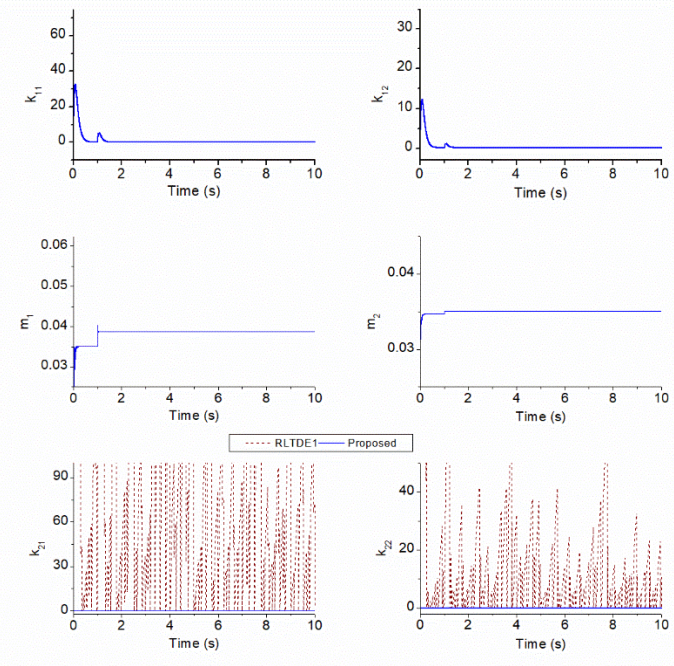


Figure 8: Comparative gains learning in the second simulation.

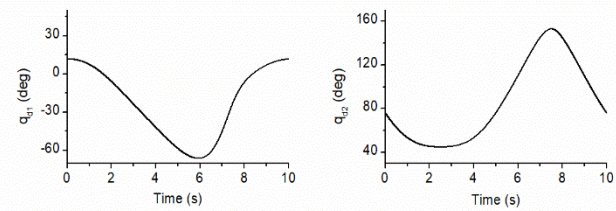


Figure 9: Desired joint angles computed from the eclipse trajectory of the end-effect in the third simulation.

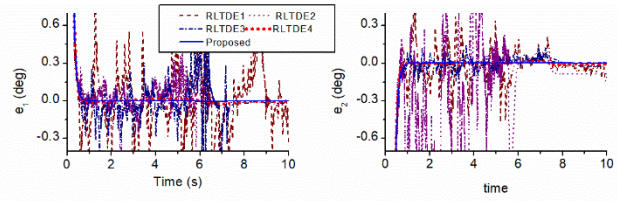


Figure 10: Control errors obtained in the third simulation.

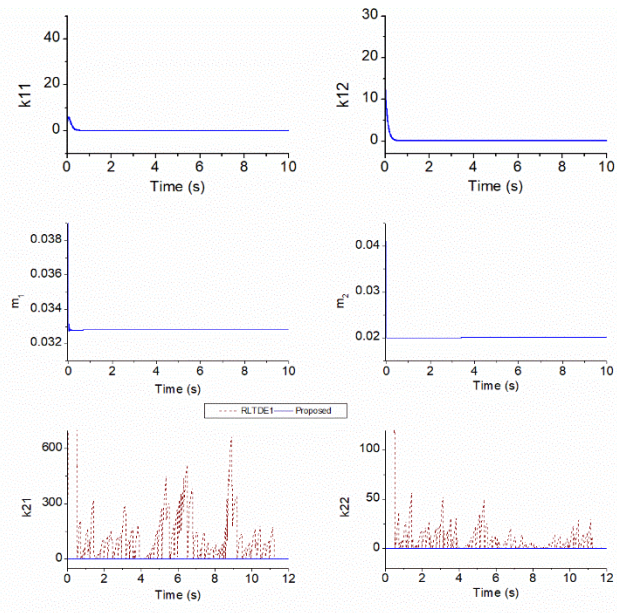


Figure 11: Comparative gains learning in the third simulation.

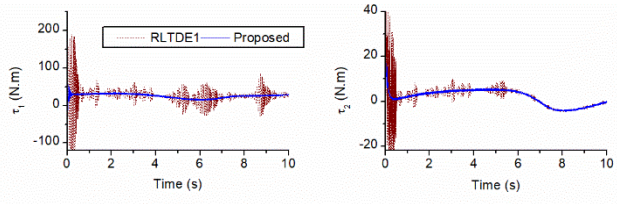


Figure 12: Control signal generated in the third simulation.

As shown in Figure 7, the control performances of the RLTDE1, 1.1 (MA deg) and 0.73 (MA deg) of the control errors at joints 1 and 2, were degraded as comparing to those of the other controllers. An improper selection of the nominal inertia matrix would make the systematic perturbance ($\boldsymbol{\delta}$) become large. Meanwhile, the driving gain ($\mathbf{K}_{c1}$) was fixed with a small value, it means that the control performance was covered only by the robust gain ($\mathbf{K}_{c2}$). Another reason came from the special learning mechanism of the RLTDE: the robust gain ($\mathbf{K}_{c2}$) was respectively increased and decreased as the sliding surfaces were outside and inside of specific regions. It implies that the RLTDE control outcome in this case depended on the regions predefined. New features integrated in the proposed controller could make the system avoid the aforementioned problems: as observed in Figure 8, the adaptation laws of the inertia matrix and control gains worked well to ensure the closed-loop system stabilize at origin or force the control error to be as small as possible. Hence, outperformance of the proposed controller, 0.005 (MA deg) and 0.01 (MA deg) of the control errors at joints 1 and 2, over the RLTDE1, 2, and 3 has been exhibited in this simulation scenario. Interestingly, once the perfect values of the driving gain ($\mathbf{K}_{c1}$) and nominal inertia matrix ($\bar{\mathbf{M}}$) were used, the RLTDE4 obtained an excellent control accuracy: 0.002 (MA deg) and 0.0004 (MA deg) of the control errors at joints 1 and 2.

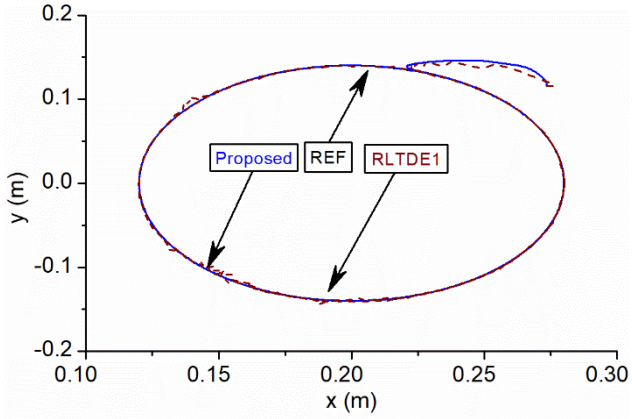


Figure 13: Comparison of control results in Cartesian space

In the third verification case, the end-effector of the robot was controlled to track an eclipse trajectory. By applying inverse kinematic computation for the testing robot, the desired joint angles are displayed in Figure 9. Tracking control errors of the controllers for the new trajectory are plotted in Figure 10. The control results indicate that the RLTDE controllers tried to automatically adjust their gains to minimize the control error. However, using only one degree-of-freedom adaptation of the control gains was not adequate to result in the minimal sliding manifold even though they were supported with various control gains and nominal inertia matrices fine-tuned. The best control quality, 0.024 (MA deg) and 0.015 (MA deg) of the control errors at joints 1 and 2, were accomplished by the RLTDE4 which had just shown the outstanding performance in the second simulation. The control accuracy was significantly improved, 0.004 (MA deg) and 0.005 (MA deg) of the control errors at joints 1 and 2, by using three degree-of-freedom in the gain learning proposed by this paper.

Indeed, control input data generated by the controllers shown in Figure 12 reveal that sharing the robustness burden of the control mission into three learning spreads provided a faster response and a smaller control error. Its effectiveness and feasibility are demonstrated more clearly by plots of the comparative end-effector positions obtained by the proposed controller and a certain RLTDE one, as shown in Figure 13.

4.3. Discussion

The key properties of the proposed adaptation laws could be evidently expressed by data shown in Figure 5, 8 and 11: in transient time, the closed-loop system needed strong power to drive the control objective back to desired position regardless of uncertain nonlinearities and disturbances. To serve this mission, the driving and robust gains were promptly activated under the designed constraint (9), and the optimal nominal gain matrix was also searched according to the strategy (10). After that, the whole system would relax in steady-state time with the small-activated gains. By further considering these figures, much high-frequency oscillation could be observed in the learning gains of the RLTDE1

controller. In real-time operation with smooth control signals, the proposed controller could maintain the lifetime of the actuators.

Table 3: Statistical Evaluation of the Control Performances in the Simulation Scenarios

Control Error		Joint 1		Joint 2	
		MA	RMS	MA	RMS
Test 1	RLTDE1	2.87	0.53	1.98	0.44
	RLTDE2	0.6	0.17	0.26	0.12
	RLTDE3	0.9	0.2	0.38	0.11
	RLTDE4	0.22	0.09	0.11	0.06
	Proposed	0.05	0.03	0.04	0.014
Test 2	RLTDE1	1.1	0.25	0.73	0.2
	RLTDE2	0.09	0.03	0.024	0.01
	RLTDE3	0.48	0.13	0.18	0.06
	RLTDE4	0.002	0.0004	0.0004	0.0001
	Proposed	0.005	0.001	0.01	0.007
Test 3	RLTDE1	2.02	0.4	0.47	0.09
	RLTDE2	0.39	0.08	1.17	0.2
	RLTDE3	2.2	0.27	0.13	0.03
	RLTDE4	0.024	0.01	0.015	0.009
	Proposed	0.004	0.0018	0.005	0.002

Table 3 summarizes static values of the control errors using maximum absolute (MA) and root-mean-square (RMS) computation for a specific time (5s to 12s). As seen in the table and by recalling the second simulation results, if the optimal set of the control gains and nominal inertia matrix were known, an exceptional control performance would be achieved. However, the optimal one was not same in different working conditions. Instead of searching the optimal values with time consuming, the designed controller integrated an automatic gain-tuning algorithm for suboptimal ones. The statistical data in Table 3 show that the proposed controller yielded the excellent MA and RMS results as comparing to the RLTDE ones in most of the simulation cases. Hence, the influence of such the suboptimal gains and inertia matrices on the control performance was acceptable. Here, advantages of the proposed controller are confirmed throughout both the analytical and validation results.

5. Conclusion

In this paper, a new adaptive robust control method is developed for position tracking-control problems of robot manipulators. The controller design is mainly based on the sliding mode control framework. Effect of the systematic uncertain nonlinearities and disturbances are well treated by an adaptive time-delay estimator. The driving-robust control signal is then employed for realizing the control objective. Fast gain-learning laws are properly designed to increase flexibility of the controller in various working conditions. The control gains are automatically tuned to provide both the desired transient performance and the steady-state behavior. Effectiveness of the closed-loop system is intensively investigated by theoretical proofs and comparative simulations.

Conflict of Interest

The author declares no conflict of interest.

References

[1] D.X. Ba, M.-H. Le, "Gain-learning sliding mode control of robot manipulators with time-delay estimation," in 2019 International Conference on System Science and Engineering (ICSSE), 478–483, 2019.

[2] Y. Park, I. Jo, J. Lee, J. Bae, "A Dual-cable Hand Exoskeleton System for Virtual Reality," Mechatronics, **49**(November 2017), 177–186, 2018, doi:10.1016/j.mechatronics.2017.12.008.

[3] T.L. Nguyen, S.J. Allen, S.J. Phee, "Direct torque control for cable conduit mechanisms for the robotic foot for footwear testing," Mechatronics, **51**(August 2016), 137–149, 2018, doi:10.1016/j.mechatronics.2018.03.004.

[4] G. Chen, B. Jin, Y. Chen, "Nonsingular fast terminal sliding mode posture control for six-legged walking robots with redundant actuation," Mechatronics, **50**(August 2016), 1–15, 2018, doi:10.1016/j.mechatronics.2018.01.011.

[5] M. Tavakoli, P. Lopes, L. Sgrigna, C. Viegas, "Motion control of an omnidirectional climbing robot based on dead reckoning method," Mechatronics, **30**, 94–106, 2015, doi:10.1016/j.mechatronics.2015.06.003.

[6] J.J. Craig, Introduction to Robotics: Mechanics and Control, 3rd ed., Pearson Prentice Hall, USA, 2005.

[7] L. Wang, T. Chai, L. Zhai, "Neural-network-based terminal sliding-mode control of robotic manipulators including actuator dynamics," IEEE Transactions on Industrial Electronics, **56**(9), 3296–3304, 2009, doi:10.1109/TIE.2008.2011350.

[8] Z. Liu, G. Lai, Y. Zhang, C.L.P. Chen, "Adaptive Fuzzy Tracking Control of Nonlinear Time-Delay Systems with Dead-Zone Output Mechanism Based on a Novel Smooth Model," IEEE Transactions on Fuzzy Systems, **23**(6), 1998–2011, 2015, doi:10.1109/TFUZZ.2015.2396075.

[9] W.-H. Zhu, Virtual Decomposition Control: Toward hyper degrees of freedom robots, Springer-Verlag Berlin Heidelberg, 2010.

[10] S. Islam, X.P. Liu, "Robust Sliding Mode Control for Robot Manipulators," IEEE Transactions on Industrial Electronics, **58**(6), 2444–2453, 2011.

[11] W. Kim, C.C. Chung, "Robust output feedback control for unknown non-linear systems with external disturbance," IET Control Theory and Applications, **10**(2), 173–182, 2015, doi:10.1049/iet-cta.2014.1299.

[12] T.C. Hia, S. Jung, "A Simple alternative to Neural Network Control Scheme for Robot Manipulators," IEEE Transactions on Industrial Electronics, **42**(4), 414–416, 1995.

[13] S. Kim, J. Bae, "Force-mode Control of Rotary Series Elastic Actuators in a Lower Extremity Exoskeleton using Model-inverse Time Delay Control (MiTDC)," IEEE/ASME Transactions on Mechatronics, **22**(3), 1392–1400, 2017, doi:10.1109/TMECH.2017.2687979.

[14] T.C. Hia, "A New Technique for Robust Control of Servo Systems," IEEE Transactions on Industrial Electronics, **36**(1), 1–7, 1989.

[15] K. Youcef-toumi, O. Ito, "A Time Delay Controller for Systems With Unknown Dynamics," Journal of Dynamic Systems, Measurement, and Control, **112**(1), 133–142, 1990.

[16] Y. Wang, D. Yu, Y. Kim, "Robust Time-Delay Control for the DC – DC Boost Converter," IEEE Transactions on Industrial Electronics, **61**(9), 4829–4837, 2014.

[17] M. Jin, J. Lee, N.G. Tsagarakis, "Model-free Robust Adaptive Control of Humanoid Robots with Flexible Joints," IEEE Transactions on Industrial Electronics, **64**(2), 1706–1715, 2016, doi:10.1109/TIE.2016.2588461.

[18] F. Mazenc, P.-A. Bliman, "Backstepping Design for Time-Delay Nonlinear Systems," IEEE Transactions on Automatic Control, **51**(1), 149–154, 2006.

[19] C. Hu, B. Yao, Q. Wang, "Performance-Oriented Adaptive Robust Control of a Class of Nonlinear Systems Preceded by Unknown Dead Zone With Comparative Experimental Results," IEEE Transactions on Industrial Electronics, **18**(1), 178–189, 2013.

[20] V.I. Utkin, "Sliding Mode Control Design Principles and Applications to Electric Drives," IEEE Transactions on Industrial Electronics, **40**(1), 23–36, 1993.

[21] J. Zhang, W.X. Zheng, "Design of Adaptive Sliding Mode Controllers for Linear Systems via Output Feedback," IEEE Transactions on Industrial Electronics, **61**(7), 3553–3562, 2014.

[22] M. Zhihong, A.P. Paplinski, H.R. Wu, "Robust MIMO Terminal Sliding Mode Control Scheme for Rigid Robotic Manipulators," IEEE Transactions on Automatic Control, **39**(12), 2464–2469, 1994.

[23] A. Pisano, M. Rapaic, E. Usai, Sliding modes after the first decade of the 21st century, Springer-Verlag Berlin Heidelberg: 412, 2012.

[24] J. Baek, M. Jin, S. Han, "A New Adaptive Sliding Mode Control Scheme for Application to Robot Manipulators," IEEE Transactions on Industrial Electronics, **63**(6), 3628–3637, 2016, doi:10.1109/TIE.2016.2522386.
[25] Y. Shtessel, M. Taleb, F. Plestan, "A novel adaptive-gain supertwisting sliding mode controller: Methodology and Application," Automatica, **48**(5), 759–769, 2012, doi:10.1016/j.automatica.2012.02.024.
[26] C. Xia, G. Jiang, W. Chen, T. Shi, "Switching-Gain Adaptation Current Control for Brushless DC Motors," IEEE Transactions on Industrial Electronics, **63**(4), 2044–2052, 2016.
[27] M.H. Khooban, T. Niknam, F. Blaabjerg, M. Dehghani, "Free chattering hybrid sliding mode control for a class of non-linear systems : electric vehicles as a case study," IET Science, Measurement and Technology, **10**(7), 776–785, 2016, doi:10.1049/iet-smt.2016.0091.
[28] N.M. Tri, D.X. Ba, K.K. Ahn, "A gain-adaptive intelligent nonlinear control for an electrohydraulic rotary actuator," Internation Journal of Precision Engineering and Manufacturing, **19**(5), 665–673, 2018.
[29] M.H. Khooban, N. Vafamand, T. Niknam, T. Dragicevic, F. Blaabjerg, "Model Predictive Control based on T-S Fuzzy model For Electrical Vehicles Delayed Model," IET Electric Power Applications, **11**(5), 918–934, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 912-919 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Investigation of the Effect of FBG Profiles, Temperature and Transmission Distance for Environmental Sensing & Monitoring

Muhammad Arif Riza, Yun Ii Go*

School of Engineering & Physical Sciences, Heriot-Watt University Malaysia, Putrajaya, 62200, Malaysia

ARTICLE INFO

Article history:
Received: 29 August, 2020
Accepted: 19 November, 2020
Online: 08 December, 2020

Keywords:
Apodization
Environmental
FBG
Sensor
Simulation
Temperature

ABSTRACT

Optical sensors exist in various forms and one of them is Fiber Bragg Gratings (FBG). Performance evaluation was simulated for FBG sensing. The simulation involves different grating profiles tested with complete optical sensing system. The system involves a broadband light source, FBG of various profiles and optical spectrum analyzer for data interpretation. Different profile of gratings imprinted on an FBG outputs wavelength spectra and simulated results satisfy this theory obtained from literatures. Temperature sensing was also simulated for the FBG at various temperature ranges that suit different industrial demands. At low temperatures FBG does give noticeable sensing capabilities. When reaching 400°C and above, the FBG was still capable of providing response, however when in real environments the fiber may suffer from thermal damage. FBG performance when integrated within a system with long distances of fibers were also simulated. Power attenuation was noticed at the reflectivity spectra.

1. Introduction

This paper is an extension of work originally presented in "IEEE International Conference on Sensors and Nanotechnology 2019" [1]. Environmental monitoring have been a main topic in most fields of study. Conventional sensors nowadays such as microelectronic mechanical sensors (MEMS) or probe sensors are vulnerable to electromagnetic (EM) interference and wiring complications. This phenomena prevents its use in long-range sensing applications [2]. Fiber optic sensors is a promising alternative to conventional sensors due to its compact nature, EM proofing and multiplexing capabilities [3]. The simplicity of optical based sensors allows various sensing configurations such as absorption based sensors (interferometers) and grating based sensors. Fiber Bragg Grating (FBG) sensor is an example of an optical sensor that operates by creating a periodic perturbation of light signals within the refractive index of its core. The gratings serves as wavelength filter that permits only light with wavelengths that dissatisfy the Bragg's equation to permeate through while the light waves that satisfy the Bragg equation will be reflected back. Therefore, strains acting on the gratings with a set periodicity can induce a shift in the wavelength spectra. The shift in wavelength defines the condition of the parameter being sensed [4]. Commonly, FBG sensors are used for pressure and thermal sensing. However, FBG sensors can be modified to sense other forms of parameter such as humidity by introducing a sensitive coatings onto the sensitized regions of the sensor.

Operations of FBG sensors require a source of propagated broadband light. In some cases, FBG sensors are positioned in remote areas where power input availability is scarce. Normally, the light source are powered by conventional batteries which could have a short lifespan and can be inconvenient maintenance wise. Hence, a sustainable power source is desired. A promising source of green energy is from ambient vibrations within the earth's crust [5]. The concept of vibration energy harvesting involves converting amplified vibrations at resonance into useful electrical power by means of kinetic energy distribution. Ambient vibration frequencies generally fall below 100.0 Hz. However, in civil structures, the frequency can drop to less than 10.0 Hz [6]. The acceleration level of these vibrations are very low being less than 1.0 g where g = 9.81 ms^{-2}, making it more suitable for low-power electronic application such as powering a light source.

This paper focuses on reporting simulated results of FBG tested with various different grating profiles, temperature sensing and transmission distance. Whilst simulating the FBG in different grating configurations, basis of the system which will be explained further sections will be tailored to a vibrational energy harvester power output. Since the proposed harvester design provides a low but steady power output to act as a light source, the whole FBG system will be run under low power to be simulated within various

*Corresponding Author: Yun Ii Go, Email: y.go@hw.ac.uk

conditions including thermal and attenuation. It is an opportunity to simulate how well the system of FBG can be run with sufficient power with different types of gratings which differs from most other closely similar articles involving FBG simulation [7-9].

1.1. Optical vs Conventional Sensor

Environment sensors that are still in use comes in various types. Electrical based sensor types such as the resistive probes, p-n junction diodes and MEMS sensors have been the staple sensor devices used over the years in monitoring various parameters [10]. Some sensors that have a very low power consumption of <1.0 mW and are made to be self-sustained by equipping the sensor with a solar panel harvester [11]. However, the operation these conventional sensors require electrical flow on the active sensitized regions of the sensor. This will render the sensors to be susceptible to electromagnetic interference and damage due to short circuit. Besides having a high maintenance cost, the complexity of the sensing mechanism in electrical based sensors available in market limits its capability to be embedded within a structure [12]. These drawbacks propel researchers to find alternatives such as optical fiber sensors.

Optical sensors alone can have many different configurations based on how the modification is formed on the surface, cladding, core or tip of the fiber. In addition to FBG sensors, other optical sensors include interferometers and absorption based sensors, which function based on the principal of the energy absorption spectrum and refractive index changes. However, grating based sensors possess an advantage over other optical based sensors in terms that they are more compact and ease of modification. In addition, FBG sensors have low noise, high sensitivity and a high multiplexing imprint capabilities via phase masking, which helps provide a large mapped sensing area within an environment [13].

1.2. FBG Sensing Principle

FBG are made typically from phase masking, where periodic modulations of the refractive index are imprinted within the fiber core by intense UV laser, exposed perpendicularly to a single mode optical fiber. The modulated refractive indexes within the fiber core that has different refractive index values than the core and cladding acts as the wavelength filter for the incident propagated light. Based on Figure 1, light signals at wavelengths other than the Bragg wavelength would pass through whereas light signals with wavelength equal to the Bragg wavelength that would be reflected back. Therefore, the transmitted ray that passes the grating will have the Bragg wavelength absent in the main spectrum [14]. FBGs can be serially imprinted into one fiber at multiple sensitized regions (multiplexing). This allows the FBGs to have multiple functional sensor nodes along a fiber length which is essential for area mapping.

On their own, FBGs are strain, pressure and temperature sensitive. In temperature sensing, different temperatures would impose axial strain upon the FBG fiber based on coefficient of thermal expansion of the FBG material (which would normally be silica). This would cause a change in the pattern spectrum of the reflected light, resulting in a shift in the reflected wavelength. This shift in wavelength can be calibrated to identify temperature change [15]. Detection of non-physical parameters requires the FBG to be modified by coating the sensitized region with a reversibly reactive material nanostructures that can induce mechanical strain or reaction to the FBG when exposed to a measurand [16]. Generally, it is possible for a FBGs to sense chemical parameters such as salinity concentration [8], chemical gas concentration [17] and relative humidity of the environment [18]. Ease of modification of FBG to suit various environmental sensing applications makes FBG a viable sensing device to be studied and optimized.

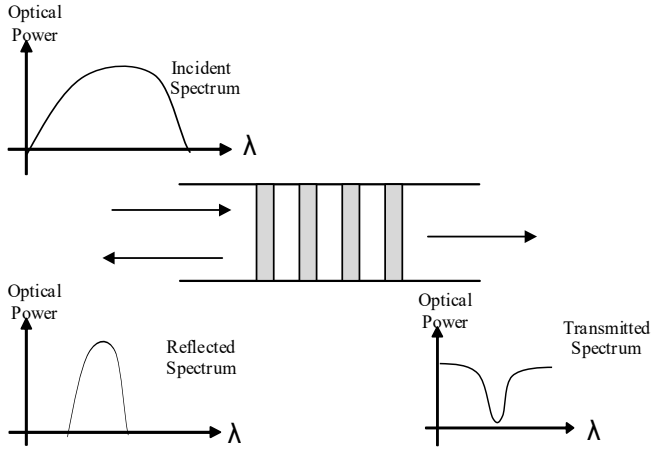


Figure 1: Principle of FBG.

2. Methodology

In this study, the behaviour of FBG fiber is simulated with different grating profiles. Different grating profiles can have an effect to the output transmission and reflectivity of the fibers. The behaviour of the FBG as a temperature sensor are simulated in temperature ranges that relates to certain industries. Connection of fibers in a long range fiber setup are also simulated to show the viability of FBG sensor array assembly in a long-range network.

2.1. Overall System Description

FBG sensors typically involve several components to be available for environmental monitoring. The following subsections explain each of the components in Figure 2 that are mandatory for FBG sensor implementation. Table 1 presents the constant parameters used in this study.

Table 1: System Setting.

Source Spec.	Optical Fibers	FBG Spec.	OSA Spec.
Broadband white light source Frequency: 193.1 THz Power: 130 dBM	Silica SMF Length: 10 cm	Grating profile: Uniform Bragg wavelength: 1550 nm IM: 0.0002	Resolution: 0.001 nm

2.2. Source

Optical based devices require sources in form of propagated photons. Light source of broad spectrum is preferred when using FBG. Broad spectrum light source examples include white light source, amplified spontaneous emission (ASE) source and broadband light source.

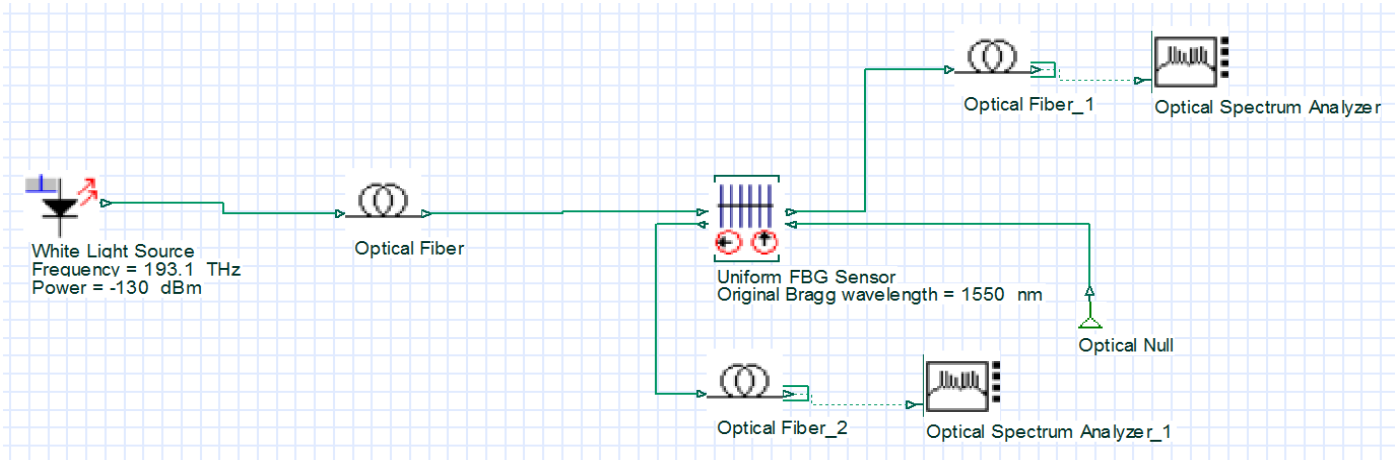


Figure 2: Arrangement of FBG setting.

2.3. FBG

The FBG component is the main component in the simulation that has variables manipulated. The Bragg grating wavelength chosen is 1550 nm since it is among the commonly used wavelength for FBG sensing. The grating length of fiber is set constant at 10 mm throughout the study. Index modulation is set at 0.0002 for the FBG and is also a controlled parameter. The grating profile will be varied in order to observe the effects of different grating profiles on the transmission and reflectivity spectra in the simulation.

2.4. OSA

Optical spectrum analyser or optical receiver is mandatory for any optical based sensors including FBG as it detects the incoming light propagation. The optical receiver then sends the received light signals to be converted into readable data. In lab-based study, OSA typically combine these functions of receiving and signal conversion into transmission and reflectivity spectrum. The OSA settings used in this simulation is with resolution of 0.001 m. The resolution value allows for more data points to be recorded brings more accurate results.

3. Results and Discussion

3.1. Effect of Grating Profiles

There are many different grating profiles of an FBG. These grating profiles may affect the spectrum of both the transmission and reflectivity. The grating profiles are made different with the periods or arrangement of the grating follows different model. Different phase mask can be used for the imprinting of different grating profiles. In this study, Uniform grating along with several other grating profiles is tested.

3.1.1. Uniform Profile

Uniform grating profiles is the most common imprinted FBG profiles that has seen its use in sensing. The uniform periodicity profile affects light modulation and follows the Braggs equation which is:

$$\lambda_B = 2n_{eff} \cdot \Lambda_G \quad (1)$$

where, λB is the Bragg wavelength, n_{eff} is the refractive index of the propagating light and ΛG is the index modulation periodicity [19]. In this study, light modulation across a uniform profile FBG has been simulated. In Figure 3 transmission and reflectivity of light propagated pass the profiles has been recorded. As predicted, the response have showed that the peak power exists on the 1550 nm wavelength. This is the wavelength at which the grating was imprinted on. Figure 2(a) showed the transmission power trough occurring and on Figure 2(b), reflectivity spectra showed peaks at the same Bragg wavelength. This simulation has demonstrated the regular uniform FBG profiles connected with regular power source. There appears to be minor side lobes occurring on Figure 2(b). These side lobes may not be favourable for FBG sensors that are tailored for high precision sensing. However, the side lobes does not hinder the uniform profiles from entering the market as region or array monitoring system.

3.1.2. Chirped Profile

A different profile than the typical uniform profile of FBG is the chirped FBG. The gratings in a chirped FBG varies in terms of periodicity along the grated regions depending on the chirp pattern on the phase mask. This produces spectrums that are different than the original, uniform and non-apodized FBGs. The wavelength spectrums produced via chirped FBGs are much broader than uniform profiled FBGs which can range between several nanometers or even thousands of nanometers depending on the chirp rate, grating length and phase mask used for writing the FBG in a fiber [20].

Based on Figure 4, chirped FBG with uniform apodization and follows similar Bragg wavelength, index modulation and grating length as table 1, has been simulated. It was seen that both Figure 4(a) and Figure 4(b) showed broader spectrum that uniform FBG. No shifts of transmission peaks and reflectivity troughs was seen which is expected. The breadth of the peaks is observed to be approximately 4 nm. Both transmission power and reflectivity shows no deterioration when chirp profile is applied without altering other parameters. Figure 4 has successfully demonstrated a chirp profile via simulation that satisfy with other studies [21, 22]. While having wider Bragg peaks in both reflectivity and transmission, chirped FBGs can have certain suitable purposes.

However, if high accuracy detection is preferred, chirped FBG may not be the best grating profile to be utilised. This is due to small shifts that occur upon detection may not be visible or detected and could lead to misinterpretation at the end signal-to-data conversion.

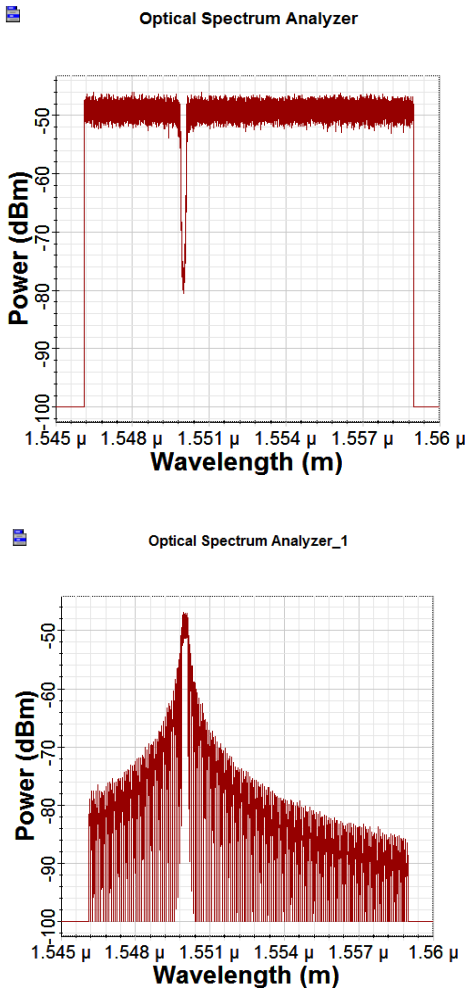
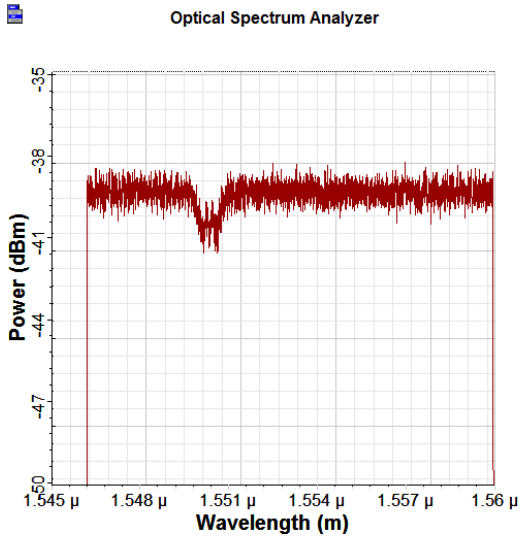
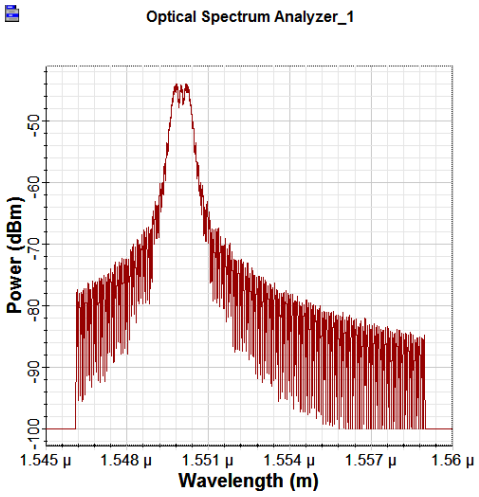


Figure 3: Optical power spectrum of uniform FBG (a) Transmission and (b) Reflectance.

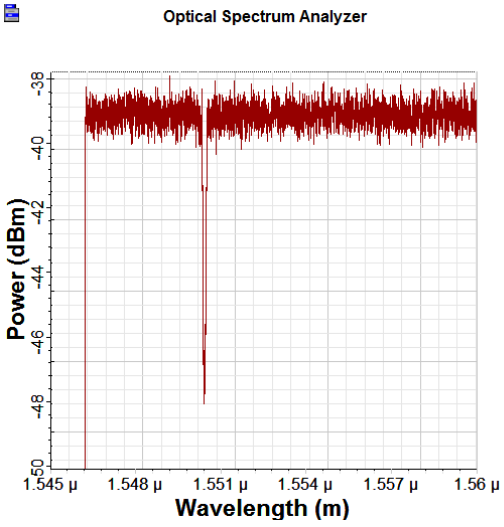


(a)

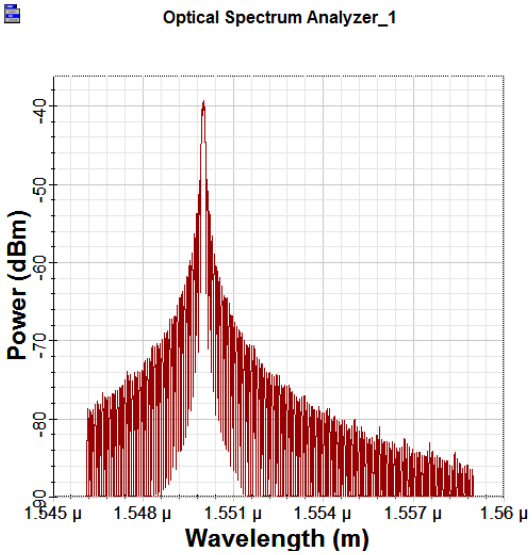


(b)

Figure 4: Optical power spectrum of linear chirped FBG (a) Transmission and (b) Reflectivity.



(a)



(b)

Figure 5: Optical power spectrum of Gaussian apodized FBG (a) Transmission and (b) Reflectivity.

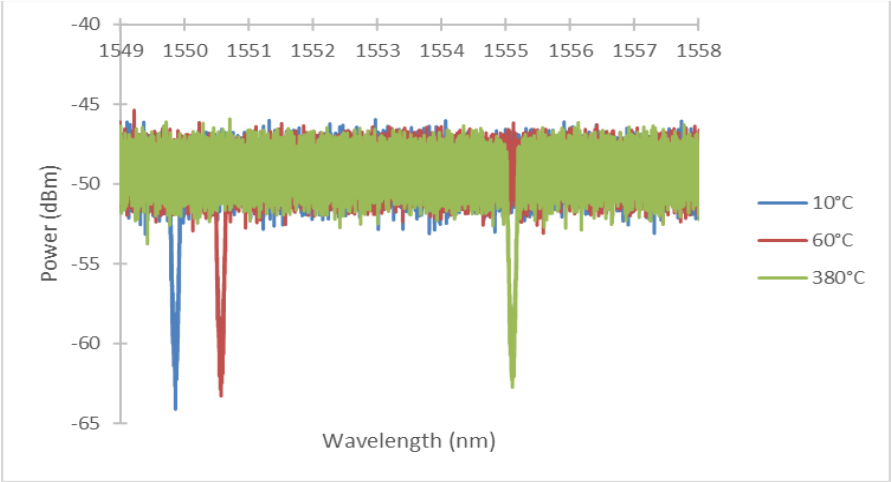
3.1.3. Gaussian Apodized Profile

Another variant of FBG profile other than chirped and uniform are apodized FBGs. These FBGs differ in terms of the way the gratings are imprinted on to the fiber. They are written in a manner that the grating intensities gradually change along the grated region until it becomes similar to the rest of the fiber. Specifically, in gausian apodized FBG, the grating intensity change along the grating length follows Gaussian modelling variant of the original Bragg wavelength equation. The emergence of this type of grating profile serves to reduce side lobes that may appear on uniform grated FBG [8]. Apodized FBG are also known to result in more narrow peaks of both reflectivity and transmission spectra. Besides gaussian model for index modulation and grating intensity determination, other models include Nuttal model and Pi (π) apodized model [23].

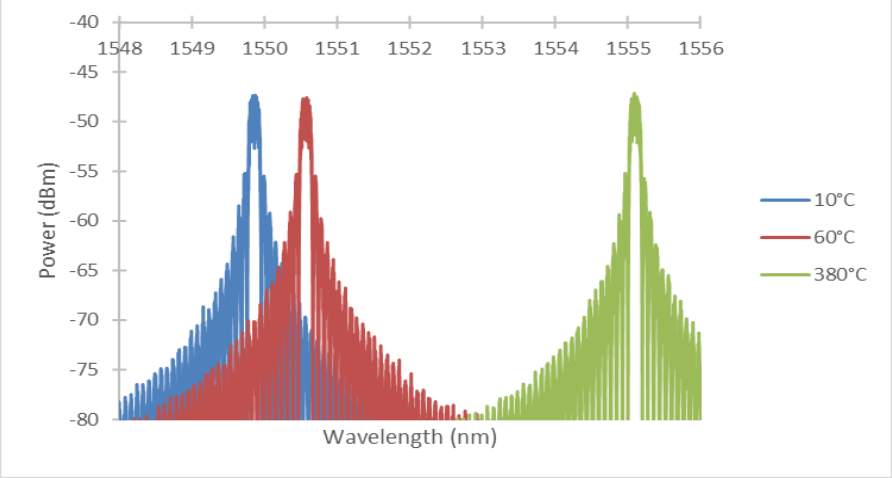
In this study, Gaussian apodized profile was simulated based on table 1 properties of the FBG. Figure 5 shows the simulated transmission and reflectivity spectra of the aforementioned FBG. Side lobe reduction occured and is observable on both of the spectra when compared to a uniform FBG. The peaks reflectivity and transmission trough are slightly narrower if compared with uniform profiles. Reflectivity and transmission power however, does suffer loss of ~10 dbm in comparison with uniform profile FBG. With the presence of side lobe reduction and narrow Bragg peak, utilization of apodized FBG is favourable for accurate sensing. Even though Gaussian apodized FBG may have lower side lobe appearance and narrow peaks, the low transmission power may be a hindrance if the FBG system is to take part in longer distances.

3.2. Effect of Temperature with FBG

FBG type sensors are used in wide variety of temperature ranges depending on the type of industries that utilize it. Simulation of FBG in different range of temperatures is shown on Figure 6. When used in oceanography for monitoring ocean temperatures, a sensor should be at least sensitive at temperature ranges of 10˚C [24]. This posed no problem for FBG sensors array as simulated in Figure 6. At higher than ambient temperatures of around 60˚C, A notable shift of the spectrum from its original wavelength of 1550 nm is detectable. At the said temperature, is observed when an FBG sensor is used in conjunction of a surgical tool to ablate cancer cells [25]. Upon higher temperature use for nuclear reactor temperature monitoring, an FBG may require modifications or special temperature proof casing. As mentioned by Mohamed, et al. [26], temperatures can reach up to 1600˚C. When using normal SMF made of silica, melting point of 400C should be taken into account. This limits the use of FBG if it is grated within a standard SMF down to only ~380˚C. While 390-400˚C is still possible to simulate, in reality, engineering constraints have to be considered for safety.

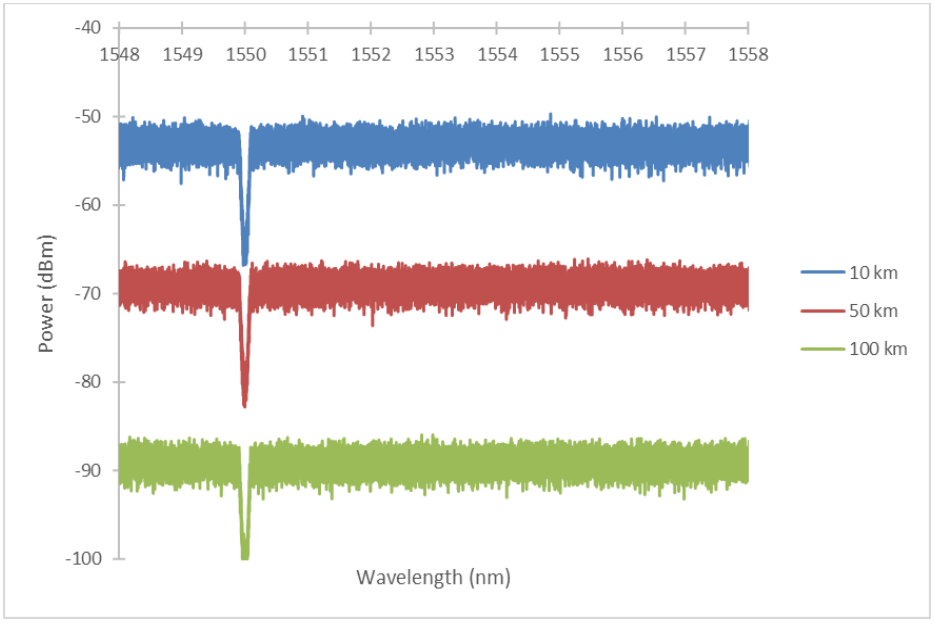


(a)

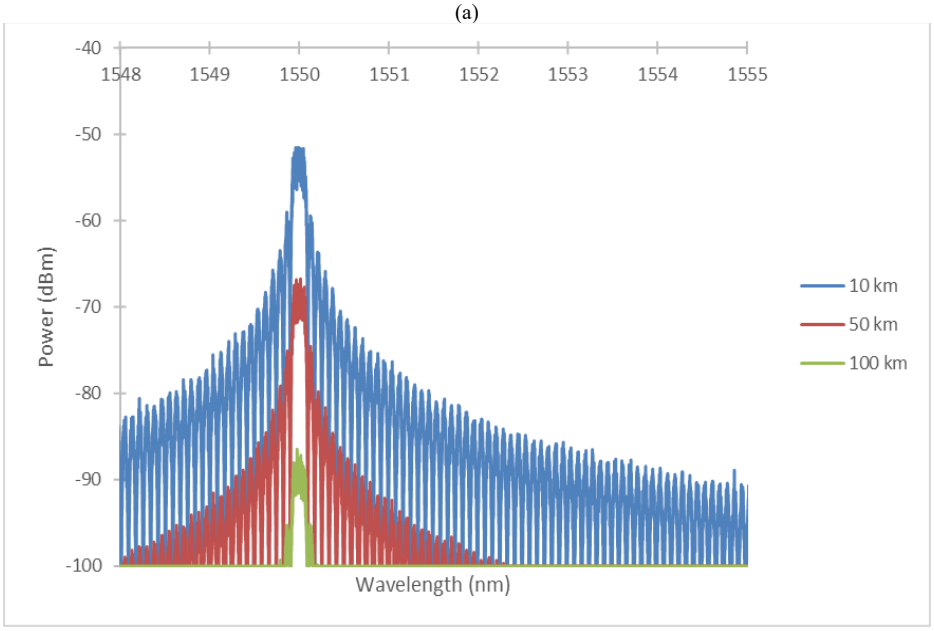


(b)

Figure 6: Optical power Spectra of FBG when exposed to various temperatures (a) Transmission and (b) Reflectivity.



(a)



(b)

Figure 7: Spectra of FBG utilized within various system distance (a) Transmission spectra. (b) Reflectivity spectra.

3.3. Effect of Transmission Power and Distance

Transmission power is at its ideal value when it is simulated or close to its ideal value when tested within lab scale lengths (<1m). When incorporating FBG within an array where the source is further than 1km in distance, power attenuation will occur. The reason for simulating this is to show that at what distance will the transmission and reflected power will deteriorate. The chosen length was based on different uses in the industry. At a distance of 10 km chosen due to most small and medium industries or premises require this distance of FBG array for environment monitoring [27]. For larger area coverage such as land and oceanic surveying applications, 50 km distance may be favourable therefore it is chosen as one of the simulation parameter for fiber distance [28]. For remote sensing in rural areas where the source and signal receiver has to be far from the area of monitoring due to power source availability, may require fiber length of up to 100 km.

Figure 7 (a) and (b) are the spectra of FBG sensor simulated within an optical circuit over various distances. The FBG used for simulation is a uniform profile variant due to it being the most stable in terms of providing low power loss due to attenuation. Furthermore, uniform FBGs are much easier to mass produce due to uniform phase mask availability. Within 10 km distance, there are hardly any power loss for both reflectivity and transmission spectra as shown on Figure 7 (a) and (b). Reflectivity on Figure 7 (b) shown significant attenuation power loss when distance of FBG optical circuit is extended to distances of 50 km and 100 km. Transmission spectra in Figure 6 (a) shown a minimal and slightly noticeable power loss. This is due to that reflected power may experience loss when reflected by the Bragg gratings due to leakage of evanescent waves occurring on FBG grated regions.

3.4. Sustainable Power Source

In environmental sensing, it is essential for light sources of sensors to have stable power input and output. While access to power may be available in established areas and urban environments, some other areas may not connected to main power grid. To remedy this issue, a light source powered by sustainable and renewable energy is favoured. A good candidate for powering light source can be a vibrational sensor which rely on the continuous minute vibration of the earth's crust [1]. In powering light source for optical sensing applications, at least 25 - 50 mW power is required. Vibrational sensors that can resonate with frequencies of <10 Hz at acceleration level (g) by <1.0 can supply enough power to FBG sensors [1]. To achieve this, vibrational sensor proposed within our previous paper [1] was equipped with rectangular clamp-free cantilever beams with attached magnet at the tips. The magnets will vibrate at the low acceleration level causing change in magnetic flux and inducing voltage in the coil. This in turn can produce electrical DC power which can be use with various low power consumption devices. By having a renewable energy source powering the FBG sensors, environmental condition mapping in remote locations is possible.

4. Conclusion and Recommendation

Several forms of FBG grating profiles has been reported via simulation of its transmission and reflectivity spectra. Temperature sensing of FBG has been simulated at various temperatures that are used in different industries ranging from as low as 10˚C to as high as 400˚C. Higher temperatures of >400˚C may require modifications that protects the FBG thermally and mechanically for certain rough applications such as monitoring greater nuclear reactor or volcanic core. When used within a high distance optical circuit, reflectivity was seen to suffer greatly with increasing distance and nearly not usable at distances spanning >100 km. Therefore, it is advisable to rely on remote optical source that is powered by renewable energy source within a FBG sensing system or array that spans greater distances. Multiple FBG of 50 km length should be optimal for use with one remote optical source. For greater coverage, multiple system of FBG and power source each with only 50 km distances between each other may be required for monitoring over a very large area. Furthermore, when an independent and self-powering optical broadband source is used, more remote areas can be monitored such as in vast open seas or subterranean regions. It has been demonstrated that FBG sensor can is capable to operate in power as low as 130 dBm but still maintains proper functions as shown in the obtained results. Although power loss due attenuation was apparent, it does not limit the sensing capability of the FBG to other environmental parameters such as temperature to some extent. This paper provided performance evaluation for FBG of various grating profiles and satisfied the theory of each grating performance.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

The funding provider of this study is the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS), No. FRGS/1/2018/TK10/HWUM/02/2.

References

[1] M. A. Riza, F. M. Foong, Y. I. Go, C. K. Thein, "Development of Sustainable FBG Sensor for Environmental Monitoring Through Vibration Energy Harvesting," in 2019 IEEE International Conference on Sensors and Nanotechnology,1-4,2019,doi: 10.1109/SENSORSNANO44414.2019.8940082.

[2] M. Amanzadeh, S. M. Aminossadati, M. S. Kizil, A. D. Rakić, "Recent developments in fibre optic shape sensing," Measurement, **128**, 119-137, 2018, doi: 10.1016/j.measurement.2018.06.034.

[3] H. Ahmad, M. T. Rahman, S. N. A. Sakeh, M. Z. A. Razak, M. Z. Zulkifli, "Humidity sensor based on microfiber resonator with reduced graphene oxide," Optik, **127**(5), 3158-3161, 2016, doi: 10.1016/j.ijleo.2015.11.184.

[4] X. Jin, S. Yuan, J. Chen, "On crack propagation monitoring by using reflection spectra of AFBG and UFBG sensors," Sensors and Actuators A: Physical, **285**, 491-500, 2019,doi: 10.1016/j.sna.2018.11.052.

[5] C. B. Williams, R. B. Yates, "Analysis Of A Micro-electric Generator For Microsystems," in Proceedings of the International Solid-State Sensors and Actuators Conference - TRANSDUCERS '95, **1**, 369-372, 1995, doi: 10.1109/SENSOR.1995.717207.

[6] M. Rhimi, N. Lajnef, "Tunable Energy Harvesting from Ambient Vibrations in Civil Structures," Journal of Energy Engineering, **138**, 185-193, 2012, doi: 10.1061/(ASCE)EY.1943-7897.0000077.

[7] A. N. Anoshkin, V. Y. Zuiko, K. A. Pelenev, P. V. Pisarev, G. S. Shipunov, "Numerical simulation for development of methodology of stress-strain state control of composite bulkhead for aviation application with the usage of FBG sensors", PNRPU Mechanics Bulletin, **4**, 47-57, 2018, doi: 10.15593/perm.mech/2018.4.04.

[8] N. A. Mohammed, T. A. Ali, M. H. Aly, "Performance optimization of apodized FBG-based temperature sensors in single and quasi-distributed DWDM systems with new and different apodization profiles," AIP Advances, **3**(12), 122125, 2013, doi: 10.1063/1.4859115.

[9] P. Xia, L.-H. Zhang, Y. Lin, "Simulation Study of Dispersion Compensation in Optical Communication Systems Based on Optisystem," Journal of Physics: Conference Series, **1187**(4), 042011, 2019, doi: 10.1088/1742-6596/1187/4/042011.

[10] G. Dastgeer, M. F. Khan, G. Nazir, A. M. Afzal, S. Aftab, B. A. Naqvi, J. Cha, K. A. Min, Y. Jamil, J. Jung, S. Hong, J. Eom, "Temperature-Dependent and Gate-Tunable Rectification in a Black Phosphorus/WS2 van der Waals Heterojunction Diode," ACS Applied Materials & Interfaces, **10**(15), 13150-13157, 2018, doi: 10.1021/acsami.8b00058.

[11] C. C. Wu, W. Y. Chuang, C. D. Wu, Y. C. Su, Y. Y. Huang, Y. J. Huang, S. Y. Peng, S. A. Yu, C. T. Lin, S. S. Lu, "A Self-Sustained Wireless Multi-Sensor Platform Integrated with Printable Organic Sensors for Indoor Environmental Monitoring," Sensors, **17**, 715, 2017, doi: 10.3390/s17040715.

[12] S. F. H. Correia, P. Antunes, E. Pecoraro, P. P. Lima, H. Varum, L. D. Carlos, R. A. S. Ferreira, and P. S. André, "Optical Fiber Relative Humidity Sensor Based on a FBG with a Di-Ureasil Coating," Sensors, **12**(7) 8847-8860, 2012 doi: 10.3390/s120708847.

[13] D. Luo, J. Ma, Z. Ibrahim, Z. Ismail, "Etched FBG coated with polyimide for simultaneous detection the salinity and temperature," Optics Communications, **392**, 218-222, 2017, doi: 10.1016/j.optcom.2016.12.068.

[14] V. G. Annamdas, "Review on Developments in Fiber Optical Sensors and Applications", International Journal of Materials Engineering, **1**(1), 2166-5389, 2012, doi: 10.5923/j.ijme.20110101.01.

[15] A. Boersma, R. Cremers, R. Jansen, "Fiber Bragg Grating Distributed Chemical Sensors," Procedia Engineering, **168**, 1245-1248, 2016, doi: https://doi.org/10.1016/j.proeng.2016.11.435.

[16] Z. L. Poole, P. R. Ohodnicki, A. Yan, Y. Lin, K. P. Chen, "Potential to Detect Hydrogen Concentration Gradients with Palladium Infused Mesoporous-Titania on D-Shaped Optical Fiber," ACS Sensors, **2**(1), 87-91, 2017, doi: 10.1021/acssensors.6b00583.

[17] S. J. Mihailov, "Fiber Bragg Grating Sensors for Harsh Environments," Sensors,**12**(2), 1898-1918, 2012, doi: 10.3390/s120201898.

[18] T. L. Yeo, T. Sun, K. T. V. Grattan, "Fibre-optic sensor technologies for humidity and moisture measurement," Sensors and Actuators A: Physical, **144**(2), 280-295, 2008, doi: 10.1016/j.sna.2008.01.017.

[19] K. O. Hill, Y. Fujii, D. C. Johnson, B. S. Kawasaki, "Photosensitivity in optical fiber waveguides: Application to reflection filter fabrication," Applied Physics Letters, **32**(10), 647-649, 1978, doi: 10.1063/1.89881.

[20] D. Tosi, "Review of Chirped Fiber Bragg Grating (CFBG) Fiber-Optic Sensors and Their Applications," Sensors, **18**(7), 2147, 2018, doi: 10.3390/s18072147.

[21] E. Vorathin, Z. M. Hafizi, A. M. Aizzuddin, M. K. A. Zaini, K. S. Lim, "Temperature-independent chirped FBG pressure transducer with high

sensitivity," Optics and Lasers in Engineering, **117**, 49-56, 2019, doi: 10.1016/j.optlaseng.2019.01.012.
[22] D. M. Wang, L. S. Wang, Y. Y. Guo, Y. C. Wang, A. B. Wang, "Key space enhancement of optical chaos secure communication: chirped FBG feedback semiconductor laser," Optics Express, **27**(3), 3065-3073, 2019, doi: 10.1364/OE.27.003065.
[23] M. A. Riza, Y. I. Go, S. W. Harun, R. R. J. Maier, "FBG Sensors for Environmental and Biochemical Applications—A Review," IEEE Sensors Journal, **20**(14), 7614-7627, 2020, doi: 10.1109/JSEN.2020.2982446.
[24] L. Wang, Y. Wang, J. Wang, F. Li, "A High Spatial Resolution FBG Sensor Array for Measuring Ocean Temperature and Depth," Photonic Sensors, **10**, 57-66, 2020, doi: 10.1007/s13320-019-0550-0.
[25] M. Jelbuldina, A. Korobeinyk, S. Korganbayev, D. Tosi, K. Dukenbayev, V. J. Inglezakis, "Real-Time Temperature Monitoring in Liver During Magnetite Nanoparticle-Enhanced Microwave Ablation With Fiber Bragg Grating Sensors: Ex Vivo Analysis," IEEE Sensors Journal, **18**(19), 8005-8011, 2018, doi: 10.1109/JSEN.2018.2865100.
[26] A. E. N. A. Mohamed, A. N. Zaki Rashed, M. Z. Elsiefy, M. M. Zaky, A. I. Elsaket, "Characterization of Fiber Bragg Grating for applications in Nuclear Research Reactors," The Egyptian International Journal of Engineering Sciences & Technology, **27**, 2019, doi: 10.21608/eijest.2019.97276.
[27] J. Nedoma, M. Fajkus, R. Martinek, "Fiber-Optic Breath Sensors: A Comparison Study," Journal of Biomimetics, Biomaterials and Biomedical Engineering,**40**,56-63,2019, doi: 10.4028/www.scientific.net/JBBBE.40.56.
[28] C. R. U. Kumari, D. Samiappan, K. R, and T. Sudhakar, "Fiber optic sensors in ocean observation: A comprehensive review," Optik, **179**, 351-360, 2019 doi: 10.1016/j.ijleo.2018.10.186.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 912-921 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Procrustes Dynamic Time Wrapping Analysis for Automated Surgical Skill Evaluation

Safaa Albasri[1,*], Mihail Popescu[2], Salman Ahmad[3], James Keller[1]

[1]*Electrical Engineering and Computer Science, University of Missouri, Columbia, 65211, USA*

[2]*Health Management and Informatics, University of Missouri, Columbia, 65211, USA*

[3]*Acute Care Surgery, University of Missouri, Columbia, 65211, USA*

ARTICLE INFO

Article history:
Received: 16 September, 2020
Accepted: 21 January, 2021
Online: 12 February, 2021

Keywords:
DTW
Procrustes
k-NN
Accelerometer
Surgery skills evaluation

ABSTRACT

Classic surgical skill evaluation is performed by an expert surgeon examining an apprentice in a hospital operating room. This method suffers from being subjective and expensive. As surgery becomes more complex and specialized, there is an increase need for an automated surgical skill evaluation system that is more objective and determines more exactly the skills (or lack thereof) the apprentice has. The main purpose of our proposed approach is to use an existing skill database with known proficiency levels to evaluate the skills of a given apprentice. The skill of the apprentice will be assessed to be similar to the closest skill example found in the database (case-based reasoning). A key element of the system is the skill distance measure employed, as each skill example is a multidimensional time series (sequence) with widely varying values. In this paper, we discuss a new surgery skill distance measure denoted as Procrustes dynamic time warping (PDTW). PDTW integrates the search for exact alignment between two skill sequences using DTW and Procrustes distance as a measure for the similarity. The Procrustes approach is a shape distance analysis that involves rotation, scaling, and translation. We evaluated our proposed distance on three surgical motion data, a widely used JIGSAWS robot surgery dataset, a wearable sensor dataset, and a Vicon motion system dataset. The results showed that the proposed framework produced a better performance for surgeon skill assessment when PDTW was used compared to other time series distances on all three datasets. Also, some experimental results for the JIGSAWS dataset outperformed existing deep learning-based methods.

1. Introduction

This paper is an extension of work initially presented in the E-Health and Bioengineering Conference (EHB) [1] . Recently, the need for objective surgical skills assessment has captured the interest of practitioners and medical institutions due to the ever-increasing complexity and degree of specialization of the surgical procedure [2]. Traditionally, a senior expert surgeon performs direct observation, scores, assess, and gives feedback to the trainee surgeon (apprentice) with less practice in the hospital operating room. This traditional surgical proficiency evaluation approach is problematic due to its subjectivity, time consumption and cost. Furthermore, it is prone to errors and sometimes insufficient as lacking details related to deficiencies. To address these difficulties, an automated skill assessment procedure is needed for an objective and detailed measure of proficiency levels [3, 4].

As any healthcare domain, surgery is continuously changed by technological advances and medical innovations that alter everyday surgical procedures. The challenge is to assist surgical procedure via the quantifiable data analysis to a better understanding of the surgical operating and to obtain more knowledge about human activities during surgery for advance and further study [5]. A reasonable solution to these challenges is to use technological advances like Robotic Minimally Invasive Surgery (RMIS) that improve overall operating room efficiency [5]. For instance, da Vinci surgical technology provides data-driven that potentially helps optimize and develop training skills

* Corresponding Author: Safaa Albasri, University of Missouri, USA. saaxfc@mail.missouri.edu

for surgeons [6]. This information includes kinematic and video data that conduct a useful resource of quantifiable human motion during surgical operating [7, 8]. Wearable sensing devices that provide detailed motion information for surgical activities are a further example [9]. These recorded data give spacious resources to assess surgical proficiencies by modeling and analyzing descriptive mathematical approaches. The emergence of using machine learning methods with recent robotic surgery systems such as da Vinci and wearable sensing devices via data-driven enable and encourage developers to build and analyze automatic models for evaluating surgeon expertise and may help better coaching potential apprentices [10–12] .

Different earlier works focused on the automated surgical assessment seen good progress. The current techniques for objective surgical evaluation can be divided into three main research areas [10, 13]: 1) surgeon skill assessment, 2) surgical task analysis, and 3) surgemes recognition. These methods considered the surgeon movement using either: 1) kinematic information recorded by a robotic surgical system, 2) video records and 3) wearable sensors data. In this paper, we focused on the surgical skill evaluation based on kinematic and wearable sensors information. One of the initial works used Hidden Markov models (HMM) [14] to evaluate the surgical skills. This approach is structured-based and depends on the number of training samples, tuning parameters and it takes massive pre-processing. This type of model needs complicated preprocessing [3] and leads to low performance with a low number of samples [14]. Another method was proposed by [3] to predict the surgeon skill level (expert and novice) based on movement features of the surgical arms using logistic regression (LR) and support vector machines (SVM) classifiers for suturing surgical task. They extended their work to include eight global movement features (GMF) in [15], they applied LR, SVM, and kNN classifier to distinguish between the previous expertise levels for suturing and knot tying surgical tasks. In [16], a framework based on trajectory shape using DTW and k-nearest neighbor classifier proposed for surgical skill evaluation. This model can also provide online performance feedback through training. More recently, [13] proposed an approach based on symbolic aggregate approximation (SAX) and vector space model (VSM) to identify distinctive patterns of surgical procedure. They used the SAX to obtain the sequence of letters by discretizing the time series first. Then they utilize the VSM to find the discriminative patterns that represent a surgical motion which finally used them to be classified. A variety of holistic analysis features and a weighted features integrated approach proposed by [9] for automated surgical skill evaluation and GRS score prediction. These holistic features include approximate entropy, sequential motion texture, discrete Fourier and discrete cosine transform. They used the nearest neighbor as a classifier and linear support vector regression (SVR) for prediction. The works of literature mentioned above used the kinematic data information obtained from RMIS for surgical skill assessment. However, none of these methods were applied to the wearable sensors data like accelerometer which might give more information about the surgeon's motion during a surgical practice.

Recently, several advanced techniques applied the convolution neural network and deep learning methods for automated surgical skill evaluation. A parallel deep learning framework was proposed by [17] to identify the surgeon skill and task recognition. In their approach, they used a fusion technique between convolution neural networks and gated recurrent networks. Alternative deep convolution neural architecture based on ten layers proposed by [12] for surgical expertise evaluation. Another parallel deep learning approach was proposed in [18] by combining the LSTM recurrent network and CNN to indicate the skill levels. Additionally, recent studies have suggested approaches that use motion from videos [19,20] and wearable sensors to evaluate surgical skills [21,22]. These methods platform various features to perform Objective Structured Assessment of Technical Skills (OSATS) assessments. An approach proposed for surgical skill assessment is based on the acceleration data of both hands performing a basic surgical procedure in dentistry [2]. Also, an entropy-based features technique that utilizes both video and accelerometer data proposed for surgical skill assessment [4]. Despite these techniques which are building the basis and inspire performance results in the surgical skill area, however, some limits and drawbacks occur for the existing methods. some methods need predefined boundaries of the surgemes which done usually by a chief surgeon, i.e., consuming a large time. In other methods, decomposing the motion sequence requires a massive and complicated preprocessing in addition to a deficiency of robustness. Alternatively, the need to developing a new distance measure might have an advantage to a more robust and accurate assessment framework.

In this paper, our contribution to this work can be abridged as follows: 1) we defined a new surgical skill distance combined the best alignments between two multidimensional signals using DTW and measuring the distance between the two aligned sequences using Procrustes analysis 2) we proposed an automated skill classification framework based on using PDTW and kNN technique in the proposed framework to distinguish between the expertise levels focusing on overall performance 3) we investigated the proposed framework on a wearable sensor data for a surgical task. The purpose of this work is to present a technique that handles different kinds of sensor data in addition to the existing public JIGSWAS dataset. Some surgery motion results obtained by a Vicon camera with a 3D marker-based system and wearable device data are examples of the data we use.

2. Methodology

In this section, we illustrate the main components of our proposed framework, which are: motion alignment, Procrustes distance, and skills classifier, as shown in Figure 1. First, DTW is used to align two multidimensional time series performed by surgeons, while the Procrustes distance calculates similarity measure. Lastly, the skill levels of the surgeon are classified by kNN.

2.1. Similarity Measure

To obtain a useful classification, defining a reasonable distance is a crucial element to measure between two surgery tasks. Each

surgery task is represented by a set of features obtained from the traces (time series) of the motion capture sensors. One possible method is the Euclidean distance.

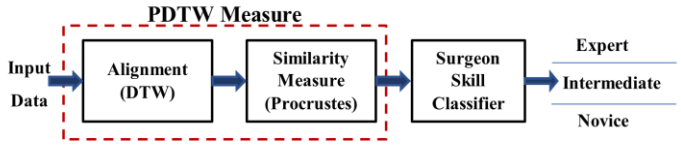


Figure 1: kNN based PDTW evaluation Framework

Euclidian distance is simple and widely used, whereas, it has some limitations and disadvantages. The Euclidean method is very sensitive to outlier and it is suffering from noise, shifting, and requires both signals to have the same length. Thus, we need a measure that can handle sequences with different lengths because the same surgery task might have different lengths even when operated by the same surgeon. A warping distance measure such as the Dynamic Time Warping (DTW), is one solution to do the job. The DTW can process time series with different lengths, it expands or contracts both signals (aligns them) such that their length becomes equal [23].

Let $X_{n\times v} = [X_1, X_2, \dots, X_n]$ and $Y_{m\times v} = [Y_1, Y_2, \dots, Y_m]$ be two sequences having v features and of length n and m respectively. To align X and Y, we form a two-dimensional $(n \times m)$ grid distance. Each point d_{ij} of the grid corresponds to the distance measure (usually Euclidean) between every possible combination of two instances x_i from X and y_j from Y of the same features length (v) as follow [24]:

$$d_{ij}(x_i, y_j) = \sqrt{\sum_{k=1}^{v}(x_{ik} - y_{jk})^2} \quad (1)$$

The next step is to find the warping path through the grid, the path that attempts to minimize the total distance (warping cost) and give the best match between two signals and satisfy boundary conditions, continuity, and monotonicity constraints. It is usually achieved by using a dynamic program to calculate the cumulative distance $\gamma(i,j)$, which is the distance of the current cell (d_{ij}) and the minimum of the cumulative distance of the adjacent cells [24]:

$$\gamma(i,j) = d_{ij} + \min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (2)$$

Despite the wide use of DTW in many applications and is a more robust distance measure than Euclidean distance, it fails for complex multidimensional signals. Also, when the unevenness occurred in the Y-axis, DTW can produce singularities by warping the X-axis. Inflection points, valleys, and peaks features can cause DTW to fail to align two signals properly [24].

The Procrustes analysis is a standard method in statistical analysis to compare the similarity of shape objects [25, 26]. The Procrustes distance is a shape metric that involves matching two shapes using similarity transformations (rotation, reflection, scaling, translation) to be as close as possible in the least-squares sense [27]. The Procrustes analysis also can estimate the mean shape to examine the shape variability in a dataset [28].

Assume X_1 and X_2 be two configuration matrices of the same $k \times m$ dimension (k points in m dimensions) that can be centered (normalized) using the following equation [28]:

$$(X_i)_c = CX_i \quad , \; i = 1,2 \quad (3)$$

$C = H^T H$ is the centering matrix and H is the Helmert submatrix, let Z_1 and Z_2 be the pre-shapes unit size of X_1 and X_2 respectively, where the original configuration is invariant under the scaling and translation with the pre-shape [28]:

$$Z_i = \frac{(X_i)_H}{\|(X_i)_H\|} = \frac{H(X_i)}{\|H(X_i)\|} \quad , \; i = 1,2 \quad (4)$$

$$(X_i)_H = HX_i \quad , \; i = 1,2 \quad (5)$$

The full Procrustes distance between X_1 and X_2 is achieved by fitting the pre-shape Z_1 and Z_2 as closely as possible as the following [25]:

$$D_P(X_1, X_2) = \inf_{s,a,b,\theta,} \left\| Z_1 - Z_2 \, s \, e^{j\theta} - (Z_2 + jb)1_k \right\| \quad (6)$$

where $\|.\|$ is the Euclidean norm, s is the scale, Θ is the rotation, and $(a + jb)$ is the translation, 1_k is a k-dimensional vector of ones.

This work presents a distance measure PDTW based on a pairwise synchronization between two time series by utilizing a combination of Procrustes distance and DTW to overcome the drawbacks of using DTW alone. First, we use DTW as an alignment approach and then use Procrustes as a distance measure. DTW is used to locate the best matching between two signals, whereas Procrustes is used to minimize the distance.

2.2. Classification

The simplicity of the k-Nearest Neighbors (kNN) method and its reasonable results made it a handy feature classifier. It predicts the new unlabeled query point by using the labels of training data based on their similarity measure. kNN classifier assigns a label for the test point to the majority label of the k-closet neighborhoods [29]. We found k = 3 is a reasonable value and the one we utilize in this paper.

3. Experimental Evaluation

We used three datasets to evaluate the proposed PDTW-kNN model on the public surgical data JIGSAWS [7], and our two data MU-EECS [30], and EM-Cric. The JIGSAWS is a minimally invasive surgical skill assessment working set consist of various fundamental surgical tasks. Each task performed by a surgical surgeon with a different proficiency degree; an expert surgeon who performs the da Vinci Surgical System (dVSS) more than 100 hours of training, a novice surgeon who practice less than 10 hours on dVSS, and an intermediate surgeon (practice on dVSS between 10 and 100 hours). A motion capture based on markers, a Vicon system is used to collect the data from a resident surgeon in the MU-EECS data. The surgeon presented a tracheostomy surgery performed the same procedure six times. The EM-Cric data includes data from four surgeons with different expertise levels who performed the Emergency Cricothyrotomy task. Each surgeon performs the task four times, where the wrist wearable sensors are used to capture both hand motions. More details about the three datasets in the following parts:

3.1. JIGSAWS Data

We evaluate the proposed PDTW-kNN method for surgical proficiency assessment on a public widely used JIGSAWS dataset [7]. Moreover, we use this dataset for direct comparisons with other state-of-the-art approaches for surgical skill evaluation. MIS surgeons performed many types of elementary procedures on Da Vinci robotic systems because it gives confidence, precision, and real-time feedback to improve overall surgical treatment for the patient in the operation room [31].

JIGSAWS dataset consists of kinematic and video data collected from surgical surgeons with various surgical robotic skills performing basic surgical training curricula. All surgeons were right-handed: two expert surgeons (E) with > 100 robotic surgical practice hours, four novice trainee surgeons (N) having < 10 practice hours, and four intermediate surgeons (I) reported between 10 and 100 surgical robotic experience practice hours. The dataset provides two types of data: video and kinematic records for each trail get done by a subject in each task. All the subjects were required to do three fundamental surgical tasks five times repetitively. In this work, we use only kinematic data captured as 76-dimensional time series at 30 Hz from the da Vinci Surgical System (dVSS) using its Application Programming Interface (API). The three elementary surgical tasks are identified as suturing (SU), knot-tying (KT), and needle-passing (NP). Figure 2 presented sample frames of the three surgical tasks achieved by a surgical surgeon and defined them as follows [7]:

- Suturing: the surgeon picks the needle up, first and advances it to the bench-top model toward the incision. Then, the subject stitches up the needle through a dot-marked tissue on one aspect of the incision and extracts it out from the corresponding dot-marked on the other part of the incision. Lastly, the surgeon passes it to the right-hand and repeats the same process till the surgeon gets four times in total.
- Knot Tying: the surgeon makes one tie after selecting one side of a stitch that is tied to an elastic tube connected by its rims to the surface of the bench-top model.
- Needle Passing: the surgeon selects the needle. Then, passes the needle from the right side to the left through 4 tiny metal hoops that are placed over the surface of the bench-top model.

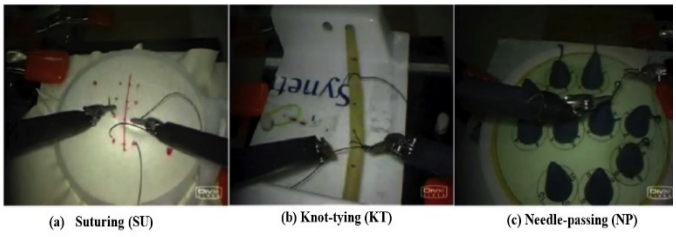


Figure 2: RMIS basic surgery tasks [7].

This dataset consists of a surgical manual annotation for the surgical skill of each trial. An annotating surgeon, with extensive robotic surgical experience, watched the entire trial and appointed a score based on a modified global rating score (GRS). GRS is the measure of the surgical technical skill of the surgeon who performed the trial. GRS presents the total score of six elements illustrated in Table I. Where each component rating scale is between 1 and 5 and the best with a higher total score [7].

Table 1: Elements of Global Rating Score (GRS) [7]

Element	**Rating scale**		
Respect for tissue	Force on tissue	Careful tissue handling	Consistent handling
Suture/needle handling	Poor knot tying	Majority appropriate	Excellent suture
Time and motion	Unnecessary moves	Efficient time/ unnecessary moves	Economy moves/ Max efficiency
Flow of operation	Frequent interrupted	Reasonable progress	Planned operation/ efficient transitions
Overall performance	Very poor	Competent	Superior
Quality of the final product	Very poor	Competent	Superior
Rating score	**1**	**3**	**5**
Min. score =	$\sum = 6$	Max. score =	$\sum = 30$

3.2. MU-EECS Vicon Data

In this dataset, a Vicon system and IR reflective markers were used synchronously to trace and visualize the arms movement of the surgeon while carrying out a surgical procedure. Ten IR reflective markers were placed in different positions on both surgeon's arms as displayed in Figure 3 (a). Also, we can see seven Vicon cameras were located inside the lab to capture the resident surgeon's motions. The MU-EECS includes data presented by a resident surgeon who performed the same tracheostomy surgical procedure six times repeatedly. The earliest three procedures repeat in a consistently appropriate manner, whereas the remaining practices were performed with inaccurately way. This working set was collected through a project at the Center for Eldercare and Rehabilitation Lab in the Dept. of EECS at the University of Missouri Columbia [30].

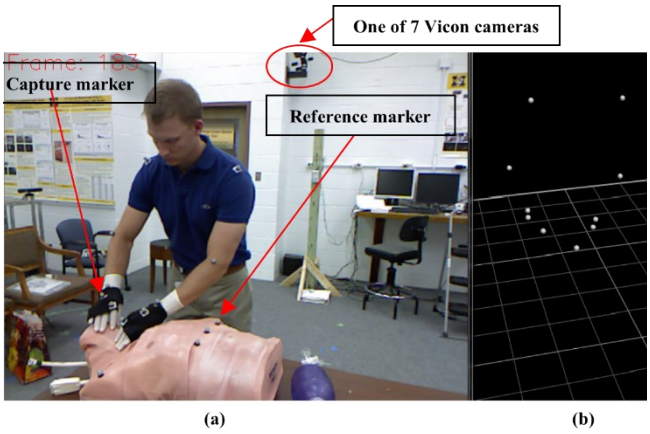


Figure 3: Tracheostomy surgery with Vicon Camera

3.3. EM-Cric Dataset

Emergency Cricothyrotomy (Cric) is a procedure for potentially lifesaving a human being under a high-stress situation, it happens when a person fails to restore enough oxygenation. Cric

is an incision through the skin and cricothyroid, which results in a better patient airway [32]. There are three main steps of the surgical Cric procedure skin incision, incision cricothyroid, and endotracheal tube placement membrane [33].

The EM-Cric dataset includes data from four surgical surgeons (subjects) who performed the Cric procedure with varying expertise levels to study skilled surgical human motion. Two residents reported as Novice (N) surgeon, one intermediate (I) surgeon, and one expert (E) surgeon, respectively. All surgeons are reportedly right-handed except one lefty hand. All surgeons perform the Cric procedure five times on a Trauma Man Surgical Simulator at the Medical Intelligent System Laboratory (MISL) in the Medicine School at the University of Missouri-Columbia. We placed the wristband sensors on both wrists of the surgeon's hands to capture the data, as shown in Figure 4. We use low cost synchronized data transmission MetaMotionR (MMR) sensors introduced by MbientLab. MMR is a 9-axis IMU wearable device that provides continuous monitoring of movement and real-time sensor data [34].

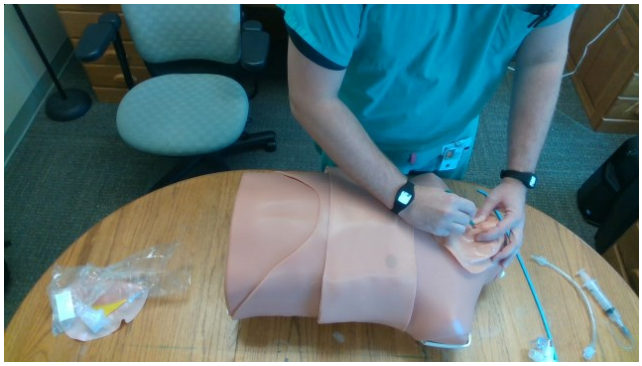


Figure 4: Cric surgical operation on TraumaMan Simulator by a medical surgeon.

The data was conducted for a total of three male right-handed, and one female left-handed participants with different expertise levels were recruited for this study. Two MMR sensors were used for the Cric procedure task, one attached to each wrist of the surgeon's hand. The captured data consists of three-dimensional acceleration with respect to time for each accelerometer, and result in 6-dimensional time series for both sensors. For this study, we use only raw accelerometer data which range was set to ±16 g. The sampling rate of data collection was set to 100Hz.

3.4. Performance Evaluation

We used different cross-validating schemes to evaluate our skill assessment framework on both kinematic and accelerometer data to compare our results with other approaches.

- Leave-One-Trial-Out (LOTO): For each surgical task, training all the trials except one i-th trial reserved for testing ($i = 1,..,N$). N is the total number of trials in a task.
- Leave-One-Supertrial-Out (LOSO): Different from LOTO setup, where we created five folds ($j = 1,2..5$). The j-th fold combines all the j-th trials from all the surgeons for a given surgical task. Then, we repetitively training on four sets and keeping a single set for testing and reporting the average classifying results. The fold j-th is known as supertrial j-th. In this scheme, the robustness of a technique can be assessed by keeping a supertrial out each time [7]. Also, repeating the task in a row can possibly impact the performance of the surgical apprentice in terms of boredom or tiredness, hence keeping the supertrial out perhaps catch that effect on the surgeons.

To evaluate the performance of our proposed technique and to quantitatively compare with other methods, we used the mean accuracy of surgical classification for each output class on the data-driven to validate the performance. The average accuracy, defined in (10), is the percentage of the sum of accurately predicted (TP+TN) over the total number of predictions (TP+TN+FP+FN) [35]:

$$\mathrm{ACC} = \frac{\mathrm{T_P}+\mathrm{T_N}}{\mathrm{TP+TN+FP+FN}} \tag{10}$$

where T_P, T_N, F_P, and F_N represent the number of true positive (predicted correctly belong to the target class), true negative (correctly classified not belong to the target class), false positive (incorrectly predicts to the target class), and false negative (incorrectly predict not belong to the class level) respectively [35].

4. Results and Discussions

In this part, the proposed approach and evaluation metrics described in the preceding sections were evaluated on kinematic and accelerometer data. Also, the results for all the datasets that were explained previously were reported in the following sections, respectively.

4.1. JIGSAWS Dataset

For JIGSAWS data, we perform two sets of experiments for the LOSO validation set up to identify the three expertise levels (E, I, and N) on our proposed approach. For the first assortment, we made use of all the 76-dimensional movement features of the time series. Whilst, in the second set we utilized just the coordinates features (x, y, z) of the two hands.

Figure 5 (a) illustrates the comparison of classification accuracy for surgical expertise levels versus k (the number of neighborhoods) in each task using all kinematic information. For the LOSO scheme, the improvement in accuracy for almost all cases of k of our kNN classifier based PDTW for all surgical tasks. e.g., the mean accuracy for all tasks at k = 3 is 95.7%. Also, kNN-PDTW provide an advantage over the traditional method (DTW) with a reduction in sensitivity to changing the number of neighbors (k) in k-NN.

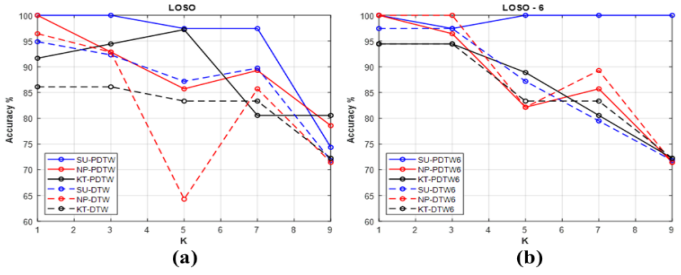


Figure 5: Accuracy of the proposed approach using PDTW and DTW as a function of k.

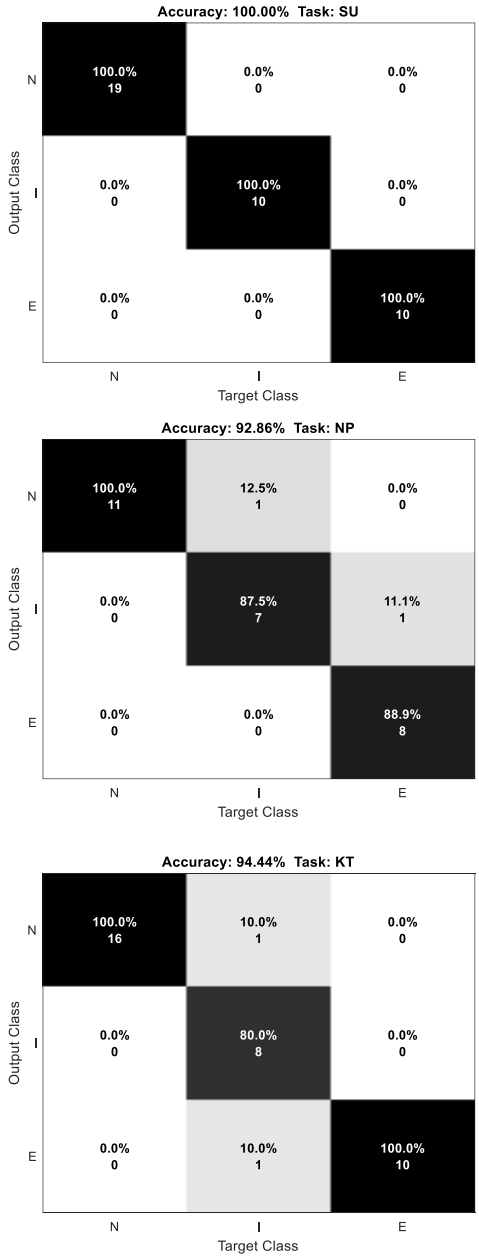


Figure 6: kNN-PDTW Confusion matrix of the three tasks SU, NP, KT for LOSO at k=3.

We also perform another experiment by using only 3D location information of the two hands for the LOSO scheme. Some interesting intuitions results can be seen in Figure 5(b). The accuracy results of the proposed kNN-PDTW6 using the Cartesian coordinates almost achieved the same results as using all the 76-dimensional motion data. This can be explained by the fact that Procrustes analysis works on the similarity of shapes and the motion data are traces in three dimensions space, which encourages us to use the wearable sensors later.

For a further comprehensive comparison, the confusion matrices result for each task is shown in Figure 6 at k=3. For the suturing task, surgeon expertise levels are 100% correctly classified. However, for the other tasks, the misclassifying happened when distinguishing between intermediate level and other levels which in turn reduced the average accuracy to about 94% and 93% for knot tying and needle passing tasks, respectively. We must put into our perspective that each surgeon performs the task in a different style from other surgeons, even within the same expertise level regardless of the hours spent on practice. Because individual surgeons like to improve their proficiencies following their mentor. Thus, small differences between an intermediate surgeon and an expert make the classifier to introduce an error to recognize their skill levels and vice versa. The same case between intermediate and novice surgeons happened.

Another interest intended of our analysis, that we calculate the pairwise PDTW distance inside a group of expert-expert, expert-intermediate, and expert-novice surgeons, separately for each task. Figure 7 illustrates the boxplot of each group distance in each task. From the results, it is clear that the smallest distance is among expert surgeons, and then between expert-intermediate surgeons followed by the expert-novice group for each task. Also, we can see that the differentiating among expert-intermediate surgeons is more complicated in needle-passing than other tasks. one explanation is the needle-passing might be more challenging to learn or more complicated than suturing or knot tying. This might be related to the complication level of the task as can be seen in Figure 6 for the needle-passing task where an expert surgeon classified as intermediate surgeon mistakenly.

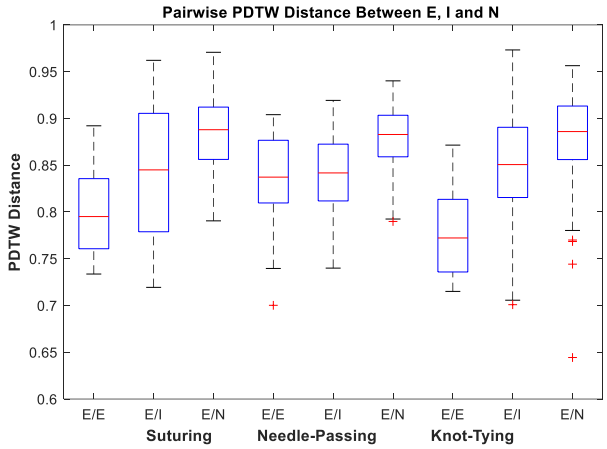


Figure 7: PDTW-distance within E/E, E/I, and E/N surgeons in each task.

Table 1 shows the classification accuracy results of our proposed skill assessment for the JIGSAWS dataset using the kinematic data only. Also, we report the state-of-the-art results for comparative intent under the LOSO validation scheme for each task separately. The results show that the proposed kNN-PDTW properly recognizes the surgeon skill levels and matched the work from CNN [36] for suturing. From Figure 7 we can see that it is straightforward to differentiate between the expertise levels with the help of using PDTW measure. Additionally, the NN classifier learned the dynamic information which already comes from various motion patterns of the surgeons that might benefit this result. In knot-tying, our proposed kNN-PDTW approach outperforms both CNN [36] and Deep Learning [12] approaches in terms of accuracy. Also, our results were near the CNN+LSTM+SENET method [18]. Our results were improved more for suturing and knot-tying tasks than the needle-passing

task, and we did slightly better than [12] in this task. The small distinctions between intermediate surgeons with other surgeons in this task illustrated in Figure 7 might explain the less performance on the needle-passing task. Furthermore, we can notice from Table I that no technique is suitable for the three tasks. In other words, an integration methodology of various approaches is needed for surgical proficiency assessment purposes for these tasks.

Table 2: Skill Assessment Classification Comparative of kNN-PDTW Performance using LOSO for JIGSAWS Data.

Approach		Accuracy		
		SU	NP	KT
Farad [15]	kNN	89.7%	-	82.1%
	LR	89.9%	-	82.3%
	SVM	75.4%	-	75.4%
Wang [12]		93.4%	89.8%	84.9%
Forestier [13]		89.7%	96.3%	61.1%
Fawaz [36]		**100**%	**100**%	92.1%
Anh [18]		98.4%	98.4%	**94.8**%
kNN-PDTW (proposed)		**100**%	92.8%	94.4%

As mentioned previously in section 3.1, the modified global rating score measures the surgical technical skill done by the annotation surgeon for the entire trial provided in the JIGSAWS dataset. Figure 8 presents the boxplot of the surgeons' GRS scores for each task. We can see from this figure, the consistency of the expert surgeons compared to the novice and intermediate surgeons in all tasks. Where the lowest variance the expert surgeons have ultimately implied their steadiness. Another interesting viewpoint from Figure 8, that we can see the scores challenge to differentiate among the surgeon's proficiency in the needle-passing task, which produces the misclassifications. One more thing to be observed in Figure 8, some intermediate subjects score better than expert subjects. This means that these surgeons might be eligible to be in a higher skill level or position.

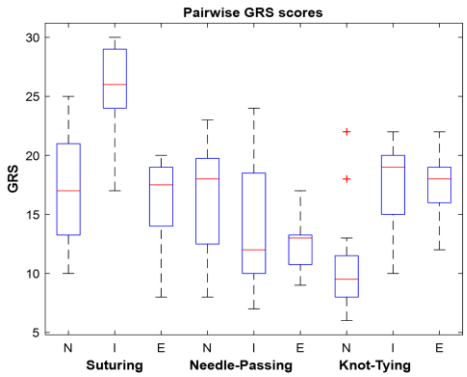


Figure 8: Boxplot of GRS scores for each task.

4.2. MU-EECS dataset

We experiment on the tracheostomy dataset to classify the trial level as either *Good* or *Bad*. In this experiment, we calculate the pairwise PDTW distance among the six trials that operated by a resident surgeon [30]. Figure 9 presents the resulting distance of this experience for the MU-EECS dataset, where the yellow color is the farthest and the closer trials to each other are in darker blue.

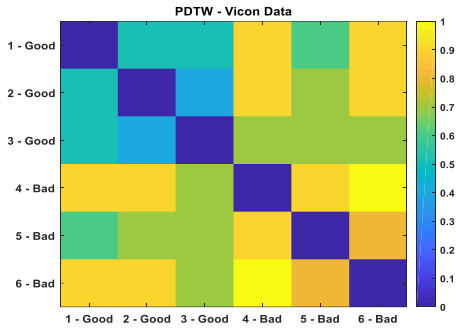


Figure 9: PDTW distance matrix for MU-EECS data.

Overall, the *Good* trials, which are the first three trials in Figure 9, has a similarity less than or equal to 0.5. e.g., about 0.3 is the difference between trials 2 and 3. On the other hand, the pairwise distance between *Bad* procedures, the last three trials, is greater than 0.7 in distance to each other. Also, we can see those *Good* procedures are nearly 0.7 far away from *Bad* trials except among trial-Good 1 and trial-Bad 5 about 0.55 difference.

Another insight from Figure 9, it is straightforward to cluster the trials into *Good* (the upper left corner) and *Bad* (in the lower right corner). That means the PDTW distance helps accurately to identify between the trials in this task where each group looks to cluster together. Finally, the boxplot of the PDTW measure among the Good and Bad trials separately is presented in Figure 10. In this figure and from a statistical viewpoint comparison, the mean and variance of the *Good* procedures ($\mu_{G-G} = 0.12,\ \sigma_{G-G} = 0.08$) is less than the *Bad* procedures ($\mu_{B-B} = 0.21,\ \sigma_{B-B} = 0.16$) which is consistent along with prior results.

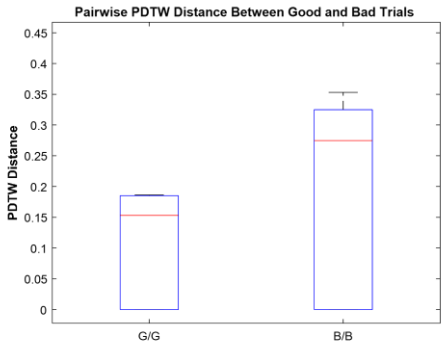


Figure 10: Boxplot of PDTW distance for MU-EECS dataset.

4.3. EM-Cric dataset

For the EM-Cric dataset, we performed two sets of cross-validation schemes, the LOTO for the trial level and the LOSO to identify the surgical proficiency levels (Expert, Intermediate, or Novice) of the subjects. As we mentioned previously in section 3.3, this dataset includes accelerometer data collected from four surgeons (expert, intermediate, and two novices) who performed the same task five times repetitively.

Before evaluating the classification accuracies, we calculate the pairwise distance among all the collected trials. Figure 11 (a) and (b) illustrate pairwise distance matrices comparison between DTW and PDTW measures, respectively. The first five trials represent the expert surgeon procedures, the second five stand for the intermediate surgeon trials, and the remaining ten trials are for the two novice surgeons, all performing the same task. Where the similar performances made by participants are indicated in strong blue squares in this figure. Also, the three separate square blocks in Figure 11 (b) give a visual insight for the possibilities of clustering expertise levels where the task is performed by different surgeons for this data using only the accelerometer data. Also, we can notice from this figure that PDTW distance separates well between expertise levels better than using DTW distance alone. The results in Figure 11 (b) shows that the expert surgeon has a dissimilar pattern to both intermediate and novice surgeons. Moreover, novice surgeons themselves are quite like each other.

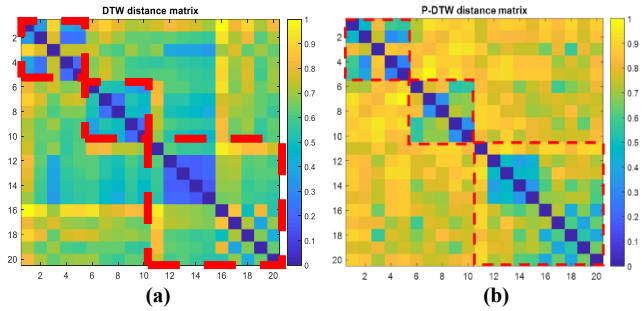


Figure 11: The pairwise distance for each trial on EM-Cric using (a) DTW and (b) P-DTW

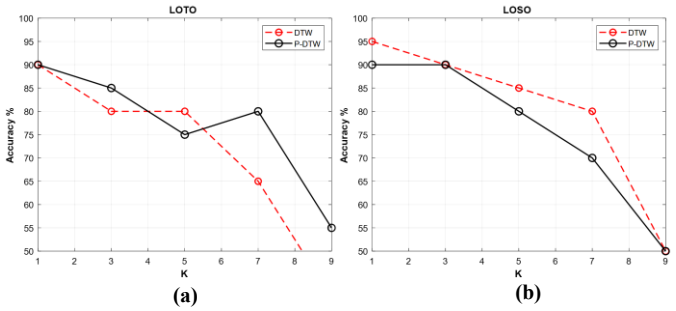


Figure 12: Classification accuracy as a function for k (a) LOTO and (b) LOSO cross-validation for Cric data

First, we performed experiments to compare how DTW and PDTW perform for classifying surgeon levels on Cric data using both LOTO and LOSO configurations. Figure 12 presents comparisons of the classification accuracy results of the proposed model for different values of K (number of neighbors) using LOTO and LOSO cross-validations, respectively. Figure 12 (a) shows that the results of our method based on PDTW performs better compared to using only DTW distance. These results indicate that our approach can identify the surgical skill levels well at trial levels because it utilizes the Procrustes analysis. Secondly, Figure 12 (b) presents the kNN-PDTW performance for the LOSO setup for the Cric dataset. The kNN based DTW approach performs slightly better for the accelerometer data. Whereas our approach results were improved, and the performance was reasonably well and still having a higher classification accuracy of 90% at k = 3.

Figure 13 shows the confusion matrix of our kNN based PDTW for surgeon expertise at k = 3 for Cric data using LOSO configuration. We can see that the intermediate surgeon was classified correctly, whereas both expert and novice surgeons were misclassified in one trial. From Figure 11 (b), we can notice that there is one trial (#3) from the expert surgeon that seems far from other trials with Expert trials and the same for novice surgeons with the trial (#11) in the same figure. The average classification accuracy was 90%.

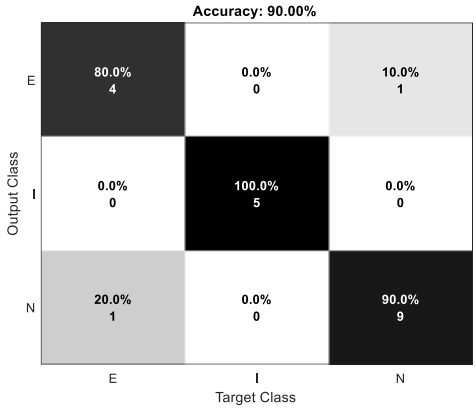


Figure 13: kNN-PDTW Confusion matrix for LOSO at k=3 for Cric data

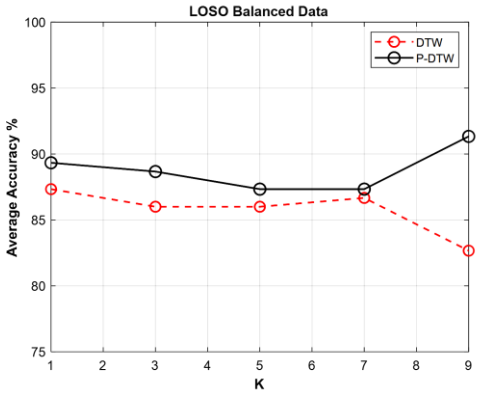


Figure 14: Balanced data classification results for the Cric data

Lastly, for a more thorough comparison, we perform another experiment for Cric data by using balanced data and evaluating using LOSO with a k-fold cross-validating scheme. The balanced data was obtained by having equal trials from each surgeon level. The reason we chose the balanced data experiment because we had two novice surgeons, one expert, and one intermediate surgeon. In this conduct experiment, we pick five trials randomly from a total of ten novice surgeon's trials and put them together with other trials from the expert surgeon and the intermediate surgeon trials. Then repeat the process ten times and report the average classification accuracy. Figure 14 shows the comparison classification accuracy as a function of k between PDTW and DTW based kNN classifier. Furthermore, Figure 15 presents the confusion matrix of kNN-PDTW predictions of the surgical skill classes. We can see from both above figures that the average accuracies of using PDTW much better than using DTW for all

values of k. Also, our approach using balanced data achieved average classification accuracy about 3% higher than using unbalanced data. the balancing data helps classified the novice surgeon's skill correctly with 100%.

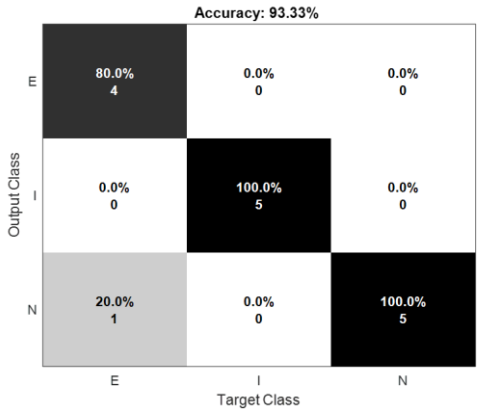


Figure 15: Balanced data confusion matrix for the Cric data

5. Conclusions

In this paper, we define a new surgery skill distance measure PDTW. It incorporates the exploration for best alignment using DTW and the similarity measure using Procrustes distance among two multidimensional time series. We show that the proposed framework based PDTW can enhance the overall performance for surgical proficiency evaluation. We attain an average accuracy of 97% for the JIGSAWS dataset and the results outperform most state-of-the-art methods using kinematic data and are comparable to techniques based on deep schemes.

Also, here we have examined the use of wearable motion sensor devices in proficiency assessment to achieve an entirely objective evaluation. Although our results are encouraging, there are quite a few limitations. The number of subjects is relatively small, not as desired. Furthermore, only one surgical task the subjects were asked to work on and there is no break between the trials which might impact the performing of the trials. Despite the limitations, our results indicate that PDTW distance can be used by classifying techniques to categorize the expertise levels accurately. In the future, we plan to increase the number of participants with a variety of expertise which might have the potential to give more information and robustness to our method. Also, more tasks to be utilized instead of only a given surgical task. Furthermore, consider using another or a combination of classifiers to improve the overall classification accuracy for skill assessment.

References

[1] S. Albasri, M. Popescu, J. Keller, "A novel distance for automated surgical skill evaluation," in 2019 7th E-Health and Bioengineering Conference, EHB 2019, 2019, doi:10.1109/EHB47216.2019.8970029.

[2] G. Arbelaez-Garces, D. Joseph, M. Camargo, N. Tran, L. Morel, "Contribution to the objective assessment of technical skills for surgery students: An accelerometer based approach," International Journal of Industrial Ergonomics, **64**, 79–88, 2018.

[3] M.J. Fard, S. Ameri, R.B. Chinnam, A.K. Pandya, M.D. Klein, R.D. Ellis, "Machine learning approach for skill evaluation in robotic-assisted surgery," Lecture Notes in Engineering and Computer Science, **2225**, 433–437, 2016.

[4] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," International Journal of Computer Assisted Radiology and Surgery, **13**(3), 443–455, 2018.

[5] F. Lalys, P. Jannin, "Surgical process modelling: A review," International Journal of Computer Assisted Radiology and Surgery, **9**(3), 495–511, 2014, doi:10.1007/s11548-013-0940-5.

[6] Intuitive Surgical, 2020, doi:https://www.intuitive.com/en-us.

[7] Y. Gao, S.S. Vedula, C.E. Reiley, N. Ahmidi, B. Varadarajan, H.C. Lin, L. Tao, L. Zappella, B. Béjar, D.D. Yuh, C.C.G. Chen, R. Vidal, S. Khudanpur, G.D. Hager, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, 1–10, 2014.

[8] C.E. Reiley, H.C. Lin, D.D. Yuh, G.D. Hager, "Review of methods for objective surgical skill evaluation," Surgical Endoscopy, **25**(2), 356–366, 2011.

[9] A. Zia, I. Essa, "Automated surgical skill assessment in RMIS training," ArXiv, **13**(5), 731–739, 2017.

[10] M. Jahanbani Fard, Computational modeling approaches for task analysis in robotic-assisted surgery, Ph. D. Thesis, Wayne State University, 2016.

[11] H.C. Lin, I. Shafran, D. Yuh, G.D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," Computer Aided Surgery, **11**(5), 220–230, 2006, doi:10.1080/10929080600989189.

[12] Z. Wang, A.M. Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," ArXiv, **13**(12), 1959–1970, 2018.

[13] G. Forestier, F. Petitjean, P. Senin, F. Despinoy, A. Huaulmé, H.I. Fawaz, J. Weber, L. Idoumghar, P.A. Muller, P. Jannin, "Surgical motion analysis using discriminative interpretable patterns," Artificial Intelligence in Medicine, **91**(July), 3–11, 2018, doi:10.1016/j.artmed.2018.08.002.

[14] L. Tao, E. Elhamifar, S. Khudanpur, G.D. Hager, R. Vidal, "Sparse hidden Markov models for surgical gesture classification and skill evaluation," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **7330 LNCS**, 167–177, 2012, doi:10.1007/978-3-642-30618-1_17.

[15] M.J. Fard, S. Ameri, R. Darin Ellis, R.B. Chinnam, A.K. Pandya, M.D. Klein, Automated robot-assisted surgical skill evaluation: Predictive analytics approach, International Journal of Medical Robotics and Computer Assisted Surgery, **14**(1), 2018, doi:10.1002/rcs.1850.

[16] M.J. Fard, S. Ameri, R.D. Ellis, "Skill Assessment and Personalized Training in Robotic-Assisted Surgery," CoRR, 2016.

[17] Z. Wang, A.M. Fey, "SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE: 1793–1796, 2018.

[18] X.A. Nguyen, D. Ljuhar, M. Pacilli, R.M. Nataraja, S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," Computer Methods and Programs in Biomedicine, **177**, 1–8, 2019, doi:10.1016/j.cmpb.2019.05.008.

[19] Y. Sharma, V. Bettadapura, T. Ploetz, N. Hammerla, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, I. Essa, "Video Based Assessment of OSATS Using Sequential Motion Textures," in Fifth Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), Georgia Institute of Technology, 2014.

[20] Y. Sharma, T. Plötz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, I. Essa, "Automated surgical OSATS prediction from videos," in 2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014, IEEE: 461–464, 2014, doi:10.1109/isbi.2014.6867908.

[21] N. Ahmidi, M. Ishii, G. Fichtinger, G.L. Gallia, G.D. Hager, "An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data," in International forum of allergy & rhinology, Wiley Online Library: 507–515, 2012.

[22] A.L. Trejos, R. V. Patel, M.D. Naish, A.C. Lyle, C.M. Schlachta, "A sensorized instrument for skills assessment and training in minimally invasive surgery," in Journal of Medical Devices, Transactions of the ASME, IEEE: 965–970, 2009, doi:10.1115/1.4000421.

[23] F. Gómez-Vela, F. Martínez-Álvarez, C.D. Barranco, N. Díaz-Díaz, D.S. Rodríguez-Baena, J.S. Aguilar-Ruiz, "Pattern recognition in biological time series," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **7023 LNAI**, 164–172, 2011, doi:10.1007/978-3-642-25274-7_17.

[24] E.J. Keogh, M.J. Pazzani, "Derivative dynamic time warping," in Proceedings of the 2001 SIAM international conference on data mining, SIAM: 1–11, 2001.

[25] J.T. Kent, "New directions in shape analysis," The Art of Statistical Science,

115, 1992.
[26] K. V Mardia, P.E. Jupp, Directional statistics, John Wiley & Sons, 2009.
[27] M.B.B. Stegmann, D.D.D. Gomez, "A brief introduction to statistical shape analysis," Informatics and Mathematical …, (March), 1–15, 2002.
[28] I.L. Dryden, K. V. Mardia, Statistical shape analysis, with applications in R: Second edition, John Wiley & Sons, 2016, doi:10.1002/9781119072492.
[29] J.M. Keller, D. Liu, D.B. Fogel, Fundamentals of computational intelligence: neural networks, fuzzy systems, and evolutionary computation, John Wiley & Sons, 2016.
[30] M. Popescu, C.J. Cooper, S. Barnes, "Automated Operative Skill Assessment Using IR Video Motion Analysis," in AMIA, 2014.
[31] B.J. Dlouhy, R.C. Rao, "Surgical skill and complication rates after bariatric surgery," The New England Journal of Medicine, **370**(3), 285, 2014.
[32] M.G. Katos, D. Goldenberg, "Emergency cricothyrotomy," Operative Techniques in Otolaryngology-Head and Neck Surgery, **18**(2), 110–114, 2007.
[33] A. MacIntyre, M.K. Markarian, D. Carrison, J. Coates, D. Kuhls, J.J. Fildes, "Three-step emergency cricothyroidotomy," Military Medicine, **172**(12), 1228–1230, 2007.
[34] INC, MBIENTLAB, 2020, doi:https://mbientlab.com/metamotionr/.
[35] T. Fawcett, "ScienceDirect.com - Pattern Recognition Letters - An introduction to ROC analysis," Pattern Recognition Letters, **27**(8), 861–874, 2006.
[36] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, "Evaluating surgical skills from kinematic data using convolutional neural networks," in arXiv, Springer: 214–221, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 913-919 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Effect of Cover Number on Distilled Water Production of Distillers with a Novel Water Feeding

Mirmanto Mirmanto*, Made Wirawan, I Made Adi Sayoga, Abdullah Abdullah, Muhamad Faisal

Jurusan Teknik Mesin, Universitas Mataram, Mataram, 83125, Indonesia

ARTICLE INFO

Article history:
Received: 03 September, 2020
Accepted: 29 September, 2020
Online: 12 October, 2020

Keywords:
Distiller
Glass cover
Seawater
Distilled water

ABSTRACT

An experimental study on the effect of a cover number of solar distillers with a continuous seawater feeding system was conducted. The seawater feeding in this study was a continuous feeding, which was not utilized yet in the previous studies. Three identical distillers i.e. single caver distiller, double cover distiller, and triple cover distiller were designed and examined. The material tested in this study was seawater taken from Tanjungkarang beach, Mataram, NTB, Indonesia, and converted into distilled water through distillation processes. The overall size of the distillers used was 1136 mm x 936 mm x 574 mm (outer dimension), while the absorber plat size was 0.8 m x 1 m. The experiment was performed in July 2019 from 09.00 to 16.00 local time. The results showed that increasing the number of glass cover decreased the amount of distilled water. The single cover distiller resulted in 949 ml a day, while the double and triple cover distillers resulted in 260 ml and 88 ml a day. The distiller with 3 glass covers was hotter than others so that the seawater vapour could not condense on the glass cover. Therefore, the distiller with a single glass cover was recommended.

1. Introduction

Many regions in Indonesia always have freshwater shortages in the dry season; e.g. Gunung Kidul, East Jawa, Central and East Lombok, and East Nusa Tenggara. Meanwhile, freshwater is essential substantive for a human being. The human being cannot live without fresh water. Therefore, efforts to fulfil the demands freshwater are crucial. However, to have freshwater requires much cost. Is there a cheap device to produce freshwater?

Indeed, devices to make freshwater have been available and they are quite sophisticated. One of the devices is reverse osmosis (RO). Researches on RO have been rigorously conducted by many previous researchers, e.g. [1-2]. Nevertheless, the price of the RO is still expensive and its operational cost is high enough. Also, it needs electrical energy. Then, RO is suitable for the production of large quantities of water and applicable electrical energy. However, for the community who have no electrical power and they are in low economic class, RO may not be a choice. Is there a cheaper way of producing freshwater?.

Yes, there is a cheaper way for this; that is a distilling system. The distilling system can be performed for several cases; e.g. distilling dirty water, distilling seawater, distilling rainwater. However, the energy used for distilling is high and expensive. When the distilling system is run using liquid fuel or wood or gases, then this method is quite expensive. Consequently, this method is not suitable for people with low economic levels. The heat used for the distilling system can be from many sources such as solar energy (solar thermal distilling), mechanic energy, electrical energy, chemical energy and so on. Meanwhile, the process of distilling is also various; e.g. evaporation-condensation, filtration, and crystallization as reported by in [3].

Some distilling techniques that are being developed are solar chimneys, greenhouses, natural vacuum, adsorption desalination, membrane distillation (MD), membrane bioreactors (MBR), forward osmosis (FO), and ion exchange resins (IXR). RO followed by multistage flashing (MSF) and multi-effect distillation (MED) systems are the most widely applied desalination. According to the 2015 International Desalination Association (IDA), more than 300 million people depend on water produced by 18426 destinations in 150 countries, which provide more than 86.8 million cubic meters per day as explained in [4].

Western countries and developed countries prefer RO systems because of efficient power consumption, while Middle Eastern and Gulf countries prefer MSF and MED systems because oil sources are abundant. The largest desalination plant which started operating at the end of 2014 was Ras Al-Khair in Saudi Arabia. The plant produces around 728,000 cubic meters of desalination

* Corresponding Author: M. Mirmanto, Jl. Majapahit no. 62, Mataram, NTB, 83125, Indonesia, +6282111738971, m.mirmanto@unram.ac.id

water per day by implementing MSF and RO technology as revealed in [5]. In [6], it was stated that the second-largest desalination plant was Carlsbad in California. United States produces around 190,000 cubic meters of desalination water per day by applying RO technology, opened in December 2015. Nevertheless, that's not the case for low-income communities.

The simplest desalination technology is the solar still distillation (SD) system, which is suitable for remote areas with low water demand because this system has low productivity, [7-9]. However, the problem for this distiller model is that its water production capacity is very low because the productivity depends on the device areas and environmental conditions.

In reference [9], the supply of freshwater is only 0.5 litres per day, and a pyramid-shaped distiller with a size of 1 m x 1 m and a total height of about 50 cm was used. In [8], the yield of freshwater obtained was only 15 ml/hour, with a tool area of 1 m². For this reason, some studies modify solar still distillation with various absorber models and multilevel models. Multilevel models have been studied in [6], nevertheless, the effect of the cover number on freshwater production was not yet investigated.

Recently, research on solar distillation using three different forms of the absorber was conducted in [10]. The form of absorber used was a flat plate, stringy absorber, and finned absorber. He explained that the fin absorber produced larger freshest water. Unfortunately, he did not examine the cover number in his study. Meanwhile, increasing the cover number raises the water temperature in the basin of the distiller. The more amount of glass is expected to be able to evaporate more seawater so that the distilled water becomes more numerous. Besides that, this distillation tool uses a continuous system, unlike distillers that have been studied by previous researchers. The continuous feed water creates the seawater level inside the basin constant. Consequently, the effect of water thickness in the basin can be eliminated and the water production continuously occurs. To analyze experimental data, equations about heat transfer, evaporation and condensation are taken from [7-9]. The equations determine the heat input, heat use and heat loss to the environment. The long term aim of this study is obtaining the cheapest, simplest, easy to operate, distiller, but the distiller has high water productivity.

2. Research Method

The research apparatus comprised of three identical distillers. One distiller had one cover, another distiller contained 2 covers and the rest included 3 covers. The absorbers of the three distillers were made of a galvalume flat plate. The apparatus is indicated in Figure 1. The distiller with a single cover is case A, the distiller with double covers is case B, and the distiller with triple covers is case C. The dimension of the distillers was 1136 mm x 936 mm x 574 mm, while the areas of absorbers were 1 m x 0.8 m.

The seawater was poured down in the top reservoir, and then it flowed forward the distiller chamber. After the level of the seawater inside the distiller reached a certain height, the valve with a floating ball closed by its self. However, when the vaporization took place so that the seawater level decreased, the valve opened automatically and the new seawater came into the distiller basin and so on until the complete experiments were done for one day. Using this method, the seawater level remained constant. This was called a novel feeding seawater or a continuous seawater feeding. This system, however, was not utilized yet in the previous studies. The seawater level in the basin was kept at approximately 4 cm. The temperatures in this current study were verified by K type thermocouples that were calibrated with an accuracy of ± 0.5°C. The solar power was measured using a pyranometer Lutron SPM-1116SD with accuracy was employed to record the solar power with an accuracy of ± 5% reading (stated in the device specification). Meanwhile, the wind velocity around the distillers was predicted using an anemometer Benetech GM8901 with an accuracy of ± 3% reading as stated in the specification brochure. The experiments were conducted at the Universitas Mataram at positions of 08° 33' - 08° 38' South Latitude and 116° 04' - 116° 10' East Longitude. The freshwater obtained was weighed at every hour using a digital scale. The density of the seawater was measured by weighing it on the scale and putting it in the glass volume. It was of approximately 1111 kg/m³.

Parameters tested in this study were temperatures, solar power, wind velocity, and distilled water. The solar power, temperatures, and distilled water were measured directly, while other parameters such as useful heat, the heat lost and efficiency were calculated. Some equations employed to analyze the parameter data were taken from [7-9].

Heat input is heat from the sun (solar power) multiplied by aperture area and glass cover transmissivity. The heat passes through the glass cover can be estimated using Eq. (1) that was taken from [11].

$$Q_{in} = \tau I A \tag{1}$$

Q_{in} is the heat input (W) absorbed by the absorber plate in the distiller basin; τ is the transmissivity of the glass cover 0.88 as reported in [12-13]. For the double cover distiller, the transitivity was obtained to be 0.8^2, and for the triple cover distiller, the emissivity was determined as 0.8^3. Meanwhile, the heat that is used for heating the seawater from the initial temperature to the final temperature is expressed in Eq. (2). Equation (2) can be obtained in [14].

$$Q_s = mc_p(T_f - T_i)/t \tag{2}$$

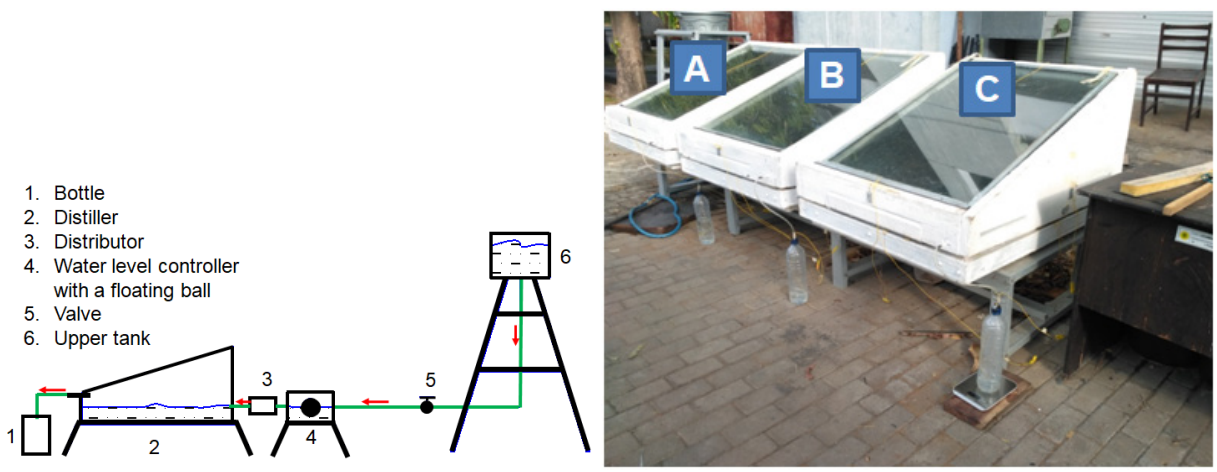
Q_s is the sensible heat (W), m indicates the seawater mass in the basin (kg), c_p represents the specific heat (J/kg°C), T_f and T_i are the final and initial temperatures (°C), while t is the time (s). Besides sensible heat, as expressed in Eq. (2), there is latent heat. The latent heat is the heat that is used for evaporating the seawater inside the distiller and written in Eq. (3) that can be found in [14].

$$Q_l = m_l h_{fg}/t \tag{3}$$

Q_l is the latent heat (W), m_l is the mass of the distilled water (kg), h_{fg} is energy for evaporation (J/kg). For that, according to [14], the total heat can be written as:

$$Q = Q_s + Q_l \tag{4}$$

Efficiency parameter is a parameter that shows the distiller performance. The efficiency can be expressed as Eq. (5) that can be attained in [14].



(a) (b)

Figure 1: Schematic diagram of the apparatus; (a) process line schematic, (b) photograph, Faisal [10].

$$\eta = \frac{Q}{Q_{in}} \quad (5)$$

3. Results and discussion

Figure 2 shows the solar power that was recorded during experiments for case A, case B, and case C at three days in June 2019 at Jurusan Teknik Mesin, Universitas Mataram, Indonesia. The solar power recorded was not influenced by the cases because all the distillers were close together; therefore, the solar power for the three distillers was the same. In figure 2, the solar power increases with the observation time and then it reaches a peak at about 13.00 (local time). After 13.00, it decreases continuously. However, the trends were also found in the previous studies conducted by Faisal [10], and Mirmanto et al. [15]. Nevertheless, on the third day, at around 10.00 (local time), solar power decreased. That was caused by the temporary cloud covering the sun. The solar power was recorded at every hour for figure 2, therefore, the effect of the cloud can be seen clearly. Then, the solar power was averaged and the results of recording solar power were 414.5 W/m² (day 1), 530.4 W/m² (day 2), and 537.2 W/m² (day 3). Of course, the results of recording solar power from day to day were different, they depended on the brightness of the sun. However, in general, solar power has a maximum value at around 13.00 (local time), because at 13.00 the sun shines brightly and also has maximum heat radiation. After 13.00 the solar power decreases further till at 16.00 (local time). Based on solar power, the heat input can be estimated using Eq. (1) and the calculation result is presented in figure 3.

Meanwhile, the heat input is different for each distiller because of the number of the glass cover. For the single cover, the transmissivity is approximate 0.88, while for the double and triple covers; the transmissivities are 0.7744 and 0.681 respectively. Hence, increasing the number of covers decreases transmissivity.

Figure 3 shows the trend of solar power, in general, increases with the time until it reaches a peak value at about o'clock local time. After that, the solar power decreases continuously until the experiment is complete. This phenomenon was due to the sun cycles. The sun rises, afternoon, and then sunset.

The experimental heat input is shown in figure 3. The heat input was calculated using Eq. (1) on the experimental data basis. The trend of this parameter is almost the same as that of solar power trends. However, the magnitudes for the three cases are different. This was due to the different number of the cover. The higher the number of covers, the lower the transmissivity.

Recorded absorber temperatures are presented in figure 4. The trends of the temperature are different from that of solar power. The temperatures increase with the observation time, then after at around 13.00 o'clock, they are flat until the experiment is finished at about 16.00 o'clock. For the single cover, the absorber temperature at after 13.00 o'clock is constant, while that of double and triple covers still increases fairly. This trend indicates that adding the number of covers elevates the absorber temperature. This could happen because the heat inside the distiller was difficult to be emitted back to the ambient. It was trapped by the number of glass covers and the gaps between the cover. Nevertheless, distillers with multiple covers had higher heat losses than that with a single cover. The heat loss occurred due to the high different temperatures between the absorber and the ambient. Mostly the heat loss happened through the conduction mode. The heat flowed through the distiller walls and bottom.

To know further the effect of cover numbers, let's see Figure 5. Figure 5 shows the recorded seawater temperatures. Figure 5 indicates that trends of the seawater temperatures are almost the same as the trends of the absorber temperatures. The seawater temperatures increase with time and then get constant. Moreover, the magnitudes of the temperatures are nearly the same. This was caused by the direct contacts of absorber and seawater. Previous studies also showed similar findings, e.g. in [8, 10, 15].

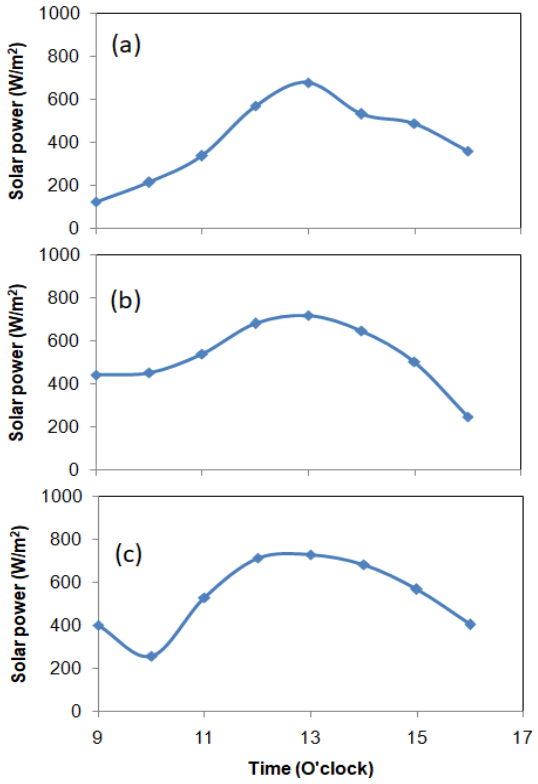


Figure 2: Recorded solar power: (a) day 1, (b) day 2, and (c) day 3

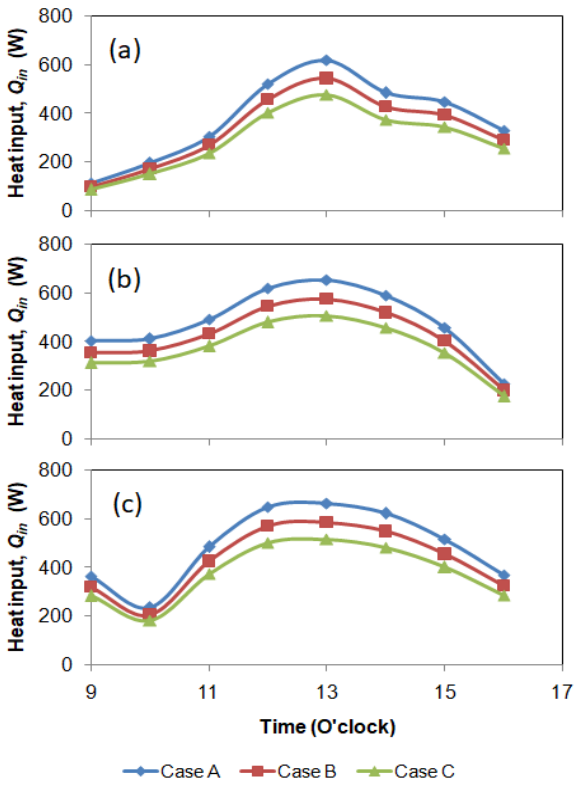


Figure 3: Calculated heat input (Q_{in}) for the three cases: (a) day 1, (b) day 2, and (c) day 3.

Another parameter recorded in this study was the amount of distilled water. Distilled water production per hour is presented in Figure 6, and the accumulated distilled water production is given in Figure 7, while the distilled water daily production is shown in Figure 8. The distilled water density was measured to be of approximately 995.2 kg/m³ at 30°C. The hourly distilled water production seems irregular, sometimes it is much and other times it is less. This was due to the sky and environmental conditions. At the time where the sky was clear, the distilled water production was high and vice versa. This phenomenon was also found in [10, 16]. The distilled water production obtained in [10] was greater than that of [16]. This was due to the different day.

Figures 6 and 7 show that the distiller with a single cover produces larger hourly distilled water. This phenomenon was due to the cover temperatures. The cover of the distiller that contacts directly to the vapour for the single cover distiller was lower than that of double and triple covers. Therefore, when the vapour came up to the cover, it could be easily condensed. However, for the double and triple cover distillers, the lower glass cover that touched the vapour had a higher temperature, as a result, the vapour could not condense. When the glass covers were removed from the distillers at the end of the experiment, the vapour coming out from the distillers with double and triple covers was much larger than that from the single cover distiller. Moreover, the absorber and seawater temperatures for the distiller with double and triple covers were higher than that of the distiller with a single cover. Hence, why did the distillers with double and triple covers result in less distilled water? This was due to the less condensed vapour on the glass cover. Furthermore, the results of this study were less than that of [17, 18]. This was due to the simple shape of the distillers utilized in this current study.

Different from hourly distilled water production, accumulated distilled water production increased with time, see figure 7. This finding was also found by previous researchers; e.g. [15, 17]. In [16], the total distilled water production was almost 1500 ml a day. They started their experiment at 9.00 local time and finished at 16.00 local time. Meanwhile, this study produced maximum distilled water of 949 ml.

Heat transfer calculation results are presented in figure 8. The heat transfer here is called useful heat; Q. Q can be computed using Eq. (4) that comprises sensible heat and latent heat. The useful heat increased drastically at the beginning of the experiments. After at 10.00 am, the useful heat decreased continuously till the experiments were finished. The trend of the useful heat was due to heat absorption. At the beginning of the experiment, much heat could be absorbed because the seawater temperature was still low. At this level, the heat transfer occurring was sensible heat. The heat was used for heating the seawater inside the distiller. As the seawater temperature increased, the heat absorbed decreased, and the heat transfer mode changed to a combined heat transfer; e.g. sensible and evaporation heat transfer. The evaporation heat transfer is called latent heat transfer. For this case, the seawater changed to vapour.

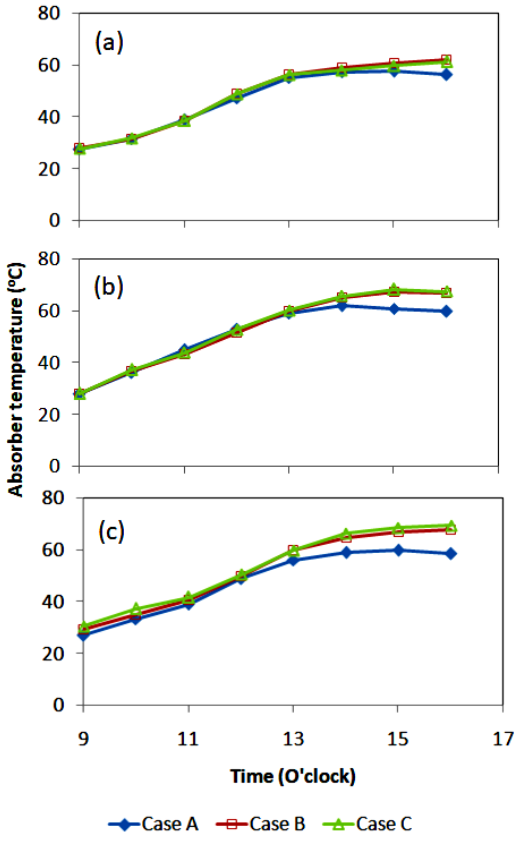


Figure 4: Trends of the absorber temperatures for three cases: (a) day 1, (b) day 2, and (c) day 3.

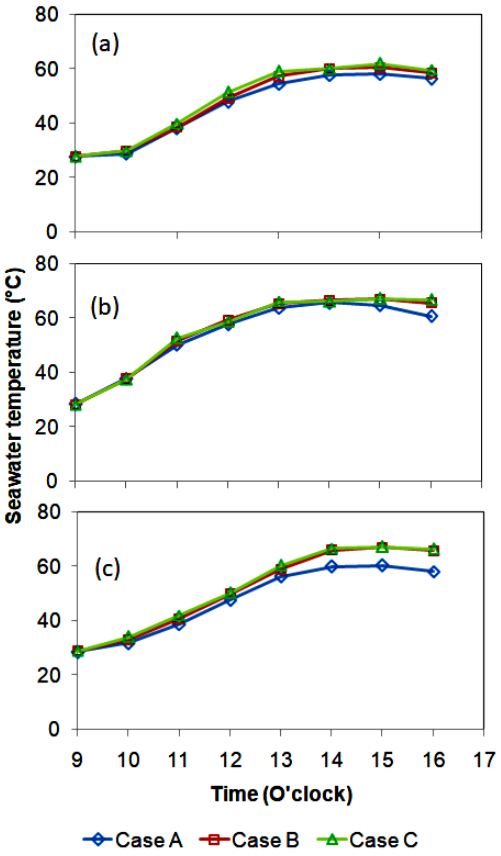


Figure 5: Seawater temperature for the three distillers; (a) day 1, (b) day 2, and (c) day 3.

In figure 8, the useful heat contained sensible and latent heat. The useful heat was predicted using equations (2) and (3). The mass of the vapour condensed was used to estimate the latent heat. For case A, the distilled water production was larger than for cases B and C, as a result, the useful heat obtained for case A also was higher than for the other two cases. This was due to the larger condensation for case A.

Another variable that indicates the performance of the distiller is efficiency. Efficiency is a comparison of useful energy to heat input. The useful heat is Q and the heat input is Q_{in}. Q_{in} is estimated using Eq. (1). Therefore, efficiency can be predicted using Eq. (5) and the calculated efficiency is presented in figure 9.

Figure 9 indicates that the trend of efficiency is almost the same as the trend of the useful heat. At the beginning of experiments, the efficiency increased sharply, and it reached that peak value. After the peak value was achieved, the efficiency decreased continuously until the experiments finished. However, for case A, the efficiency trend was a little bit different. At around 15.00 o'clock, it increased again. This occurred because the production of distilled water was still ongoing, but the heat input was down. For cases B and C the distiller still generated much vapour, but the lower layer of the glass cover was still hot, then the vapour could not condense, as a result, the useful heat (latent heat) was low. Consequently, the efficiency for case B and C was lower than that for case A.

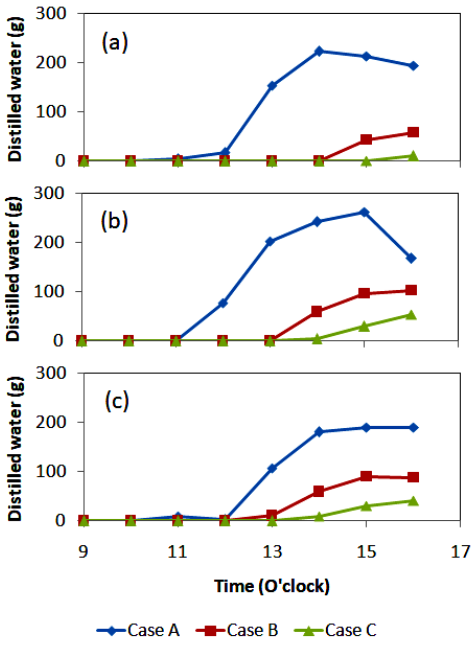


Figure 6: Hourly distilled water production obtained for three days. (a) day 1, (b) day 2, and (c) day 3.

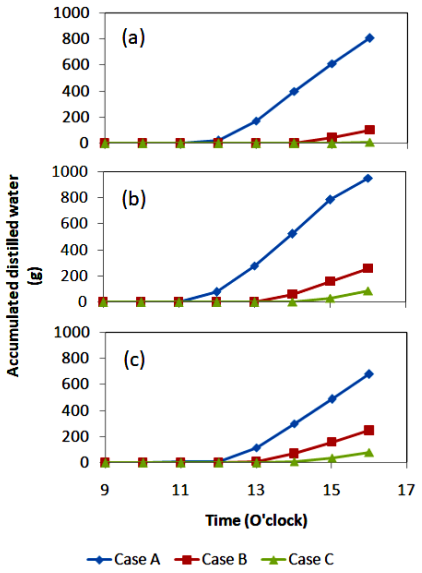


Figure 7: Accumulated distilled water production; (a) day 1, (b) day 2, and (c) day 3.

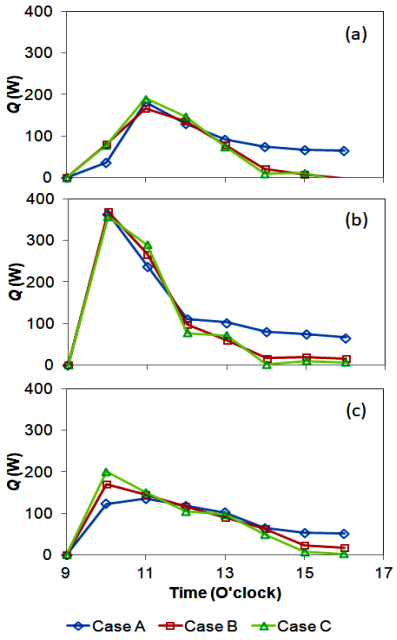


Figure 8: Useful heat transfer in the distillers for the three cases and three days.

4. Conclusions

The study to know the effect of glass cover number on the distilled water production was conducted with several tested parameters and continuous feeding seawater. Some findings that can be stated are: the solar power in Indonesia has a large potential; the hourly distilled water productions are scatter depending on sky conditions. The major remark in this study is that much-distilled water production can be achieved just using single cover; less vapour can be condensed in the double and triple cover distillers; the largest efficiency is obtained for the single cover distiller. A single cover distiller is recommended to be utilized for producing freshwater.

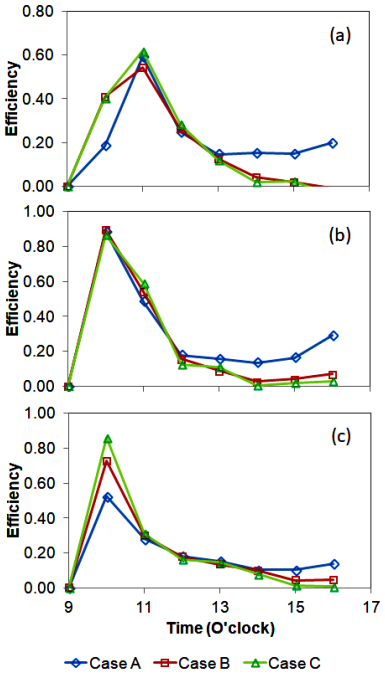


Figure 9: Distiller efficiency for the three cases and three days.

Acknowledgement

The author would like to thank the DRPM for the funding and Jurusan Teknik Mesin, Universitas Mataram for the facility.

Nomenclature

A	=	Aperture area (m²).
c_p	=	Specific heat (J/kg K).
I	=	Solar power (W/m²).
m	=	Seawater mass (kg)
Q	=	Useful heat (W).
Q_{in}	=	Heat input (W).
Q_l	=	Latent heat (W).
Q_s	=	Sensible heat (W).
t	=	Time (s).
T_i	=	Initial seawater temperature (°C).
T_f	=	Final seawater temperature (°C).
η	=	Efficiency

References

[1] A.H. Slocum, M.N. Haji, A.Z. Trimble, M. Ferrara, S.J. Ghaemsaidi, "Integrated pump hydro reverse osmosis system", Sustainable Energy Technologies and Assessments, **18**, 80-89, 2016. doi:10.1016/j.seta.2016.09.003

[2] N.K. Khanzada, S.J. Khan, P.A. Davies, "Performance evaluation of reverse osmosis (RO) pre-treatment technologies for in-land brackish water treatment", Desalination, **406**, 44-50, 2017. doi:10.1016/j.desal.2016.06.030

[3] A. Alkaisi, R. Mossad, A. Sharifian-Barforoush, "A review of the water desalination systems integrated with renewable energy", Energy Procedia, **110**, 268 – 274, 2017. doi:10.1016/j.egypro.2017.03.138.

[4] M. Baawain, B.S. Choudri, M. Ahmed, M., A. Purnama, "An Overview: desalination, environmental and marine outfall systems", Recent Progress in Desalination, Environmental and Marine Outfall Systems, **1**, 3-10, 2015. doi:10.1016/j.egypro.2017.03.138.

[5] S.M. Cheong, G.W. Choi, H.S. Lee, "Barriers and solutions to smart water grid development", Environmental Management, **57(3)**, 509-615, 2016. doi:10.1007/s00267-015-0637-3

[6] N. Heck, A. Paytan, D.C. Potts, B. Haddad, "Coastal residents' literacy about seawater desalination and its impacts on marine ecosystems in California", Marine Policy, **68**, 178-86, 2016. doi:10.1016/j.marpol.2016.03.004

[7] Water and waste water tratement spesialist yang dipercaya dalam proses pengolahan air, tirtamandiri.com.

[8] A. Mukadim, M. Wirawan, I.B. Alit, "Analisa pengaruh variasi bentuk absorber pada alat destilasi air laut terhadap kenaikan suhu air dalam ruang pemanas dan jumlah penguapan air yang dihasilkan", Dinamika Teknik Mesin, **3(2)**, 127-135, 2013.

[9] Mirmanto, Destilasi surya pengubah air laut menjadi air tawar. Laporan Penelitian DPP/SPP, Universitas Mataram, 2003.

[10] M. Faisal, Pengaruh variasi absorber alat distilasi air laut tenaga surya terhadap produksi air tawar, Skripsi, Jurusan Teknik Mesin, Universitas Mataram, 2019.

[11] J.A. Duffie, W.A. Beckman, "Solar Engineering of Thermal Processes", 4th Edition y John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.

[12] B.Y. Idi, D.K. De, "Transmissivity of the glazing surface of a solar flat plate collector based on the metrological parameters of Yola, Nigeria", Journal of American Science, **7(1)**, 639-643, 2011.

[13] M. Syahri, "Rancang bangun sistem desalinasi energi surya menggunakan absorber bentuk separo elip melintang", Prosiding Seminar Nasional Teknik Kimia "Kejuangan", Pengembangan Teknologi Kimia untuk Pengolahan Sumber Daya Alam Indonesia, Yogyakarta, 22 Februari, 2011.

[14] J.P. Holman, "Heat Transfer", 8th Edition. McGraw-Hill Inc., USA, 1997.

[15] M. Mirmanto, M. Wirawan, I.M.A. Sayoga, S. Syahrul, M. Faisal, A. Abdullah, "Effect of absorber types of conventional distillers on the amount of distilled water production", Frontiers in Heat and Mass Transfer, **13**, 1-7, 2019. doi:10.5098/hmt.13.10.

[16] Abdullah, "Pengaruh jumlah *cover* alat distilasi air laut tenaga surya terhadap poduksi air tawar", Skripsi, Jurusan Teknik Mesin, Universitas Mataram, 2019.

[17] A.A. Kabeel, "Performance of solar still with a concave wick evaporation surface", Energy, **34**, 1504–1509, 2009. doi:10.1016/j.energy.2009.06.050.

[18] Mulyanef, Marsal, R. Arman, K. Sopian, K., "Sistem distilasi airlaut tenaga surya menggunakan kolektor plat datar dengan tipe kaca penutup miring", Teknik Mesin, Universitas Bung Hatta, Padang, Indonesia, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 920-925 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Barriers and Supports in Engineering Career Development: An Exploration of First-Year Students

Rosmery Ramos-Sandoval*, Jano Ramos-Diaz

Centre for Interdisciplinary Science and Society Studies, CIICS, Universidad de Ciencias y Humanidades, Lima, Lima-39, Peru

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 18 November, 2020
Online: 08 December, 2020

Keywords:
Social Cognitive Career Theory
Barriers
Supports
Engineering student retention

ABSTRACT

Previous research has found that first-years in college is challenging due to changes in academic demands and the adaptation process. In addition to this, research on career development has found that barriers and supports may influence career interest, motivation to continue and student retention. This study explores the influence of specific supports and barriers among first-year engineering college students. To reach this goal, a case study was conducted with 425 engineering students at three universities in Lima, Peru. Based on Social Cognitive Career Theory (SCCT), we performed a Structural Equation Model (SEM) to assess effects of contextual factors such as barriers and supports on self-efficacy, coping self-efficacy, and goals. Results showed that certain supports and barriers influence self-efficacy, cope self-efficacy and goals. Effects of these factors in first-year students and possible strategies for retention in STEM careers are discussed.

1. Introduction

Science and technology are critical subjects for the advancement in societies because they create prosperity, wellbeing and more opportunities [1, 2]. Despite its increasing importance in daily life, the necessary number of science, technology, engineering and mathematics (STEM) workers is not being produced [3]. As a result, this gap has become a major problem, especially in developing countries, given the pressure to become competitiveness in the global economy. Therefore, increase and retention of STEM students, graduates, and workers have vital importance for Latin-American organizations, governments and educational institutions.

Accordingly, Peru showed an increase in higher education enrollments, however, this tend to vary significantly in specific careers. A recent report showed that between 2008 and 2019, 771 900 and 1'509 400 students were registered at national level [4]. Although, the same report showed that in 2017, more than a half (52%) were in social sciences and only 27.8% were in engineering and technology. Nonetheless, the rates of registered students in STEM fields indicated that there was a decline between 2018 and 2019 [4]. Besides, previous official reports indicated that economic issues (31%), job responsibilities (28%), and family issues (14%) [5] were the main reasons to suspend higher education in STEM. Based on the above-mentioned reports, it seems that several factors are involved in Peruvian STEM students' choice to stay or leave university. In addition to this, these rates and factors may contribute to the gap in STEM careers in Peru. As a result, this may have economic and social consequences for the country development.

Therefore, exploring how environmental and behavioral process interact in engineering career behavior may help to understand the STEM gap and first-year Peruvian students experiences in STEM careers. What is more, research on these factors is scarce in Peruvian STEM careers. Therefore, by using a survey design, the present study aims to explore how specific barriers and supports may influence self-efficacy, cope self-efficacy and goals.

1.1. Who leaves and who stays?

First-years in higher education are critical for students due to changes in study style and the time that it takes to adapt from high school to college. Previous research suggested that it is critical to focus on factors related to students retention in early years of STEM careers [6–9]. This advice is particularly relevant especially in certain areas of Peru where a lack of knowledge from parents about higher education constitute an additional challenge for students. Moreover, lack of support in first-year students may generate a biased perception of higher education that may result in withdrawal. Therefore, it is crucial to get insights based on theory and evidence to understand first-year engineering career development in the Peruvian context.

*Corresponding Author: Rosmery Ramos-Sandoval, Email: rramos@uch.edu.pe

Consequently, the literature on engineering career behaviors has highlighted several factors involved in career progression. Previous studies identified that lack of enjoyment, interest in engineering courses, demotivation and unpreparedness were the main reasons to leave engineering careers [8, 10]. In addition to this, Snyder and colleagues [7] found that self-perception, beliefs and motivations may influence competence and expectations for achievement in engineering career. Based on this, psychological factors seems to have effects in how students behave and take decisions as well as their drive to continue engineering careers.

1.2. Self-efficacy, coping self-efficacy and goals

One particular theory that has been useful in the explanation of career development is Social Cognitive Career Theory (SCCT) [11]. SCCT model hypothesizes the association between self-efficacy, outcome expectation, interest, goals, barriers and supports [11, 12]. SCCT proposed that when someone feels competent concerning a specific activity (self-efficacy), they will anticipate positive results (outcome expectation) and will develop interest in that particular domain. As a result, subjects will set goals that later will become in actions, however, these will be affected by barriers and supports. However, this study will focus in how specific barriers and supports affect certain variables of the model. For instance, several models have been constructed in order to explore in more detail specific SCCT variables [13, 14].

The term self-efficacy is defined as beliefs that people has regarding their capabilities to achieve successfully a specific task [15, 16] whereas goals is defined as student's intentions to persist in their career. The term cope self-efficacy is defined as beliefs about one's ability to manage environmental obstacles in a particular domain [12]. In this regard, previous reports have found that specific contextual factors may influence student's self-efficacy, career goals and cope self-efficacy [6, 17, 18]. In addition to this, relationships between barriers and supports may be indirectly related through coping self-efficacy [19]. Similarly, recent reports indicated that goals may affect determination to engage in career behaviors and that it is mediated by barriers and supports [11, 17, 20]. Taken together, these studies support the notion that certain individual variables of SCCT are suitable to understand career progress. Therefore, using certain core variables from SCCT seems suitable to understand in more detail how specific barriers and supports (e.g. Economics, Institutional and Family) interact with self-efficacy, cope self-efficacy and goals.

1.3. Barriers

Much of the current literature on contextual factors in career development pays particular attention to barriers and supports. Barriers refer to negative contextual influences that hinder different aspects of career progress [12, 21]. In addition to this, previous research identified that barriers are associated with career-related behaviors, engineering college student's progress and attrition [22–24]. Moreover, recent studies found that barriers associated with family, social networks and lack of social support may influence career choice and decision to stay or leave current careers [25, 26]. These studies suggest that there is an interaction between internal and contextual factors that may affect student's decision to continue engineering degrees and possible consequences such as withdrawal. Besides, specific differences in culture and context may change how these barriers are expressed in the Peruvian context. Therefore, exploring in more detail how these barriers interact could provide useful insights into Peruvian universities and possible interventions to reduce possible dropouts in engineering college students.

1.4. Support

On the other hand, support is defined as an environmental variable that facilitates the formation and pursuit of an individual's career choice [12]. Previous research has established that parents or family support have a critical role on vocational behavior and educational outcomes in undergraduate students [22]. In addition to this, peer social support was also positively associated with higher levels of aspiration and expectation in high school students [27]. Furthermore, Lent et al. [28] proposed that certain supports such as teaching staff or family are relevant in career decision-making and expectancies of career success. Therefore, by identifying specific supports, we could explore to what degree these may affect students career behavior (e.g. self-efficacy, coping self-efficacy and goals). In addition to this, by exploring these factors, Peruvian educational institutions may minimize the pre-existent barriers that exist in state engineering faculties such as the lack of resources, materials, equipment and quality standards. Moreover, it may contribute to develop strategies for education as a whole and increase student's retention.

Therefore, the present study aimed to explore the relationship between self-efficacy, cope self-efficacy and goals and how they interact with specific barriers and supports. Based on previous research [21–23, 29], the following hypotheses are proposed:

- H1. Social acceptance support have a direct effect on self-efficacy, goals, and cope self-efficacy.
- H2. Family support have a direct effect in self-efficacy, goals, and cope self-efficacy
- H3. Economic support have a direct effect on self-efficacy, goals, and cope self-efficacy.
- H4. Teacher support has a direct effect on self-efficacy, goals, and coping self-efficacy.
- H5. Social barriers have negative effects on engineering student's self-efficacy, goals, and their cope self-efficacy.
- H6. Economic barriers have negative effects on self-efficacy, goals, and cope self-efficacy.
- H7. Gender barriers have negative effects on self-efficacy, goals, and cope self-efficacy.
- H8. Teacher barriers have negative effects on self-efficacy, goal, and cope self-efficacy.
- H9. Racial barriers in have negative effects on self-efficacy, goal, and cope self-efficacy.

2. Methodology

2.1. Data collection

This research was carried out by contrasting the reference sample universe (N=40962; Male=77%; Female=23%) in students population enrolled at three university campuses in the North of Lima, Peru. This information was obtained from government public information [30]. To calculate the sample size, we used the finite population analysis formula which is considered to have a 95% confidence level. The final sample was composed of 425 engineering students (77.4% males; 22.6% females) from ten engineering fields. The mean age of participants was 21 years

(SD=4.02; Min=16; Max=38). As mentioned before, female students in this sample were low (< 23%) and this was aligned with national reports for female engineering students (25%) [30]. Before data collection, ethical standards were considered such as the right to withdrawal, inform consent and anonymity protection.

2.2. Survey characteristics

The Spanish version from the Engineering Field Questionnaire [31] was used from the original version of Social Cognitive Career Theory proposed by Lent, Brown & Hackett [11,12]. This study employed the following SCCT core variables to model the multiple relationships in the hypotheses mentioned before.

The first construct was designed to test participant's self-efficacy. This scale had 4-items with a 9-point Likert scale that ranges from 0 (no confidence at all) to 9 (complete confidence). Example items included: "How confident are you about performing well in your training over the next two courses?" To find out participants goals, 4 items were used to measure student's intention to persist. Each item was scored on a 5-point Likert scale (strongly disagree, disagree undecided, agree, strongly agree). An example item included: "I expect to remain enrolled in this course for the next year". To assess coping self-efficacy, a 7-item scale was used with a 9-point Likert scale that ranges from 0 (no confidence at all) to 9. Some of the questions were: "I feel confident to be able to continue my studies, even if I feel that the classroom environment was not welcoming".

To explore specific supports, a 5-item scale was used. These items were related to positive environmental cues that students received from friends, family, teachers as well as economic factors that facilitate their career. The response format was a 4-point Likert that ranges from 1 (No at all likely) to 5 (Extremely likely). Example items included: "During the first year, did you get financial support to continue your studies?" To identify barriers, a 6-item scale was used. This scale described obstacles that students receive during their career such as discrimination, lack of teacher support and issues with peers. The response format was in Likert format that ranges from 1 (No at all likely) to 5 (Extremely likely). Example items included: "During the first year, have you felt discouraged by the lack of social support?". This questionnaire was adapted using specific procedures to adapt psychological tests [32].The engineering field questionnaire has shown good psychometric properties and it has been used in engineering students [21,33].

2.3. Statistical analysis

Since the study aim was to confirm and explore effects from supports and barriers in first-year students, we designed a Structural Equation Model (SEM) [34]. In addition to this, data were analyzed in SPSS© IBM 23 and modelling was executed in AMOS© 21 [35].

3. Results

3.1. Independent Variables

Barriers and supports were tested as independent variables, individual scores were obtained through exploratory factor analysis (Table 1).

Table 1: Independents variables from barriers and support in first-year

Factors	Independent variable	Number of categories	Estimate
First-year Support	Social Acceptance Support	4	0.705
	Family Support		0.716
	Economic Support		0.632
	Teacher Support		0.620
First-year Barriers	Social Barriers	5	0.606
	Economic Barriers		0.420
	Teacher Barriers		0.367
	Gender Barriers		0.870
	Racial Barriers		0.843

Estimate > 0.50.

3.2. Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was implemented in order to confirm correlation relationships between the three factors proposed (self-Efficacy, goals & cope self-efficacy). A maximum likelihood (ML) estimation method was performed, providing acceptable model fit to the data: X^2= 216.806; df = 83; p<0.001; CFI = 0.963; TLI = 0.954; RMSEA [90% CI] = 0.062 (0.052; 0.072); SRMR = 0.059.

As shown in Table 2, validity measures and the internal consistency were explored through Cronbach's alpha (α), Composite reliability (CR) and the Average Variance Extracted (AVE). These were confirmed in reliability and convergent validity measures.

Table 2: Cross-construct correlation and validity measures

Factors	α	CR	AVE	Goal	Self-Efficacy	Cope Self-Efficacy
Goal	0.800	0.834	0.568	**0.754**		
Self-Efficacy	0.878	0.866	0.624	0.406	**0.790**	
Cope Self-Efficacy	0.896	0.887	0.531	0.365	0.540	**0.728**

Reliability measures: α > 0.7; CR > 0.7; Convergent validity: AVE > 0.5.

3.3. The validity of the causal structure

When CFA was confirmed, the path diagram modelled was tested in this phase. In order to determinate the predictive effects of barriers and supports on cognitive variables (self-efficacy, goals & cope self-efficacy), a SEM analysis was conducted using ML estimation. The SEM estimation including ML produced the following goodness of fit indices: X^2= 405,946; df = 193; p<0.001; CFI = 0.957; TLI = 0.938; RMSEA [90% CI] = 0.051 (0.044; 0.058); SRMR = 0.079.

The results of path relationships confirmed the model proposed (Fig. 1), indicating a significant standardized regression coefficient and a positive relationship between family supports and self-Efficacy (SE), goals (GL) and cope self-efficacy (CSE), thus, confirming H2a (β=0.192; ρ<0.05), H2b (β=0.166; ρ<0.05),

and H2c (β=0.135; ρ<0.05); at the same time, a relationship between teacher support and CSE was identified, confirmed H4c (β=0.186; ρ<0.05).

Accordingly, negative effects were found on individual variables because barriers were perceived during first year; social barriers on SE & CSE was established as a significant negative condition, confirming H5a (β=-0.185; ρ<0.05) and H5c (β=-0.129; ρ<0.05). Also, teacher barriers had negative effects on SE & CSE, confirming H8a (β=-0.161; ρ<0.05) and H8c (β=-0.300; ρ<0.001); while racial barriers had negative effects on Goals, confirming H9b (β=-0.212; ρ<0.05). In relation to economic barriers, although effects on CSE were identified as significant, the effects were not recognized as negative, thus, rejecting H6c (β=0.11; ρ<0.05).

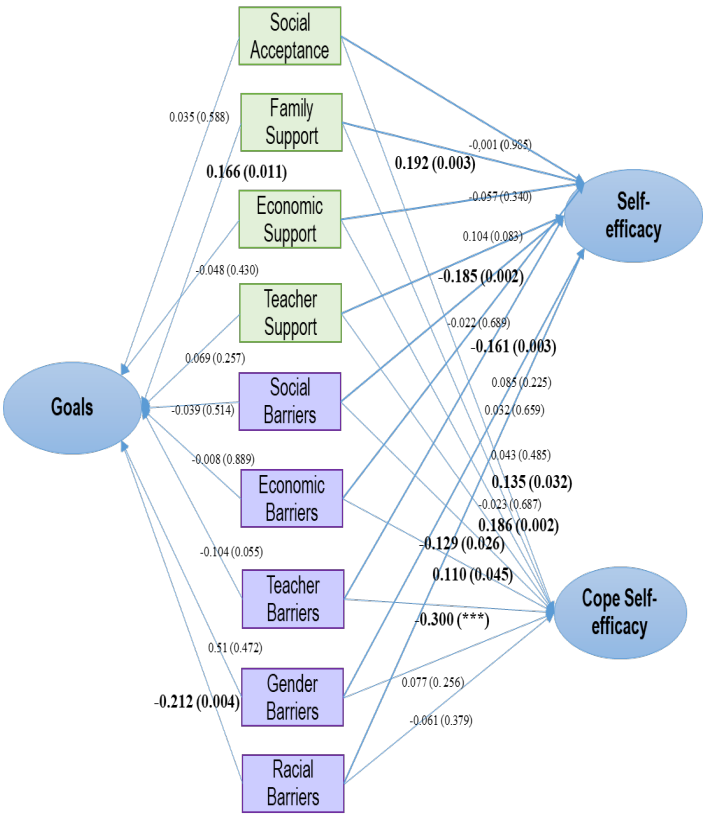


Figure 1: Path relationships confirmed the model proposed. *** p < 0.001; **p < 0.01; *p < 0.05

4. Discussions and Conclusions

The present research aimed to generate insights into contextual factors that may hinder or promote Peruvian engineering student's career progress during their first-year. The present study identified that family support was the major factor that contributes to self-efficacy, goals and cope self-efficacy. A possible explanation for this is that family support may represent acceptance among students because it guarantees economic support for other family members. This study confirms the findings from Peña-Calvo and colleagues [21], who found that family has a strong direct and indirect influence on SCCT factors in undergraduate engineering students. This finding reflects the importance of the effective role that family play in student's college life. For instance, parents may need to understand that students not only face academic challenges but also social and personal challenges (making friends, extracurricular activities, take own responsibility, time management, etc.). Future studies should assess and intervene family dynamics (e.g. authoritarian, permissive parenting style) and misconceptions that parents of first-year students may have about college life.

One interesting finding was the positive and negative effects of teacher's support on cope self-efficacy. This result may be explained by the fact that most of the students in our sample came from families who only had basic education and low income. In addition to this, Peruvian national state universities lack specialized online resources for engineering students, thus, making the role of teacher support critical for students coping and success. These conditions may limit students to find extra academic support and make the role of teachers critical to influence student's capacity to solve issues during their career. These results further support the idea that contextual factors among Peruvian students may influence the decision to continue or dropout higher education [4]. This finding also accords with earlier reports that identified that poor quality in teaching and advice influenced the decision to leave engineering careers [9]. These results have practical implications for Peruvian engineering faculties and teachers because first-year courses are challenging and there are high expectations of academic performance among students [9,10]. Teachers may have the potential to boost self-efficacy on students by providing adequate supports such as positively interpreting their failures, reduce career biases and minimize possible withdrawal.

Another interesting finding was the effect of social barriers (e.g. family, friends) on self-efficacy and coping self-efficacy. It seems possible that negative comments, a sense of not belonging to that particular group and family pressure may diminished students repertoire to cope with adverse situations as well as their judgment on their own capabilities. This result was also reported by Peña-Calvo and colleagues [21] who concluded that peer barriers is a predictor of self-efficacy and goals, which may be determinant for students career development. This finding further support the idea of Wilcox and colleagues [25] who showed the importance of family and friends support in first-year students. This finding highlights implications for student's social relationships and family dynamics. With intention or due to a lack of knowledge, family members and students may negatively influence in the decision to continue engineering degree. As a result, this situation may increase engineering career gaps in Peruvian engineering faculties. Therefore, strategies to enhance positive student's relationships and provide training to families on how to support first-year students may be beneficial.

In this study, racial discrimination had effects on goals. One possible explanation is that racial discrimination may be perceived as a long-term issue and this may affect student's determination to continue engineering career. Another possible explanation is the lack of policies related to race/ethnicity discrimination in student's universities. This finding is consistent with previous studies that found that racial discrimination influence student's decisions to leave engineering career [14, 36, 37]. Further research should be undertaken to investigate other

types of aversive interactions among students such as bullying, sexual orientation and sexual harassment.

The aim of the present research was to examine specific barriers and supports and possible effects on self-efficacy, coping-self efficacy and goals among first-year Peruvian engineering college students. This study may lay the foundation for further exploration in college dropouts, teacher support, racial discrimination and the role of peers in continuing engineering careers. For instance, further studies may explore these issues using qualitative research or longitudinal designs.

A limitation of this study is the low number of women, which may biased results. However, this gender proportion was aligned with national statistics reports [4]. The major limitation of this study is the low number of items to explore specific barriers and supports. Despite its exploratory nature, this study offers some insights into environmental factors that influence engineering career development in the Peruvian context. Further research might explore these weaknesses by using robust psychometric questionnaires and/or conducting multi-group analyses. These findings suggest several courses of action at many levels in Peruvian universities with engineering careers. For instance, policies that systematically assess these specific barriers and supports among students may be helpful to promote engineering persistence in first-year students. In addition to this, economic incentives from government agencies to universities who minimize these barriers (e.g. by creating students unions) may be useful and beneficial to students and society as a whole. In addition to this, training psychologists and tutors with SCCT framework may boost the quality of staff and student's relationships.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We acknowledge the support of the Project 104-2017-FONDECYT funded by *Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica – CONCYTEC* and *Consorcio de Investigación Económica y Social - CIES.*

References

[1] R. Atkinson, M. Mayo, Refueling the U . S . Innovation Economy : Fresh Approaches to Science, Technology, Engineering and Mathematics (STEM) Education, 2010.

[2] R.D. Atkinson, How ICT Can Restore Lagging European Productivity Growth, 2018.

[3] A. Fiszbein, C. Cosentino, B. Cumsille, El desafío del desarrollo de habilidades en América Latina: Un diagnóstico de los problemas y soluciones de política pública, Washington, DC, 2016.

[4] Superintendencia Nacional de Educación Superior Universitaria SUNEDU, II Informe Bienal sobre la realidad universitaria en el Perú, 2020.

[5] Asamblea Nacional de Rectores, Datos Estadísticos Universitarios, Lima, 2011.

[6] R.W. Lent, M.J. Miller, P.E. Smith, B.A. Watford, K. Hui, R.H. Lim, "Social cognitive model of adjustment to engineering majors: Longitudinal test across gender and race/ethnicity," Journal of Vocational Behavior, **83**(1), 22–30, 2013, doi:10.1016/j.jvb.2014.11.004.

[7] K.E. Snyder, S.M. Barr, N.B. Honken, C.M. Pittard, P.A.S. Ralston, "Navigating the First Semester: An Exploration of Short-Term Changes in Motivational Beliefs Among Engineering Undergraduates," Journal of Engineering Education, **107**(1), 11–29, 2018, doi:10.1002/jee.20187.

[8] E. Godfrey, T. Aubrey, R. King, "Who leaves and who stays? Retention and attrition in engineering education," Engineering Education 2010: Inspiring the Next Generation of Engineers, EE 2010, **0052**, 2010, doi:10.11120/ened.2010.05020026.

[9] R.M. Marra, K.A. Rodgers, D. Shen, B. Bogue, "Leaving engineering: A multi-year single institution study," Journal of Engineering Education, **101**(1), 6–27, 2012, doi:10.1002/j.2168-9830.2012.tb00039.x.

[10] B.N. Geisinger, D.R. Raman, "Why they leave: Understanding student attrition from engineering majors," International Journal of Engineering Education, **29**(4), 914–925, 2013.

[11] R.W. Lent, S.D. Brown, G. Hacket, "Toward Unifying Social Cognitive THoery of Career and Academic Interest, Choice, and Performance," Journal of Vocational Behavior, **45**, 79–122, 1994.

[12] R.W. Lent, S. Brown, G. Hackett, "Contextual supports and barriers to career choice: A social cognitive analysis," Journal of Counseling Psychology, **47**(1), 36–49, 2000, doi:http://dx.doi.org/10.1037/0022-0167.47.1.36.

[13] R.W. Lent, T.R. Morris, L.T. Penn, G.W. Ireland, "Social-cognitive predictors of career exploration and decision-making: Longitudinal test of the career self-management model," Journal of Counseling Psychology, **66**(2), 184–194, 2019, doi:10.1037/cou0000307.

[14] R.W. Lent, H. Bin Sheu, M.J. Miller, M.E. Cusick, L.T. Penn, N.N. Truong, "Predictors of science, technology, engineering, and mathematics choice options: A meta-analytic path analysis of the social-cognitive choice model by gender and race/ethnicity," Journal of Counseling Psychology, **65**(1), 17–35, 2018, doi:10.1037/cou0000243.

[15] A. Bandura, "Self-Efficacy Mechanism in Human Agency," American Psychologist, **37**(2), 122–147, 1982.

[16] A. Bandura, "Perspectives on Psychological Science Toward a Psychology of Human Agency," 2013, doi:10.1111/j.1745-6916.2006.00011.x.

[17] R.W. Lent, S.D. Brown, "On Conceptualizing and Assessing Social Cognitive Constructs in Career Research : A Measurement Guide," **14**(1), 12–35, 2006, doi:10.1177/1069072705281364.

[18] R.W. Lent, S.D. Brown, J. Schmidt, B. Brenner, H. Lyons, D. Treistman, "Relation of contextual supports and barriers to choice behavior in engineering majors: Test of alternative social cognitive models.," Journal of Counseling Psychology, **50**(4), 458–465, 2003, doi:10.1037/0022-0167.50.4.458.

[19] I. Fort, A. Murariu, "The Paths Between Gender, Barriers, Social Support, Coping Efficacy, and Educational Goals," Journal of Career Assessment, **26**(1), 68–76, 2018, doi:10.1177/1069072716679924.

[20] M.F. Shoffner, D. Newsome, C.A. Barrio Minton, C.A. Wachter Morris, "A qualitative exploration of the STEM career-related outcome expectations of young adolescents," Journal of Career Development, **42**(2), 102–116, 2015.

[21] J. Peña-Calvo, M. Inda-caro, C. Rodriguez-Menèndez, C. Fernandez-Garìa, "Perceived supports and barriers for career development for second-year STEM students," Journal of Engineering Education, **105**(2), 2016, doi:10.1002/jee.20115.

[22] J.J. Dahling, M.N. Thompson, "Contextual supports and barriers to academic choices: A policy-capturing analysis," Journal of Vocational Behavior, **77**(3), 374–382, 2010, doi:10.1016/j.jvb.2010.07.007.

[23] N.A. Fouad, G. Hackett, P.L. Smith, N. Kantamneni, M. Fitzpatrick, S. Haag, D. Spencer, "Barriers and Supports for Continuing in Mathematics and Science: Gender and Educational Level Differences," Journal of Vocational Behavior, **77**(3), 361–373, 2010, doi:10.1016/j.jvb.2010.06.004.

[24] M.C. Cadaret, P.J. Hartung, L.M. Subich, I.K. Weigold, "Stereotype threat as a barrier to women entering engineering careers," Journal of Vocational Behavior, **99**, 40–51, 2017, doi:10.1016/j.jvb.2016.12.002.

[25] P. Wilcox, S. Winn, M. Fyvie-Gauld, "'It was nothing to do with the university, it was just the people': The role of social support in the first-year experience of higher education," Studies in Higher Education, **30**(6), 707–722, 2005, doi:10.1080/03075070500340036.

[26] K.A. Tate, W. Caperton, D. Kaiser, N.T. Pruitt, H. White, E. Hall, "An Exploration of First-Generation College Students' Career Development Beliefs and Experiences," Journal of Career Development, **42**(4), 294–310, 2015, doi:10.1177/0894845314565025.

[27] M.E. Kenny, D.L. Blustein, A. Chaves, J.M. Grossman, L.A. Gallagher, "The role of perceived barriers and relational support in the educational and vocational lives of urban high school students," Journal of Counseling Psychology, **50**(2), 142, 2003.

[28] R.W. Lent, S.D. Brown, B. Brenner, S.B. Chopra, T. Davis, R. Talleyrand, V. Suthakaran, "The Role of Contextual Supports and Barriers in the Choice of Math / Science Educational Options : A Test of Social Cognitive Hypotheses," Journal of Counseling Psychology, **48**(4), 474–483, 2001, doi:10.1037//0022-0167.48.4.474.

[29] M. Inda, C. Rodríguez, J.V. Peña, "Gender differences in applying social

cognitive career theory in engineering students," Journal of Vocational Behavior, **83**(3), 346–355, 2013, doi:10.1016/j.jvb.2013.06.010.
[30] Superintendencia Nacional de Educación Superior Universitaria SUNEDU, Informe Bienal sobre la realidad universitaria en el Perú, 2018.
[31] M. Rodriguez-Menendez, M. Inda-Caro, J. Peña-Calvo, "Validacion de la teorìa cognitivo social de desarrollo de la carrera con una muestra de estudiantes de ingenierìa.," Educacion **18**(2), 257–276, 2015, doi:10.5944/educXX1.14018.
[32] J. Muñiz, P. Elosua, R.K. Hambleton, "Directrices para la traducción y adaptación de los tests: segunda edición [Guidelines for the translation and adaptation of test:]," Psicothema, **25**(2), 151–157, 2013.
[33] M. Inda-Caro, C. Rodríguez-Menéndez, J.V. Peña-Calvo, "Spanish High School Students' Interests in Technology: Applying Social Cognitive Career Theory," Journal of Career Development, **43**(4), 291–307, 2016, doi:10.1177/0894845315599253.
[34] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, R.L. Tatham, Multivariate data analysis, Prentice hall Upper Saddle River, NJ, 1998.
[35] B. Byrne, Structural equation modeling with AMOS: basic concepts, applications, and programming, 2010, doi:10.4324/9781410600219.
[36] L.Y. Flores, R.L. Navarro, S.R. Ali, "The State of SCCT Research in Relation to Social Class: Future Directions," Journal of Career Assessment, **25**(1), 6–23, 2017, doi:10.1177/1069072716658649.
[37] R.L. Navarro, L.Y. Flores, H.S. Lee, R. Gonzalez, "Testing a longitudinal social cognitive model of intended persistence with engineering students across gender and race/ethnicity," Journal of Vocational Behavior, **85**(1), 146–155, 2014, doi:10.1016/j.jvb.2014.05.007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 926-936 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Prediction of Vessel Dynamic Model Parameters using Computational Fluid Dynamics Simulation

Nu'man Amri Maliky[1,*], Nanda Pratama Putra[2], Mochamad Teguh Subarkah[3], Syarif Hidayat[4]

[1]*Faculty of Mechanical and Aerospace Engineering, Aerospace Engineering, Institut Teknologi Bandung, 40132, Indonesia*

[2]*Faculty of Mechanical and Aerospace Engineering, Mechanical Engineering, Institut Teknologi Bandung, 40132, Indonesia*

[3]*School of Electrical Engineering and Informatics, Electrical Engineering, Institut Teknologi Bandung, 40132, Indonesia*

[4]*School of Electrical Engineering and Informatics, Electrical Power Engineering, Institut Teknologi Bandung, 40132, Indonesia*

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 17 November, 2020
Online: 08 December, 2020

Keywords:
Hydrodynamic Forces
Dynamic Positioning
Computational Fluid Dynamics
Control System Parameters
Autonomous Boat
Computational Simulation

ABSTRACT

Dynamic positioning, a system to maintain vessel position and heading, is a technology that is used in many vessels and being intensively used as research topic in marine engineering. In order to make this system work properly, an accurate parameters value is needed. This research focuses on finding several parameters needed in this control system, which are resistance and added mass. These parameters are identified using CFD simulation. The method has an advantage of being fairly high accuracy and has a lower cost than the experimental method. The results released from this simulation verified by several empirical methods, namely Holtrop and Mennen for resistance and Ellipsoid for added mass. Baruna Jaya III is used as an implementation object for the simulation. The simulation resulted in a small error compared to the verification method. So, this computational fluid dynamics simulation can be an alternative method for obtaining resistance and added mass values on ships.

1. Introduction

This paper is an extension of work originally presented in ICEVT 2019 [1]. Vessel is still considered as one of the most important vehicles for transportation as it is the largest vehicle for transporting objects in the world. Although it maneuvers at a slow speed, shipping transportation is the most effective mode for moving objects in large quantities and masses. In terms of cost, freight transport by sea also known to be efficient. Not only moving objects, several jobs, such as oil refineries, cable planting, system checking, and welding on the seabed, to monitoring marine areas; require underwater vehicles to facilitate. Recent research in this field has led to a vessel that is more efficient, sustainable, and reduces many human errors while working.

Research that is being intensively currently in progress is the autonomous control system, a control system that allows a vehicle to operate automatically without human intervention. One of the control systems that currently being developed is dynamic positioning. Dynamic Positioning (DP) is a computer-controlled system that automatically maintain the position and the direction of vessel using existing thrusters. It takes many components that must be considered to create a DP control system, such as: computers, sensors, drivers, etc. The determination of the hydrodynamic parameters of the vessel is considered to be important in designing the DP control system.

When talking about vessel, certainly it is necessary to also discuss hydrodynamics. Hydrodynamics is the study of fluid motion with media/substances in the form of a liquid, where the substance cannot be pressed (incompressible), which are influenced by external and internal forces. In the DP control system, there are two important parameters that must be extracted from hydrodynamics, which are drag and added mass. These hydrodynamic forces values can be obtained in several ways, including analytical, experimental, and numerical.

Computational fluid dynamics is the commonly used numerical to obtain the hydrodynamic forces that occur in a vessel. This method is used because it has a fairly high accuracy for initial predictions compared to analytics, but it doesn't require much cost than experiments. To be able to use this method, it is necessary to do some modeling from the real problem to the equations that can be understood by the computer system that will calculate the model.

*Corresponding Author: Nu'man Amri Maliky, maliky.numan@gmail.com

This paper presents the work of CFD simulation that composed into 8 sections. Section 1. Introduction, talking about the review of the CFD simulation research. Section 2. Literature Reviews, mentioning some literatures that support the work. Section 3. Materials and Methodologies, containing some materials like the system, equation, parameters, vessel dimension, and methodology to conduct the simulation. Section 4. Result and Analysis, presenting the simulation result and analyze it. Section 5. Conclusion and Future Works, presenting the conclusion of the work and review what needed to be developed on next research. Section 6. Conflict of Interests, statement from the author for submitting the paper. Section 7. Acknowledgements, gratitude from the authors to whom revise or give any critics to this work. And last Section 8. Reference, presenting the list of reference used in this work.

2. Literature Reviews

There are several literatures that serve as references and paradigms in this research and paper. Basically, this work is referring to [1], which discuss the use of computational fluid dynamics on a barge to find dynamic positioning control system parameters. This publication is the continuation and the expansion of reference [1].

Vessel modeling and control system design is explained further in [2]. This thesis discusses the equation of motion of a ship, where there are several basic parameters that affect the moving ship, namely: mass, Coriolis, and drag/resistance. These parameters will be sought from a ship design to design its dynamic positioning control system [2].

One of the parameters to obtain is added mass. Paper that discuss further about added mass identification is found on [3]. The method to design using computational fluid dynamics to obtain the added mass value is also stated on the reference. The other parameter is drag or resistance. The presence of ship resistance is caused by a combination of several factors. These factors are heavily influenced by a dimension of the ship [3].

Validation is carried out to ensure that the simulation that has been done correctly. The validation uses the empirical calculation of Holtrop and Mennen's Method described in [4]. The calculation for the parameters needed. For added mass calculation, validation uses the ellipsoid and/or strip theory method. These methods explained well in [4], [5].

For the basic theory of computational fluid dynamics and the additional theory are explained well in [6], [7].

Another reference used is the proceedings "Preliminary Computational Fluid Dynamics (CFD) Simulation of EIIB Push Barge in Shallow Water". This paper describes a method for finding the resistance of a ship using computational fluid dynamics. The formation of ship geometry to support the method discussed is described in "Analysis of Hull Resistance of Pushed Barges in Shallow Water". Then, the required mesh is discussed in "Viscous Drag Calculations for Ship Hull Geometry" [8]-[10].

This research main goal is to obtain the hydrodynamic parameters from a vessel through CFD simulation. The detailed parameters can be found from mentioned reference above which are in [1], [2]. The parameters are added mass and vessel drag, where the calculation explained in [3]. This simulation result needs to be verified with the others method. The verification can be done by a method mentioned in [4], [5]. The correct way to do a proper CFD simulation can be found in [8]-[10].

3. Materials and Methodologies

3.1. Dynamic Positioning Control System

DP System consists of the position reference system, propulsion system, and computer control at the deck room. Position reference system acting as a navigation and guidance system for DP. The sensors commonly used for this subsystem are GPS and gyro-compasses. This will inform the vessel position and heading and act as state feedback for control calculation as well.

The control calculation is conducted on the DP computer, which generates a thrust command to each thruster every defined sampling time. The signal from feedback control calculation is converted through a certain allocation algorithm. The full block diagram of the control system shown in Figure 1.

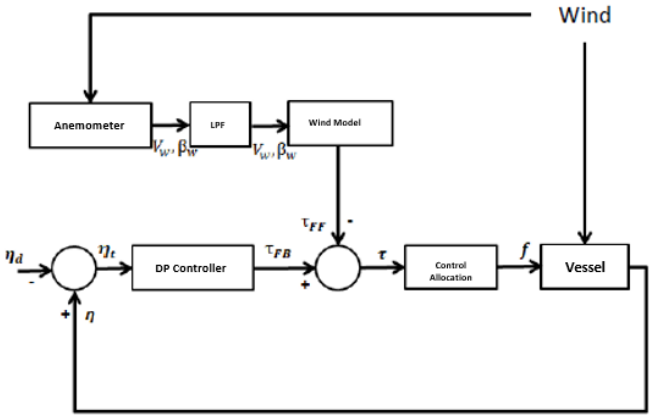


Figure 1: Dynamic Positioning Control System

A wind feedforward control is a common algorithm implemented on DP nowadays. It prevents a bigger deviation of position earlier. The wind signal is filtered with low pass filter to prevent a big oscillation of the control signal. Feedback, feedforward, control allocation, and full control system implementation are explained in [2].

In order to design or implement the controller, a vessel model is required. The model is derived from its dynamic equation of motion which consists of mass, Coriolis, and drag. The equation is can be found in equation 1 with a further detail explanation of the equation can be found in [2].

$$\mathbf{M\dot{v}} + \mathbf{C(v)\dot{v}} + \mathbf{D(v)\dot{v}} = \boldsymbol{\tau} + \boldsymbol{\tau_e} \quad (1)$$

M, C(v), and D(v) stand for mass, Coriolis, and Drag. While τ and τ_e are internal force and external force respectively. DP controls the 3 DoF motion of the vessel, surge, sway, and yaw. Equation 1 is modeled with respect to the motion so the fixed variable, mass, drag, and Coriolis are 3x3 matrix while the rest is a 3x1 matrix.

Mass and Coriolis consist of a rigid body and added mass. While there are linear and nonlinear drags. The variable of the parameters is shown in equation 2 to 10.

$$\mathbf{M} = \mathbf{M_{RB}} + \mathbf{M_{AM}} \tag{2}$$

$$\mathbf{M_{RB}} = \begin{bmatrix} m & 0 & -m\,y_g \\ 0 & m & m\,x_g \\ -m\,y_g & m\,x_g & I_z \end{bmatrix} \tag{3}$$

$$\mathbf{M_{AM}} = \begin{bmatrix} -X_{\dot{u}} & 0 & 0 \\ 0 & -Y_{\dot{v}} & -Y_{\dot{r}} \\ 0 & -N_{\dot{v}} & -N_{\dot{r}} \end{bmatrix} \tag{4}$$

$$\mathbf{C(v)} = \mathbf{C_{RB}} + \mathbf{C_{AM}} \tag{5}$$

$$\mathbf{C_{RB}} = \begin{bmatrix} 0 & 0 & -m(x_G r + v) \\ 0 & 0 & -m(y_g r - u) \\ m(x_G r + v) & m(y_g r - u) & 0 \end{bmatrix} \tag{6}$$

$$\mathbf{C_{AM}} = \begin{bmatrix} 0 & 0 & Y_{\dot{v}} v + \frac{Y_{\dot{r}} + N_{\dot{v}}}{2} r \\ 0 & 0 & -X_{\dot{u}} u \\ -Y_{\dot{v}} v - \frac{Y_{\dot{r}} + N_{\dot{v}}}{2} r & X_{\dot{u}} u & 0 \end{bmatrix} \tag{7}$$

$$\mathbf{D(v)} = \mathbf{D_{linear}} + \mathbf{D_{non-linear}} \tag{8}$$

$$\mathbf{D_{linear}} = \begin{bmatrix} X_u & 0 & 0 \\ 0 & Y_v & 0 \\ 0 & 0 & N_r \end{bmatrix} \tag{9}$$

$$\mathbf{D_{non-linear}} = \begin{bmatrix} X_{u|u|} & 0 & 0 \\ 0 & Y_{v|v|}|v| + Y_{v|r|}|r| & Y_{r|v|}|v| + Y_{r|r|}|r| \\ 0 & N_{v|v|}|v| + N_{v|r|}|r| & N_{r|v|}|v| + N_{r|r|}|r| \end{bmatrix} \tag{10}$$

where

- M_{RB} : mass of constant variable (rigid body)
- M_{AM} : mass of mass dependent variable (added mass)
- C_{RB} : coriolis of constant variable (rigid body)
- C_{AM} : coriolis of mass dependent variable (added mass)
- $X_{\dot{u}}$: surge added mass dependent to surge acceleration
- $Y_{\dot{v}}$: sway added mass dependent to sway acceleration
- $N_{\dot{r}}$: yaw added mass dependent to yaw acceleration
- $Y_{\dot{r}}$: sway added mass dependent to yaw acceleration
- $N_{\dot{v}}$: yaw added mass dependent to sway acceleration
- X_u : surge drag dependent to surge velocity
- Y_v : sway drag dependent to sway velocity
- N_r : yaw drag dependent to yaw velocity
- $X_{u|u|}$: surge drag dependent to surge velocity
- $Y_{v|v|}$: sway drag dependent to sway velocity
- $N_{r|r|}$: yaw drag dependent to yaw velocity

The rigid body of the vessel could be measured. While the added mass and drag parameters should be derived from a computational simulation through Computational Fluid Dynamics (CFD) simulation. Therefore, further hydrodynamic analysis is required to generate an exact model of the vessel.

3.2. Hydrodynamic Forces on a Ship

Hydrodynamic forces on ships can be classified into two forces, which are drag or resistance and added mass. Cited from [3], for an object which moves through a surrounding fluid, the force balance would be denoted by the equation of motion as equation 11. The separated component of hydrodynamics forces can be seen in equations 12 and 13.

$$M\mathbf{a} = \mathbf{F} - \mathbf{F_H} \tag{11}$$

where

m = Mass of the object

a = Acceleration of the object

F = Driving forces of the object

F_H = Hydrodynamic forces acting on the object

$$\mathbf{F_H} = \mathbf{F_D} + \mathbf{F_A} \tag{12}$$

$$\mathbf{F_A} = m_A\mathbf{a} \tag{13}$$

where

F_D = Drag forces acting on the object

F_A = Hydrodynamic reaction forces

m_A = Proportionality constant (added mass)

a = Acceleration of the object

The main idea was to accelerate the fluid that surrounds the body and measures the forces that are acting on the surface of the body. By subtracting the drag F_D from the total forces, we could obtain the hydrodynamic reaction forces F_A that is proportional to the acceleration of the fluid (i.e. the body). [3]

- Resistance of a ship

As the ship moves through the water, the ship would experience opposing forces which are often referred to as drag or resistance. The presence of the drag is caused by a combination of several factors. Some major components of the total drag can be denoted as written in equation 14.

$$R_T = R_V + R_W + R_{AA} \tag{14}$$

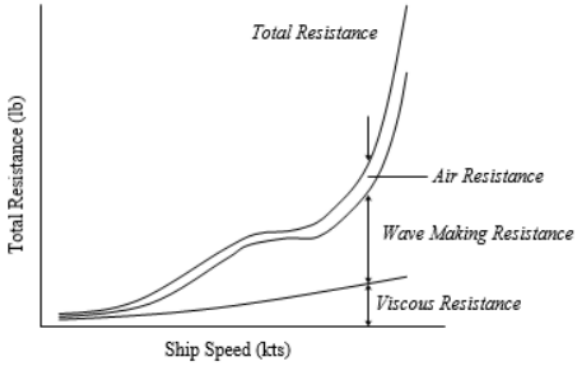
where

R_T = Total hull resistance

R_V = Viscous (friction) resistance

R_W = Wave making resistance

R_{AA} = Air resistance caused by ship moving through calm air



As portrayed in Figure 2, the effect of the wave making resistance tends to dominates as the speed of the vessel increases.

In order to quantify this effect, a dimensionless parameter, "Froude number" (F_n) is used. The Froude Number defined as equation 15.

$$F_n = V/\sqrt{gL} \quad (15)$$

where

V = Velocity

g = Gravity

L = L_{PP}, L_{BP} or L_{WL} of the ship

Sketch of L_{PP}, L_{BP} and L_{WL} of ship can be seen on Figure 3.

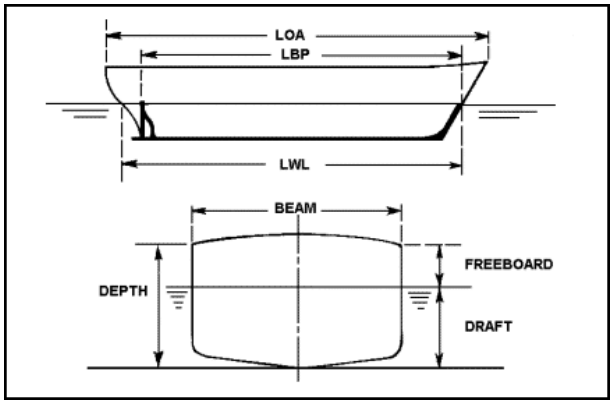


Figure 3: Ship Parameter

L_{PP} or L_{BP} is the length between perpendicular, a longitudinal distance between the forward and aft perpendiculars, and L_{WL} is the length of the hull that is intersecting with the surface of the water. The relation between this Froude number and the effect of the generated wave can be seen in Figure 4. Typically, at service speed, the effect of surface dynamics on the vessel cannot be ignored.

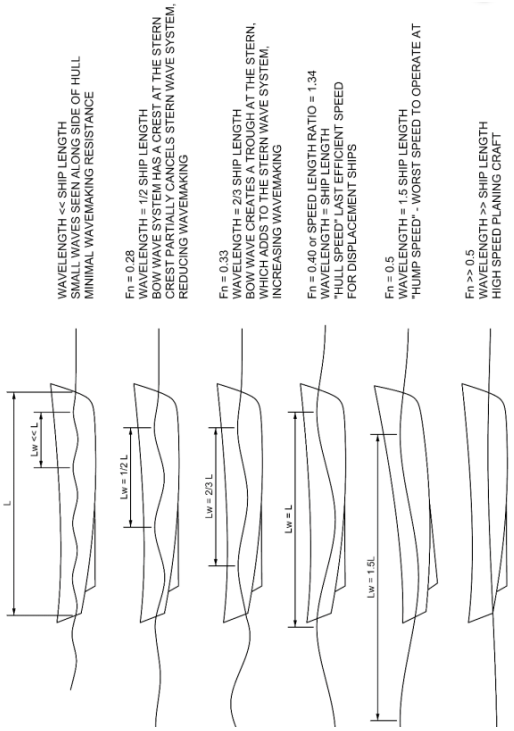


Figure 4: The Effect of Vessel's Speed on Generated Wavelength

The drag or resistance of the ship can be calculated using analytical, computational, and experimental method. Each method has its own advantages and disadvantages. One of the analytical methods that commonly used is Holtrop and Mennen's method. This method used a regression analysis of a vast range of model tests and trial data. However, recommended that, in order to obtain reasonable results from Holtrop's method, the ship should fulfill the following criteria, written in equation 16 to 18. [4]

$$F_n \leq 0.45 \quad (16)$$

$$0.55 \leq C_p \leq 0.85 \quad (17)$$

$$3.9 \leq L/B \leq 9.5 \quad (18)$$

where

F_n = Froude number

C_P = Prismatic coefficient

L/B = Length to breadth ratio

The prismatic coefficient is a dimensionless number from the ratio of the volume of the ship to the volume of the prism, a sketch of the picture can be seen in Figure 5. The equation can be written as in equation 19. [4]

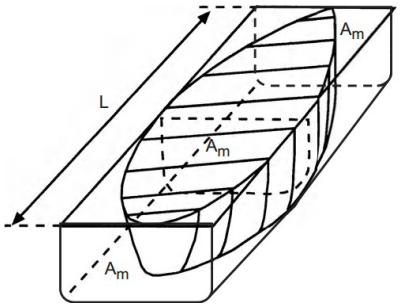


Figure 5: The Prismatic Coefficient

$$C_p = \frac{V}{A_m L} \quad (19)$$

where

V = Submerge volume of the ship

A_m = Maximum cross-sectional submerge area of the ship

L = L_{PP}, L_{BP} or L_{WL} of the ship

Holtrop method computes the total drag by dividing it into several components that can be calculated easily. Equation 20 provides the components contained in the Holtrop calculation. The calculation for each component is shown in equation 21 to 44. [4]

$$R_T = (1 + k)R_F + R_{APP} + R_W + R_B + R_{TR} + R_A + R_{AA} \quad (20)$$

where

R_F = frictional resistance

R_{APP} = appendage resistance

R_A = model–ship correlation resistance

R_W = wave resistance

R_B = resistance due to bulbous bow near the water surface

R_{TR} = pressure resistance due to immersed transom

R_{AA} = air resistance

$$R_F = \frac{1}{2}\rho v^2 S C_F \quad (21)$$

$$C_F = \frac{0.075}{[\log_{10}(Re)-2]^2} \quad (22)$$

$$Re = \frac{\rho v L_{WL}}{\mu} \quad (23)$$

$$k = -0.07 + 0.487118\, c_{14}\, [a_1\, a_2\, a_3\, a_4\, a_5] \quad (24)$$

$$a_1 = \left(\frac{B}{L_{WL}}\right)^{1.06806} \quad (25)$$

$$a_2 = \left(\frac{T}{L_{WL}}\right)^{0.46106} \quad (26)$$

$$a_3 = \left(\frac{L_{WL}}{L_R}\right)^{0.121563} \quad (27)$$

$$a_4 = \left(\frac{{L_{WL}}^3}{V}\right)^{0.36486} \quad (28)$$

$$a_5 = (1 - C_P)^{-0.604247} \quad (29)$$

$$R_{APP} = \frac{1}{2}\rho v^2 (1 + k_2)_{eq} C_F \sum_i S_{APP_i} + \sum R_{TH} \quad (30)$$

$$(1 + k_2)_{eq} = \frac{\sum_i (1+k_{2i}) S_{APP_i}}{\sum_i S_{APP_i}} \quad (31)$$

$$R_{TH} = \rho v^2 \pi d^2 C_{D_{TH}} \quad (32)$$

$$R_W(Fr) = \begin{cases} R_{W_a}(Fr) & \text{if } Fr \le 0.4 \\ R_{W_{a,b}}(Fr) & \text{if } 0.4 < Fr \le 0.55 \\ R_{W_b}(Fr) & \text{if } Fr \ge 0.55 \end{cases} \quad (33)$$

$$R_{W_a} = c_1 c_2 c_5 \rho g V \exp\left[m_1 Fr^d + m_4 \cos(\lambda Fr^{-2})\right] \quad (34)$$

$$R_{W_b} = c_{17} c_2 c_5 \rho g V \exp\left[m_3 Fr^d + m_4 \cos(\lambda Fr^{-2})\right] \quad (35)$$

$$R_{W_{a,b}} = R_{W_a}(0.4) + \frac{20Fr-8}{3}\left[R_{W_b}(0.55) - R_{W_a}(0.4)\right] \quad (36)$$

$$R_B = 0.11\rho g\left(\sqrt{A_{BT}}\right)^3 \frac{{Fr_i}^3}{1+{Fr_i}^2} e^{-3.0{P_B}^{-2}} \quad (37)$$

$$P_B = 0.56\frac{\sqrt{A_{BT}}}{T_F - 1.5h_B + h_F} \quad (38)$$

$$Fr_i = \frac{v}{\sqrt{g(T_F - h_B - 0.25\sqrt{A_{BT}} + h_F + h_W)}} \quad (39)$$

$$R_{TR} = \frac{1}{2}\rho v^2 A_T c_6 \quad (40)$$

$$R_A = \frac{1}{2}\rho v^2 (C_A + \Delta C_A)[S + \sum S_{APP}] \quad (41)$$

$$C_A = 0.00546(L_{WL} + 100)^{-0.16} - 0.002 + 0.003\sqrt{\frac{L_{WL}}{7.5}}\, {C_B}^4 c_2 (0.04 - c_4) \quad (42)$$

$$\Delta C_A = \begin{cases} 0 & \text{if } k_s = 150\ \mu m \\ \dfrac{0.105 k_s^{\frac{1}{3}} - 0.005579}{L_{WL}^{\frac{1}{3}}} & \text{if } k_s > 150\ \mu m \end{cases} \quad (43)$$

$$R_{AA} = \frac{1}{2}\rho v^2 A_V C_{DA} \quad (44)$$

where

ρ = density of the fluid

μ = viscosity of the fluid

v = ship velocity

L_{WL} = length in waterline

B = beam

T = mean draft

T_A = draft at aft perpendicular

T_F = draft at forward perpendicular

V = volumetric displacement

C_P = prismatic coefficient (based on L_{WL})

C_M = midship section coefficient

C_{WP} = waterplane area coefficient

C_{stern} = stern shape parameter

c_i, m_i = specific coefficient, detail explanation in [4]

A_V = area of ship and cargo above waterline

A_T = immersed transom area

A_{BT} = transverse area of bulbous bow

h_B = height of center of A_{BT} above basis

D = propeller diameter

S = wetted area of the hull

S_{APP} = wetted surface of appendages

- Added mass of a ship

As the vessel accelerates through the surrounding fluid, the vessel would experience an opposing force that is proportional to the magnitude of acceleration. Recalling the general equation of motion for the vessel is on equation 1, The matrix **M** could be separated into 2 components; the inertia matrix of the rigid body, and the added mass, can be seen in equation 2 to 4.

The computational fluid dynamics simulation can be a solution for obtaining the value of added mass. Another approach to calculate the added mass is explained in [5]. The paper compared the ellipsoid and Lewis transformation method to the experimental value and gives recommendations on which method to use.

The equivalent ellipsoid method is used to estimate the amount of added mass on the ship by modeling the ship as an ellipsoid form. The modeling sketch can be seen in Figure 6. Thus, the added mass value in the surge direction can be calculated using equation 45 to equation 48. [5]

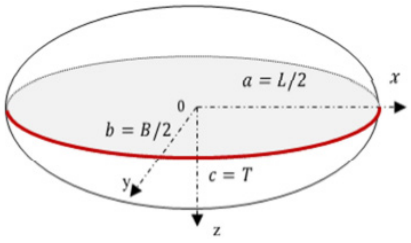


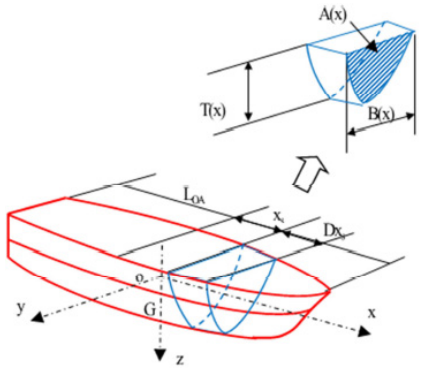
Figure 6: The ship is modeled as an ellipsoid

$$m_{11} = mk_{11} \quad (45)$$

$$k_{11} = \frac{A_0}{2-A_0} \quad (46)$$

$$A_0 = \frac{2(1-e^2)}{e^3}\left(\frac{1}{2}\ln\left(\frac{1+e}{1-e}\right) - e\right) \quad (47)$$

$$e = \sqrt{1-\frac{b^2}{a^2}} \quad (48)$$



Meanwhile, the strip theory method divides the ship into several pieces that resemble the hull shape of the ship. The sketch of this method can be seen in Figure 7. Due to the complexity of each hull cut to shape on the ship, the Lewis transform is used to simplify calculations. However, to simulate the added mass of ships in the surge direction, it is recommended to use the ellipsoid method. [5]

3.3. Computational Fluid Dynamics

Computational Fluid Dynamics is a method that uses numerical analysis and data structures to analyze and solve problems involving fluid flow. Computers are used to carry out the calculations needed to simulate fluid flow and fluid interactions with surfaces determined by boundary conditions. Analysis using the CFD method requires some modeling which is the conversion of physical data in the real world to a property understood by the software. Therefore, a technical approach is needed for each variable to be analyzed. Figure 8 shows the process of the computational fluid dynamics method [6].

Turbulence in CFD can be modeled with several governing equations. Each model has advantages and disadvantages. The choice of turbulent models is determined by the flow phenomenon to be simulated or analyzed. For obtaining the resistance and added mass, this research is using k-epsilon for the turbulence model. This model was chosen because it has the advantage of being robust and good for getting initial values based on what ANSYS has written. It also tends to require less computer memory [7].

Because the effect of surface dynamics on the vessel cannot be ignored, in order to obtain the whole value of drag force on the ship, it is necessary to conduct a two-phase simulation. The two-phase simulation consists of water on the bottom and air on the top of the computational domain. To perform a two-phase simulation, a computational domain that can capture the phenomena is needed. A suitable domain for the simulation described in [8], the sketch can be seen in the Figure 9.

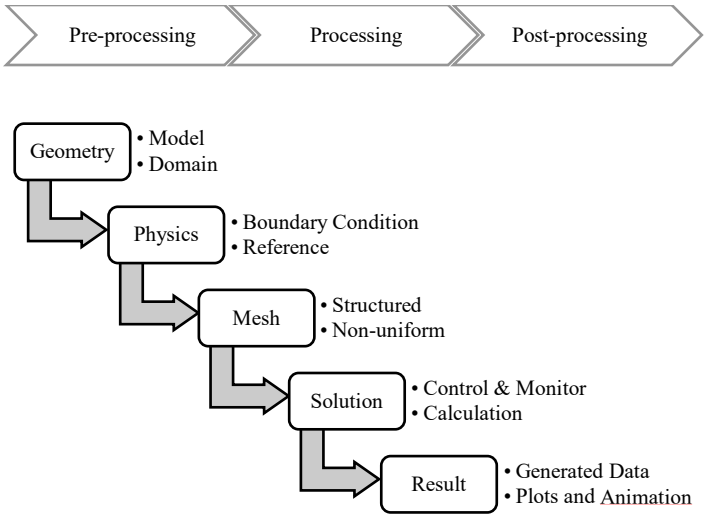
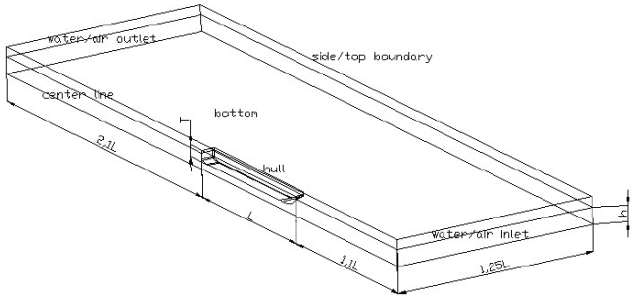
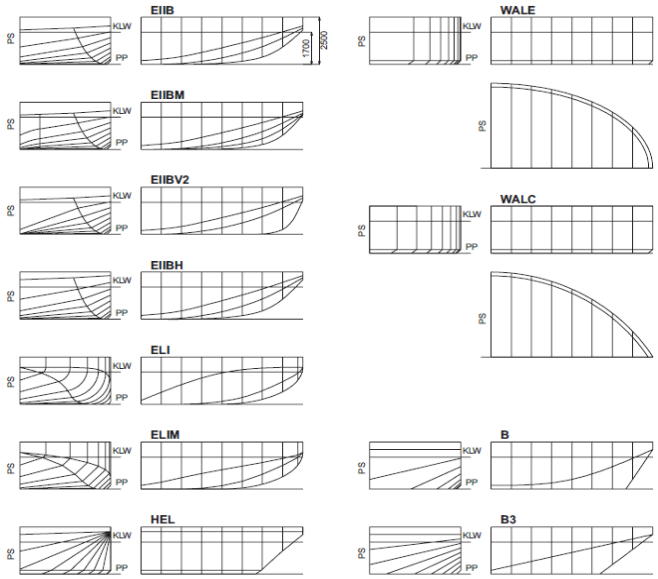


Figure 8: CFD Simulation Flowchart



The model of the ship should also be created in a way that would reflect the original ship. Oversimplification of the hull shape may result in a significant difference in the CFD results. For simulations in shallow water, the results might be significantly different from the ones that are conducted in the deep-water computational domain. As a rule of thumb, for barge ship, there exist several types of ship hulls discussed in [9]. These types of hulls can be seen on Figure 10.



For free surface flows, the meshing process needs to be able to discretize the boundary layer and the free surface. Scaling and the usage of structured mesh might be used to save computational time [8]. The computational domain for two-phase simulation could be created based on flat-plate boundary layer theory and setting the parameter $y^+ = 1$. Y^+ can be approached by using equation 49 [10].

$$y^+ = 0.172 \left(\frac{y}{L}\right) Re^{0.9} \quad (49)$$

where

y = Distance from wall surface

L = Body length

Re = Reynold numbers

3.4. Methodology

The methodology is divided into two parts: obtaining the value of resistance and added mass. Although using the same method, computational fluid dynamics, these two components have quite different steps for obtaining them because of the unique characteristics of each component. Flowcharts can be seen in Figure 11 and Figure 12 respectively.

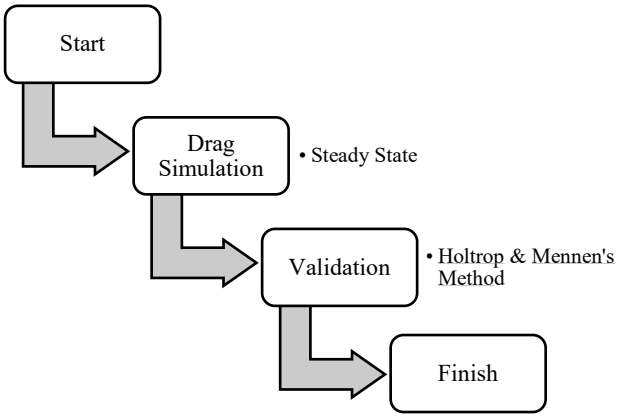


Figure 11: Drag Simulation Flowchart [8]

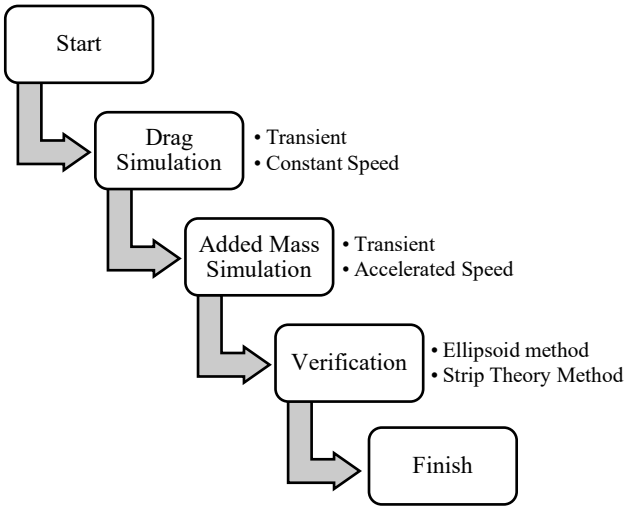


Figure 12: Added Mass Simulation Flowchart [3]

The CFD was conducted by using ANSYS FLUENT software. Both obtaining value of resistance and added mass use the same simulation settings and boundary conditions. Simulation settings and boundary conditions are listed in Table 1 and Table 2 respectively. The boundary condition for resistance is similar to the one that is listed in [8] and for added mass is similar to [3]. For the calculation of added mass, it is necessary to define the speed change of the ship in the simulation.

Table 1: Simulation Settings

Parameters	Settings
Solver	Pressure based, steady/transient
Model	Multiphase, VOF
Vol. Fraction Parameters	Open channel flow - implicit
Interpolation method	Presto! QUICK

Table 2: Boundary Condition

Region	Settings
Ship	Wall
Top/Side/Bottom	Wall
Inlet	Pressure inlet, open channel, vessel speed, free-surface and bottom level defined.
Outlet	Pressure outlet, open channel, free-surface and bottom level defined.
Symmetry	Symmetry

Air (FLUENT database) and sea water (user defined, $\rho = 1025$ kg/m^3, $\mu = 0.00109$ Pas) were used as the fluids in the analysis. The choice of viscous model such as k-epsilon or SST would significantly affect the result as demonstrated in [8]. As the symmetry method were used in the analysis, the results obtained to describe forces must be multiplied by two for the full-scale result. The value of coefficient might result in wrong amount if the references values were not defined. The simulation should be iterated until convergence criteria are achieved. Typically, the criteria for convergence are as follow.

- RMS residual values are below 10^{-4}
- The solution imbalance is less than 1%
- Quantity of interest shows a steady value

To make sure that the result obtained is mesh independent, the simulation above should be repeated several times with multiple number of mesh. The solution is considered to be mesh independent if the increase on number of mesh doesn't affect the result by significant values.

The result should be validated with data in order to make sure that the simulation reflect the actual situation. The easiest method would to compare the results with experimental data obtained from the vessel. However, the experimental data is unavailable, so, this research will use Holtrop and Mennen's method [4] for the validation of the drag and approximation for added mass.

3.5. Implementation

This particular research used ship called Baruna Jaya III as the geometric model and calculation references. Baruna Jaya III is one of the former TNI-AL warships built in 1990. This ship is used for mapping tasks for areas in Indonesia. Figure 13 shows the photo of the Baruna Jaya III ship. Technical drawings of Baruna Jaya III can be seen in Figure 14 to Figure 17.



Figure 13: Baruna Jaya III: Ship Photo

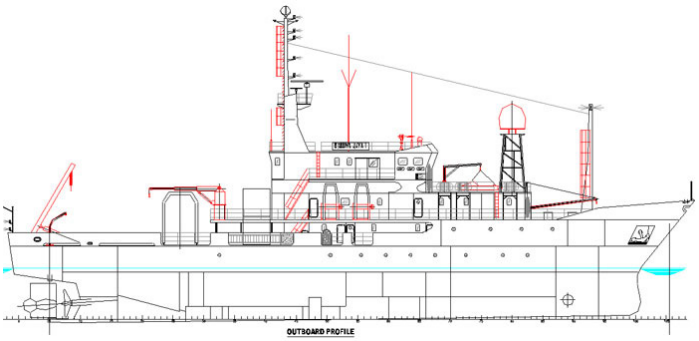


Figure 14: Baruna Jaya III: Outboard Profile

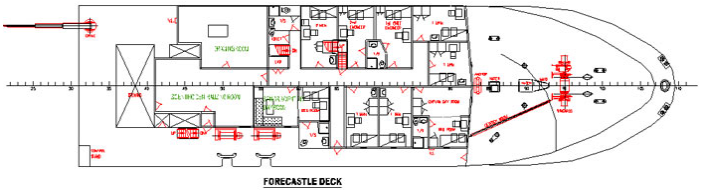


Figure 15: Baruna Jaya III: Wheelhouse deck dan Forecastel deck

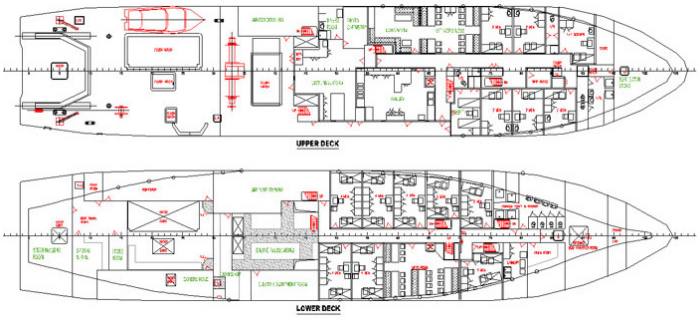


Figure 16: Baruna Jaya III: Upper deck dan lower deck

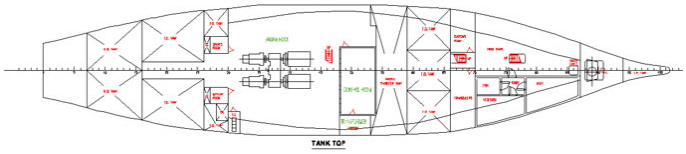


Figure 17: Baruna Jaya III: Tank Top

Table 3: Baruna Jaya III Dimension

Dimension	Value
LOA	60.40 m
LBP	52.39 m
Breadth	11.60 m
Draft Mean	5.70 m
Speed	8 knots

The pictures and dimensions listed in Table 3 will be used as references for modeling the ship to be carried out the simulation and calculating the analytical solution.

Geometry

The geometry of the ship was modelled to be able to represent its original shape while considering the approximation as discussed earlier. The geometry of the ship made must be able to represent its original shape. Therefore, the approximation discussed earlier is needed. Figure 19 shows the CAD of the Baruna Jaya Ship which has been approximated from its original form and ready for further processing. Detailed approximations are required on the bottom of the vessel which is subject to the seawater phase. This section needs to be quite detailed because even a few changes will affect the final result.

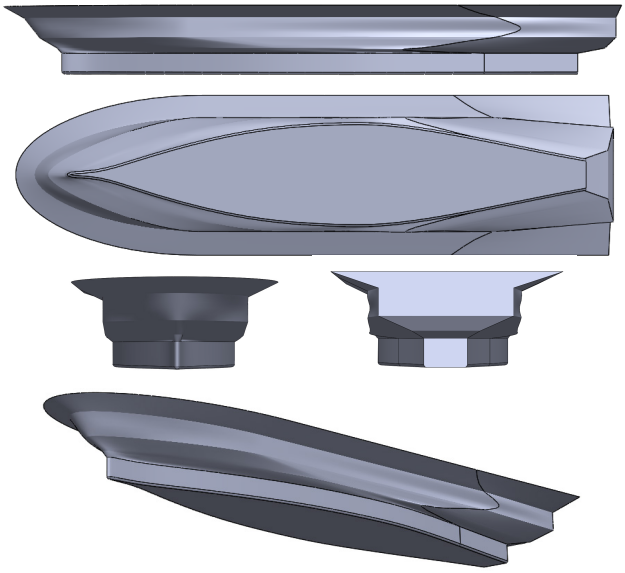
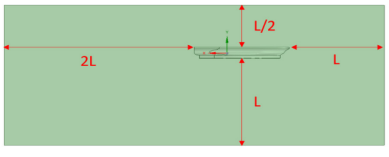
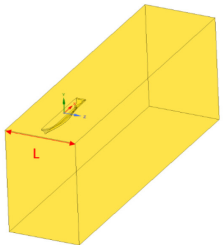


Figure 18: CAD for CFD Simulation of Baruna Jaya III

Meshing

After obtaining a suitable geometry to represent the ship, the next process is meshing the ship's far-field. The computational domain must be defined here. Figure 19 gives the computational domain used by this implementation.





Additional geometry in the form of blocks was added in the computational domain to form a body of influence in the computing domain. This is done to form a body of influence in the computing domain. Thus, the mesh will be smoother on that part. The computational domain for body of influence can be seen on Figure 20.

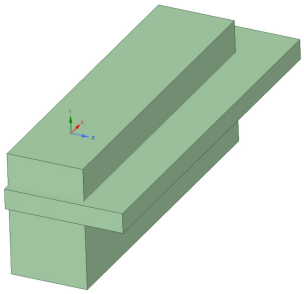


Figure 20: Computational Domain for Baruna Jaya III

The meshing process or discretization of the model is done using face sizing on the surface area of the ship with the dimensions of the mesh is 0.84 m. Meanwhile, for other computing domains, body sizing with a size of 1.2 m is used with the choice of the body of influence activated. The total number of mesh is 1,701,030 elements. The result mesh can be seen on Figure 21 and Figure 22.

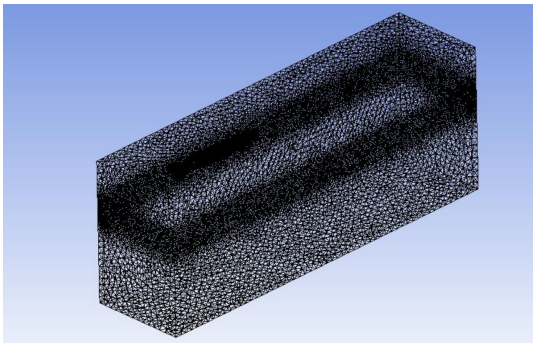


Figure 21: Mesh for Baruna Jaya III

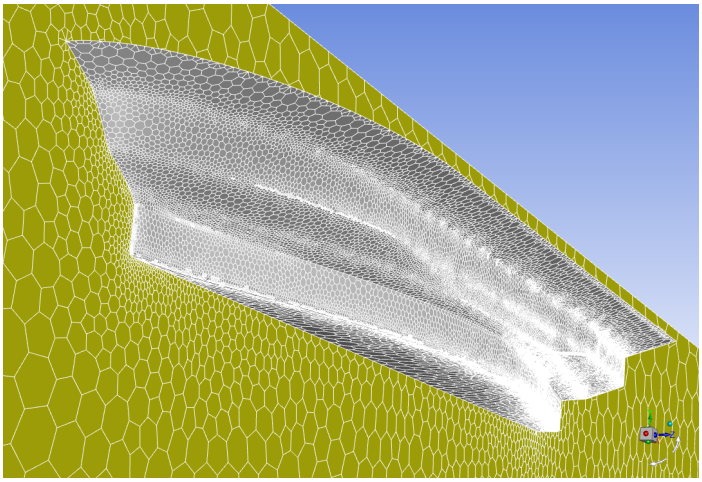


Figure 22: Symmetry and Hull Mesh of Baruna Jaya III

Simulation Setting

As mentioned before, simulation settings will follow [8] for resistance and [3] for added mass. Time setting should be steady for resistance and transient for added mass. Figure 23 shows the simulation settings, the other settings will follow the default one.

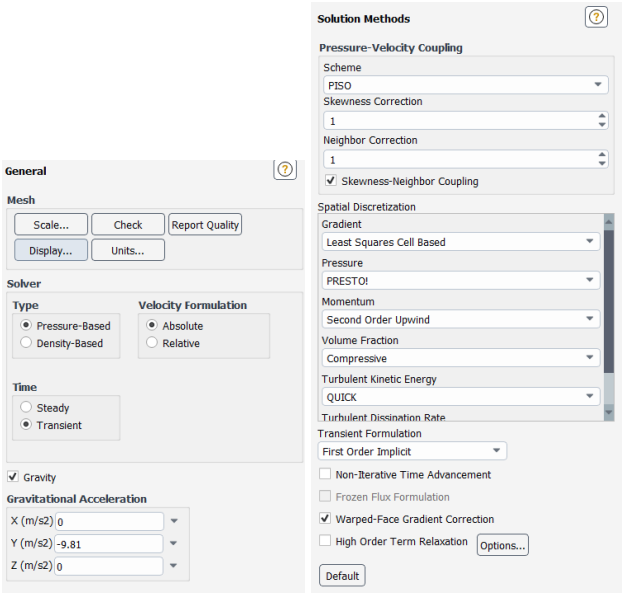


Figure 23: Simulation Settings on Fluent: General and Solution Method

The multiphase model used is the volume of fluid (VOF). The formulation is set to implicit form and the open channel flow option is activated. For the turbulence model, the simulation uses Realizable k-epsilon. These settings can be seen on Figure 24.

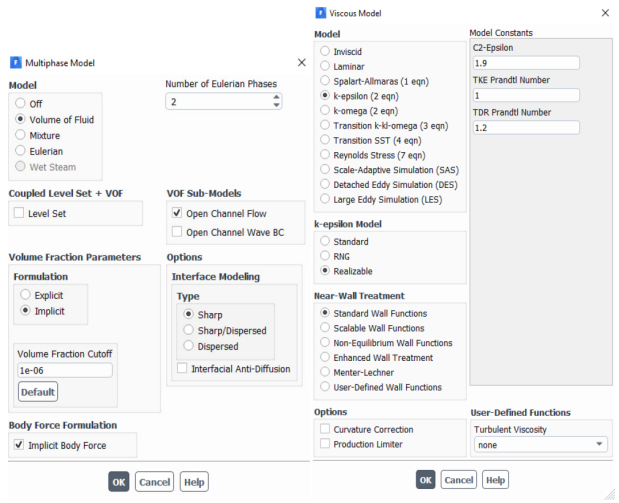


Figure 24: Simulation Settings on Fluent: Multiphase and Turbulence Model

The inlet portion of the computational domain is defined as the pressure inlet. In the inlet, the open channel option is activated and defined based on the entry speed and the location of the bottom and free surface in the compute domain. Based on measurements in the model, -62.5 m refers to the lower part of the computational domain and 2.2 m is equivalent to a draft height of 5.7 m. The input settings are listed in Figure 25.

The simulation is carried out until the convergence criterion is reached. Based on previous experiments, the simulation will converge before 900 iterations. The timestep is set to automatic with a scale factor of 0.2. The convergence criterion is set to obtain loose convergence, or the value of each residue is below 10^{-4}.

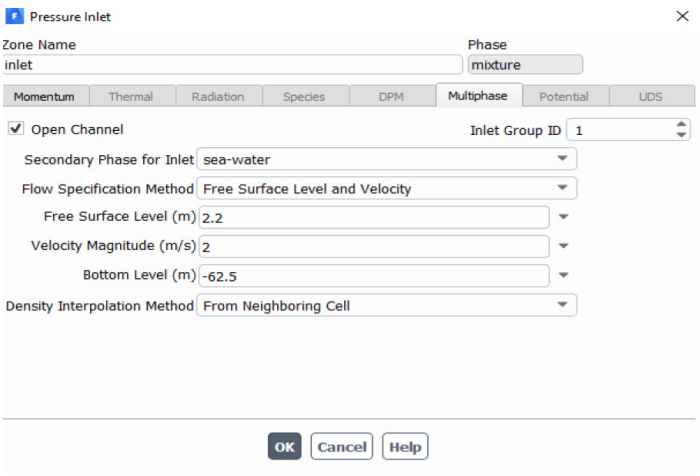


Figure 25: Simulation Settings on Fluent: Inlet Input

For the added mass, there is one simulation setting that has to be defined, which is the function of velocity magnitude to get the effect of added mass. The simulation setting should be defined as Figure 26.

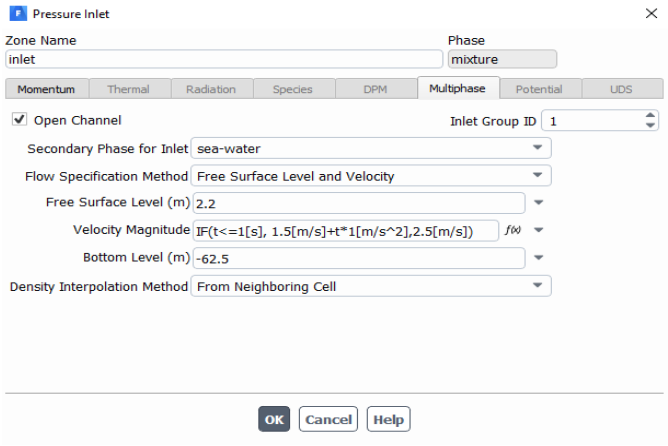


Figure 26: Simulation Setting Addition for Added Mass

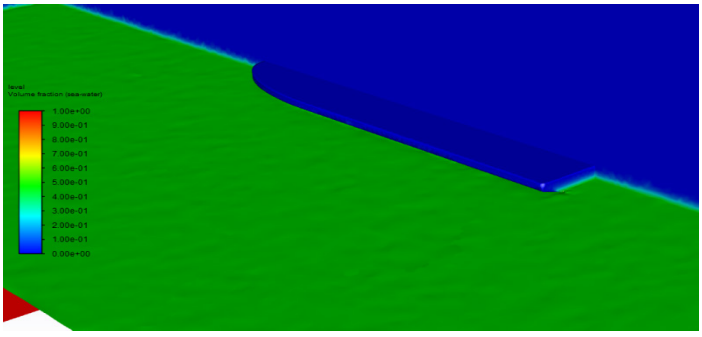


Figure 27: Visualization of CFD Simulation

4. Result and Analysis

The results obtained from the simulations that have been carried out are the value of drag in Newton units and the value of added mass in kilograms which is derived from the force by the acceleration. Visualization of both can be seen in Figure 27.

The validation of the drag simulation results on the ship is done by comparing the values obtained from CFD, with the theoretical values proposed by Holtrop and Mennen. The geometric parameters of the ship are calculated first to check whether this method is valid for use on the ship that is tested. The geometric model parameters of the ship that obtained from the CAD can be seen in Table 4 and the prerequisite geometry table can be seen in Table 5.

Table 4: Geometric Model Parameters

Parameter	Symbol	Value	Unit
Waterline Length	L_{WL}	57.576	m
Length Overall	L_{OA}	60.4	m
Beam/Breadth	B	12.379	m
Depth	d	6.5	m
Draft/Draught	T	5.7	m
Displacement	∇	2580.207	m^3
Wetted Surface Area	S	1170.01	m^2
Cross Section Area	A_M	61.23	m^2
Immersed Transom Area	A_T	38.819	m^2

Table 5: The Geometric Prerequisites for the Holtrop and Mennen Method

Parameter	Symbol	Value
Froude Number	F_n	0.084
Prismatic Coefficient	C_p	0.732
Length to Breadth Ratio	L/B	4.651

Because the values obtained are based on the calculations in Table 5 in correspondence with the criteria, Holtrop and Mennen's method is suitable for use as a verification method. Calculation of Froude number used equation 15 and prismatic coefficient using equation 19. The ship parameters used for this calculation use the specifications previously given.

In this verification, the components that are considered to have a large contribution to the total drag on the ship are drag due to friction, drag due to wave generation, and additional pressure drag due to the submerged transom area of the ship. The other components are considered to have no significant effect on the total drag, so they are assumed to be zero. Based on the calculation method written in [4], equation 21 to 44, it is found that the value of each component is listed in Table 6.

Table 6: Resistance Component of Baruna Jaya III

Resistance Components	Value
Friction	7003.9 N
Wave Generation	0.012 N
Submerge Transom Area	14830.2 N
Total	21834.14 N

For added mass, validation uses the ellipsoid method. The calculation can be done by substituting the ship dimension parameters into equations 19 to 20. From these calculations, the results are shown in Table 7.

Table 7: Surge Added Mass (Ellipsoid Method)

Parameter	Symbol	Value	Unit
Ellipticity	e	0.9850246	
	A_0	0.0907254	
	k_{11}	0.0475182	
Surge Added Mass	m_{11}	123547.40	kg

Recaps for the simulation results of resistance and added mass data can be seen in Table 8 and Table 9, combined with the method values for validation.

Table 8: Result for Drag Simulation

Method for Resistance	Result
CFD Simulation	21030.37 N
Holtrop and Mennen	21834.14 N
Error	-3.681%

Table 9: Result for Added Mass Simulation

Method for Added Mass	Result
CFD Simulation	115719.91 kg
Ellipsoid	123547.40 kg
Error	-6.764%

The simulations carried out can capture the phenomena that occur well and have a value that is quite close to the analytical results. The error has a value below 5% for resistance or drag simulation and has 6% error for added mass simulation. It means this CFD simulation method is accurate in analyzing a ship, especially Baruna Jaya III.

The difference between the simulation and analytical results could occur due to modelling and analytical result value. Error in modelling could happen due to simplification of the original shape that was modelled into 3D Model, the convergence of the mesh, or the simplified equations used. If the modeling of the CFD simulation is the problem, it is necessary to check the step-by-step modeling to be able to trace what went wrong. In terms of analytical result value, further comparison with experimental results must be done do verify the results. Further research should be taken to solve this problem.

5. Conclusion and Future Works

Drag and added mass estimation using computational fluid dynamics has been carried out and gives a drag simulation result of 21030.37 N, verified using the Holtrop and Mennen method, and added mass value of 115719.91 kg, verified by the Ellipsoid method. These estimated values have a very low error compared to the verification value. This means that the CFD simulation can be an alternative method to obtain drag and added mass. The methodology can be used to capture the hydrodynamic forces that occur on the ship, especially Baruna Jaya III, and complete the parameters required for a dynamic positioning control system. With a cost lower than the experimental method, computational fluid dynamics could be considered as the best option for getting the hydrodynamic parameters of the dynamic positioning control system.

In the future, it is necessary to investigate the added mass method in the CFD simulation further so that it can have a value closer to the analytical solution. Validation also needs to be compared with the experimental results so that values are more convincing because they are compared with actual values. The values of drag and added mass that occurs in the sway and yaw direction of the ship also will be further investigated in order to complete all the parameters needed to create a dynamic positioning control system.

Conflict of Interest

The authors: Nu'man Amri Maliky, Nanda Pratama Putra, Mochamad Teguh Subarkah, and Syarif Hidayat; declare that we have no conflict of interest for writing this paper to be submitted to ASTES Journal: Special Issue Paper (Special Issue on Multidisciplinary Innovation in Engg Science & Tech).

Acknowledgement

The authors are very grateful to all reviewers who have checked and criticized the papers that have been written to become a great paper in this journal. Thank you also to all ASTES committees, especially those who take care of the Special Issue Paper (Special Issue on Multidisciplinary Innovation in Engg Science & Tech) section who has invited us to submit our paper to the journal here.

References

[1] N.A. Maliky, M. Teguh Subarkah, S. Hidayat, "Added Mass and Drag Prediction using CFD Fluent Simulation for an Autonomous Barge Parameters," in ICEVT 2019 - Proceeding: 6th International Conference on Electric Vehicular Technology, 2019, doi:10.1109/ICEVT48285.2019.8993988.

[2] H. Qu, E.I. Sarda, I.R. Bertaska, K.D. Von Ellenrieder, "Wind feed-forward control of a USV," in MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World, 2015, doi:10.1109/OCEANS-Genova.2015.7271438.

[3] V. Mishra, S. Vengadesan, S.K. Bhattacharyya, "Translational added mass of axisymmetric underwater vehicles with forward speed using computational fluid dynamics," Journal of Ship Research, 2011.

[4] L. Birk, Holtrop and Mennen's Method, 2019, doi:10.1002/9781119191575.ch50.

[5] D.T. Sen, T.C. Vinh, "Determination of Added Mass and Inertia Moment of Marine Ships Moving in 6 Degrees of Freedom," International Journal of Transportation Engineering and Technology, 2016.

[6] J. Anderson, "Computational Fluid Dynamics: The Basics with Applications. 1995," McGrawhill Inc, 1995.

[7] ANSYS, "ANSYS - Turbulence Modelling and the Law of the Wall: Tutorial," ANSYS User Manual, 2014.

[8] P. Beneš, R. Kollárik, "Preliminary Computational Fluid Dynamics (CFD) Simulation of EIIB Push Barge in Shallow Water," Scientific Proceedings Faculty of Mechanical Engineering STU in Bratislava, 2012, doi:10.2478/v10228-011-0012-1.

[9] T. Tabaczek, J. Kulczyk, M. Zawiślak, "Analysis of hull resistance of pushed barges in shallow water," Polish Maritime Research, 2007, doi:10.2478/v10012-007-0002-4.

[10] C. Lin, S. Percival, E.H. Gotimer, Viscous Drag Calculations for Ship Hull Geometry, Virginia, 1995.

[11] A. Fluent, "ANSYS Fluent 12.0 user's guide," Ansys Inc, 2009.

[12] A. Fluent, "Ansys Fluent Theory Guide," ANSYS Inc., USA, 2013.

[13] C. Chas, R. Ferreiro García, "Introduction to ship dynamic positioning systems," Journal of Maritime Research, 2008.

[14] J.D. Anderson, Fundamentals of Aerodynamics SI, 2011.

[15] D.L. Logan, A First Course in Finite Element Method, Biomedicine, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 944-954 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Dense SIFT–Flow based Architecture for Recognizing Hand Gestures

Suni S S[*,1] K Gopakumar[2]

[1]*LBS Centre for Science & Technology, University of Kerala, Kerala, 695033, India*

[2]*TKM College of Engineering, Kollam, Kerala, 691005, India*

ARTICLE INFO

Article history:
Received: 04 July, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Human-computer interaction
Hand segmentation
Scale Invariant Feature Transform
Support Vector Machine
K-Nearest Neighbour
Hand gesture recognition

ABSTRACT

Several challenges like changes in brightness, dynamic background, occlusion and inconsistency of camera position make the recognition of hand gestures difficult in any vision-based method. Diversity in finger shape, size, distribution and motion dynamics is also a big constraint. This leads to the motivation in developing a dense Scale Invariant Feature Transform (SIFT) flow based architecture for recognizing dynamic hand gestures. Initially, a combination of three frames differencing and skin filtering technique is used for hand detection to reduce the computational complexity followed by a SIFT flow technique to extract the features from the detected hand region. SIFT flow vectors obtained from every pixel can lead to overfilling, data redundancy and dimension disaster. A dual layer belief propagation algorithm is utilized to optimize the feature vectors to resolve the dimensionality problem. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers are used to evaluate the performance of the developed framework. Experiments were conducted on hand gesture database for HCI, Sebastien Marcel Dynamic Hand Posture Database and RWTH German finger spelling database. The simulation results demonstrate that the developed architecture has excellent performance on the uneven background and varying camera position and it is robust against image noise. A comparative analysis with the state of the art methods illustrates the effectiveness of the architecture.

1. Introduction

Man's innate, innovative and intentional desire to grab and cope up with the smart devices and interactive systems in the modern era is fathomless. Recent studies show that 80% of the population uses smart decision making systems in one or another way [1]. For example, a person who is doing multiple tasks can change display of computer using a hand movement or gesture, while reading newspaper. Artificial intelligence leaps to the zenith that it becomes an art in presenting artistic perfection and accuracy increasing the percentage of digitalized people resulting restless efforts to propose new algorithms for interacting with machines. Numerous approaches are there to tap and convert gesture movement captured using vision sensors to decision support systems [2,3]. The work of Liu and Zhang was a human computer conversation based on hand gestures from the video streams using Hu invariant moments and support vector machine [2]. A system to recognize the basketball referee signals from videos using image segmentation and support vector machine classifier was build [3].

In human computer interaction systems, the present and widely used trend is dynamic texture based techniques [4,5] and multimodal feature based frameworks [6,7] because of its powerful representation. One of the most broadly used features is local binary pattern (LBP) introduced by Ojala [8] for texture classification, that proved to be useful in computer vision applications. The authors of [4] was developed a dynamic texture based technique for human-computer interaction using volumetric spatiograms of local binary patterns. The approach is congenial to detect static and dynamic gestures with the expense of high dimensional feature vectors. A spatio temporal descriptor called directional local ternary pattern from three orthogonal planes was introduced by the authors of [5] to classify the dynamic scenes and utilized convolutional neural network for detection. However, performance of these systems are found to be affected by external environment such as complex background and image noise and most of the systems perform well only for specific dataset. Human

[*] Corresponding Author: Suni S S, suni.ss@gmail.com

gestures are unique in velocity and motion scale. Another work was to build a multimodal feature based framework that fuses shape and movement information for recognizing facial expressions. The design uses histogram of gradients and optical flow features to obtain the video descriptor [6]. The authors of [7] developed a framework that uses combination of moments and directional wavelet local binary pattern for recognizing human actions. These methods give higher recognition rate at the cost of high dimensionality of feature vectors. The present attempt is to reduce the changes between distinct people performing the same gesture and to exaggerate the coherence of the gestures. Some of the constraints observed in the above mentioned literature are summarized as follows:

- Most feature vectors are calculated on limited number of pixels in an image. So it will not take the micro scale variations of movements (eg: finger movements). This will affect the overall accuracy of the system.
- Suggested methods are not strong enough to avoid image noise.
- High dimensionality of feature vectors leads to increase in computational time.
- Most of the systems perform well only for specific dataset. Robustness is needed.

To solve the above-mentioned drawbacks, a newly refined descriptor, dense SIFT flow that preserves the local appearance features of the image and the temporal variations (movement) of the video can be utilized. Dense SIFT flow algorithm calculates the flow vectors of the SIFT image. The proposed architecture is invariant to rotation, scale, illumination, view point and noise.

The main contributions of this work are,

- An architecture that accounts the spatial and temporal consistency, and can be manage to adapt itself to diverse lighting environments. Moreover, it can improve the system accuracy when connected with gesture identification methods.
- In the pre-processing stage to reduce the computational complexity and dimensionality of feature vector a hand segmentation method is applied.
- A sift flow algorithm that captures features from video frames in terms of spatial coherence and motion. It improves the discriminative power and effectiveness of the feature vector.
- System is tested for three different databases with varying background and viewpoints. The results obtained indicates the robustness of the system.

The rest of the article is framed as follows: Related works in the field of gesture recognition are detailed in Section 2. Theoretical framework of the proposed approach is given in section 3. The classifiers and its detection process is explained in section 4. Simulations, performance evaluations, comparative study of proposed algorithm with other existing methods are illustrated in section 5 followed by conclusion in section 6.

2. Related works

Attention has been focused in describing hand gesture using different feature based approaches, i.e., local and global feature based approaches. Recognition of hand gestures involves segmentation, preprocessing, feature extraction, gesture description and representation & classification.

Mostly, the state of movement of objects in the whole frame is represent as optical flow at a given time. Optical flow enables the foreground object motion study under static background. The methods based on the optical flow [9-11] are mostly used one to find the displacement of an entity in a image sequence. Estimation of optical flow captures local dynamics in the temporal domain and is computatinally efficient. The authors of [9] improved the classical optical flow algorithm by adding median filtering during optimization, that help to preserve better motion details. Dense optical flow algorithm is applied to label data for convolutional neural networks for motion detection in one work [10]. Another work [11] which utilized optical flow to detect moving objects and to estimate depth using triangulation. Optical flow algorithm with Kalman filter is successfully used for tracking multiple objects [12]. The authors of [13] proposed a solution with very low time complexity and accuracy for the estimation of dense optical flow. To detect obstacles at the time of landing of a moving air vehicles, a self-supervised learning (SSL) technique based on optical flow cues was developed [14]. All these approaches, however, do not take into account the characteristic temporal dynamics of all micro variations in movements and are limited to preserve local appearance representations.

A new system is introduced by authors of [15] based on a Single-Shot Multibox Detector (SSD) deep learning algorithm for hand gesture recognition. The system identifies the hand gestures in complex scenes with limited test cases. The authors of [16] build a framework that uses knowledge from multiple modalities by utilizing separate 3D convolutional neural networks (3D-CNNs) for recognizing dynamic hand gestures. The authors of [17] integrated density clustering and image entropy to obtain the key frames to enhance the performance of fast and robust gesture recognition and feature combination technique is also introduced. They also created two datasets named Action3D and Hand Gesture datasets. A dynamic hand gesture recognition system using skeletal representation of hand introduced in [18]. The system used a hand topology technique called later hand skeletal data to capture kinematic descriptors from the gesture frames and were encoded in using a multi-level temporal pyramid and a Fisher kernel. Mahayuddin and Saif [19] has described a method to improve learning process of mathematics using modified extrusion method for recognizing hand gestures which converts 2D shape to 3D shape. The method achieved good accuracy, but robustness of the system has to tested. Handling the large dataset is also a big concern.

The idea of using local appearance based feature like SIFT was introduced by Lowe [20]. SIFT algorithm converts an image into a set of local feature vectors. Each of these feature vectors is is distinctive and invariant to any image translation, rotation or scaling. After introduction of this feature, it has been applied in the different areas of video processing problems like gesture recognition, image registration and image retrieval. Hossein-Nejad and Nasri utilized SIFT features and RANdom SAmple Consensus (RANSAC) transform in image registration for extracting and matching features [21]. SIFT features are applied in gender and gaze recognition problem and produced good results in [22]. One

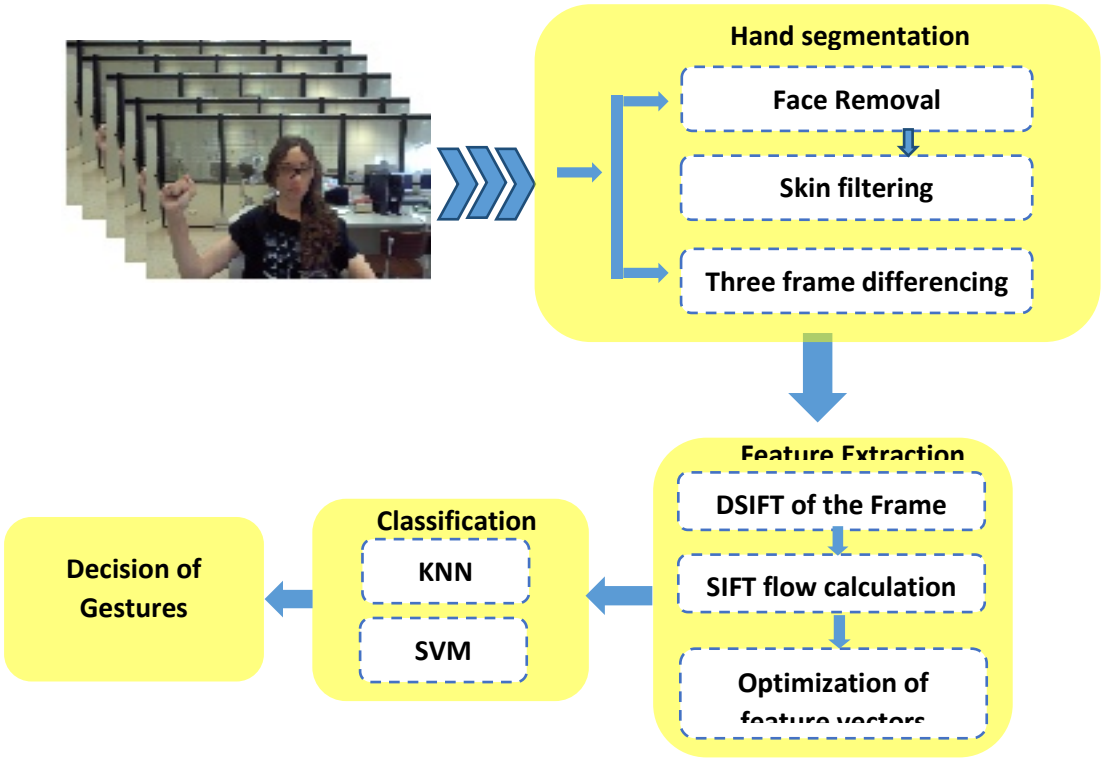


Figure 1: Proposed architecture of hand gesture recognition

disadvantage of using this feature matching method is that they become computationally expensive. Furuya and Ohbuchi [23] introduced the dense sampling of SIFT and a fast encoding algorithm for the application of 3D model retrieval. The dense SIFT algorithm takes the local gradient information of all pixels in an image and has the limitation for capturing motion dynamics of the video. This leads to the motivation in developing a SIFT flow based framework that captures spatial and motion features for detection.

SIFT flow vectors generated are high dimension. Optimization is necessary to reduce the dimensionality of feature vectors. Numerous methods are used for optimizing the SIFT descriptors. Daixian adopt key point location optimization using Chamfer distance [24]. The technique is proved to be reduce computation time and achieves good accuracy. Here the optimization was done before descriptor generation. VSA(Vector Symbolic Architecture) is an encoding technique which converts the high dimensional vectors into the sparse discrete representation [25]. The mapping is one of the key constraints when working on real time data. The double layer belief propagation was proved to be a good optimization technique for SIFT flow vectors [26].

A framework is developed for dynamic hand gesture recognition, wherein multimodal behavior of the image sequences in terms of spatial coherence and motion is extracted. In this application, a descriptor called dense sift flow is utilized which takes flow vectors of the local gradients. Dense SIFT flow refers to taking dense sampling in space of all image sequences. An algorithm is applied that performs gesture analysis to recognize and take decisions to interact with machines. It yields dense sift flow fields that are robust to image noise. Thus, the final video descriptor achieves highly discriminative hand gesture representation. The method is invariant to scale, rotation, illumination, view point and noise. It is observed that the developed architecture gives a better recognition rate.

3. Our Architecture

The prime motivation of the proposed work is to build an efficient framework which can capture shape and motion features from image frames for recognizing hand gestures. There are four phases in the present approach: pre-processing, feature extraction, optimization and detection. In the first stage, the hand is segmented by using three frame differencing and skin filtering [24]. The feature vectors are evaluated using dense SIFT flow algorithm and create a descriptor representation that can express the particular gesture. Classifiers are trained with feature vectors and then assign the new descriptors into various categories of gestures. The proposed architecture of a hand gesture recognition system is shown in Figure 1. The subsections will describe the SIFT flow algorithm and its optimization.

3.1. Hand Segmentation

Hand segmentation is the first step in any hand gesture recognition process. The work flow of hand segmentation process is shown in Figure 2. There are three main processes that combined together to achieve the segmented image. They are hand region tracking, skin filtering, and three frame differencing of both color and grayscale image frames [27].

Initially, the face of the person performing gesture is identified and eliminated from the second image of the video using the Viola–Jones algorithm [28, 29]. Skin filtering operation [30] is performed after removing the face to retrieve only the skin

coloured objects from the image frames. In the meantime, three-frame differencing is carried out with the first three frames of both coloured and grayscale sequences. Morphological operations (AND/OR) are done as shown in Figure 2 to attain the expected results. The desired hand is obtained from the combined results of the three-frame differencing and skin filtering. But, it has been notifying that some small skin coloured objects are also present with the desired hand. Thus, the largest binary linked object is selected and is treated to be the required hand in the scene.

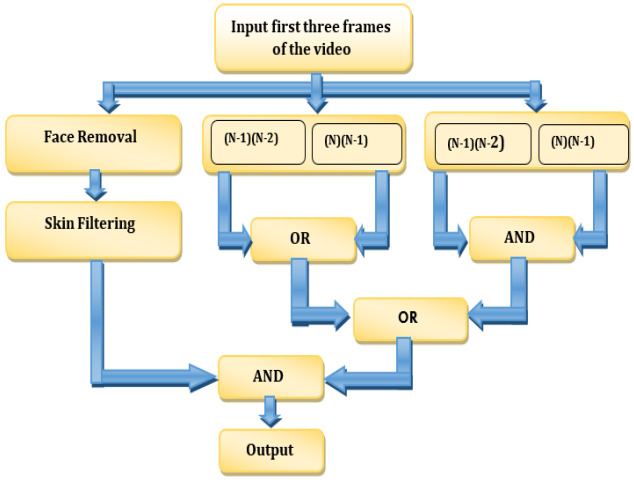


Figure 2: Work flow of Hand segmentation process

Algorithm for Hand Segmentation

Input :*Colour image frame from the video*
Output: *Segmented Hand Image*

1. For n =1 to 3 do
2. I(n) = input colour image frame
3.%%%Colour to gray scale conversion of image
4. I1(n) = rgb I(n) to the grayscale image conversion
5. if n =2 then
6.%%%Face detection of image
7.J = face detection using Viola – Jones algorithm
8.%% Removing face
9.J1 = Remove face J from frame I(n)
10.J2 = Skin filtering of J1 image
11. end if
12. if n =3 then
13.K1 = Skin filtering of (I(n-2) – I(n-1))
14. K2 = Skin filtering of (I(n-1)-I(n))
15.L = OR (K1, K2)
16.M1 = Grey to binary (I(n-2) – I(n-1))
17.M2 = Grey to binary (I(n-1)-I(n))
18.P = AND(M1,M2)
19.Q = OR(P,L)
20.R = AND (Q,J2)
21. %% Biggest binary linked object is selected
22.Seg_image = BBLOB(R)
23.end if
24.end for



Figure 3: Result of hand segmentation algorithm

Figure 3 shows the result of hand segmentation algorithm. The yellow box is the segmented hand region.

3.2. SIFT flow algorithm

3.2.1. Dense SIFT

Object tracking problems are usually solved using visual extraction algorithms. Dense SIFT (Dense scale invariant feature transform) is an innovative step in this field conducive even there is change in the background and appearance of the tracked object during tracking [31]. The process of extraction of dense SIFT is done by taking SIFT histogram for all pixels at a single scale having overlapping patches. A dense sampling assures a good characterization of foreground gestures as well as the consistency of features in the adjacent image frames. The following steps have been used to obtain dense SIFT descriptor. For each pixel of a segmented grey scale image the derivative value ρ and the derivative direction θ can be represented as

$$\rho(x,y) = \sqrt{dx^2 + dy^2}$$
$$\theta(x,y) = \tan^{-1}\frac{x}{y} \quad (1)$$

SIFT descriptor indicates the neighbourhood gradient information. As such, every pixel in an image, select 16 x16 neighbourhood, into 4 x4 cell array. The histogram of derivative direction comprising each block of size 4x4 is obtained and quantizing the orientation into 8 bins. Thus obtaining 4 x4 x 8 = 128 dimensional SIFT description for a pixel. Figure 4 shows the principle of SIFT local feature generation.

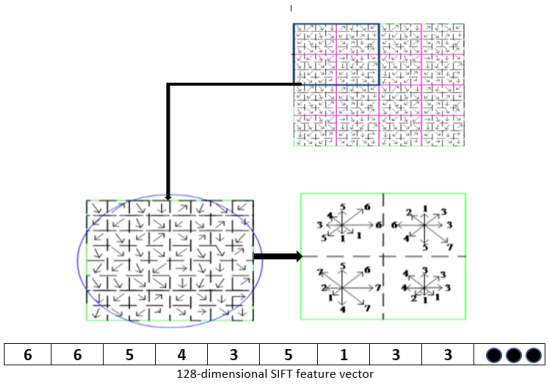


Figure 4: Principle of SIFT feature generation

3.2.2. SIFT flow algorithm

The Dense optical flow [32] opened the avenue of tracking the flow fields very densely from two consecutive SIFT images. SIFT flow is the latest innovative step by authors of [26] and is formulated to adjust an image frame to its nearest neighbor for huge image changes in scale. SIFT flow vector is obtained as follows. Assume $p = (x, y)$ be the grid coordinates of the image frames and $\omega(p) = (v(p), u(p))$ be the flow vector at Where $v(p)$ and $u(p)$ are integers with L possible states. s_1 and s_2 be the two SIFT images and ϵ contains all spatial neighborhoods (4 neighbourhood). The objective energy function of SIFT flow is designed as,

$$E(\omega) = \sum_p \min(\|s_1(p) - s_2(p + \omega(p))\|_1, t) \quad + \tag{3}$$

$$\sum_p \eta(|v(p) + u(p)|) \quad + \tag{4}$$

$$\sum_{(p,q)\in\varepsilon} \min(\alpha|v(p) - v(q)|, d) \quad + \quad \min(\alpha|u(p) - u(q)|, d) \tag{5}$$

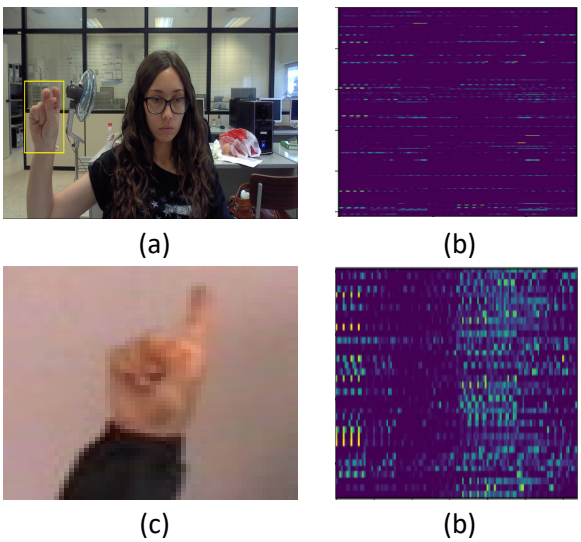


Figure 5: Sample image and its dense SIFT image (a) shows the sample image from hand gesture database for HCI. (b) shows its dense SIFT image. (c) shows sample image from Sebastien Marcel Dynamic Hand Posture Database (d) its dense SIFT image

The data term in (3) is a SIFT descriptor match constraint to be matched via the flow vector $\omega(p)$. The small displacement constraint in (4) allows the flow vector to be as small as possible when no other information is available. The smoothness term inhibits (5) the neighboring pixels to have similar displacement. In this objective energy function, truncated L1 norm is used as a matching threshold in both the smoothness and data term with $d$and t respectively (d=0.1, t=0.01). Figure 5 shows the sample image and its dense SIFT image.

3.2.3. Optimization of feature vector

A dual layer loopy belief propagation method is utilized for optimizing the objective energy function (5) [31]. Figure 6 shows the components in horizontal and vertical direction using dual layer loopy belief propagation. The direct application of SIFT flow vectors to the high dimensional image will result in poor performance. Hence, a coarse to fine SIFT flow matching method is used to formulate the correspondence between two consecutive frames, to accelerate the speed and matching performance.

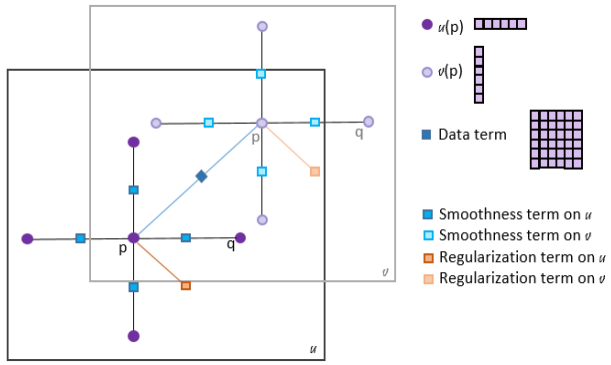


Figure 6: Illustration of Dual-layer belief Propagation [32]. Vertical and horizontal components of objective function

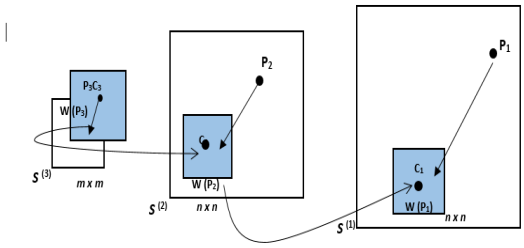


Figure 7: Coarse-to-fine SIFT flow matching process on a pyramid. Blue Square shows the searching window for P_k at pyramid level k.

The foremost aim is to calculate the flow at the coarse level of image patch, constantly resulting and the flow is refined from coarse to fine level [26]. The process is demonstrated in Figure 7. The SIFT images s_1 and s_2 are represented as 's'. A sift pyramid $\{s(k)\}$ is created where $s(1) = s$ and $s(k+1)$ is smoothed and down sampled from $s(k)$. The pyramid level at K, Pk is the coordinate of the pixel to match. Ck is the centroid or offset of the searching window and the right match is $\omega(Pk)$. At the loop pyramid levels $s(2), P3$ is the centroid of the searching window with size mxm. Then the belief propagation converges, at this

level. The system transfers the optimized flow vector $\omega(P3)$ to the next finer level $C2$, where searching window of $P2$ is centered. The process continues from $s(3)$ to $s(1)$ till the flow vector $\omega(P1)$ is obtained. The computational complexity is reduced to a factor of $O(h2logh)$. Thus, the algorithm become very energy efficient by doubling η and retain α and d. Figure 7 shows the process of matching method from coarse to fine SIFT flow.

4. Recognition of hand gestures

4.1. KNN classifier

The SIFT flow vectors generated in the feature extraction stage is used to train the classifiers to assess the performance of the designed system. KNN classifier is selected here, because of its capability of handling large training data. Training was done in the system for two values of k such as 3 and 5. The draw votes in recognition process can be eliminated by selecting odd values of k. 'Leave one group out cross validation' strategy is utilized for testing.

4.2. Support Vector Machine Classifier

The aim of the detector stage is to create a SVM classifier with all data samples of extracted feature vectors. This supervised learning classifier splits the data samples of various categories by calculating a maximum-margin boundary between them. The system is trained using feature vectors of videos from selected database with quadratic, radial, linear and polynomial kernel based SVM classifiers. The classifier was selected because it has been tested successfully in various image classification and object detection tasks. SVM with four kernels is trained with different training videos in hand gesture database. It is seen that radial function based SVM performs better than others. Hence SVM with RBF kernel is used for testing. 'Leave one group out cross-validation' technique is used for testing.

5. Results and Discussion

All test procedures are conducted using Matlab 2014a on a i5 2.7GHz computer with 8.00GB RAM. To analyze the proposed algorithm, we performed the set of experiments with three databases such as hand gesture database for HCI [4,33], Sebastien Marcel Dynamic Hand Posture Database [34] and RMTH German finger spelling database [35]. The performance parameters of the system are analyzed on the basis of certain metrics such as sensitivity, specificity and accuracy [36,37]. The parameters are defined as follows.

$$\text{Sensitivity} = TP/(TP+FN) \quad (6)$$

$$\text{Specificity} = TN/(TN+FP) \quad (7)$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (8)$$

$$\text{F1 Score} = 2TP/(2TP+FP+FN) \quad (9)$$

$$\text{Negative predictive value} = TN/(TN+FN) \quad (10)$$

where TN, TP, FN, and FP are true negatives, true positives, false negatives and false positives respectively. A comparative analysis has been done with different state of art methods.

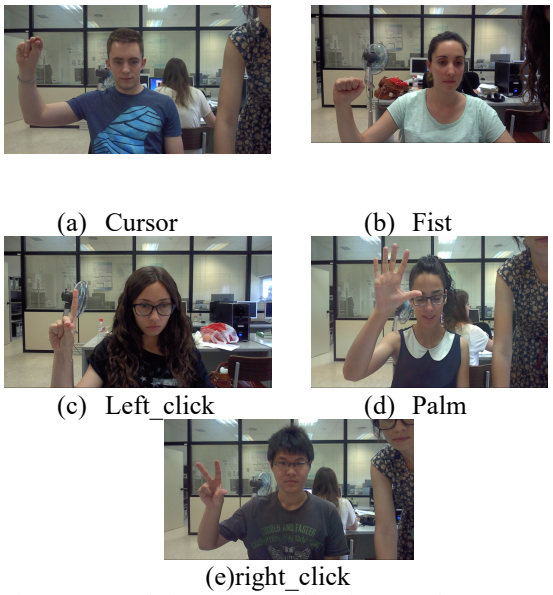
5.1. Performance Analysis

5.1.1. Performance Analysis with Hand gesture Database for HCI

The database contains a group of $1280\,x\,720$ pixel resolution color image frames that are generated for human computer interaction applications using Senz3D sensor [4]. This was generated for testing the mouse functionalities such as left and right click, curser, fist and palm. The samples were captured in realistic way with uneven environment. It composed of two sets of image frames. Five various gestures performed by distinct people are contained in Set 1 and Set 2. For testing, individual videos were created. Samples of hand gesture image frames are shown in Figure 8.

In the first stage, process of hand detection is done and the resolution (width is reduced by a factor of 5 and height is reduced by a factor of 3) of the image frames is reduced. Thus the process will lessen the computational complexity and dimension of the feature vectors and saves the computational time. The feature vectors are generated with the Dense SIFT flow algorithm and are tested with KNN and SVM classifiers.

Evaluation is done using KNN classifier with two values of K=3, K=5. The performance analysis based on recognition rate is provided in the Fig .9. It shows that the highest recognition rate of 87.8% with K = 5 is obtained for the testing video of seq 5 and the average recognition rate is 85.9%.



(a) Cursor (b) Fist

(c) Left_click (d) Palm

(e)right_click

Figure 8: Sample image frames of hand gestures for HCI

The system is tested for using KNN classifier with K =3 and K =5. The performance analysis is provided in Table. 1. The highest training and testing accuracy were achieved for k = 5.

The system is validated using SVM classifier with various kernel functions such as polynomial, linear, radial basis and quadratic in hand gesture database and the performance analysis is shown in Figure 10. The performance level shows SVM classifier with radial basis kernel produces better results than the other kernels and it gives 97.6%.

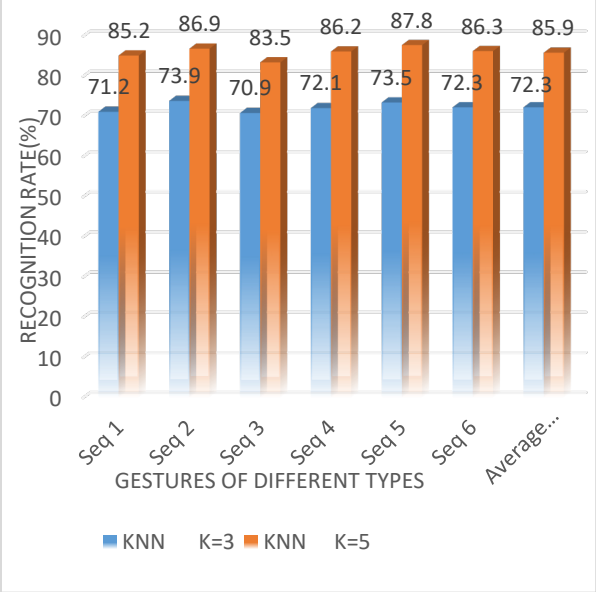


Figure 9: Performance of the system for various hand gestures on hand gesture database using KNN with K=3 and K=5

Table 1: Accuracy of algorithm on hand gesture database using KNN classifier

Expt	K = 3		K = 5	
	Training accuracy	Testing accuracy	Training Accuracy	Testing Accuracy
Subject 1	0.768	0.712	0.873	0.852
Subject 2	0.769	0.739	0.891	0.869
Subject 3	0.748	0.709	0.852	0.835
Subject 4	0.751	0.721	0.878	0.862
Subject 5	0.772	0.735	0.902	0.878
Subject 6	0.758	0.723	0.889	0.863
Mean Accuracy	**0.761**	**0.723**	**0.881**	**0.859**

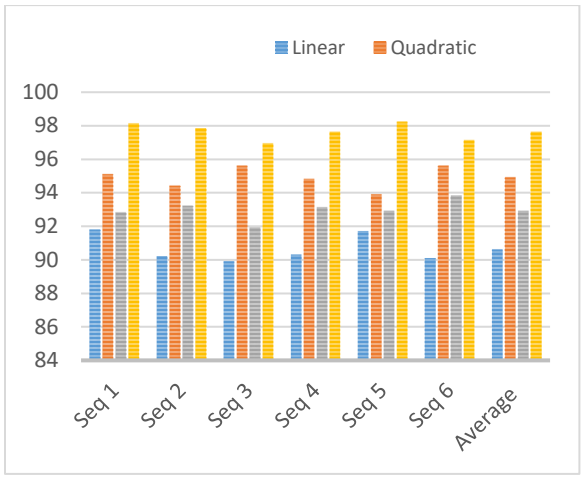


Figure 10: Performance of the system using SVM classifier with different kernels

Group of experiments were done using the multiclass SVM classifier with rbf kernel. Table. 2 shows the evaluation based on accuracy in both training and testing phase. The mean accuracy is obtained as 0.976 which is better as compared with the existing methods.

Table 2: Accuracy of algorithm on hand gesture database using SVM classifier

Gestures	Training accuracy	Testing accuracy
Subject 1	0.998	0.981
Subject 2	0.982	0.978
Subject 3	0.978	0.969
Subject 4	0.989	0.976
Subject 5	0.991	0.982
Subject 6	0.986	0.971
Mean Accuracy	**0.987**	**0.976**

5.1.2. Performance analysis with Sebastien Marcel Dynamic Hand Posture Database

The dynamic hand posture database [34] consists of image frames for four hand gestures such as Rotate, Click, No, and Stop-Grasp-Ok from 10 different persons with various resolution ranges from 50 x 50 pixels to 80 x 80 pixels. Figure 11 shows the Sample image frames.



(a)Clic (b)StopGraspOk (c)No (d)Rotate

Figure 11: Sample frames from Sebastien Marcel Dynamic Hand Posture Database

Table 3 and Table 4 show confusion matrix obtained for the four dynamic hand gestures using KNN classifier with K =3 and K = 5. We observed that the size of hand gestures ranges from 50 × 50 pixels to 80 × 80 pixels, didn't affect the efficiency of the framework in terms of recognition rate and computational time. Some of the frames 'clic' and 'no' hand gestures seem to be same that is why the recognition rate is less for these gestures while using KNN classifier.

Table. 3: Confusion matrix for hand gesture database using KNN with K=3

Gesture	Clic	No	Rotate	StopGraspOk	Recognition Rate(%)
Clic	11	3	0	1	73.3
No	3	10	1	0	71.4
Rotate	1	1	9	2	69.2
StopGraspOk	1	0	3	11	73.3

Table. 4. Confusion matrix for hand gesture database using KNN with K=5

Gesture	Clic	No	Rotate	StopGraspOk	Recognition Rate(%)
Clic	12	3	0	0	80.3
No	2	12	0	0	85.7
Rotate	1	0	11	1	84.6
StopGraspOk	1	0	2	12	80.0

For efficient classification, multi SVM classifier is used and got the recognition rate of 100%. Table. 5 shows performance matrix.

Table. 5: Confusion matrix for hand gesture database using SVM with rbf kernel

Gesture	Clic	No	Rotate	StopGraspOk	Recognition Rate(%)
Clic	15	0	0	0	100
No	0	14	0	0	100
Rotate	0	0	13	0	100
StopGraspOk	0	0	0	15	100

Figure 12 shows the system performance analysis for different types of classifiers based on true positive rate and false positive rate. From the observation, we can say that the SVM classifier with radial basis kernel gives better results. The experiments reveals that the developed framework gives good results for dynamic gestures with different illuminations and resolution levels.

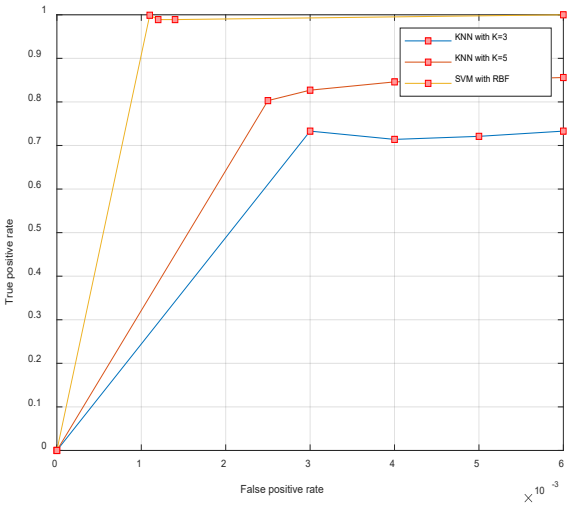


Figure 12: System performance with classifiers

5.1.3. Performance Analysis with RMTH German finger spelling Database

The database [35] consists of gestures of 35 types denoting 1 to 5 numbers, the symbols A to Z, 'SCH and the German umlauts Ä, Ö, Ü. It contains five dynamic gestures captured in uneven environment with varying camera positions. Among 1400 image sequences in the dataset, 700 videos are used for testing. The videos are captured at the rate of 25 frames per second with 320*240 pixels of resolution. The sample image sequences are shown in Figure 13.



Figure 12: Sample frames from RWTH German finger spelling database

The testing and evaluations are performed on five dynamic gestures of RWTH gesture database and confusion matrix obtained is shown in Table 6. The recognition rate achieved is 93% with SVM classifier (rbf kernel). The performance analysis in Figure 13 reveals that the dense SIFT flow-based framework provides better results for gestures from uneven background.

Table 6: Confusion matrix attained for five gestures in RMTH gesture Database

Gesture types	J	Z	Ä	Ö	Ü	Recognition Rate in %
J	18	0	0	2	0	90
Z	0	18	0	0	2	90
Ä	0	1	19	0	0	95
Ö	0	0	1	19	0	95
Ü	0	0	1	0	19	95

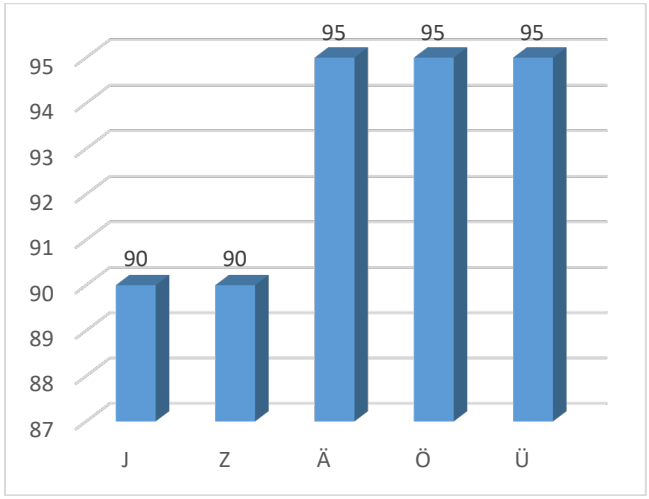


Figure 13: Performance of the system using SVM classifier with rbf kernel

Table 7: Performance Metrics of SVM (rbf) classifier on three different databases

Metric	Hand gesture Database for HCI	Sebastien Marcel Dynamic Hand Posture Database	RMTH German finger spelling Database
Accuracy(%)	97.6	100	93
Sensitivity(%)	97.6	100	93.1
Specificity(%)	98.8	100	94.7
Negative Predictive Value(%)	99.1	100	95.8
F1 Score(%)	97.8	100	92.9
Response time(Sec)	34	28	37

Table 7: Performance comparison of the proposed architecture with the state of the art methods on Hand gesture database

Method	Subject1	Subject2	Subject3	Subject4	Subject5	Subject6	Mean Accuracy
VS-LBP[4]	0.951	0.962	0.939	0.931	0.801	0.959	0.927
LBP+ Moments [7]	0.942	0.918	0.937	0.923	0.851	0.923	0.915
TPM-SLB$_s$P$_{8,1,10,8,1}$ [33]	0.974	0.961	0.976	0.960	0.926	0.975	0.965
TPM-SLBP$_{8,1,10,8,3}$ [33]	0.980	0.957	0.981	0.972	0.924	0.974	0.962
PHOG_TOP+ Optical flow based framework [39]	0.967	0.971	0.942	0.939	0.892	0.962	0.946
Dense SIFT flow based framework	**0.981**	**0.978**	**0.969**	**0.976**	**0.982**	**0.971**	**0.976**

Table 8: Comparative analysis of the proposed method with existing techniques on RMTH German Finger spelling database

Study	Technique	Error rate
Reference [35]	Thresholding with skin color probability+ camshift tracking	35.7%
Reference [38]	features from hand movements and contour shape	27.6%
Reference [39]	PHOG_TOP+ Optical flow based framework	11%
Proposed architecture	Dense SIFT flow	**7%**

From Table. 7, the accuracy of the proposed architecture on the selected datasets come out to be well (97.6% for hand gesture database for HCI, 100% for Sebastien Marcel dynamic hand posture database and 93% for RMTH German finger spelling database). The negative predictive value and specificity records higher values. Higher sensitivity is recorded. Response time of the proposed architecture is very good (< 37). So, the system can work well in real time applications.

5.2. Comparative Analysis

The developed Dense SIFT flow-based framework for hand gesture recognition differs from other methods, in terms of various preprocessing and filtering techniques. Therefore, it is hard to make a quantitative analysis with the state of the art methods. However, we have analyzed performance based on accuracy with existing approaches on the hand gesture database and tabulated on Table 7. **T**he global and local spatial information with temporal data utilizes in volumetric spatiograms of local binary pattern (VS-LBP), achieves the accuracy of 0.927 [4]. The authors of [7] developed an algorithm that fuses directional wavelet Local Binary Pattern (LBP) and moments and obtained the accuracy of 0.914. But it is sensitive to dynamic backgrounds. The Temporal Pyramid Matching of Local Binary Pattern (TPM-LBP) algorithm created by authors of [33] achieved good recognition rate with the expense of computational complexity. In one of recent work [39], hand gesture recognition based on pyramid histogram of gradients (PHOG) and optical flow achieves a recognition rate of 94.6% with the expense of high dimensional feature vectors. From this analysis, we can found that proposed architecture shows considerable improvement in recognition rate and work well in non-uniform background and varying viewpoints.

Table. 8 illustrates the comparative analysis of the proposed architecture with existing techniques on RMTH German finger spelling database based on error rate. The authors of [35] generated a hidden Markov model emission probabilities based on the appearance features to detect the hand gestures and was achieved an error rate of 35.7%. The features from contour shape (orientation, seven hu moments area and perimeter) and motion of hand (angle and velocity of movement) are used to detect hand gestures and achieved a better error rate 27.6% [38]. Another work [39] obtained 11% of error rate by using a multiple feature based framework. The proposed SIFT flow based architecture attained an error rate of 7%. The outcome of the framework is promising in the light of the fact that some of the symbols are very identical in database, for example the gestures for M, N and A.

6. Conclusions

In this paper, we have proposed a novel architecture for hand gesture recognition based on dense SIFT flow. A hand segmentation method is introduced in the pre-processing stage, which utilizes the three frame differencing and skin filtering to reduce the dimensionality of feature vectors. SIFT flow algorithm calculates the displacement between SIFT features of all pixels in

two consecutive sequences in a video. The flow vectors generated are invariant to image rotation and scale, image noise, illumination and camera view point. The KNN classifier and Support vector machine classifier is used for testing effectiveness of features. The present design can further benefit from improvements in SIFT flow estimation and hand segmentation technique. This approach considers the spatial and temporal consistency, and it is capable of accommodating itself to various illuminations. The evaluation results showed that the presented architecture can achieve high accuracies even for non-even background and image noise. As a result of comparison, we determined that our architecture has better recognition accuracy (0.976 for hand gesture database for HCI) than the state of the art methods.

However, a few aspects remain untapped, which may lead to further investigation. Even though our framework provides good results in various levels, it is computationally a little bit expensive. Hence developing faster and more accurate framework will certainly receive more attention in the future. In hand segmentation stage, adding a good key frame extraction technique may well significantly improve both speed and accuracy. And this problem is definitely worth further enhancement.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] T.H. Davenport, "From analytics to artificial intelligence", Journal of Business Analytics, **1**(2), 73 -80, 2018. https://doi.org/10.1080/2573234X.2018.1543535

[2] Y. Liu, P. Zhang, "Vision-Based Human-Computer System Using Hand Gestures", Proceedings of International Conference on Computational Intelligence and Security, Beijing, China, 2009.

[3] J. Zemgulys, V. Raudonis, Rytis Maskeliūnas and Robertas Damaševičius, "Recognition of basketball referee signals from videos using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM)" Procedia Computer Science **130**, 953–960, 2018.

[4] I. Ana, R Carlos "Human-computer interaction based on visual hand- gesture recognition using volumetric spatiograms of local binary patterns", Computer Vision and Image Understanding, Special Issue on Posture & Gesture, **141**, 126-137, 2015.

[5] M. Azher Uddin, Mostafijur Rahman Akhond and Young-Koo Lee, "Dynamic Scene Recognition Using Spatiotemporal Based DLTP on Spark", IEEE Access, 6, 66123 – 66133, 2018.

[6] X. Fan and T. Tjahjadi, "A spatial – temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences", Pattern Recognition, **48**(11), 3405-3416, 2015. https://doi.org/10.1016/j.patcog.2015.04.025

[7] M.N. AI-Berry, A.M. Mohammed "Fusing directional wavelet local binary pattern and moments for human action recognition", IET ComputerVision, **10**(2), 1-10, 2015.

[8] T. Ojala, M. Pietikainen and Topi Maenpaa, "Multiresolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns", IEEE Transactions on Pattern Analysis & Machine Intelligence, **24**(7), 971-987, 2002.

[9] D. Sun, Roth, S. and Black, M.J., "A quantitative analysis of current practices in optical flow estimation and the principles behind them", International Journal of Computer Vision, **106**(2), 115–137, 2014.

[10] J. Walker, "Dense optical flow prediction from a static image", Proceedings of IEEE International Conference on Computer Vision (ICCV), 2443–2451, 2015.

[11] S.-Lara, L., Sun, D., Jampani, V. and Black, M.J., "Optical flow with semantic segmentation and localized layers", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3889–3898, 2016.

[12] S. Shantaiya, "Multiple object tracking using Kalman filter and optical flow", European Journal of Advances in Engineering and Technology, **2**(2), 34–39, 2015.

[13] T. Kroeger et al., "Fast optical flow using dense inverse search", Proceedings of European Conference on Computer Vision, 471–488, 2016.

[14] H.W. Ho et al., "Optical-flow based self-supervised learning of obstacle appearance applied to MAV landing", Robotics and Autonomous Systems, **100**, 78–94, 2018.

[15] P. Liu, Xiangxiang Li, Haiting Cui, Shanshan Li and Yafei Yuan, "Hand Gesture Recognition Based on Single-Shot Multibox Detector Deep Learning", Mobile Information Systems, Article ID 3410348, 2019.

[16] M. Abavisani, Hamid Reza Vaezi Joze and Vishal M. Patel, "Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1165-1174,2019.

[17] H. Tang et al., "Fast and Robust Dynamic Hand Gesture Recognition via Key Frames Extraction and Feature Fusion", Neurocomputing, **331**, 424-433, 2019.

[18] Q.D. Smedta et al., "Heterogeneous hand gesture recognition using 3D dynamic skeletal data", Computer Vision and Image Understanding, **181**, 60-72, 2019.

[19] Z. Rasyid et al., "Efficient Hand Gesture Recognition Using Modified Extrusion Method based on Augmented Reality", Test Engineering and Management **83**(5-6), 4020-4027, 2020.

[20] D. G. Lowe, "Distinctive image feautres from scale-invariant keypoints," Int. J. Comput. Vis., **60**(2), 91–110, 2004.

[21] Z. Hossein-Nejad and Mehdi Nasri, "An adaptive image registration method based on SIFT features and RANSAC transform", Computers & Electrical Engineering, **62**(C), 524-537, 2017. https://doi.org/10.1016/j.compeleceng.2016.11.034

[22] W. Zhang et al., "Gender and gaze gesture recognition for human-computer interaction", Computer Vision and Image Understanding, **149**(C), 32-50, 2016. https://doi.org/10.1016/j.cviu.2016.03.014

[23] T. Furuya and R. Ohbuchi, "Dense sampling and fast encoding for 3D model retrieval using bag-of-visual features", Proceedings of the ACM International Conference on Image and Video Retrieval, Article 26., 2009.

[24] Z. Daixian, "SIFT algorithm analysis and optimization," 2010 International Conference on Image Analysis and Signal Processing, Zhejiang, 2010, 415-419, doi: 10.1109/IAS2010.5476084.

[25] D. Simon, " A New Building Material for Artificial General Intelligence", Proceedings of the 2008 conference on Artificial General Intelligence , 414–418, 2008.

[26] C. Liu et al., " SIFT Flow: Dense Correspondence across Scenes and Its Applications", IEEE Transactions on Pattern Analysis and Machine Intelligence, **33**(5), 978 – 994, 2011. DOI: 10.1109/TPAMI.2010.147

[27] J. Singha, Amarjit Roy and Rabul Hussain Laskar, "Dynamic hand gesture recognition using vision-based approach for human–computer interaction", Neural Comput & Applic, **29**, 1129–1141 ,2018.

[28] P. Viola, "Rapid object detection using a boosted cascade of simple features", Proceedings of the IEEE conference on computer vision and pattern recognition, **1**, 511–518, 2001.

[29] P. Viola "Robust real-time face detection", International Journal of Computer Vision, **57**(2), 137 -154, 2004.

[30] D. Chai "Face segmentation using skin-color map in videophone applications", IEEE Trans Circuits and Systems for Video Technology, **9**(2), 551–564, 1999. DOI: 10.1109/76.767122

[31] Y. Liu, "Multi-focus image fusion with dense SIFT", Information Fusion, **23**, 139 -155, 2015.

[32] H. Wang, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", International Journal of Computer Vision, **103**(1), 60–79, 2013.

[33] I. Ana et al., "Temporal Pyramid Matching of Local Binary Subpatterns for Hand-Gesture Recognition", IEEE Signal Processing Letters, .**23**(8), 1037 -1041, 2016. DOI: 10.1109/LSP.2016.2579664

[34] S. Marcel, O. Bernier, J-E. Viallet and D.Collobert, "Hand gesture recognition using Input/Output Hidden Markov Models", Proceedings of 4th International Conference on Automatic Face and Gesture Recognition (AFGR), 2000.

[35] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. DOI: 10.1109/ICECCE47252.2019.8940756

[36] P. Baldi et al., "Assessing the accuracy of prediction algorithms for classification: an overview", Bioinformatics Review, **5**(5), 412-424,2000.
[37] T. Raghuveera, R Deepthi, R Mangalashri and R Akshaya. A depth-based Indian Sign Language recognition using Microsoft Kinect. Sadhana; 2020:45:34.
[38] M.S. Abdalla, "Dynamic Hand Gesture Recognition of Arabic Sign Language using Hand Motion Trajectory Features", Global Journals Inc. (USA), 13, 2013.
[39] S. S. Suni and K. Gopakumar, "Fusing Multimodal features for Recognizing Hand Gestures," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 1-6, 2019. doi: 10.1109/ICACC2019.8882910.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 955-959 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Scattering of H-Wave by a Moving Dispersive Conducting Complex Object

Esmail Mohamed Abuhdima*, Gurcan Comert, Ahmed El Qaouaq, Ashleigh Nicole Reeves

Computer Science, Physics and Engineering, Benedict College, Columbia, SC 29204, USA

ARTICLE INFO

Article history:
Received: 04 July, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Moving conducting cylinder
Scattered phase and magnitude
Dispersive conducting cylinder

ABSTRACT

The effect of the moving dispersive conducting complex scatterer on the scattering of incident H-wave is investigated herein. In this research, a simulation shows how the scattered phase and magnitude of a moving (rotating and translating) circular cylinder with higher conductivity, made of dispersive material, is affected in the case of incident H-wave (TE-mode) polarization. The Franklin transformation is applied to study the scattering of incident wave by a rotating dispersive higher conductivity cylinder, and then the effect of translating dispersive higher conductivity cylinder is investigated by applying the Lorentz transformation. This effect is studied at different speed of rotation. Also, this work shows that the pattern of scattered phase and magnitude are changed in terms of incident frequency. Moreover, a created model is used to demonstrate the impact of moving dispersive higher conductivity cylinder using backscattered static data which is generated using a comprehensive computational electromagnetics software (FEKO). The comparison between patterns of scattering of incident H-wave by a moving dispersive and non-dispersive higher conductivity cylinder is considered to show clear behavior of scattering patterns, in terms of the material of scatterer.

1. Introduction

The research is an expansion of the previous study published in 2019 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE) [1]. In previous work, the scattering of incident E-wave and H-wave by a rotating dispersive higher conductivity cylinder was investigated. The simulation result of this previous work shows a periodic pulse during rotation of dispersive conducting cylinder. It is found that the duration of these pulses is directly proportional to the speed of rotating dispersive higher conductivity cylinder, but inversely proportional to radius of dispersive higher conductivity cylinder. The difference between backscattered field of stationary case and backscattered field of rotating case is clear explicit in the incident H-wave polarization. For this reason, the incident H-wave is considered herein to study the effect of moving (rotating and translating) higher conductivity cylinder on the characteristic of the backscattered field. First of all, the Lorentz conversion is applied to figure out the effect of dispersive translating dispersive higher conductivity cylinder on the characteristics of the backscattered field. Secondly, Franklin conversion and Lorentz conversion are applied to study the characteristics of the backscattered phase and magnitude of moving dispersive conducting cylinder. Moreover, a created model is used to demonstrate the movement (rotation and translation) of dispersive higher conductivity cylinder using backscattered static data that is generated using FEKO. This data inserts into proposed model to demonstrate rotation and translation.

2. Scattered phase and magnitude

The effect of translation and rotation of a dispersive higher conductivity cylinder on scattered field (phase and magnitude) is investigated. The mathematical expression of scattered phase and magnitude is found by applying Lorentz and Franklin conversion [2]. Figure 1 shows the block diagram of the research problem.

It is assumed that a rotating dispersive higher conductivity cylinder is translated in the x-direction. Also, the incident wave propagates in the x-direction. Since, the magnetic field is in z-direction, so the electric filed is in y-direction by applying Maxwell's equations. Also, the direction of rotation is in the φ direction.

*Corresponding Author: Esmail Mohamed Abuhdima, Faculty of Computer Science, Physics and Engineering, Benedict College, Columbia, SC, USA. Email: esmail.abuhdima@benedict.edu

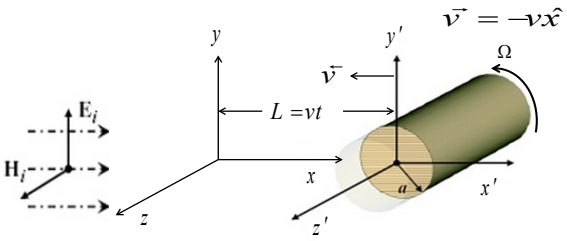


Figure 1: H-wave of moving dispersive conducting cylinder inrotatranslation

2.1. The effect of translating dispersive conducting cylinder

In this part, Lorentz conversion is applied to discuss the scattering of incident H-wave by a translating dispersive higher conductivity cylinder. The incident H-wave is defined as [3]

$$H_z^i = e^{j(\omega t - kx)} \tag{1}$$

where ω is the angular frequency and k called wavenumber. The electrical field is computed by

$$\bar{E}^i = \frac{1}{j\omega\varepsilon_o}\nabla\times\bar{H}^i = \hat{y}\frac{jk}{j\omega\varepsilon_o}\left[e^{j(\omega t-kx)}\right] = \hat{y}z_c e^{j(\omega t-kx)}, \tag{2}$$

where $z_c = \sqrt{\frac{\mu_o}{\varepsilon_o}}$ .These fields, which are defined by (1) and (2), are in laboratory frame. The Lorentz conversion is applied to compute fields in the co-moving frame using [4]

$$\bar{H}'_{\perp} = \gamma\left(\bar{H}_{\perp} - \bar{v}\times\bar{D}\right), \tag{3}$$

Where $\gamma = \left[1-\left(\frac{v}{c}\right)^2\right]^{-1/2}$, D is an electrical flux density, and v is a speed of translation.

The expression of scattered magnetic field in co-moving frame can write as [5]

$$\bar{H}_z^{sc\prime} = e^{j\omega' t'}\sum_{-\infty}^{+\infty} C_n H_n^{(2)}\left(k'r'\right)e^{jn\phi'} \tag{4}$$

where C_n is the unknown scattered coefficient. The unknown amplitude coefficient can be found by applying the boundary condition as $C_n = -\gamma\left[1+\frac{v}{c}\right]j^{-n}\frac{J'_n(k'a)}{H'^{(2)}_n(k'a)}$ [5].

The expression of magnetic field in the laboratory frame is computed using [6]

$$\bar{H}^{sc} = \gamma\left(\bar{H}_{\perp} + \bar{v}\times\bar{D}\right) \tag{5}$$

The backscattered magnetic field is given by

$$\bar{H}_z^{sc} = -\frac{e^{-jkr\left(1+2\frac{v}{c}\right)}}{\sqrt{r}}\sqrt{\frac{2j}{\pi k'}}\left(1+2\frac{v}{c}\right)e^{j\omega t\left(1+2\frac{v}{c}\right)}\sum_{-\infty}^{+\infty}\frac{J'_n\left(ka\left(1+2\frac{v}{c}\right)\right)}{H'^{(2)}_n\left(ka\left(1+2\frac{v}{c}\right)\right)}e^{jn\pi} \tag{6}$$

Equation (6) defined the backscattered magnetic field by translating very good dispersive conducting circular cylinder. It is seen that when $v = 0$ (stationary case), (6) leads to (11-115) that generated by Balanis [7], in the case of scattering of incident H-wave by a stationary perfect conducting cylinder.

2.2. Effect of movement

The scattering of incident H-wave by a rotating dispersive higher conductivity cylinder is investigated by taking the Franklin conversion. The expression of the magnetic field is written as [3]

$$H_z^i = e^{-jkx} = \sum_{-\infty}^{+\infty} j^{-n}J_n(kr)e^{jn\varphi}. \tag{7}$$

Fields inside the dispersive conducting cylinder is computed using Maxwell's equations as following [9],

$$H_Z = \sum_{-\infty}^{+\infty} A_{hn}J_n\left(\gamma_n r\right)e^{jn\varphi}, \tag{8}$$

where A_{hn} is the unknown transmission coefficient. The expression of scattered magnetic field can define as [9]

$$H_z^{sc} = \sum_{-\infty}^{+\infty} B_{hn}H_n^{(2)}\left(kr\right)e^{jn\varphi}, \tag{9}$$

where B_{hn} is the unknown scattering constant. This unknown scattering coefficient is found using the boundary conditions on the surface of the dispersive conducting cylinder [10] and [11]. These coefficients are defined by

$$B_{hn} = \frac{j^{-n}\left[kJ'_n(ka)J_n(\gamma_n a)\left(\alpha^2\dot{o}_r(\omega)+n\frac{\Omega\dot{o}_r(\omega)}{\omega}\right)-c\gamma_n J_n(ka)J'_n(\gamma_n a)\left(1+n\frac{\Omega}{\omega c\dot{o}_o}\right)\right]}{c\gamma_n J'_n(\gamma_n a)H_n^{(2)}(Ka)\left(1+n\frac{\Omega}{\omega c\dot{o}_o}\right)-kJ_n(\gamma_n a)H'^{(2)}_n(ka)\left(\alpha^2\dot{o}_r(\omega)+n\frac{\Omega\dot{o}_r(\omega)}{\omega}\right)} \tag{10}$$

$$A_{hn} = \frac{j^{-n}\left[J_n(ka)H'^{(2)}_n(ka)\left(\left(1+n\frac{\Omega}{\omega c\dot{o}_o}\right)\right)-J'_n(ka)H_n^{(2)}(Ka)\left(1+n\frac{\Omega}{\omega c\dot{o}_o}\right)\right]}{H'^{(2)}_n(ka)-J'_n(\gamma_n a)H_n^{(2)}(Ka)\left(1+n\frac{\Omega}{\omega c\dot{o}_o}\right)\left(\frac{c\gamma_n}{k\alpha^2\dot{o}_r(\omega)}\right)} \tag{11}$$

where $\gamma'^2_n = k'^2N^2\left(1-jq\right)+\frac{n\omega'\Omega}{2c^2}\left(N^2-1-j\frac{N^2q}{2}\right)$. The scattering of incident H-wave by a rotating dispersive higher conductivity cylinder is computed using (9). An approximation to Lorentz transformation $\omega' = \omega\left(\frac{c+v}{c-v}\right)\square\ \omega\left(1+2\frac{v}{c}\right)$ and $k' = k\left(\frac{c+v}{c-v}\right)\square\ k\left(1+2\frac{v}{c}\right)$ is substituted into (9) to find the

mathematical model of the effect of movement of dispersive higher conductivity cylinder on the backscattered field. The formula of a backscattered field is given by

$$H_z^{sc} = \sum_{-\infty}^{+\infty} B_{mn} H_n^{(2)}(kr) e^{jn\varphi}, \qquad (12)$$

where B_{mn} is the unknown moving scattering constant. It is written as

$$B_{mn} = \frac{j^{-n}\left[kJ_n'(k'a)J_n(\gamma_n'a)\left(\alpha^2\dot{o}_r'(\omega')+n\frac{\Omega\dot{o}_r'(\omega')}{\omega}\right)-c\gamma_n'J_n(k'a)J_n'(\gamma_n'a)\left(1+n'\frac{\Omega}{\omega' c\dot{o}_o}\right)\right]}{c\gamma_n'J_n'(\gamma_n'a)H_n^{(2)}(K'a)\left(1+n'\frac{\Omega}{\omega' c\dot{o}_o}\right)-k'J_n(\gamma_n'a)H_n^{(2)}(k'a)\left(\alpha^2\dot{o}_r'(\omega')+n'\frac{\Omega\dot{o}_r'(\omega')}{\omega'}\right)},$$

where $\varepsilon_r'(\omega') = 1-\frac{\omega' p^2}{\omega'^2 + j\omega'\Gamma}$, $n'(\omega')^2 = 1+j\frac{\sigma(\omega')}{\omega'\varepsilon_o}$, $\sigma'(\omega') = \varepsilon_o \frac{\omega' p^2}{-j\omega'+\Gamma}$, $\omega p^2 = \frac{nq^2}{m_e\varepsilon_o}$, q is the electron charge, m_e called mass of charge, Γ called damping coefficient, and ε_o is vacuum electric permittivity [1]. The backscattered field (far field), can be defined as

$$\mathrm{H}_m^{\mathrm{sc}} = -\hat{z}\sqrt{\frac{2j}{\pi k'}}\sum_{-\infty}^{+\infty} B'_{mn} e^{jn\pi}. \qquad (13)$$

The new proposed model is shown in Figure 2. The characteristic of this model is computed as $\mathrm{H}_m^{\mathrm{sc}} = \mathrm{H}_s^{\mathrm{sc}} \times A(s)$, so

$$A(s) = \frac{\mathrm{H}_m^{\mathrm{sc}}}{\mathrm{H}_s^{\mathrm{sc}}} = \frac{-\sqrt{\frac{2j}{\pi k'}}\sum_{-\infty}^{+\infty} B'_{mn} e^{jn\pi}}{\frac{1}{\sqrt{\pi k}}\sum_{-\infty}^{+\infty}\frac{J_n'(ka)}{H_n^{(2)\prime}(Ka)} e^{jn\pi}}, \qquad (14)$$

where $\mathrm{H}_s^{\mathrm{sc}}$ is scattered magnetic field of stationary perfect conducting cylinder [7].

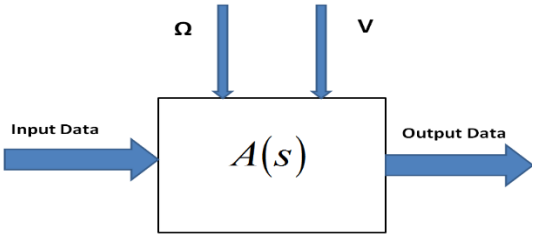


Figure 2: Proposed model

3. Numerical result

Equation (13) defines the far field (backscattered phase and magnitude) of moving dispersive higher conductivity cylinder. It is considered that the dispersive higher conductivity cylinder has a radius equal to $0.03m$ and the incident frequency changes from 2GHz to 20GHz. In this simulation, the conductivity of cylinder's martial is $\sigma = 5.76\times10^7\ s/m$ (Copper). This simulation is taken place using a specific values of $\beta_a = \Omega a/c = 0.009$ and $v/c = 0.02$.

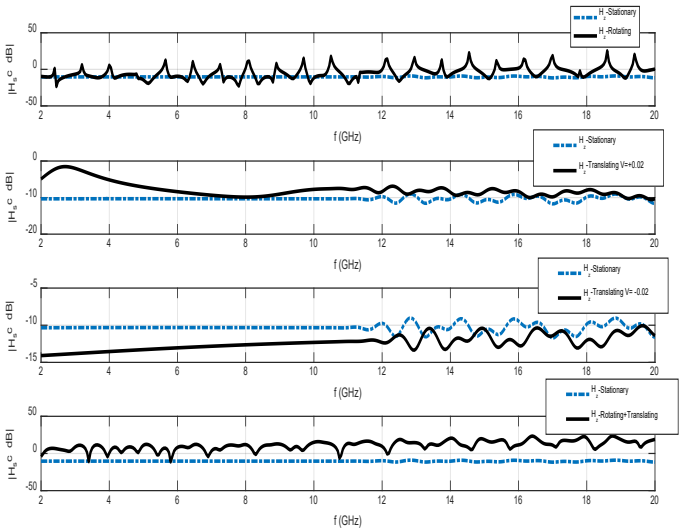


Figure 3: Magnitude of backscattered H-wave

It is shown the backscattered field (phase and magnitude) of the rotating, translating and moving dispersive higher conductivity cylinder cases in comparison with stationary case are shown in Figure 3. The graphs' y-axis and the x-axis represent the magnitude the operating frequency of backscattered magnetic field respectively. The first graph is a comparison of the backscattered magnitude of a stationary and a rotating dispersive higher conductivity cylinder. The blue line represents the stationary case, and the rotating case is represented by black line. Visibly it is observed that the black line (rotating case) shows pulses as frequencies increases and the blue line (stationary case) shows tiny ripples in higher frequencies, as shown in the first graph.

The second and the third graph show the comparison of the backscattered magnitude of a stationary and translating very good dispersive conducting cylinder, that is heading in a positive direction (away from the source) and the negative direction (towards the source). When the dispersive higher conductivity cylinder is heading in a positive direction, the backscattered magnitude is shifted in the higher frequencies in comparison with the stationary case. On the contrary, the backscattered magnitude is shifted in the lower frequencies when the dispersive higher conductivity cylinder is heading in the negative direction. The effect of a moving dispersive higher conductivity cylinder on the backscattered magnitude is shown in the fourth graph. The backscattered magnitude shows a pulses and negative slop in the moving case.

Figure 4 shows the backscattered phase of rotating, translating and moving cases. The first graph shows the difference between the backscattered phases of stationary and rotating. This difference shows up as a periodic pulse. In the translating case, the difference between the backscattered phase of stationary and translating cases is shown in the second and third graphs. When the dispersive higher conductivity cylinder is heading in a positive direction, the phase difference appears as a series of pulses. The widths of these

pulses increase as the dispersive higher conductivity cylinder heads in a positive direction. Also, slope of the pulses in the case of a positive translation is negative. On the contrary, the width of the pulses decreases when the dispersive higher conductivity cylinder heads in the negative direction. In this case, the slope of the pulses is positive. In comparison to the past work, the backscattered phase and amplitude of the moving non-dispersive conducting cylinder with high conductivity showed up a sinusoidal behavior, especially in higher frequencies [12]. The effect of movement of dispersive higher conductivity cylinder is shown in graph four.

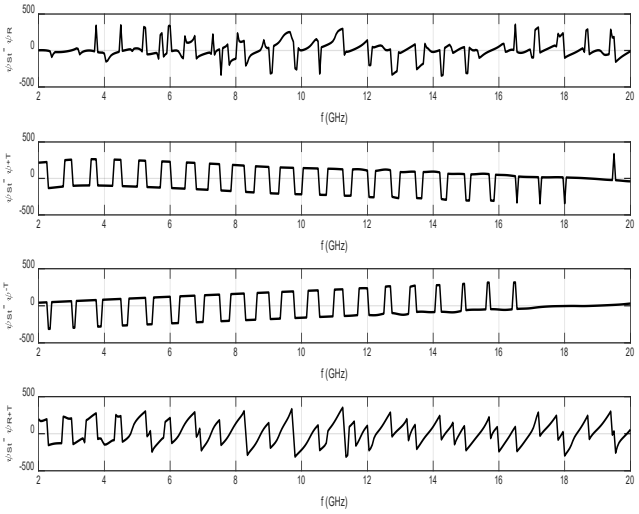


Figure 4: Phase of backscattered H-wave

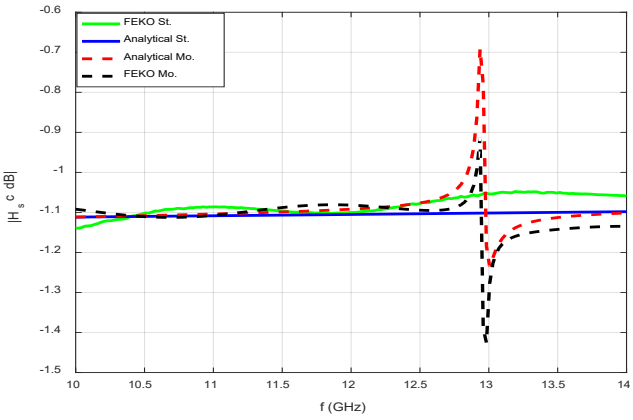


Figure 5: Magnitude of backscattered H-wave

The dispersive higher conductivity cylinder that is displayed in Figure 1 constructed by using a comprehensive computational electromagnetics software (FEKO). To construct the geometry of problem (Figure 1), It is assumed the radius, the height and the conductivity of cylinder are $a = 0.025\,m$, $h = 0.1\,m$ and $\sigma = 5.76\times10^{7}\ s/m$ respectively. Also, the range of simulation frequency is from 10 GHz to 14GHz.

In this part, the generated stationary static data is inserted into proposed model that is sown in Figure 2. Figure 5 and Figure 6 show that the analytical and simulation results of the backscattered phase and magnitude show a similar behavior in the stationary (St) and moving (Mo) cases.

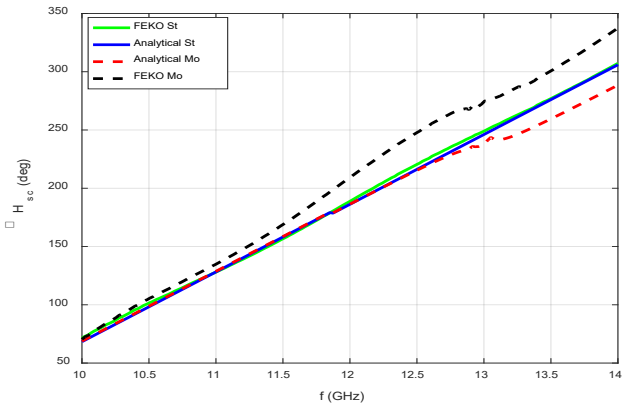


Figure 6: Phase of backscattered H-wave

4. Conclusion

In this research, the backscattered phase and magnitude of movement dispersive conducting cylinder is studied in the case of incident TE mode. The effect of the movement of the dispersive higher conductivity cylinder on the scattering of incident H-wave is studied. In this research, a mathematical model that represents the moving backscattered field (magnitude and phase) is created using Franklin and Lorentz conversion. It is found that the effect of the rotation shows as a periodic pulse on the patterns of backscattered phase and magnitude. Also, simulation result shows that the duration of pulses is affected when a speed and radius of dispersive higher conductivity cylinder are changed. In the translating case, the backscattered magnitude of an incident H-wave is shifted in the direction of higher frequencies when the dispersive higher conductivity cylinder moves away from the source of incident wave. Conversely, the backscattered magnitude of the incident H-wave is shifted in the lower frequencies when dispersive higher conductivity cylinder moves towards the source of incident wave. The phase difference of a backscattered field appears as series pulses in the translating mode. The slop of pulses is negative and the width of pulses increases when the dispersive higher conductivity cylinder moves away from the source of incident wave. When the dispersive higher conductivity cylinder moves towards the source of incident wave, the slop of pulses is positive and the width of pulses decreases. The demonstration results of the proposed model and an analytical results gives a similar backscattered phase and magnitude patterns. In comparison to the past work, the pattern of backscattered phase and amplitude of dispersive higher conductivity cylinder showed a periodic pulse. However, the pattern of backscattered phase and amplitude of non-dispersive higher conductivity cylinder showed a sinusoidal behavior, especially in higher frequencies. These phenomena display the impact of the electrical properties of material of scatterer on the scattering pattern of an incident H-wave.

Acknowledgment

The research is supported by the Center for Connected Multimodal Mobility (C2M2). Authors would also like to acknowledge U.S. Department of Homeland Security (DHS) Summer Research Team Program Follow-On, FY19 US

Department of Education MSEIP Grant Award P120A190061, and National Science Foundation (NSF, No. 1719501, 1436222, and 1400991) grants. Any ideas, results, epilogue or commendation that are expressed in this study are for those author(s), and do not necessarily reverberate the opinions of (C2M2), USDOT, DHS, U.S. Department of Education, or NSF. The U.S. Government suppose no liability for contents or utilize therefrom.

References

[1] E. Abuhdima, Gurcan Comert, Ahmed Elqaouaq, Ashleigh Reeves and Williams Kellen, "Scattering of EM waves from a Rotating Dispersive Very Good Conducting Cylinder," in 2019 IEEE International conference on wireless for space and extreme environments, October, 2019.

[2] Philip Franklin, "The meaning of rotation in the special theory of relativity," Proc. Nat. Acad. Sci., **8**(9), 1922.

[3] Harrington, Time Harmonic Elecromagnetic Fields, McGraw-Hill, 234, 1961.

[4] E. Abuhdima and R. Penno, "A New Model for Simulation of Scattered EM Fields from a conducting Cylinder in Rotation and Translation using Static Data," in IEEE NAECON-OIS, 2017.

[5] Abuhdima, Esmail, "Simulation of the Scattered EM Field of a Moving Dynamic Object Using Static Data," Ph. D Thesis, University of Dayton, 2017.

[6] H. Gholizade, "Radar Cross Section of Moving Objects," The European physical journal, 2013.

[7] C. A. Balanis, Advanced Engineering Electromagnetics, John Wiley &Sons, Inc, 1989.

[8] E. Abuhdima and R. Penno, "Simulation of the scattered EM fields from a rotating conducting cylinder," in IEEE International Radar, 2015.

[9] P. Hillion, "Scattering by a rotating circular conducting cylinder I," Mathematical physics, **41**, 1998.

[10] P. Hillion, "Scattering by a rotating circular conducting cylinder II," Mathematical physics, **41**, 1998.

[11] D. De Zutter, "Scattering by a rotating circular cylinder with finite conductivity, " IEEE Transactions on Antenna and Propagation , **AP-31**(1), 166-169, 1983.

[12] E. Abuhdima and R. Penno,"The Effect of Rotation and Translation upon the Scattered EM Fields of a Conducting Cylinder," in NAECON-OIS, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 984-990 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Impact of Innovation on the Performance of Manufacturing Enterprises in Vietnam

Thi Anh Van Nguyen*, Khac Hieu Nguyen

Department of Industrial Management, HCMC University of Technology and Education, Ho Chi Minh City, 700000, Vietnam

ARTICLE INFO

Article history:
Received: 23 August, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Product Innovation
Technology Innovation
Organization Innovation
Firm Performance
Manufacturing Enterprises

ABSTRACT

This paper examines the impact of innovation on the performance of manufacturing enterprises in Vietnam. Innovation is measured by product innovation (3 observed variables), technology innovation (8 observed variables), and organization innovation (6 observed variables) while firm performance is measured by revenue and profit. The OLS regression model was used with data collected from 806 enterprises in four industrial sectors. The results show that innovation has a positive effect on firm performance. From the results, some implications are proposed to improve the performance of manufacturing enterprises in Vietnam.

1. Introduction

Nowadays, with the changing business environment, innovation is the key to decide the success or failure of many businesses. In the world, innovation takes place quickly and strongly. Many studies show that innovation (technological innovation, product innovation, organizational innovation) has a positive effect on the performance of enterprises. However, most of the studies in Vietnam focusing on small and medium-sized enterprises and these studies mostly ignored the impact of the innovation factor on the performance of enterprises. Ho Chi Minh City is the largest economic center in Vietnam, and a place to welcome investment waves as well as technological innovations. In recent years, the city has had a policy of investing in the development of four key industries: the Food processing industry; Pharmaceutical-rubber industry; Mechanical industry and Electronics & Information technology. However, measuring the impact of policies as well as the innovation on performance is an issue of little concern. This is a research gap for the authors to implement the study **"The impact of innovation on the performance of manufacturing enterprises in Vietnam".** This paper examines the relationship between innovation (including technological innovation, product innovation, organizational innovation) and business performance (including revenue and profits). From the results, the paper proposes some recommendations to improve the performance of enterprises in four key industries in Ho Chi Minh City.

In this article, innovation includes technological innovation, product innovation, and organizational innovation. Business performance is analyzed and measured based on revenue and profit.

2. Literature review and research model

This section presents related concepts such as innovation, product innovation, technological innovation, organizational innovation, and the four key industries. Studies of the impact of innovation on the performance will be reviewed in the following part. Finally, a research model will be proposed based on the previous researches.

2.1. Related concepts

2.1.1 Innovation:

Innovation has become the main driving force of economic strength and national competitiveness. Most people think that innovation only brings technological aspects to create outstanding new products, such as Apple's iPad, Sony's PlayStation, or applying artificial intelligence to business, etc. Others argue that innovation is only associated with research and development (R&D) activities conducted at national universities and laboratories. This concept is too limited because innovation has a much broader meaning [1].

Innovation nowadays is understood as "a new idea, creative thought, or new imagination in the form of equipment or methods". Innovation is often seen as the application of better solutions to meet new requirements and existing market needs. The innovation takes place through the provision of new products, new processes, new services, new technologies or new efficient business models

*Corresponding Author: Thi Anh Van Nguyen, Email: anhvan@hcmute.edu.vn

[2]. Innovation is related invention, but not like it, because innovation is most likely related to the actual implementation of an invention (i.e. new/improved ability) to create a meaningful impact in the market or society, and not all innovations require an invention [3], [4].

A fairly complete definition of the Organization for Economic Co-operation and Development (OECD)'s innovation is "implementing a new product or a significant improvement (for a particular product or service), a new process, marketing method, or a new organizational method in business practices, workplace organization, or external relations " [5]. From here, we can see the concept of innovation including technology and non-technology. In this study, the concept of innovation used includes product/service innovation; technological innovation and organizational innovation.

2.1.2. Product innovation

Product innovation refers to the creation and introduction of new products and services, whereby the direction of innovation is associated with the speed of innovation (necessary time to develop new products), the ability to replace regular products with improved versions and the ability to launch new products to new markets [5] [6].

2.1.3. Technology innovation

Technology innovation is the replacement of a major part or the whole of an existing technology with more advanced and more efficient technology. Technology innovation aims to increase the productivity, quality and efficiency of the production process or create a new product or service to serve the market.

Technological innovation includes new products and processes and significant technological changes of products and processes [7]. An innovation has been made if it is launched in the market. Technological innovation is inevitable for companies that want to develop and remain competitive or achieve the goal of entering new markets [8]. There are many variables that represent technological innovation such as: technology infrastructure (computers, internet, R&D costs, etc.); government support for technology; enterprise competencies (enterprise training, technology strategic plan, ...) and corporate culture (human resource level, management support, ...).

2.1.4. Organization innovation

Organization innovation involves upgrading management processes through new business approaches. Organization innovation can improve business performance by supporting necessary changes, reducing transaction and administrative costs, improving workplace satisfaction, reducing input costs [9]. Organization innovation includes training programs, product design programs, and management group creativity in problem-solving [10]. In 2006, authors proposed that organizational innovation involves management efforts to refresh organizational habits, procedures, mechanisms or systems to create ultimate ecological innovation.

2.1.5. Four key industries

According to the Ho Chi Minh City General Statistics Office [11], the codes for the four key industries are shown in Table 1 as follows:

Table 1: Industries belong to 4 key industries groups

Key industries group	Industry code	Industry name
Foods (1)	11	Beverage manufacturing
Pharmaceutical chemicals - Rubber (2)	20	Chemicals and chemical products manufacturing
	21	Medicine, pharmaceutical chemicals and pharmacognosy manufacturing
	22	Rubber and plastic products manufacturing
Mechanical (3)	24	Metal production
	25	Manufacturing products from prefabricated metal (except machinery and equipment)
	28	Producing machines and equipment not classified in any category
	29	Motor vehicles manufacturing
	30	Other means of transportation manufacturing
	27	Electrical equipment manufacturing
Electronics and information technology (4)	26	Manufacturing electronic products, computers and optical products

2.2. Study overview

2.2.1. Overview of related research works in the world

Currently, the issue of innovation is being strongly concerned in the world, especially with the strong development of the industrial revolution 4.0. Therefore, there are a lot of researches related to this field. In the context of the topic, the author reviews a number of studies related to the question of how innovation affects the company's operating results such as revenue, profits, and costs.

In 2008, the researcher collected data from 7302 companies in Europe to test the relationship of using technology via the internet, different types of innovation and the performance of enterprises. Research results indicated that all types of product and process innovation (including products related to e-commerce and traditional products) have positive effects on revenue and labor force growth [12]. Another research studied small and medium-sized companies in Portugal on the relationship between Total Quality Management (TQM), customer orientation, innovation and the operation of businesses [13]. The author concluded that most of the components of TQM affect customer orientation and business performance. Meanwhile TQM does not affect innovation. However, innovation affects business performance [13].

In 2010, the research on 600 small and medium-sized companies in Australia was done and proposed that the enterprise formal structure and innovation strategy are the main keys affecting enterprises' operations [14].

In 2014, the relationship between innovation (process innovation, product innovation, organizational innovation) to the performance of enterprises in 121 enterprises in Taiwan were investigated in some other papers. They suggested that

organizational innovation has the most positive impact on the performance of businesses. Meanwhile, process innovation and product innovation affect organizational innovation, and organizational innovation affects the performance of businesses [15].

In 2017, Rangus and Slavec examined the relationship between organizational characteristics, firm's innovation and firm performance. After studying 421 manufacturing and service companies in Slovenian, they proposed that innovation has a positive impact on the performance of businesses [16].

Recently, a study used the Ordinary Least Squares (OLS) and simultaneous equation models to analyze large data which included 63303 firms from 13 different European Union countries. The results showed that both human capital and the number of adopted eco-innovation strategies influenced positively firm performance [17]. Another study analyzed the impact of innovation on firm performance in developing economies. The study used data from 8,551 enterprises which were surveyed between 2014 and 2016 covering 19 Sub Saharan African countries. The CDM model and ordinary least squares method were adopted to analyze the effect of domestic R&D on innovation output, and the contribution of innovation to firm performance in the sub region. The result showed that both product and process innovations had a positive and significant effect on firm performance [18].

In general, although studies are conducted at different times in many countries, the results are quite similar in testing the relationship between innovation and business performance.

In the industrial revolution 4.0 and the current globalization, innovation is really an urgent issue for businesses in Vietnam in general and Ho Chi Minh City in particular. However, according to experts, the rate of innovation in Vietnamese enterprises is very low. There are nearly 600,000 businesses in the country, most of them are using technology lagging behind the average level of the world from two to three generations.

Ho Chi Minh City is an important economic, cultural and educational center of Vietnam. The industrial production value of Ho Chi Minh City accounts for about 45% of industrial output value in the southern key economic region and contributes about 26% of the national industrial production scale [19]. The investment policy of Ho Chi Minh City is to focus on investing in four key industries: mechanical engineering, electronics, chemicals - rubber - plastic and food processing. However, studies on innovation issues in these enterprises and the impact of innovation on business performance are limited.

Most of the research on innovation and business performance in Vietnam is about small and medium enterprises (SMEs). In 2008, some researchers examined the relationship between innovation and export at Vietnamese SMEs and concluded that innovation has an impact on export [20]. For instance, some authors surveyed 389 SMEs in Can Tho City in 2011. The results showed that factors such as the level of access to government support policies, the education level of the business owner, the size of the business, the social relationship of the enterprise and the revenue growth rate affect the efficiency of business operation of SMEs in the Can Tho city [21].

One of the few studies on innovation to business results of Vietnamese enterprises was conducted by Quan Minh Nhut in 2018 [22]. The author surveyed primary information from 55 randomly selected agricultural enterprises representing the agricultural enterprises in Ben Tre province. After studying the effect of the scientific and technological investment on enterprises' business results, discriminatory analysis results showed that there are 4 statistically significant factors affecting the difference of enterprises' profit: total revenue, investment capital for application of science and technology, type of enterprises and market expansion [22].

Finally, in 2019, some authors did research on the drivers of firm productivity, focusing on the role of international management standards certification in Vietnam. They concluded that the possession of an internationally recognized standard certificate leads to significant productivity. They also proposed that the effect of certification on productivity is particularly strong for firms with technological innovation [23].

2.3. Research model

From the above results, most of the researches concluded that innovation has a positive impact on the performance of enterprises. Therefore, the author proposes a research model with the dependent variable as the business performance of the enterprise, measured by revenue and profit. Independent variables are product innovation, technological innovation and organizational innovation.

After reviewing the above studies, the author proposes the following model:

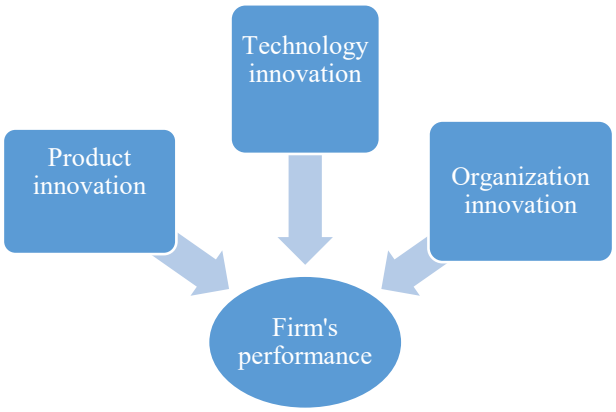


Figure 1: Proposed Research model

3. Research methodology and data

In order to test the proposed research model (Figure 1), the author uses Ordinary Least Squares (OLS) regression. There are many studies which did research about firm performance by OLS method and the results are consistent and reliable [17], [18], [24]-[26]. In this study, the authors suggest dependent variable denoted by Yi, independent variable denoted by Xi and error denoted as u_i, a constant β1 and a vector of regression coefficients β2. The regression equation is expressed as follows:

In which, Yi is a vector of dependent variables including: Revenue, Profit, Xi is an independent variable vector. Details of variable encoding and definitions are shown in the following table.

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Table 2: Variables names and definitions

Variable name	Definition
Revenue	Log10 revenue of the enterprise in 2018 (Unit: Million VND)
Profit	Log10 profit of the enterprise in 2018 (Unit: Million VND)
Product Innovation	
New products/services	Binary variable, equal to 1 if the enterprise develops new products/services or innovates the existing products/services.
Quantity of new product types	Quantity of new products/services launched to the market in 2018
Cost for testing new products/services	Cost for testing new products/services in 2018. (Unit: Million VND)
Technology Innovation	
New technology investment	Binary variable, equal to 1 if there is any technology innovation or improvement of the current technology.
Cost of hiring machines and technology	Expenses for hiring/buying specialized machines, equipment, software and new technologies in 2018. (Unit: Million VND)
Cost of improving machines and technology	Expenses for the improvement of machines, equipment, specialized software and technologies in 2018. (Unit: Million VND)
Cost of technology project research	Expenses for technology project research in 2018. (Unit: Million VND)
Level of computer equipment	Level of computer equipment is calculated by the quantity of current equipment dividing by the actual quantity of the equipment needed for work.
Using internet/computer network	Binary variable, equal to 1 if the enterprise use internet or computer network
Private information website	Binary variable, equal to 1 if the enterprise has its own portal/website.
Cloud computing model	Binary variable, equal to 1 if the enterprise uses cloud computing model for work.
Organization Innovation	
New marketing method	Binary variable, equal to 1 if the enterprise has a new marketing method, new experience for customers.
New management method	Binary variable, equal to 1 if the enterprise has implemented a new management and organization method.
Research & Development (R&D)	Binary variable, equal to 1 if the enterprise has R&D department.
Science and Technology Development Fund	Binary variable, equal to 1 if the enterprise has established Science and Technology Development Fund.
Training cost	Labor/human resource training costs in technology hiring/innovation/research activities in 2018 (Unit: Million VND).
Government's support	Binary variable, equal to 1 if the government assists in innovation.

The data was from the surveys conducted in enterprises in four key industries including: Food processing industry; Pharmaceutical-rubber industry; Mechanical industry and Electronics & Information technology. The results obtained 806 qualified answers.

4. Research results and discussion

4.1. Descriptive statistics

Survey results collected 806 qualified samples, of which enterprises in the field of mechanical manufacturing - automation accounted for 47.8%; plastic - rubber sector (29.5%); food processing (16%); and the lowest is information technology - electronics and telecommunications (6.7%).

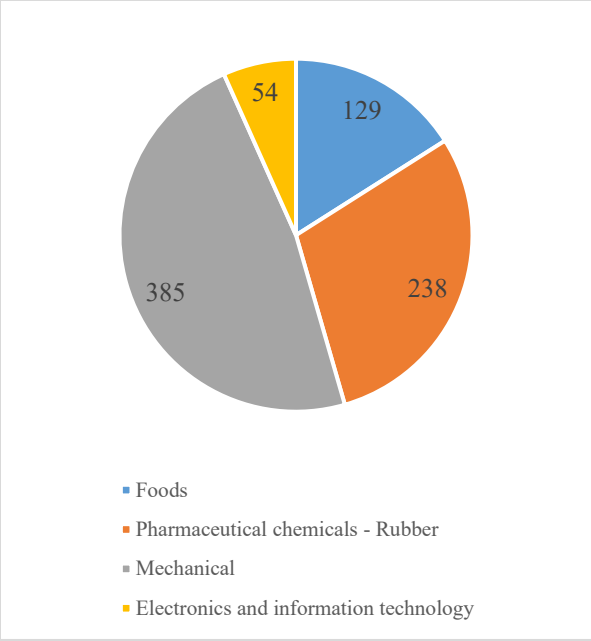


Figure 2: Number of surveyed companies in four sectors

4.2. Regression results with the dependent variable REVENUE and PROFIT.

To see the impact of innovation on the performance of enterprises, the authors analyze data by the OLS regression method. Details of regression analysis results are presented in Table 3 and Table 4. In order to ensure that the regression model is not faulty, the authors tested multi-collinearity and heteroscedasticity error of the model. The results of the multi-collinearity test show that all the magnification coefficients are less than two, hence, we can conclude that the regression equation does not occur the multi-collinearity phenomenon. The heteroscedasticity error is checked by the White test. The results show that there is no heteroscedasticity error because of the significance of Prob (F)> 10%.

Table 4: Regression results with the dependent variable is REVENUE

Observed variables	Coef	P-Value	VIF
Intercept	3.866	0.000	
Cost of hiring machines and technology	0.000001**	0.028	1.10
Expenses for improving machines and technology	0.000111***	0.000	1.07
Cost of technology project research	-0.000004	0.834	1.19

Cost for testing new service/products	0.000027***	0.004	1.06
Training cost	0.000324*	0.078	1.24
Quantity of new product types	0.00408	0.254	1.15
Level of computer equipment	0.000384	0.579	1.04
Using internet/computer network	0.367**	0.045	1.07
Private information website	0.2581***	0.000	1.18
Cloud computing model	-0.0519	0.529	1.15
Government's support	0.1935**	0.005	1.12
R&D	-0.0357	0.680	1.31
Science and Technology Development Fund	0.212	0.239	1.16
New products/services	0.0349	0.701	1.67
Investment in new technologies	0.2904***	0.000	1.61
New marketing method	-0.2525***	0.009	1.49
New management method	0.5946***	0.000	1.23
R-squared	0.302649		
F-statistic	20.06607***		
Prob(F-statistic)	0.000000		

Note: *, ** and *** represent for the level of significance 10%, 5% and 1%, respectively.

The results of the regression analysis show that the factors affecting revenue are Expenses for hiring machines and technology; Expenses for improving machines and technology; Cost for testing new service/products; Training costs; Using internet/computer network; Private information website; Government's support; Investment in new technologies; New marketing methods; New management method.

Table 5: Regression results with the dependent variable is PROFIT

Variable	Coef	P-Value	VIF
Intercept	5.1103	0.000	
Cost of hiring machines and technology	0.000001***	0.000	1.10
Expenses for improving machines and technology	0.000023***	0.000	1.07
Cost of technology project research	-0.000004	0.126	1.19
Cost for testing new service/products	0.000008***	0.000	1.06
Training cost	0.000082***	0.004	1.24
Quantity of new product types	0.001156**	0.037	1.15
Level of computer equipment	0.000172*	0.108	1.03
Using internet/computer network	0.0101	0.722	1.07
Private information website	-0.00092	0.915	1.18
Cloud computing model	-0.0051	0.691	1.15
Government's support	0.0027	0.800	1.12
R&D	0.0305**	0.023	1.31
Science and Technology Development Fund	0.1019***	0.000	1.16
New products/services	0.0143	0.309	1.67
Investment in new technologies	-0.0018	0.879	1.61
New marketing method	0.0006	0.967	1.49
New management method	0.02431***	0.007	1.23
R-squared	0.261308		
F-statistic	16.37629***		
Prob(F-statistic)	0.000000		

Note: *, ** and *** represent for the level of significance 10%, 5% and 1%, respectively.

The results of the regression analysis show that the factors affecting profit are: Expenses for hiring machines and technology; Expenses for improving machines and technology; Cost for testing new services/products; Training costs; Quantity of new product types; Level of computer equipment; R&D; Science and Technology Development Fund; New management method.

4.3. Research results discussion

Factors positively affecting both revenue and profit: Cost of hiring machines and technology; Expenses for improving machines and technology; Cost for testing new services/products; Training cost; New management method. Specifically, the greater the cost of hiring/buying specialized machines, equipment, software and new technologies in 2018, the greater the revenue and profit. The results for the cost of improvements to machines, equipment, specialized software, technology in 2018, costs for testing new products/services and training costs in 2018 are similar. These results appear to be new contributions for firm performance studies.

Normally, we assume that cost will decrease revenue and profit. However, the results in this case show that the following costs will help revenue and profit increase: expenses for hiring/buying machinery and equipment, specialized software, new technology; cost of improvements to machines, equipment, specialized software, technology; costs for testing new products/services; costs for training labors/human resources in hiring/innovating/technology research activities in 2018.
In [27], the author presented these costs to be called prevention costs or quality costs in the company. These costs are not only to ensure the creation of a product/service that meets market needs, but also to prevent customers from being dissatisfied with the product/service, thereby reducing the failure costs of the company. Therefore, these costs bring many benefits such as increasing customer satisfaction, increasing sales and profits.

In addition, the observed variable "new management method" has a positive impact on revenue and profit of businesses with a significant level of p = 1%. In an organization, management greatly affects the performance of the company. With the business environment constantly changing, it is required that the management method must be improved and refreshed to suit the general context. That is the reason why implementing a new management and organization method in businesses positively affects revenue and profit.

Factors that only affect revenue: Using internet/computer networks; Private information website; Government's support; Investment in new technologies; New marketing method. Specifically, the revenue of businesses will be higher if businesses use internet or computer network, has private information website, get support from the government, invest in new technology or innovate the existing technology. The results are same as the results of some previous studies. However, if the enterprises apply new marketing methods, new experiences for customers, the revenue will decrease. The use of supporting tools such as computers, internet, websites will help businesses operate more effectively, leading to better sales. Businesses that have support in innovation will also have a better competitive advantage, hence, the revenue is also better. However, according to the analysis results, the revenue of the businesses that have new marketing methods and new experiences for customer's decreases; This may be due to the high cost of these marketing methods or the inappropriate marketing methods, resulting in low efficiency.

Factors that only affect profit: Quantity of new product types; Level of computer equipment; R&D; Science and Technology Development Fund. The study results show that the more the new types of services/products are launched to the market in 2018, the higher profit the businesses get. Furthermore, the higher level of

computer equipment, the greater the profit. Also, if the businesses have R&D department or have Science and Technology Development Fund, the profit will increase. The results are equivalent to the results of some previous studies.

5. Conclusions and policy implications

The analysis results show that the factors of technology innovation, product innovation and organizational innovation all have observable variables affecting the revenue and profit of enterprises in four key industries in Ho Chi Minh City. The authors propose a few recommendations to improve business performance for these businesses as follows:

First, businesses should invest in prevention costs such as: innovation costs, costs of testing new products/services; training costs. These costs are essential for the sustainable development of businesses, improving operational efficiency as well as business performance[27].

Second, enterprises need to improve and innovate management methods in the organization to suit the business environment. The research results show that the "new management method" has the greatest positive impact on both revenue and profit. This reaffirms the role of management in the organization. Nowadays, there are many management methods in the world that Vietnamese enterprises can apply such as: ISO 9001, ISO 14001, ISO 22000, ISO 27000, etc. Businesses choose for themselves the most appropriate management model depending on the different industries and fields of operations. And it is greatly important to constantly improve and innovate to keep up with the changes in the business environment and organizational context.

Third, businesses should equip themselves with supporting tools such as computers, internet, private websites, etc. With the development of information technology nowadays, the above tools do not only help businesses increase revenue (according to analysis results), but also help businesses improve working efficiency, enhance brands, reach customers more easily, thereby improving the competitive advantage of businesses.

Fourth, the government needs to promote the support of innovation for businesses. The support here is not only financial but also support for skills training, management method updates to raise awareness as well as skills for workers; create a transparent startup ecosystem, a fair playing field for businesses. *Finally,* enterprises should have a research and development department or set up a science and technology (S&T) development fund. According to the analysis results in 4 key industries in Ho Chi Minh City, enterprises that have either of these two sections will have higher profits. As experts have said, the philosophy in any management system is "do it right from the beginning". R&D departments or S&T development funds are the original incubators for ideas. If this nursery produces good seeds, of course the results will be good.

Limitation of the study: Although the paper has achieved some results, the research still has some limitations. The study only analyzes data in a specific period of time, has not yet analyzed data to see the fluctuation of the dependent variable over time. The thesis has just verified the results by logit regression model but has not compared the results with other models. The study was conducted only in Ho Chi Minh City but not in other localities in Vietnam. The authors hope to carry out further studies to improve the above limitations.

References

[1] National Department of Science and Technology Information, "Report "Global innovation policy"," 2012. [Online]. Available: http://www.vista.gov.vn/.

[2] S. Maranville, "Entrepreneurship in the Business Curriculum," J. Educ. Bus., **68**(1), 27–31, Oct. 1992.

[3] P. Frankelius, "Questioning two myths in innovation literature," J. High Technol. Manag. Res., **20**(1), 40–51, 2009. https://doi.org/10.1016/j.hitech.2009.02.002

[4] K. Bhasin, "This Is The Difference Between 'Invention' And 'Innovation,'" 2012. [Online]. Available: https://www.businessinsider.com/this-is-the-difference-between-invention-and-innovation-2012-4.

[5] OECD, "Measuring Innovation: A New Perspective," 2010. [Online]. Available: http://www.oecd.org/.

[6] D. I. Prajogo and A. S. Sohal, "The integration of TQM and technology/R&D management in determining quality and innovation performance," Omega, **34**(3), 296–312, Jun. 2006.

[7] E. OECD, Presentación Oslo Manual 2018(October. 2018.

[8] V. Souitaris, "Technological trajectories as moderators of firm-level determinants of innovation," Res. Policy, **31**(6), 877–898, Aug. 2002. https://doi.org/10.1016/S0048-7333(01)00154-8

[9] M. Shafiq et al., "Towards successful global software development." In Proceedings of the Evaluation and Assessment in Software Engineering, 445-450. 2020. DOI: https://doi.org/10.1145/3383219.3383283

[10] P. Kemp, R., Pearson, "Final Report MEI Project about Measuring Ecoinnovation: Deliverable 15 of MEI Project," 2008.

[11] Ho Chi Minh City General Statistics Office, "Statistical Yearbook," 2018. [Online]. Available: http://www.pso.hochiminhcity.gov.vn/.

[12] P. Koellinger, "The relationship between technology, innovation, and firm performance-Empirical evidence from e-business in Europe," Research Policy, **37**(8), 1317–1328, 2008. https://doi.org/10.1016/j.respol.2008.04.024

[13] J. Carlos Pinho, "TQM and performance in small medium enterprises," Int. J. Qual. Reliab. Manag., **25**(3), 256–275, Mar. 2008. https://doi.org/10.1108/02656710810854278

[14] M. Terziovski, "Innovation practice and its performance implications in small and medium enterprises (SMEs) in the manufacturing sector: a resource-based view," Strateg. Manag. J., p. n/a-n/a, 2010.

[15] C. C. J. Cheng, C. Yang, and C. Sheu, "The link between eco-innovation and business performance: a Taiwanese industry context," J. Clean. Prod., **64**, 81–90, Feb. 2014.

[16] K. Rangus and A. Slavec, "The interplay of decentralization, employee involvement and absorptive capacity on firms' innovation and business performance," Technol. Forecast. Soc. Change, **120**, 195–203, 2017.

[17] M. Madaleno, M. Robaina, M. Ferreira Dias, and M. Meireles, "Dimension effects in the relationship between eco-innovation and firm performance: A European comparison," Energy Reports, **6**, 631–637, Feb. 2020.

[18] B. A. Asunka, Z. Ma, M. Li, and O. A. Anaba, "Linking Innovation to Firm Performance in Developing Countries: the Role of Trade Liberalization," Eur. J. Bus. Manag. Res., **5**(3), 2020.

[19] N. Thao, "Ho Chi Minh City focuses on developing 4 key industries," 2017. [Online]. Available: http://www.ipcs.vn/vn/tp-ho-chi-minh-tap-trung-phat-trien-4-nganh-cong-nghiep-trong-yeu-W1422.htm.

[20] A. N. Nguyen, N. Q. Pham, C. D. Nguyen, and N. D. Nguyen, "Innovation and exports in Vietnam's SME sector," Eur. J. Dev. Res., **20**(2), 262–280, Jun. 2008.

[21] Q. N. Nguyen and V. N. Mai, "Factors affecting business performance of small and medium-sized enterprises in Can Tho City," Sci. J. Can Tho Univ.**19b**, 122–129, 2011.

[22] Q. M. Nhựt, "The effects of science and technology investment on business performance of agricultural enterprises in Ben Tre province," Can Tho Univ. J. Sci., **54**(1), p. 187, 2018.

[23] E. Calza, M. Goedhuys, and N. Trifković, "Drivers of productivity in Vietnamese SMEs: the role of management standards and innovation," Economics of Innovation and New Technology, **28**(1), 23–44, 2019.

[24] M. L. Yeh, H. P. Chu, P. J. Sher, and Y. C. Chiu, "R&D intensity, firm performance and the identification of the threshold: Fresh evidence from the panel threshold regression model," Appl. Econ., **42**(3), 389–401, 2010.

[25] N. Kraiczy and N. Kraiczy, Innovation and firm performance, July. 2013.

[26] Vu, Nguyen, Ho, and Vuong, "Determinants of Vietnamese Listed Firm Performance: Competition, Wage, CEO, Firm Size, Age, and International Trade," J. Risk Financ. Manag., **12**(2), 62, Apr. 2019.

[27] J. S. Oakland, Total quality management and operational excellence: Text with cases. 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 991-996 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Matrix-based Minimal Cut Method and Applications to System Reliability

Emad Kareem Mutar*

Department of Mathematics, Directorate of Education Babylon, Babylon, 51005, Iraq

ARTICLE INFO

Article history:
Received: 23 August, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Reliability
Network model
Graph theory
Reliability method
Complex system
Optimal design

ABSTRACT

In this paper, we present a new method to deduce minimal cut sets depending on the minimal path sets of the complex systems (networks) to generate the Incidence Matrix, and then compared it with the truth table of the system. This comparison, based on some algebraic properties, gives minimal-cut sets of the complex network with an algorithm in Mathematica software. In addition, the minimal cut sets completely characterize the operating state of the system and equal to the complex system structural function information. So, the distinguish of the operational states of the system give us information about the binary operational states for some components. The system failure time is also given immediately if the failure times of the component parts are known.

1. Introduction

The main objective of a system reliability evaluation is to provide a certain probabilistic information on potential system component failures since a real design project needs to be evaluated in terms of reliability. A system can be defined as a stochastic graph G = (V, E) which contains the vertices set (nodes) and the edges set (links) of the system [1], [2]. The edges and vertices are usually prone to failures. In our system discussions, we claim that the edges go drop by design while the vertices function perfectly. The edge failures should be probabilistic [3]. A system can be described as working if a source vertex communicates with the vertex in the sink. In this case, it called reliability of two-terminal systems [4].

A cut set is a collection of components whose failure removes all the connections between the sources and the sinks of the system and thus leads to system failure, but a minimal cut set is a cut set that cannot be delivered in working condition without creating a path from the source to the sink, i.e., each minimal cut set causes system failure [5]. Conversely, the failure rate is not immediate or obvious [6].

In graph theory and computer science, a square matrix representing a finite graph is an adjacency matrix. The matrix elements indicate whether or not the vertical pairs are adjacent to the graph. The adjacent matrix for a finite simple graph is a (0,1) matrix with zeros at its diagonal [7]. The adjacent matrix is symmetrical if the graph is undirected (i.e. all its edges are bi-directional). The adjacent graph matrix should be separated from its incidence matrix [8]. We construct a set of minimal path set based on the matrix properties of the graph and then use the algebraic properties and multiplication to find the system's minimal cut set. Song and Kang [9], [10] presented a matrix-based system reliability approach which can be tested with simple matrix calculations for evaluating the reliability of complex system stats. Hassan and Mutar [11] have recently researched the architecture of electrical device reliability models (geometry point of view) used within spacecraft, which is considered a spacecraft's high-pressure oxygen supply system (HPOSS) [12]. With exception of current system reliability approaches, the complexity depends heavily on the details. The complexity of the formulas is not affected as the efficiency of the method is calculated with simple matrix calculations.

A connection matrix is a matrix with rows as minimal path sets and columns representing system components, then compared it to the system truth matrix. Based on some algebraic properties, this relation results in minimal cutting sets of the system. So the Matrix-based minimal cut structure of the system can be accomplished through algebraic manipulation of *the transpose of a* incidence matrix with the truth matrix . From that, we form a matrix in which the rows represent the minimal cut, and columns represent the components of the system then, we find the reliability of the system or obtain information about the system failure. Generally, this method is used to find the reliability of systems by using the *Mathematica* software.

*Corresponding Author: Emad Kareem Mutar, Email: emad77math@gmail.com

This paper presents a minimal cut method based on matrix and applies it to a simple system and engineering structures (looking at as networks), as seen in Figure 1. The units may be linked as a complex configuration. In a spacecraft, that reflects a HPOSS, i.e. the HPOSS in the cabin by the array of controls and valves from a high-pressure oxygen tank. This system consists of several components connected in a complex form for quantifying. The Matrix-based minimal cut analysis uses the representation of the incidence matrix whose elements indicate whether vertex–edge pairs are important in evaluating the minimal path sets of the systems. Then the matrix of **minimal cut** formulation of a system can be obtained depending on the minimal path sets of the systems. Generally, for solving complex models, we give an algorithm for calculating all minimal cuts, we use the Mathematica program to implement. We also gives powerful ideas to understand the MTTF of complex systems [13].

2. Models of network reliability

A reliability block diagram of a system is a graph whose edges are the system components, while there exists a pair of nodes called terminal nodes, in the backup power supply diagram. This describes the functional relationship between the components, and indicates if there exists a path between the terminal nodes which contains only edges with functional components (making, consequently, the entire system functional; in the contrary case, this is non-functional). It shows the functional relationship between the components, and indicates there is a path between the terminal nodes that contains only edges with functional parts. Otherwise it is not functional. The graphical model represents the reliability structure of the system like series , parallel and complex structures.

The system S consists of n components $X_1, X_2, X_3, \dots, X_n$ which can only work or fail in one of two states. Then we can define the boolean (binary) indicator variables for each component or S system for these states. The system design and connection paths reflect the reliability structure of the system which might or might not be the system's functional block diagram. As a practical illustration of engineering systems, modules can then be attached to a complicated configuration as illustrated in Figure 1. is a spacecraft's high-pressure oxygen supply system (HPOSS) [11] [12].

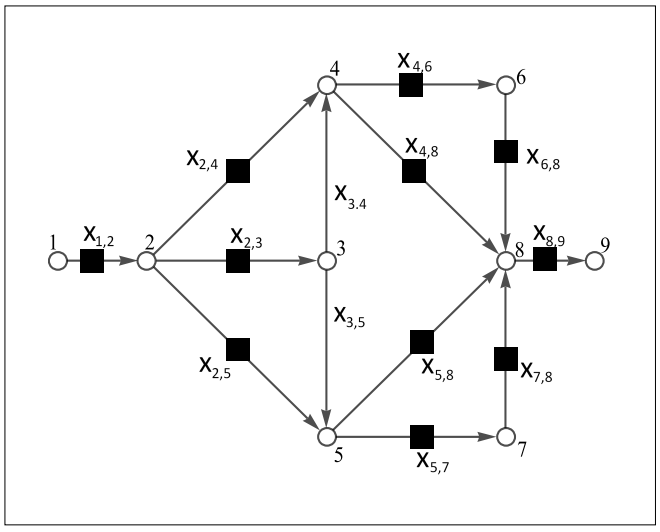


Figure 1: Reliability block diagram of system S.

As an illustration, we will study a simple model is a basic reliability block diagram with 3 subsystems as Figure 2, then we formulate a general Algorithm for complex models.

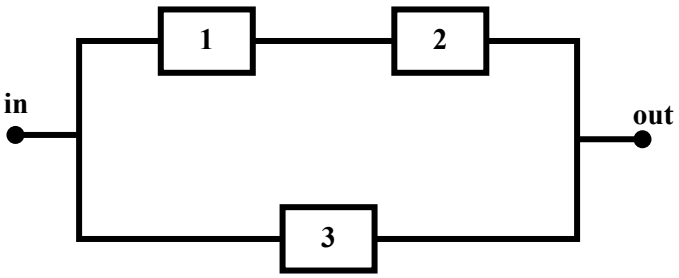


Figure 2: Reliability block diagram of system S.

3. Matrix-based minimal path deduction

The graph G = (V, E) represents the components' logical relations inside a piece of equipment. It is not necessarily the graphical representation of the system, but the system's functional components. In the graph theory, the minimal path is the problem of selecting a path from which no component can be removed without disconnecting the connections between the scour and sink node in a graph such that the sum of its constituent edges is minimized.

Consider a system that has different $\boldsymbol{X_i}$ states with its ith component, $i = 1, \dots, n$. The connection matrix (CN) is constructed in order to create the minimal paths by adding a (n × n) adjacent matrix of a simple graph with the identity matrix (unit matrix) in size n, to obtain 1's in the main diagonal, that is

$$\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Then we get the following matrix of connection

$$CN = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ a_{21} & 1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 1 \end{bmatrix} \quad (1)$$

where $\{1,2, \dots, n\}$ is the set of vertices, and $a_{ij} = (i, j)$ is the edge between vertex i and vertex j, if there is a connection between vertex i and vertex j then $a_{ij} = X_j$, otherwise $a_{ij} = 0$ .For example , in Figure 1, the simple model has the following connection matrix

$$CN = \begin{bmatrix} 1 & x_2 & x_3 \\ 0 & 1 & x_3 \\ 0 & 0 & 1 \end{bmatrix}$$

Therefore, we finding the minimal path between a two-terminal system as a special case of the paths problem, the minimal path vector for a system can be constructed by removes vertices which are either source or sink in the (CN) matrix, one at a time, until only the source vertex and sink vertex are present in the matrix [13]. When a vertex is deleted, the connection matrix inputs

are modified using the following equation for the remaining vertices:

$$a_{ij}^{1} = a_{ij} + a_{il}a_{lj}, \tag{2}$$

if vertex 1 is removed, where $i \neq j, i \neq l, j \neq l$ and $1 \leq i \leq n, 1 < j \leq n$ for $i = 1, 2, \dots, n$. Otherwise $a_{ij}^{1} = 1$ iff $i = j$.

The system shown in Figure 2 can be formulated as graph with directed edge (component) x_j of simple model is a basic RBD with 3 subsystems for i *and* j= 1,2,3. So the all minimal paths of the system are $\{X_2, X_3\}, \{X_3\}$.

Generally, for finding all minimal paths of series, parallel and complex system. A minimal path deduction algorithm to calculate all the minimal paths [14]. Based on the communication matrix and equation (1) by using the *Mathematica* software.

Algorithm 1: Minimal path deduction

```
mtrix = AdjacencyMatrix[
    G = Graph[Table[i, {i, 1, 9}], {1 ↔ 2, 2 ↔ 3, 1 ↔ 3}]];
n = VertexCount[G];
CN =
  MapIndexed[If[Equal @@ #2, 1, # Subscript[x, ## & @@ #2]] &,
     #, {2}] &@matr;
ClearAll[a]
a[m_, r_] := Module[{inds = DeleteCases[Range@Length@m, r]},
  Table[m[[i, j]] + If[i ≠ j, 1, 0] m[[i, r]] m[[r, j]],
   {i, inds}, {j, inds}]]
a[CN, 2];
ai = Nest[a[#, 2] &, CN, n - 2];
Mipaths = Expand[Part[ai, 1, 2]];
baseSet[Mipaths_] :=
 DeleteDuplicates[SortBy[MonomialList[Mipaths],
   {Head, Length}], Denominator[#2 / #1] == 1 &]
baseSet[Mipaths]
```

Applying the algorithm (1) to the complex model in Figure 1, all the minimal path sets can be obtained

$\{X_{1,2}\, X_{2,4}\, X_{4,8}\, X_{8,9}\}$, $\{X_{1,2}\, X_{2,5}\, X_{5,8}\, X_{8,9}\}$,

$\{X_{1,2}\, X_{2,4}\, X_{4,6}\, X_{6,8}\, X_{8,9}\}$, $\{X_{1,2}\, X_{2,3}\, X_{3,4}\, X_{4,8}\, X_{8,9}\}$,

The use of this approach is good practice by marking the source vertex as the first vertex and the sink vertex as the last vertex. Each intermediate node is deleted one by one until a 2×2 matrix is left with a_{12} represent summation of all minimal paths. Therefore, removing node i implies removing row i and column i from the original connection matrix where $2 < i < n - 1$ for a network with perfect n nodes [2].

4. Matrix-based minimal cut deduction

This method is based on minimal path mainly. The incidence matrix of minimal paths can be obtained from all minimal paths by use of a set of all minimal paths [3]. Assume that n minimal paths are defined by $P_1, P_2, \dots, P_n$. These paths will create a matrix while the paths represent the rows and the components represent the columns of the matrix. we determine the incidence matrix (IM) of all the minimal paths, of the form:

$$IM = \begin{array}{c} \\ P_1 \\ P_2 \\ \vdots \\ P_m \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \end{array} \tag{3}$$

where $a_{ij} \in \{0,1\}$ with $i = 1, 2, \dots, m$, $j = 1; 2, \dots, n$ and $a_{ij} = 1$ iff $X_j \in P_i$, otherwise $a_{ij} = 0$. Appling to the system illustrated in Figuer 2, we obtain incidence matrix as follows.

$$IM = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Based on the algebraic properties, we can form the truth matri x (TM) depending on the listing of a limited number of system s ates, which is the most common approach. The true matrix of the graph is an n-dimensional vector consisting of 1s and 0s in raws and the components represent the columns of the matrix. If the ith of a state is 1, component i shall be in that state; otherwise it shall be down, illustrates the system model in Figure 2 its sample space divided into $m = 2^3 = 8$. The TM matrix of the three component sates is

$$TM = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Then, we perform the multiplication process between *the transpose of a* incidence matrix (IM) and the truth matrix (TM) such that incidence matrix is $n \times m$ matrix and the truth matrix is $p \times m$ matrix, then A is an $p \times n$ matrix with entries, so the comparison matrix is formed as in the equation

$$A = TM.(IM)^T \tag{4}$$

The calculation in equation (4) can be done simply by matrix-based multiplication with improved efficiency. This is illustrated by an example in Figure 1 .

.

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 2 & 0 \\ 2 & 1 \end{bmatrix}$$

The process of this multiplication produces a new matrix that represents a fracture that helps to find parts and malfunctions in the system and the components that stop that we use the least property of Boolean algebra between the elements of the matrix rows according to the formula:

$$\boldsymbol{V} = \boldsymbol{min}\{\boldsymbol{a_{ij}} \,|\boldsymbol{j} \in \{\boldsymbol{1,2,3,\dots,n}\}\} \quad (5)$$

So for above example $v = (0,1,0,1,0,0,1)$ Whereas, this object represents the diameter elements in a matrix that create with, then we Then we make $V.I_n = diag(P_1, P_2, P_3, \dots, P_n)$ and remove the zero rows of the previous result. This is a job for logical indexing. Then we take

$$\boldsymbol{N(v)} = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{bmatrix} \quad (6)$$

The matrix above represents the process of intersecting minimal paths to find the cut sets , so we use the equation

$$CSM = N(v).IM \quad (7)$$

The CSM matrix include all cuts in rows and the components are arranged according to the columns in sequence , then the conditions of component failures are given by a single matrix.

$$CSM = \begin{matrix} \\ C_1 \\ C_2 \\ \vdots \\ C_m \end{matrix} \begin{matrix} \begin{matrix} x_1 & x_2 & \cdots & x_n \end{matrix} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \end{matrix} \quad (8)$$

Then remove any cuts that include other cuts to generate minimal cuts as following

$$MCS = min\{C_i \,|i \in \{1,2,3,\dots,m\}, \forall C_i \in CSM\} \quad (9)$$

The CSM matrix of the three component in above example in Fig. (2) is

$$CSM = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

For the edge x_j, the minimal cut is $\{x_1x_3, x_2x_3\}$.

Generally, for finding all minimal cuts of series, parallel and complex system. A minimal cut deduction algorithm to calculate all the minimal cuts [14]. Based on the CSM matrix and equation (9) by using the *Mathematica* software.

Algorithm 2: minimal cut deduction

```
M = rep@baseSet[Mipaths];
var = Variables[M];
M1 =
  var # & /@ CoefficientRules[M][[All, 1, 1]] /.
   x_i_,j_ → 1;
mat = M1;
sets = Tuples[{0, 1}, n1];
sets.Transpose[mat];
r = Pick[sets,
   Unitize[Min /@ (sets.Transpose[mat])], 1];
Sort[r.Array[Subscript[x, #] &, n1] /.
   Plus → Times];
b = DeleteDuplicates[r, #1 #2 == #1 &];
Micuts =
  Sort[b.Array[Subscript[x, #] &, n1] /.
    Plus → Times];
```

Applying the algorithm to the complex model in Figure 1, all minimal cuts can be obtained

$\{X_{1,2}\, X_{2,4}\, X_{4,8}\, X_{8,9}\}$, $\{X_{1,2}\, X_{2,5}\, X_{5,8}\, X_{8,9}\}$,
$\{X_{1,2}\, X_{2,4}\, X_{4,6}\, X_{6,8}\, X_{8,9}\}$, $\{X_{1,2}\, X_{2,3}\, X_{3,4}\, X_{4,8}\, X_{8,9}\}$,
$\{X_{1,2}\, X_{2,3}\, X_{3,5}\, X_{5,8}\, X_{8,9}\}$, $\{X_{1,2}\, X_{2,5}\, X_{5,7}\, X_{7,8}\, X_{8,9}\}$,
$\{X_{1,2}\, X_{2,3}\, X_{3,4}\, X_{4,6}\, X_{6,8}\, X_{8,9}\}$, $\{X_{1,2}\, X_{2,3}\, X_{3,5}\, X_{5,7}\, X_{7,8}\, X_{8,9}\}$

5. Minimal cat system reliability method

This method determines the reliability of a systems by using the minimal cut set and the inclusion-exclusion rule [15]. A cut Set is a set of components that fail to interrupt all the connections between source and sink, which causes device failure [16]. Then the system reliability can be written as

$$R_{System} = Pr\left(\prod_{i=1}^{n}(1 - C_i)\right) \quad (10)$$

where $C_i, i = 1,2,\dots,n$ represent all the minimal cuts, and $Pr(C_i) = Q_i$ where $Q_i = 1 - R_i$. Therefore, the probability of system failure is determined by a simple vector calculation

$$C_i = [X_1 \quad X_2 \quad \dots \quad X_n]$$

For example , in Figure 2, the system model has the following minimal cut vectors

$$C_1 = [0 \quad 1 \quad 1]$$
$$C_2 = [1 \quad 0 \quad 1]$$

So the minimal cut are $\{X_1X_3, X_2X_3\}$. These Reliability System can be generated easily by using the *Mathematica* software. Using equation (10) subsequently, the Reliability System is obtained as

$$\begin{aligned} R_{System} &= Pr(1 - C_1) \times Pr(1 - C_2) \\ &= (1 - Q_1Q_3)(1 - Q_2Q_3) \\ &= (1 - (1 - R_1)(1 - R_3))(1 - (1 - R_2)(1 - R_3)) \\ &= R_3 + R_1R_2 - R_1R_2R_3 \end{aligned}$$

So, we get

$$R_{System} = R_3 + R_1R_2 - R_1R_2R_3 \quad (11)$$

In general, the Reliability System for the union of minimal cuts is obtained using Mathematica processes

Algorithm 3: Minimal cat system reliability method

```
Micuts = Sort[b.Array[Subscript[x, #] &, n1] /. Plus → Times];
Qs = 1 - Expand[Product[1 - i, {i, Micuts}]] /. x_i_^n → x_i;
Rs = Expand[1 - Qs /. x_i_ → 1 - x_i]
Rsystem = Rs /. x_i_ → R_i
```

The minimal cut method based on the matrix has Firstly, as a result of calculating the connection matrix independently from a system definition, the reliability of a system event is calculated, the complexity of system reliability evaluations is not related to the concept of system states. Secondly, the matrix-based method helps one to understand easily and solve the failure rate of the system. Third, if component failure probabilities or statistical dependency are incomplete, the matrix-based method helps one to achieve the

shortest possible limit for all general system states. Fourth, recent developments in matrix-based computer languages and software, including Mathematica software, have made it easier and more efficient to implement matrix calculations. Finally, in this paper, the examples deal with two-state systems. The proposed approach is applicable to general multi-state component systems.

6. The Mean Time to Failure (MTTF)

The transformations we use are called linearizing transformations, which researchers have used in the past to evaluate the values of constants to best match a collection of data. The exponential function is of the form $R(t) = \lambda e^{-\mu t}$, and the ln-transformation of R(t) matches it with $lnR(t) = \mu t + ln\,\lambda$, mapping an exponential to a linear expression [17]. In particular, this approach refers to circumstances where the rate of the quantity change is directly commensurate with the radioactive decay (see [18], p.483] for further details). The exponential function also represents a curve whose form mainly depends on the exponent 's constant.

- If a > 0, then R(t) increases as t increases,
- If a < 0, then R(t) decreases as t increases.

Because of the interpretation of the Mean Time to Failure, we obtain

$$R(t) = \lambda e^{-\mu t} \tag{12}$$

where $\mu \geq 0$ and $\lambda \geq 1 > 0$: The function (15) is decreasing and satisfies the Reliability conditions. Let us also consider the polynomial of reliability (11). We assume that exponential functions correspond to $R_1, R_2 R_3$, where $\mu_i \geq 0$ and $\lambda_i \geq 1 > 0$, for $i = 1,2,3$. Then we obtain

$$R(t) = \lambda_3 e^{-\mu_3 t} + \lambda_1\lambda_2 e^{-(\mu_1+\mu_2)t} - \lambda_1\lambda_2\lambda_3 e^{-(\mu_1+\mu_2+\mu_3)t} \tag{13}$$

By using the above equation and mimicking the MTTF definition, we get

$$\begin{aligned}\text{MTTF} = \int_0^{\infty} & \boldsymbol{\lambda_3 e^{-\mu_3 t} + \lambda_1\lambda_2 e^{-(\mu_1+\mu_2)t}} \\ & \boldsymbol{- \lambda_1\lambda_2\lambda_3 e^{-(\mu_1+\mu_2+\mu_3)t}}\, dt \\ = \boldsymbol{\lambda_3} & \boldsymbol{\lim_{n\to\infty}} \int_0^n \boldsymbol{e^{-\mu_3 t}\, dt} + \boldsymbol{\lambda_1\lambda_2} \boldsymbol{\lim_{n\to\infty}} \int_0^n \boldsymbol{e^{-(\mu_1+\mu_2)t} dt} \\ & \boldsymbol{- \lambda_1\lambda_2\lambda_3} \boldsymbol{\lim_{n\to\infty}} \int_0^n \boldsymbol{e^{-(\mu_1+\mu_2+\mu_3)t} dt}\end{aligned}$$

Therefore, the integration leads to

$$MTTF = \frac{\lambda_3}{\mu_3} + \frac{\lambda_1\lambda_2}{\mu_1\mu_2} - \frac{\lambda_1\lambda_2\lambda_3}{\mu_1\mu_2\mu_3}$$

The above relationship through which it is possible to deduce the time for the failure of the system and then relay the components to give the longest time as a system.

7. The failure rate of the system reliability

In order to evaluate the system failure rates for the reliability system (11) used as a simple example with constant failure rate components. Consider the equation (12) of one component; the reliability of the component selected is given

$$R_c(t) = \lambda e^{-\mu t}$$

The failure rate is defined as the relationship between the density of probability and the reliability function

$$h(t) = \frac{f(t)}{R(t)} = \frac{-\frac{dR(t)}{dt}}{R(t)} \tag{14}$$

Then, by equation (14), we get

$$R_c(t) = \frac{\mu\lambda e^{-\mu t}}{\lambda e^{-\mu t}} = \mu$$

where a is constant. We note the failure rate of the component is constant. Generally speaking, we get the equation (13) from the equation (14), we obtain

$$h(t) = \frac{\boldsymbol{\lambda_3\mu_3 e^{-\mu_3 t} + \lambda_1\lambda_2(\mu_1+\mu_2)e^{-(\mu_1+\mu_2)t} - \lambda_1\lambda_2\lambda_3(\mu_1+\mu_2+\mu_3)e^{-(\mu_1+\mu_2+\mu_3)t}}}{\boldsymbol{\lambda_3 e^{-\mu_3 t} + \lambda_1\lambda_2 e^{-(\mu_1+\mu_2)t} - \lambda_1\lambda_2\lambda_3 e^{-(\mu_1+\mu_2+\mu_3)t}}}$$

The failure rate of the system for the failure rate of components in the complex system as constants is therefore time-related [19]. We find that if the component failure rate is constant, the system failure rate is also constant. In other words, the system's failure rate is always equal to the stable state failure rate [20].This is thus an correct representation of the state of polynomial multivariate reliability based on the values of the constants .

A practical example of engineering systems has been used with the proposed algorithm, components can be connected in a complex system, as shown in Figure 1. Including 9 nodes and 9 edges. The patterns representing system states and component problems can be defined effectively using matrix-based programming languages and applications. In addition, we obtain the failure rate equation, which results in an incremental failure of one variable being reduce by improving the pairing component, and which, obviously, provides a mathematical basis for the failure rate. This result demonstrates that our approach increases accuracy without increasing the assessment time.

8. Conclusions

This paper clearly defines a new method of deducing minimal cuts according to minimal path sets of complex systems, using the minimal cuts approach in a simplified form. The Matrix-based minimal cat approach can calculate the Reliability of complex system states by simply calculating matrix. Using matrix-based programming languages and applications, the matrices representing the system states and the component problems can be effectively defined. The components (the system) also proven to be exponentially time dependent; furthermore, we achieve a failure rate equation indicating that a gradual failure of a component can be balanced by the improvement of the pairing

component, and naturally gives a mathematical foundation for the failure rate. When the states of component failure are dependent largely. In some instances, we find a state failure rate and display certain components to be the efficient elements responsible for the device failure.

References

[1] Y.H. Kim, K.E. Case, P.M. Ghare, "A method for computing complex system reliability," IEEE Transactions on Reliability, **21**(4), 215–219, 1972, doi:10.1109/TR.1972.5215997.

[2] W. Kuo, M.J. Zuo, Optimal reliability modeling: principles and applications, John Wiley & Sons, 2003.

[3] J.-L. Marichal, "Structure functions and minimal path sets," IEEE Transactions on Reliability, **65**(2), 763–768, 2016, doi:10.1109/TR. 2015.2513017.

[4] S. Kharbash, W. Wang, "Computing two-terminal reliability in mobile ad hoc networks," in 2007 IEEE wireless communications and networking conference, IEEE: 2831–2836, 2007, doi:10.1109/WCNC. 2007.525.

[5] R.N. Allan, Reliability evaluation of power systems, Springer Science & Business Media, 2013.

[6] S.M. Alghamdi, D.F. Percy, "Reliability equivalence factors for a series–parallel system of components with exponentiated Weibull lifetimes," IMA Journal of Management Mathematics, **28**(3), 339–358, 2017, doi:10.1093/imaman/dpv001.

[7] G.T. Gilbert, "Positive definite matrices and Sylvester's criterion," The American Mathematical Monthly, **98**(1), 44–46, 1991, doi:10.1080/000 29890.1991.11995702.

[8] M.B. Javanbarg, C. Scawthorn, J. Kiyono, Y. Ono, Minimal path sets seismic reliability evaluation of lifeline networks with link and node failures, 1–12, 2009, doi:10.1061/41050%28357%29105.

[9] W.-H. Kang, J. Song, P. Gardoni, "Matrix-based system reliability method and applications to bridge networks," Reliability Engineering & System Safety, **93**(11), 1584–1593, 2008, doi:10.1016/j.ress.2008. 02.011.

[10] W.-H. Kang, Y.-J. Lee, J. Song, B. Gencturk, "Further development of matrix-based system reliability method and applications to structural systems," Structure and Infrastructure Engineering, **8**(5), 441–457, 2012, doi:10.1080/15732479.2010.539060.

[11] Z.A.H. Hassan, E.K. Muter, "Geometry of reliability models of electrical system used inside spacecraft," in 2017 Second Al-Sadiq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), IEEE: 301–306, 2017, doi:10.1109/AIC-MITCSA.2017.8722980.

[12] K.K. Aggarwal, Reliability engineering, Springer Science & Business Media, 2012.

[13] Y. Kobayashi, M. Hayashi, "A new approximation method to evaluate system reliability," in 2017 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), IEEE: 1–2, 2017, doi:10. 1109/LANMAN.2017.7972141.

[14] S.S. Skiena, The algorithm design manual: Text, Springer Science & Business Media, 1998.

[15] K. Dohmen, "Inclusion-exclusion and network reliability," The Electronic Journal of Combinatorics, R36–R36, 1998, doi:10.37236/ 1374.

[16] J.E. Biegel, "Determination of tie sets and cut sets for a system without feedback," IEEE Transactions on Reliability, **26**(1), 39–42, 1977, doi:10.1109/TR.1977.5215071.

[17] Z.A.H. Hassan, C. Udrişte, "Equivalent reliability polynomials modeling EAS and their geometries," Annals of West University of Timisoara-Mathematics and Computer Science, **53**(1), 177–195, 2015, doi:10.1515/awutm-2015-0010.

[18] T. Pyzdek, P.A. Keller, Quality engineering handbook, CRC Press, 2003.

[19] L. Xu, Y. Chen, F. Briand, F. Zhou, M. Givanni, "Reliability measurement for multistate manufacturing systems with failure interaction," Procedia CIRP, **63**, 242–247, 2017, doi:10.1016/j.procir. 2017.03.124.

[20] H. Pham, "Commentary: Steady-state series-system availability," IEEE Transactions on Reliability, **52**(2), 146–147, 2003, doi:10.1109/TR. 2003.811164.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 997-1006 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Renewable Electric Power from the Equine Treadmill: An Evaluation of the Potential

Faizan Dastgeer*, Hasan Erteza Gelani

Department of Electrical Engineering, University of Engineering and Technology, Lahore, FSD Campus, Faisalabad, 38000, Pakistan

ARTICLE INFO

Article history:
Received: 23 August, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Aerobic Power
Anaerobic Power
Distributed Generation
Efficiency
Equine Exercise
Horse

ABSTRACT

Horse power – the prime mover that has been there with humans for ages; chiefly used for transportation in the early days, and later, also used as an energy source leading to the conception of the term – horse power (hp). The current paper presents an interdisciplinary effort that brings forward an approach for evaluation of the potential of renewable power extractable from this prime mover once again. Specifically, the focus is towards extracting power from the equine treadmills (dry-type) whereby the incline of the machine is replaced by an equivalent (or pseudo-equivalent) energy generation mechanism leading to the coupling of renewable power with an equine workout. This approach comprises aerobic power (evaluated from oxygen uptake data) as well as anaerobic power (evaluated from plasma lactate data) – each of which is estimated from the difference values between equine running on a flat treadmill and when the mill is inclined. Furthermore, a literature review mentioning different inventions as well as some research efforts directed towards tapping somatic energy of animals is included in the manuscript. Also, a section dedicated towards assumptions/weak points helps judge the applicability of the presented work.

1. Introduction

For centuries, humans have benefitted from animals in one way or the other. Animals such as camels, cows and goats have provided us with milk as well as meat, hides (leather) have been used for a variety of purposes, and in the transportation department – they have been there, whether it was a draft work such as moving heavy loads by oxen carts or it was a high speed transit benefitting from the pace of Equus Caballus i.e. the horse. Despite living in close proximity for ages, the dawn of modern lifestyle has parted animals from their human beneficiaries, especially for the city dwellers who can obtain milk and meat from the market; besides getting access to a variety of honey, eggs and even leather and wool products in a superstore. As regards the transport, it has been taken up largely by fossil-fuel based automobiles. So, has this separation been for the good, or have we lost something that we shouldn't have – this may be deemed a debatable issue. Reference [1] mentions in its abstract that the positive effects of human-animal interaction have been researched by various studies and it has been shown that animal-assisted activities can be used to improve the physical and mental health of humans. The authors of [1] present a study concerning effect of equine-assisted activities (EAA) on prosocial behavior of adolescents and conclude that '… in this sense EAA has a protective effect on the behavior of adolescents'. Reference [2] identifies equine grazing, domestic biodiversity, land use, tourism and equine work as five 'green assets' of equines and mentions a case (with its certain conditions) which showed that for the same amount of carbon emissions, a donkey could prepare larger agricultural land area than a machine. The authors of [3] present a study of effects of horse riding upon children and deduce that riding a half-breed horse or a pony had a positive effect on causing the go- and no-go (self-control) reaction on children and it even improved solving of arithmetic problems.

Orienting towards the topic of the current paper, the horse in particular, has been serving humankind for ages. It offers the highest sprint speed (surpassing other rideable animals such as camels and steers) coupled with human-friendliness as well as ease of keeping. However, as the 20th century witnessed the spread of automotive industry, the horse can now be found in equestrian sports and leisure riding (exceptions may include draft work). One of the equipment used for equine exercise as well as research is the treadmill. Furthermore, some of the treadmills can have an incline function where the mill is tilted (w.r.t. ground) to increase the intensity of equine workout without reaching maximal speeds. The current paper makes the argument: why not keep the mill flat while extracting an equivalent amount of energy electrically, thus maintaining the exercise while generating clean renewable energy. To this end, the paper provides an approach for evaluating the amount of power that can be generated by an average horse.

The idea of energy generation from the horse via an equivalent (-to-incline) treadmill is unorthodox as well as interdisciplinary – involving Electrical Engineering (Power Engineering and Power

*Corresponding Author: Faizan Dastgeer, Email: faizandastgeer@uet.edu.pk

Electronics) as well as Equine Science (may include bio-mechanics as well as Equitation Science). The subsequent literature review mentions various inventions and research papers directed towards generation of energy from animate prime-movers.

2. Literature Review

This literature review is neither intended to define the scope of various inventions/research efforts nor may it be considered as a compact summary encompassing all the inner details of the inventions/research efforts.

Energy capture via treadmills is a fascinating concept with a fair number of research efforts. A US patent is presented in [4] which mentions that the invention provides a system and method for electricity generation via using animals like horses. The invention is different from treadmills; over here, the animal moves on trolleys which in turn move on a track. An elaborate system of gears, springs and levers converts the biological energy into rotational energy for an electric generator. The inventor says that this device is superior to a treadmill which only uses a half of the force of movement. He further goes on to say that it is envisioned that a treadmill like free running environment may be developed without departing from the invention. Also, it has been mentioned that the specific intention for the device is for rural areas with limited or non-existent electricity generation.

An exercise machine is presented in the patent [5], for large animals such as dairy cows. The invention is a treadmill type device and includes hoof abrasion system as well as a treadmill cleaning or flushing system. The inventor argues that as exercise gives a lot of benefit to humans, in a similar way it can be beneficial to dairy cows. A frame supports an endless track, upon which the animal will walk. One embodiment of the device includes a power generation system, wherein the walking energy of cow powers an electric generator. Energy may be stored or fed to the grid.

Continuing the treadmills, the patent [6] presents a treadmill for humans which can generate electric power. According to the inventor, a manually operated treadmill that can generate power for itself or for other electrical equipment may be beneficial for various environmental and practical reasons. In an exemplary embodiment the kinetic energy (KE) of the treadmill which the user imparted to the running belt is converted to electrical power via the power generation system.

Furthermore, the patents [7–10] also mention inventions using animate prime movers for electricity generation. The patent [7] mentions a mechanical platform comprising five stages used to convert energy from animals such as camels and horses into electrical energy. The multiple stages include a rotating base, air compression mechanism, a steam device and a dynamo for electricity generation.

Movement of livestock for energy production has been used by [8]. The patent presents equiangular radial levers which will generate moment of force when driven by livestock. Eventually, KE will be converted to electrical energy. The patent [9] uses a treadmill for harnessing animal power while the animal is walking on the belt and its head is in the feed end. It also gives the idea of getting more work from a powerful animal and vice versa; via a computer system developed for this. The patent [10] harnesses equine power where horses move in a circular path driving a central shaft via gears which drives a synchronous generator.

Looking at research publications, Fuller and Aye [11] are of the view that human and animal somatic energy is a forgotten source of renewable power. The paper mentions that on a global scale, the energy (authors are not exactly focusing on electrical energy) contributed by humans and animals is estimated to be twice that of the wind power. In the authors' words, "the paper makes the case that human and animal power be seen as renewable sources of energy"; the authors further say that the techniques for using these energy sources should be deemed "as part of the 'renewables' family." Posing the question why human and animal powers are not included in well-known renewable energy forms, the authors answer that several explanations are possible; these include human/animal powered technologies not being very fashionable and lack of support from big companies. Despite the paper not being targeted towards generation of electrical energy and being largely discussing rural habitat – it has been mentioned that the relevance of such technologies for developed countries should not be overlooked.

The authors of reference [12] give the idea of using animal draft power as well as biogas for electricity generation for rural areas. Animals are attached to rods connected to a circular plate which is connected to a chain sprocket system which transmits power to an electrical generator via a gear box. Estimated energy output (without including electrical component losses) is 18kWh/day from twelve oxen.

Energy generation via human treadmills is presented in [13–15]. Both electrical energy (generated via permanent magnet DC machine) and mechanical energy (used by a special washing machine) have been exploited in [13]. The system is not grid connected and detailed measures for maintaining (nearly constant) treadmill speed have not been mentioned. An elaborate work presenting hardware setup for a grid connected treadmill is mentioned in [14], where an infra-red sensor senses user position on the treadmill; consequently a three-phase permanent magnet synchronous machine connected to the belt is operated as either a motor (if the user leads reference) or as a generator. The synchronous machine receives/supplies power from/to a DC bus via a three phase AC/DC converter which operates the machine via field-oriented control. The DC bus is connected to the grid via a single phase bi-directional converter operated via single phased d-q frame. In [15], the authors present another human treadmill where the generated energy may be used for charging a battery or for a low voltage CFL.

In short, this section has mentioned efforts of various authors and inventors who have worked towards the idea of harnessing the energy of animate prime movers. However, the specific idea of generating electric energy from flattening the inclined dry-type treadmill for an exercising horse is rather unorthodox and thus could not be covered in literature review – although, instances of treadmills where horses walk to generate mechanical power (e.g. for wood splitting) may be found from sources other than published research literature.

3. Work, Metabolic Cost & Horse Power

3.1. Work

Encyclopedia Britannica defines work, in physics, as "measure of energy transfer that occurs when an object is moved over a distance by an external force at least part of which is applied in the direction of the displacement." Work is the product of force and the displacement: hence, if the displacement is opposite to force,

the work is negative. For example, in humans, if the triceps brachii is used for a fast accelerated extension of the fore-arm along the horizontal plane i.e. parallel to the ground surface – the muscle performs positive work, however, before reaching the straightening of the arm (i.e. the max extended position), the biceps brachii naturally exerts a decelerating force and slows down the arm – thus performing negative work. Another way to look at this is that the triceps accelerated the forearm, thus imparting KE; while the biceps decelerated the forearm thus extracting that kinetic energy. Consequently, the net work, performed by the body upon the forearm limb is zero (another way to comprehend net zero work is that the limb returns to same energy state KE = 0, Potential Energy (PE) = as before). However, internally both the muscles consumed metabolic energy (explained subsequently) of the body while applying their respective contractive forces.

3.2. Internal Work

Work performed upon the limbs w.r.t. the Center of Mass (COM) of the body may be regarded as internal work e.g. the protraction phase of an equine forelimb during a gallop, when the limb is moving forward (being protruded) may easily be comprehended as a case of internal work.

3.3. External Work

This is the work performed by the limbs to move (upward and/or forward) the COM of the body e.g. in the stance phase (hoof touching the ground) of an equine gallop, the propulsive part which accelerates the body, is performing external work.

As a summary, the stance phase includes external works – negative and then positive, when the COM is braking and then accelerating respectively, while the swing phase (hoof not touching ground) includes internal work. On a level locomotory activity such as walking, at a constant speed, for horses or humans, the *net* external work performed by the body is zero (energy expended against aerodynamic drag is not included in this). Alternative argument for the zero work concept is that the body is undergoing no *net* change in its KE or PE per stride. Reference [16] mentions this, that in level walking or running, the net work per stride is zero.

However, the body is still expending metabolic energy (discussed subsequently) which is being used up in

i. Internal work
ii. Within stride External work – this includes the decelerating and accelerating work per stride. Reference [17] gives graphs of within-stride KE variation for equine walk, trot and gallop.

3.4. Metabolic Cost

Metabolic cost for a physical activity refers to the energy expended by metabolic reactions within the body for performing that activity. For locomotion, the metabolic cost becomes Joules expended per meter of covered distance i.e. J/m; on per kg of body mass it is augmented to Jkg-1m-1. The portion of released energy which is converted to mechanical work decides the efficiency of the muscles; [18] quotes this efficiency to be in the range of 25 – 30% and mentions it to be a product of two efficiencies

- Phosphorylative coupling efficiency (around 60%)
- Efficiency of contractive coupling (around 49%)

For equine performance, [19] mentions that the energetic cost (expressed as kJ/min) is essentially linear with speed, with the condition that the animal can change gait freely. A relationship on per kg basis is provided as well, this relationship may be given as

$$Energy\ (kcal/min) = 0.0289 + 0.00074 * speed(m/min) \quad (1)$$

The author of [19] further mentions that Adenosine Triphosphate (ATP), the immediate source of energy in muscles, is quite limited (6mmol/kg of muscle mass) – it has to be replenished for muscular contractions lasting more than a second or two. The aerobic and anaerobic processes (utilizing and not utilizing oxygen respectively) are the pathways for the release of energy in the body leading to the replenishment of ATP for muscular contractions. The intensity of both the pathways can respectively be judged via

- Oxygen uptake - VO_2 (also symbolized as $\dot{V}O_2$), which is the volume of oxygen consumed given as ml.kg-1.min-1.
- Blood lactacte concentration mmol/L

3.5. Horse Power

How much power can a horse generate? 1horse-power i.e. 746 Watts? Actually, the authors of [20] say that one horse (600kg mass and assumption of 100 W/kg of muscle mass) could in theory give 24hp; it has been further mentioned that realistic estimate of peak performance is around 12 – 14.9hp. The term horse power was coined by James Watt when he was commercializing his steam engine. Watt observed that during a day's work, a horse would expend a power of 33,000 ft-lbf/min [20], equivalent to 745.7Nm/s. This unit came to be known as horse power. As this power value was intended to show a replacement potential of horses by steam engines, hence it is not an indicator of peak equine potential; rather a through-out-the-day power which a horse could deliver.

4. Treadmills

The first equine treadmill dates back to 1960 as mentioned in [21] and the initial studies were focused on metabolic variables. Treadmills were subsequently adopted in kinematic studies as well as for equine exercising. The incline function of the treadmills can increase the exercise intensity, consequently a maximal workout may be achieved at sub-maximal speeds. More recently, water treadmills have been introduced which can increase exercise intensity due to drag of water.

4.1. Scope of the current paper

The scope of the current paper is towards evaluating the potential of renewable electrical power generate-able via the equine dry type treadmill. If the treadmill is kept flat, and an adaptive/controllable energy generation unit is coupled with it that may be grid-tied or connected to a battery bank – the exercise/study intensity may be adjusted equivalent to the value that the inclined mill offered, thus producing clean renewable energy at the same time. The scope of the current paper is different from many citations quoted in literature review; over here, rather than being the primary objective, energy production may be regarded as a secondary objective, wherein the chief purpose may be the exercising of the horse.

4.2. Direct and Indirect Assessment Methods – why mgv_v is not used.

Reference [22] examines speed of racehorses on different gradients on an undulating racecourse. The authors say that the

race horses reduce speeds on inclines and the speed detriment corresponds to trade-off between metabolic costs of height gain and horizontal galloping. The paper mentions that the power required to move up an incline may be expressed as

$$power = mgv_v \; where \; v_v = v_h.s \qquad (2)$$

m and g are the body mass and gravity, v_v and v_h are the vertical and horizontal components of velocity and s is the slope of the gradient. The paper calculates that moving up a 10% incline at 8m/s i.e. at v_v = 0.8m/s equates to 8W/kg.

In the current effort, an indirect assessment of the power generatable via equine treadmill is presented. This indirect approach uses ΔVO_2 and Δlactate values (between flat and incline) for the power assessment – details of the methodology are presented in the sections 5 and 6. A direct assessment of the power actually being expended on the vertical mill, via the use of (2) has not been performed (although there are research efforts which use this approach for the treadmills). We refrain from using it because we are of the view that this approach is not exactly applicable to treadmills where the center of mass (COM) may not move absolutely the same as its movement on an actual uphill incline. A subsequent kinematic study or a prototype demonstration may be able to verify or disprove the results of this paper – in light of such results, the currently presented approach may be reworked to look for a discrepancy and readjusted as per need.

For use of ΔVO_2, we present the case of [23] where an uphill/downhill (incline of ±8%) treadmill work (for humans) was simulated on a flat treadmill, via the use of a horizontal force applied to the waist of the runner (speed 2.5m/s). For the uphill case, despite the Pearson correlation coefficient given to be 0.42 and the authors regarding the VO2 correlation between the two methods (uphill simulation on flat treadmill and actual inclined treadmill) to be not significant; still, the mean VO_2-net (ml.kg-1.min-1) values for simulated incline and actual treadmill incline running were fairly close. These values were reported to be (Mean±SD) equal to 38.6±2.19 and 37.2±4.25 for simulated and actual treadmill incline respectively. An observation of the graphical results presented in the paper shows that, out of the eleven humans tested, three gave VO_2 results for the simulated method to be very close to the actual inclined treadmill, and the remaining eight were distributed so that six gave higher VO_2 for the simulated method while two gave otherwise. For the current paper, the fair proximity in the average VO_2-net values for simulated and actual treadmill incline as mentioned earlier, may be regarded as a basis for the provided approach of evaluating the equine renewable power potential from an average horse.

Moreover, it is worth mentioning that an alteration in the ground reaction forces (GRFs) is also expected as the treadmill is changed from inclined to horizontal while maintaining exercise intensity. We deem such an alteration to GRFs beyond the scope of the current paper. However, we do like to mention that the GRFs on an inclined treadmill are already not the same as those on a flat track apparently; the change from inclined to horizontal, might just be in-favor-of, or benign-to, the exercising equine.

5. Evaluation of The Potential: The Aerobic Contribution

5.1. Evaluating the generatable renewable power

We present an indirect metabolic approach for evaluation of the power expended on inclined treadmills at different slopes. The authors of [24] have examined effect of treadmill incline on VO_2 and other parameters for five horses. The VO_2 has been plotted against treadmill speed for different inclines (reproduced here in Figure 1). Attempting to describe their data mathematically, the authors say that for 2.5% and 0% (i.e. flat), the VO_2 vs. speed relationship could be best defined by second order polynomials while for 5%, 7.5% and 10%, the relationship was described via linear regression. Table 1 gives various coefficient values for their relationships.

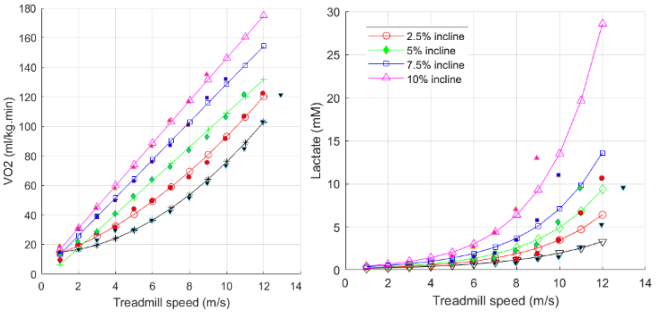
Table 1: Speed and incline relationship coefficients of [24]

Coefficients / Incline	x^2	x^1	x^0
0%	0.61	0.10	13.89
2.5%	0.43	4.08	9.29
5%	-	11.40	-5.11
7.5%	-	12.82	0.43
10%	-	14.38	2.30

'x' represents the horizontal axis quantity of speed in m/s. The y-axis quantity is VO_2 in $m,kg^{-1}min^{-1}$. Linear regression coefficient values are means only - non-inclusive of ± SE

Fig 1. plots the VO_2 and lactate concentrations (after taking antilog of $\log_{10}$ values) obtained from provided relationships as well as those obtained from the provided graphs (from [24], Figure 1A for VO_2 and Figure 4A for lactate conc. – mean values acquired via) using a MATLAB based graph digitizer program [25]. Some of the points for lactate concentration (e.g. lactate mM at 9m/s for 10% incline) with large variation between graph based and regression equation based values have been approximated to their graph digitized values. The current work takes the data of [24] and calculates the differences of oxygen uptake ΔVO_2 between each incline and the flat e.g. for the 2.5% incline $\Delta VO_{2\text{-}2.5\%}$ may be obtained as

$$\Delta VO_{2-2.5\%} = VO_{2-2.5\%} - VO_{2-0\%} \qquad (3)$$



(a) (b)

Figure 1: (a) VO_2 and (b) Plasma lactate concentrations vs treadmill speed – empty symbols are for values obtained from relationships, filled symbols are for values acquired from graph

Subsequently, the joule equivalent of oxygen ($O_{2\text{-}J\text{-}eqv}$) is taken to be 21.1 J/ml O_2 [26]. The extra power expended (not exactly expended – rather, this is the power extractable at 0% incline to produce a similar/same energetic response) for vertical motion at 2.5% incline denoted by $p_{mech\text{-}aer\text{-}2.5\%}$ may be expressed as

$$p_{mech-aer-2.5\%} = CRS_{CF} * \Delta VO_{2-2.5\%} * O_{2-J-eqv} * m_{equ} * \eta_{mscl} * (1/60) \qquad (4)$$

where η_{mscl} represents locomotor muscle efficiency, m_{equ} represents the equine mass, 1/60 converts from minutes to seconds and CRS_{CF} is the correction factor for Cardio-Respiratory system oxygen

demand explained subsequently. Furthermore, (5) provides a look at (4) in terms of dimensions (units)

$$Watts\ (J.s^{-1}) = ml.kg^{-1}.min^{-1} * J.ml^{-1} * kg * min.s^{-1} \quad (5)$$

Next we assume a value of 30% for η_{mscl} [17] and a 492 kg equine body mass (as in [24]). The factor CRS_{CF} is a rather crude approximation of the VO_2 used up by the heart and lungs. The web article [27] (although not exactly for the equine heart) provides values for the myocardial oxygen consumption (MVO_2) to be 8ml O_2 per minute per 100g, for the cardiac state of 'resting heart rate' while for the state of 'heavy exercise', the value is raised to 70ml O_2 per min. per 100g. Next we assume the lowest activity state of locomotion at 1m/s on the treadmill of 0% incline (i.e. flat mill) to be equivalent to the resting state for MVO_2. The corresponding VO_2 for this state equals to 8.8ml/kg.min (obtained via graph digitization). Furthermore, [28] presents mean heart weight of racing type (Thoroughbreds and Standardbreds) horses to be about 0.86% of bodyweight – hence, for the ongoing study mean heart mass is taken to be 0.86% of m_{equ} which equals about 4.2 kg. Consequently,

$$total\ MVO_2 = 4.2kg * 80ml.kg^{-1}.min^{-1} = 336\ ml.min^{-1} \quad (6)$$

and the MVO_2 as a percentage of total VO_2 roughly equals

$$\frac{total\ MVO_2}{total\ VO_2} = \frac{336\ ml.min^{-1}}{8.8 * 492\ ml.min^{-1}} \approx 0.078 \rightarrow 7.8\% \quad (7)$$

Performing similar calculations for the highest speed point (10m/s) for 7.5% treadmill slope (as it corresponds to the highest heart rate provided by [24]) and using the 'heavy exercise' cardiac state, total MVO_2 equals

$$total\ MVO_2 = 4.2kg * 700ml.kg^{-1}.min^{-1} = 2940\ ml.min^{-1} \quad (8)$$

and

$$\frac{total\ MVO_2}{total\ VO_2} = \frac{2940\ ml.min^{-1}}{132.2 * 492\ ml.min^{-1}} \approx 0.045 \rightarrow 4.5\% \quad (9)$$

subsequently, we approximate the total MVO_2 to be equal to 5% of the total MVO_2 being taken up for any given speed and slope. This is obviously a crude approximation, and a more detailed approach can involve use of multiple values for the *total MVO_2/total VO_2* ratio which can yield multiple values of CRS_{CF}.

Continuing towards the lungs, [29] mentions that the lung consumes 5% of whole body oxygen uptake – using this value as it is (although [29] is neither for equines nor for exercise) and combining it with the estimated 5% value of the total MVO_2, means that about 10% VO_2 is being consumed by the cardio-respiratory system for its own functionality – hence, 90% VO_2 is the value available to the rest of the body. The CRS_{CF} is then taken to be 90% (i.e. 0.9) thus showing that (only) 90% of the $\Delta VO_{2\text{-}2.5\%}$ in (4) is available for the evaluation of $p_{mech\text{-}aer\text{-}2.5\%}$.

Figure 2 shows plots of $p_{mech\text{-}aer\text{-}2.5\%}$ and the mechanical power values for other inclines. Also shown are the metabolic power values generated at different speeds and inclines. A relatively light exercise of 2.5% incline at 5m/s (3 – 6m/s has been regarded as most energy efficient speeds regardless of incline by [24]) is generating around 500 Watts of clean power – assuming a 30 minute exercise and an 80% efficiency of the power extraction and conversion system, 400W may be supplied to the grid for half an hour.

5.2. Sketching the metabolic power lines

Figure 2 also shows metabolic power $p_{met\text{-}aer}$ lines. These are sketched by looking for speed points of the same (chosen value of) metabolic power on curves of different inclines – this is achieved via a spline interpolation of speed vs. metabolic power (at different inclines) data - the metabolic power e.g $p_{met\text{-}aer\text{-}2.5\%}$ has been evaluated as

$$p_{met-aer-2.5\%} = VO_{2-2.5\%} * O_{2-J-eqv} * m_{equ} * (1/60) \quad (10)$$

Subsequently, corresponding to the speed point acquired and the $p_{met\text{-}aer}$ value in consideration, $p_{mech\text{-}aer}$ points are obtained for each incline. Connecting these points constitutes the linear joints of the $p_{met\text{-}aer}$ lines. Lastly, for the selected metabolic powers a spline extrapolation is performed using the speed points and corresponding $p_{mech\text{-}aer}$ points towards the upper end of the figure.

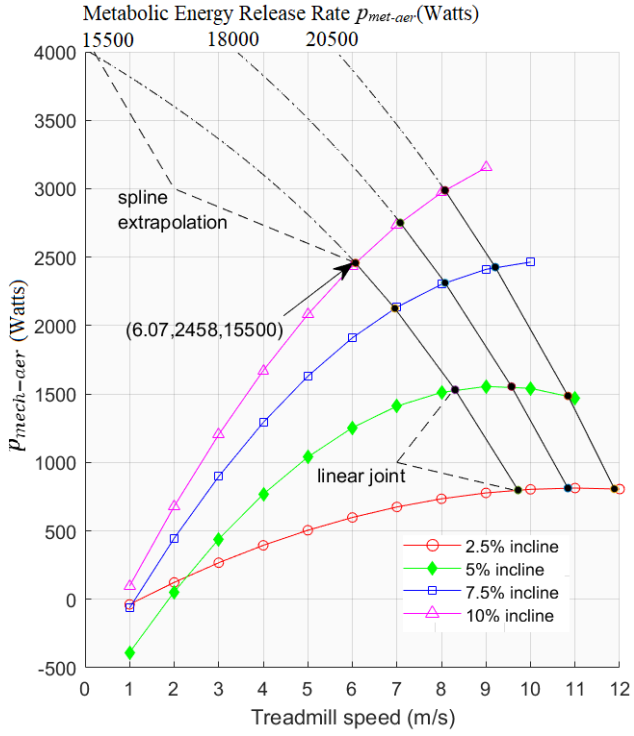


Figure 2: Power generation potential for various inclines. In general, the $p_{mech\text{-}aer}$ values beyond 3m/s may be deemed practical, while errors in speed points of 1 & 2m/s may be attributed to corresponding cluttered VO_2 data of Figure 1(a). The top horizontal axis shows metabolic power lines for different speeds and inclines. A single point (x_{bottom}, y, x_{top}) e.g. (6.07m/s, 2458W,15500W) is to be interpreted as release of 15500W metabolic energy where the horse is moving at 6.07m/s (on the incline of 10%) wherein the aerobic mechanical power generatable via flattening the mill is 2458W.

5.3. Power and Energy

For a discussion about energy, three points α, β and γ are taken from Figure 2; the coordinates are taken as (x_{bottom}, y, x_{top}) i.e. (speed, $p_{mech\text{-}aer}$, $p_{met\text{-}aer}$). These points are (approximately):

$\alpha \rightarrow$ (6.07m/s, 2458W,15500W) - 10 %incline
$\beta \rightarrow$ (8.08m/s, 2312W, 18000W) - 7.5%incline
$\gamma \rightarrow$ (8.08m/s, 2988W, 20500W) - 10 %incline

It may be seen that, while the operating point α can deliver mechanical power value of 2.46kW; the point β is delivering slightly lesser power of 2.31kW, while depleting metabolic energy at a higher rate of 18kW. Consequently, the potential of energy generation at α is higher than that of β. Furthermore, γ is showing even higher potential of delivering about 3 kW but the depletion rate of metabolic energy is also huge - a 20.5kW, hence a probable reduced time duration for maintaining a healthy workout at the point γ, may reduce its energy potential. A detailed analyses of

energy generation potentials for different equine breeds, exercise schedules for various gaits, treadmill slopes and speeds incorporating the data of on-set of fatigue/maximum exhaustion (e.g. [30] mentions Times to Fatigue of 45 and 20.2 minutes for an Arab and a Thoroughbred horse exercised at 4.5m/s, 6% incline) is deemed beyond the scope of the current which is aimed at presenting methodology for the evaluation of power that may be generated.

5.4. Point of optimum (aerobic) mechanical power generation

Looking for the point of optimum power generation potential, Figure 3 plots treadmill speed (including and beyond 3m/s) versus treadmill slope (the four inclines) versus $p_{mech\text{-}aer} : p_{met\text{-}aer}$ ratio at the different points.

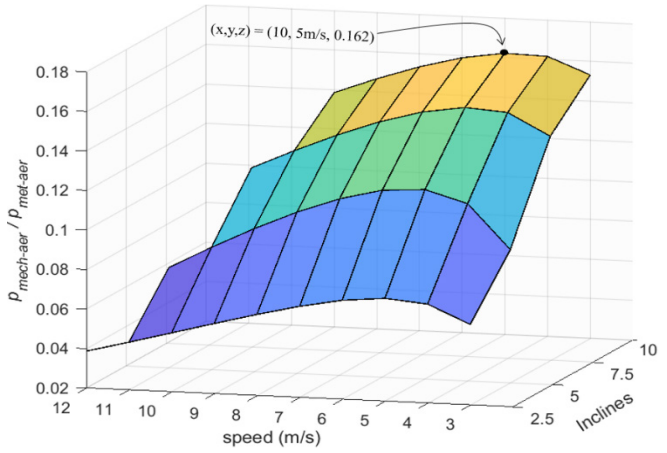


Figure 3: Treadmill speed vs. Treadmill Slope vs. $p_{mech\text{-}aer} : p_{met\text{-}aer}$ ratio. Speed points 1 and 2 m/s are ignored for reason mentioned in Figure2 caption.

The highest value of 0.162 for the $p_{mech\text{-}aer} : p_{met\text{-}aer}$ ratio was found to occur at the speed of 5m/s for the treadmill slope of 10%. With respect to the treadmill slope, Figure 3 shows a trend of increase in $p_{mech\text{-}aer} : p_{met\text{-}aer}$ ratio and w.r.t the speed, highest $p_{mech\text{-}aer} : p_{met\text{-}aer}$ ratio values are achieved around 5m/s. It may be noted that this is in line with the top values of economy of locomotion ($ml.kg^{-1}.m^{-1}$) occurring in range of 3 – 6 m/s as evaluated for the initial data (of the current study) by [24] - which further mentions high speed locomotion to be less efficient (for flat and 2.5% incline).

Lastly, towards the end of this section, it is worth mentioning that the methodology described here may be considered to be only an approach towards the evaluation of the potential. The values are not definitive – with variations arising from different factors such as equine body mass and precision in value of energy released per ml of O_2 etc. Reference [24] has itself mentioned that the regression coefficients for the speed and VO_2 relationship have range from 11.1 to 19.61 in different studies. It has been further mentioned that the variations in studies may be due to differences in methodology including different factors such as horse breed, bodyweight as well as treadmill acclimation.

6. The Anaerobic Contribution

This section provides an approach to assess the anaerobic (lactic) contribution to energy and power for equine exercise on inclined treadmill. Once again, the data of [24] has been used which has provided plasma lactate concentration graphs as well as linear regression log10 lactate concentration versus speed relationships. Mean values have been used excluding the standard deviations – the same was the case for aerobic approach. The difference in plasma lactate concentration between each incline and the flat treadmill have been evaluated. These differential lactate values represent the additional exertion expended due to the incline of the treadmill and are converted to corresponding delta (Δ i.e. difference) metabolic energetic approximations as expressed in (11) for the 2.5% incline. Differing from the approach of aerobic contribution where a portion of the VO2 was given to the cardiorespiratory system, over here, lactate value are assumed to be a result of the activity of exercising muscles only.

$$e_{met-\Delta-2.5\%} = \Delta_{lact-2.5\%} * lact_{p\rightarrow b} * lact_{O2-eqv} * O_{2-J-eqv} * m_{equ} \quad (11)$$

where,

- $e_{met\text{-}\Delta\text{-}2.5\%}$ is the Δ metabolic energy for the anaerobic pathway for the 2.5% incline
- $\Delta_{lact\text{-}2.5\%}$ is the difference in lactate concentrations between the 2.5% incline and flat treadmill – units of Δ_{lact} are milli moles per litre (the units may be symbolized as mM)
- $lact_{p\rightarrow b} = 1.5^{-1}$ is a factor to represent a 1.5 times reduced concentration of blood lactate as compared to plasma lactate as mentioned by [31]
- $lact_{O2\text{-}eqv}$ is $3ml O_2$/mmol lactate (per litre blood) – it is the oxygen equivalent of blood lactate as described in [26]
- $O_{2\text{-}J\text{-}eqv}$ is the Joule equivalent of O_2 equal to 21.1 J/ml O_2 as used by [26]; a slight variation of 20.92 J/ml O_2 has been used by [32]
- m_{equ} is the mean equine mass of 492 kg as in [24]

Subsequently, (12) shows calculation of the corresponding mechanical power generation $p_{mech\text{-}anaer}$ which is evaluated by taking difference of two consecutive energy points $\Delta e_{met\text{-}\Delta}$ and dividing this by the two minutes of interval length for lactate measurement as mentioned by [24] (which mentions blood sample collection during last 15s of each interval) – also, as for the case of aerobic power, the metabolic to mechanical conversion is achieved assuming η_{mscl} i.e. muscular mechanical efficiency of 30% as in case of aerobic power. Figure 4 presents the results of $p_{mech\text{-}anaer}$ for the different inclines and speed points.

$$p_{mec-anaer-2.5\%} = \Delta e_{met-\Delta} J/120s * \eta_{mscl} \quad (12)$$

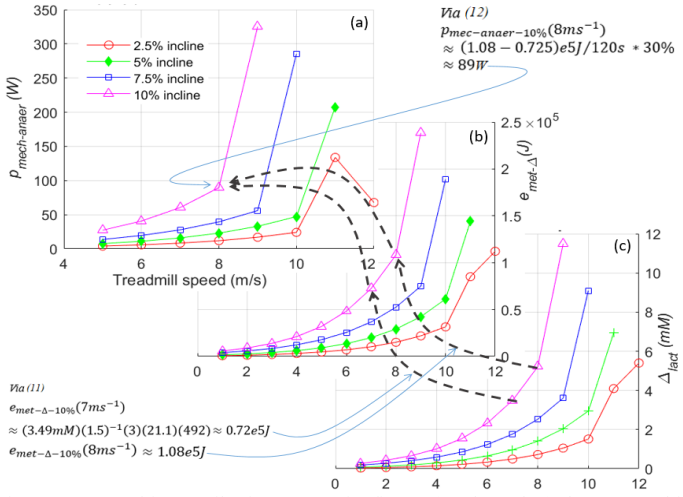


Figure 4: Anaerobic contribution composite figure. (a), (b) and (c) show anaerobic mechanical power, metabolic energy (y-axis in 10e5 Joule) and the differential lactate concentration values respectively. The horizontal axis shows treadmill speed for each subfigure. A single point of (a) e.g. (8m/s, 89W) for the 10% incline

indicates that, 89W of power generation potential is available corresponding to this point and metabolically, this power will be generated via anaerobic pathway.

Interestingly, Figure4 shows that for 12m/s at 2.5% slope, the power potential is actually reduced and this may be attributable to the fact that at this speed, the lactate threshold for the flat mill has also been crossed (as may be seen from Figure 1b) – consequently, the differential anaerobic energy is reduced and hence the power graph shows a dip. Furthermore, anaerobic alactic contributions could not be evaluated in this study because of in-availability of corresponding data in [24], however, these may be deemed minor - [32] mentions the alactic energy contributions to be around 0.1% for supramaximal exercises in their study.

7. Including the Drag

Aerodynamic drag becomes considerable as speed increases. [33] calculates this drag via $½C_D\rho Av^2/body\text{-}mass$ (with $C_D = 0.9$, $\rho = 1.29$ kg/m^3, $A = 1$m^2, $v = 12$m/s and $mass = 550$kg) to be equal to 0.15 Jkg^{-1}m^{-1}. Converting this to power translates into 990 J/s i.e. about 1kW. It has been mentioned by [33] that, while the drag is often considered negligible, it is actually a larger proportion of the external mechanical work compared to what was considered before. So, a horse galloping on the track at 12m/s is expending about 1kW of external mechanical work for countering the aerodynamic drag forces. Consequently, as the same horse is exercising at 12m/s on a high speed treadmill, extracting 1kW of power can make the equine work-out closer (in terms of power consumed) to the over track exercise. Figure 5 plots variations in the mechanical power $p_{mech\text{-}drag}$, calculated as in (13), expended (in over-ground running) for overcoming the drag

$$p_{mech-drag} = \frac{1}{2}C_D\rho Av^3 \qquad (13)$$

at various speeds while including the effect of variation in air density ρ as it varies from 1.2 kg/m3 to 1.29 kg/m3.

The two data points selected in the figure show a pmech-drag difference of about 230Watts between the extremes of ρ for the speed of 18m/s. From the figure, around 3.4kW of power generation potential is available as the same workout of 18m/s in an atmosphere of ρ = 1.29 kg/m3 is being carried out on a high speed equine treadmill.

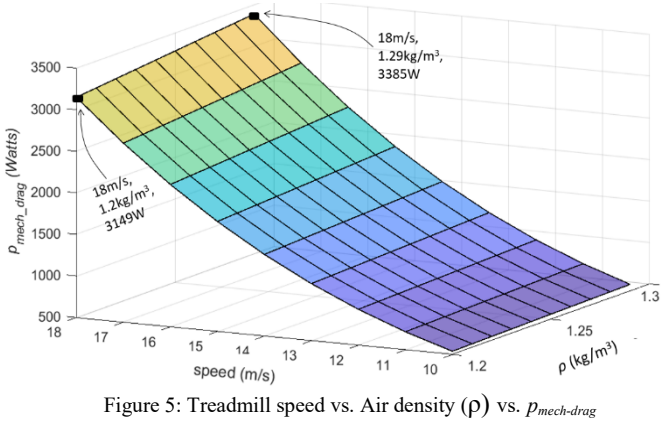


Figure 5: Treadmill speed vs. Air density (ρ) vs. $p_{mech\text{-}drag}$

Furthermore, [34] mentions a case where the exercise on flat treadmill provided a lower cardiac and blood lactate response as compared to exercise on ground at the same speeds. A treadmill incline of 3.5% was found to give heart rate closer to that of the track exercise (especially in the range of 9.1 – 10m/s). Again, if appropriate energy is being extracted from a treadmill, this might give a better replication of the over track exercise while generating renewable power at the same time.

8. Total Mechanical Power: Combining Aerobic and Anaerobic Portions

8.1. Total Generatable Power

Figure 6 presents a wholesome picture – curves for total generatable mechanical power $p_{mech\text{-}total}$ via combining the contribution from the aerobic and anaerobic mechanical powers have been sketched. Except for the last (or second last for 2.5% incline) speed points, the $p_{mech\text{-}total}$ lines are generally quite close to the pmech-aer lines, hence the anaerobic contribution is clearly far less than that of the aerobic part. Furthermore, the rise in $p_{mech\text{-}total}$ lines towards the last speed points may also be explained from Figure 4, which shows a significant rise in $p_{mech\text{-}anaer}$ for the last speed points (except for 2.5%). The dip in $p_{mech\text{-}anaer}$ for 2.5% incline is briefly explained in section 6.

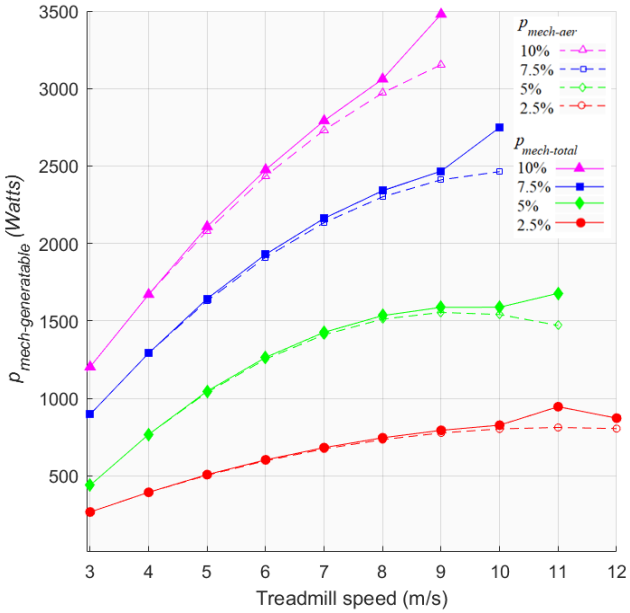


Figure 6: Total generatable power evaluated via combining aerobic and anaerobic portions. The speed points 1 and 2 m/s are ignored for reason mentioned in Figure2 caption.

8.2. Total Expended Power – Revisiting mgv_v concept

Reference [35] summarizes various studies mentioning efficiency for vertical mechanical work for different species – in case of horses the efficiency falls in range of 35 – 40%. If this is true and if a 35% value may be taken for the η_{mscl} variable ([22] mentions use of 35% for vertical efficiency for horses), then the $p_{mech\text{-}total}$ curves, now renamed as $p'_{mech\text{-}total}$, are re-plotted in Figure 7 using the altered value for η_{mscl} (for both the aerobic and anaerobic contributions). Also, plotted are the mgv_v lines for different slopes.

The results of Figure 7 showed an interesting behavior – except for the 7.5% incline (which presents an over-estimation scenario), the mgv_v lines are generally close to the evaluated $p'_{mech\text{-}total}$ curves around the range of 4 - 6m/s (the region is shaded and may be named '*High Proximity Region' (HPR)*). Now if η_{mscl} is truly equal to 35%, the results may be interpreted as suggesting that in a certain speed range, the equine locomotion is involving movement of the whole body, quite the same as it would do on an actual

incline – while at other speeds, the COM is moving lesser vertically compared to how it would have to move on an actual uphill incline.

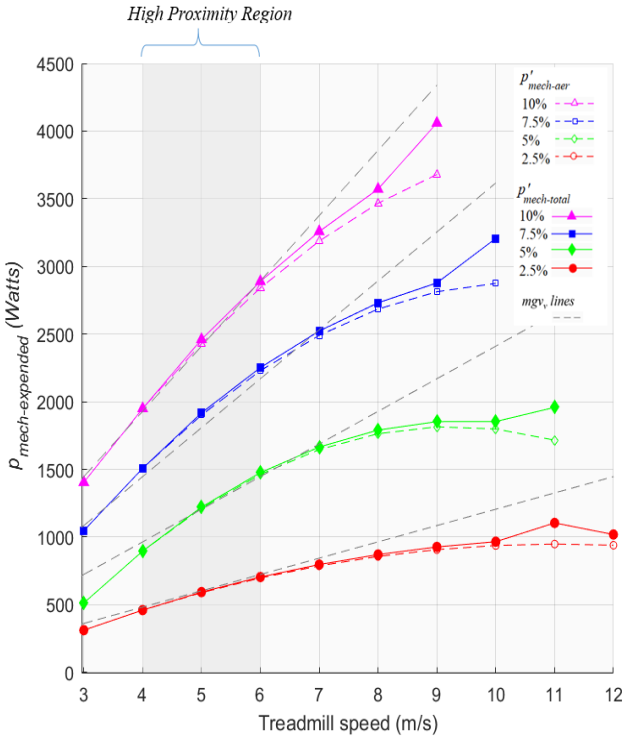


Figure 7: Total expended power evaluated via using 35% muscle efficiency. The speed points 1 and 2 m/s are ignored for reason mentioned in Figure2 caption.

9. The Electrical Power Generatable

The generatable mechanical power $p_{mech\text{-}total}$ is finally converted to electrical power which may be supplied to the grid. Different conversion schemes may be employed, however, for the sake of providing an approach for the evaluation of the potential, the current effort assumes a simple scheme. A DC generator is assumed to be coupled with the treadmill belt, thus converting $p_{mech\text{-}total}$ to p_{dc}. A fixed efficiency of 80% is taken for this conversion.

Next, a grid-tied DC/AC inverter is assumed to convert p_{dc} to AC power $p_{elect\text{-}ac}$ supply-able to distribution electrical power system. For this DC/AC conversion, the Fronius IG Plus 12 curve is selected from [36] ('Figure 4. Typical per unit efficiency curves for grid-connected solar inverters'), graph points are obtained via digitization (using 'grabit' [25]) and curve fitting is used which expresses the efficiency characteristic via (14)

$$Inverter\ Efficiency\ = ax^b + c \qquad (14)$$

where a = - 0.6217, b = -1.183, c = 96.98 and x represents the x-axis point of p_{dc} pu (the p_{dc} per unit is obtained via dividing any actual p_{dc} by 3000Watts which is taken as the rating of the inverter). Subsequently, p_{dc} is converted to $p_{elect\text{-}ac}$ according to the (14). Figure 8 presents the inverter efficiency characteristics (digitized points and the curve fitting approximation) as well as the $p_{elect\text{-}ac}$ lines.

10. Assumptions and Weak Points

The purpose of this section is to summarize some of the weak points/assumptions of the presented effort – this may lead to future improvements as well as better evaluation of the quality of results.

1. Use of ΔVO_2 and Δlactate values (section 4.2)

A justification is provided in the section. Furthermore, due to this assumption, the work may not be applicable to every single

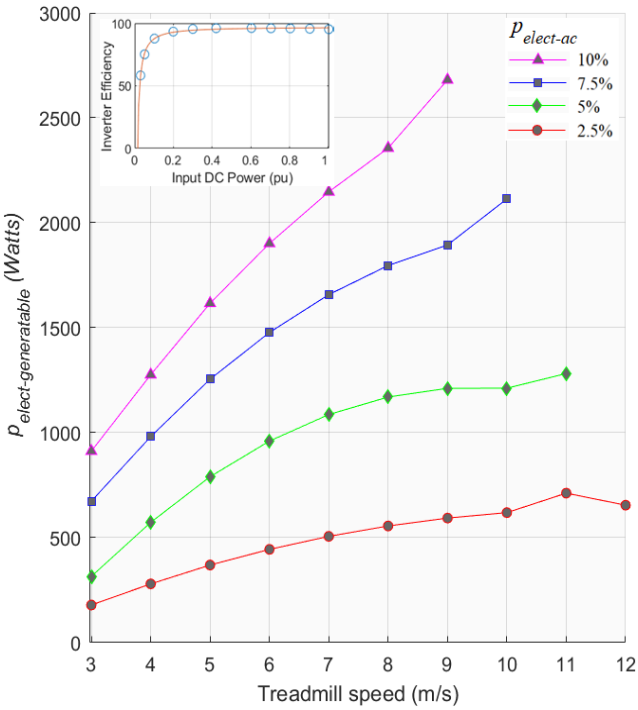


Figure 8: Total generatable (and grid deliverable) ac electrical power from the renewable exercising equine animate prime mover. Inset sub-figure presents inverter efficiency curve – circles show digitized graph points while the curve shows plot of the equation obtained via curve fitting.

individual horse, rather, it would be applicable to an average horse – in other words, if the approach of [23] (which is for humans and for ΔVO_2 only, and not for Δlactate) can be extended to this work (including Δlactate) , then the results should be applicable on the average to different equines with plus minus deviations individually.

2. Variation in $O_{2\text{-}J\text{-}eqv}$ (sections 5.1 & 6)

Depending upon respiratory exchange ratio, O2-J-eqv can vary from 19.61 to 21.1 [26]. The current study does not take any such variation into account - $O_{2\text{-}J\text{-}eqv}$ = 21.1 J/ml O_2.

3. *CRS_{CF} (section 5.1)*

The cardio-respiratory system correction factor is a crude approximation with various assumptions – derivational details are mentioned in the appropriate section.

4. *η_{mscl} (sections 5.1 & 6)*

The mechanical efficiency of muscles (for horizontal locomotion) is taken equal to 30%, also, any variation in this value has not been accounted.

5. *Using mean values only*

From the basic reference [24], which is the source of the treadmill VO_2 and lactacte data – only the mean values are taken and standard deviations have not been included.

6. *$lact_{O2\text{-}eqv}$ (section 6)*

As per our understanding, the current work is not explicitly involving any lactate clearance alongside lactate generation.

7. *Vertical η_{mscl} (section 8.2)*

A fixed value of 35% has been used.

8. *Efficiency of DC generator (section 9)*

A fixed value of 80% has been used. This is a far from true assumption, as the actual efficiency will vary with operating point. However, it may be said that the purpose of section 10 is mainly to convey the idea of $p_{mech\text{-}total}$ to $p_{elect\text{-}ac}$ conversion, so, for the DC/AC inverter, the efficiency variation with changing operating point has been used and then, to keep things simple, this variation was not used for the DC generator.

11. Contributions to Body of Knowledge

As per the understanding of the authors, the following points may be regarded as the contributions of the manuscript

- Bringing forward a novel concept of harvesting energy from an equine exercise treadmill – thus presenting a case of renewable energy generation from animate prime-movers, without compromising the original purpose i.e. exercise of the treadmill machine.
- Presenting an approach to quantify the generatable power from an exercising equine.

12. Conclusion

The paper has presented the idea of tapping renewable energy from the inclined treadmill for equine exercise. The somatic energy of the horse has long been used by humans – the current paper uses it for electricity generation, albeit maintaining exercise as the primary function of the treadmill workout. For determining the amount of the generatable power, a detailed approach has been presented which includes power of aerobic and anaerobic metabolism as well as the conversion into electrical power.

The paper does rely upon assumptions, but an effort has been made to clearly mention them. The (deemed) contributions to the body of knowledge have been mentioned as well. On the whole, the paper is an interdisciplinary effort presenting a novel concept as well as a subsequent quantitative methodology for estimation of the potential.

Acknowledgments

The paper humbly and gratefully acknowledges the contribution of Dr. Shahid in guiding towards the topic, and the contribution of Dr. Zaheer in support with biological topics.

References

[1] I.Z. Pelyva, R. Kresák, E. Szovák, Á.L. Tóth, "How Equine-Assisted Activities Affect the Prosocial Behavior of Adolescents," International Journal of Environmental Research and Public Health, **17**(8), 2020, doi:10.3390/ijerph17082967.

[2] A. Rzekęć, C. Vial, G. Bigot, "Green Assets of Equines in the European Context of the Ecological Transition of Agriculture," Animals, **10**(1), 2020, doi:10.3390/ani10010106.

[3] N. Ohtani, K. Kitagawa, K. Mikami, K. Kitawaki, J. Akiyama, M. Fuchikami, H. Uchiyama, M. Ohta, "Horseback Riding Improves the Ability to Cause the Appropriate Action (Go Reaction) and the Appropriate Self-control (No-Go Reaction) in Children," Frontiers in Public Health, **5**, 8, 2017, doi:10.3389/fpubh.2017.00008.

[4] Gomez-Nacer, "Animal Powered Electricity Generator," Patent US20050161289A1, 2005.

[5] A.R. Smith, "Bovine Treadmill," Patent US7654229B2, 2010.

[6] D.G. Bayerlein, V.E. Emons, N. Oblamski, "Power generating manually operated treadmill," Patent US9956450B2, 2018.

[7] S. Al-Zamil, "Mechanical platform for producing clean energy by means of animals منصة ميكانيكية لانتاج طاقة نظيفة تعمل بواسطة الحيوانات," Patent WO2017061906A1, 2017.

[8] T.-C. Chang, "Livestock power generation system," Patent US20160344260A1, 2016.

[9] W. Taylor, "Animal power generator," Patent US20110266091A1, 2011.

[10] A.K. Bayen, "A system for generating electricity using nonconventional source," Patent WO2019186582A1, 2019.

[11] R.J. Fuller, L. Aye, "Human and animal power – The forgotten renewables," Renewable Energy, **48**, 326–332, 2012, doi:10.1016/j.renene.2012.04.054.

[12] H.M.D.P. Wijethunge, T.G.P. Priyadarshana, "Micro hybrid power plant design with animal draft power and biogas for a rural village," in 2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS), 213–217, 2013, doi:10.1109/GHTC-SAS.2013.6629918.

[13] Sahil, P.K. Sharma, N. Hari, N. Kumar, D. Shahi, "An innovative technique of electricity generation and washing machine application using treadmill," in 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 1–5, 2016, doi:10.1109/ICPEICES.2016.7853524.

[14] H.T. Yang, T.H. Tseng, T.H. Ai, Y.H. Wu, S.H. Yeh, "A grid-connected energy conversion system for a treadmill with auto-transferring modes between a motor and a generator," in IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society, 316–321, 2015, doi:10.1109/IECON.2015.7392118.

[15] S. Ali, S. Murtaza, A. Katiyar, "Design of manual treadmill with electricity generator for energy saving," International Journal of Research in Engineering and Applied Sciences, **5**, 12–15, 2016.

[16] D.F. Preedy, G.R. Colborne, "A method to determine mechanical energy conservation and efficiency in equine gait: a preliminary study," Equine Veterinary Journal, **33**(S33), 94–98, 2001, doi:10.1111/j.2042-3306.2001.tb05368.x.

[17] A.E. Minetti, L.P. Ardigò, E. Reinach, F. Saibene, "The relationship between mechanical work and energy expenditure of locomotion in horses," Journal of Experimental Biology, **202**(17), 2329–2338, 1999.

[18] L.A. Peyré-Tartaruga, M. Coertjens, "Locomotion as a Powerful Model to Study Integrative Physiology: Efficiency, Economy, and Power Relationship," Frontiers in Physiology, **9**, 1789, 2018, doi:10.3389/fphys.2018.01789.

[19] D.F. McMIKEN, "An energetic basis of equine performance," Equine Veterinary Journal, **15**(2), 123–133, 1983, doi:10.1111/j.2042-3306.1983.tb01734.x.

[20] R. Stevenson, R. Wassersug, "Horsepower from a horse," Nature, **364**, 195, 1993, doi:10.1038/364195a0.

[21] M.M.S. Van Oldruitenborgh-Oosterbaan, H.M. Clayton, "Advantages and disadvantages of track vs. treadmill tests," Equine Veterinary Journal, **31**(S30), 645–647, 1999, doi:10.1111/j.2042-3306.1999.tb05305.x.

[22] Z.T. Self, A.J. Spence, A.M. Wilson, "Speed and incline during Thoroughbred horse racing: racehorse speed supports a metabolic power constraint to incline running but not to decline running," Journal of Applied Physiology, **113**(4), 602–607, 2012, doi:10.1152/japplphysiol.00560.2011.

[23] P. Gimenez, P.J. Arnal, P. Samozino, G.Y. Millet, J.-B. Morin, "Simulation of uphill/downhill running on a level treadmill using additional horizontal force," Journal of Biomechanics, **47**(10), 2517–2521, 2014, doi:https://doi.org/10.1016/j.jbiomech.2014.04.012.

[24] M.D. Eaton, D.L. Evans, D.R. Hodgson, R.J. Rose, "Effect of treadmill incline and speed on metabolic rate during exercise in thoroughbred horses," Journal of Applied Physiology, **79**(3), 951–957, 1995, doi:10.1152/jappl.1995.79.3.951.

[25] GRABIT - Extract (pick out) data points off image files, May 2020.

[26] C.B. Scott, A primer for the exercise and nutrition sciences: thermodynamics, bioenergetics, metabolism, Humana Press/Springer Science, 2008.

[27] R.E. Klabunde, Myocardial Oxygen Demand, Aug. 2020.

[28] H. Kline, J.H. Foreman, "Heart and Spleen Weights as a Function of Breed and Somatotype," in 3rd International Conference on Equine Exercise Physiology (ICEEP), 17–21, 1991.

[29] M.D. Loer Stephan A., M.D. Scheeren Thomas W. L., M.D.F. Tarnow Jorg, "How Much Oxygen Does the Human Lung Consume?," Anesthesiology: The Journal of the American Society of Anesthesiologists, **86**(3), 532–537, 1997.

[30] S.J. Wickler, H.M. Greene, K. Egan, A. Astudillo, D.J. Dutto, D.F. Hoyt, "Stride parameters and hindlimb length in horses fatigued on a treadmill and at an endurance ride," Equine Veterinary Journal, **38**(S36), 60–64, 2006, doi:10.1111/j.2042-3306.2006.tb05514.x.

[31] S. Franklin, K. Allen, 2 - Laboratory exercise testing, W.B. Saunders: 11–24, 2014, doi:https://doi.org/10.1016/B978-0-7020-4771-8.00002-8.

[32] S.L. Bond, P. Greco-Otto, R. Sides, G.P.S. Kwong, R. Léguillette, W.M. Bayly, "Assessment of two methods to determine the relative contributions

of the aerobic and anaerobic energy systems in racehorses," Journal of Applied Physiology, **126**(5), 1390–1398, 2019, doi:10.1152/japplphysiol.00983.2018.

[33] Z.T.S. Davies, A.J. Spence, A.M. Wilson, "External mechanical work in the galloping racehorse," Biology Letters, 2019, doi:10.1098/rsbl.2018.0709.

[34] E. Barrey, 10 - Biomechanics of locomotion in the athletic horse, W.B. Saunders: 189–211, 2014, doi:https://doi.org/10.1016/B978-0-7020-4771-8.00010-7.

[35] G.K. Snyder, C.A. Carello, "Body mass and the energy efficiency of locomotion: Lessons from incline running," Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, **150**(2), 144–150, 2008, doi:https://doi.org/10.1016/j.cbpa.2006.09.026.

[36] R.S. Faranda, H. Hafezi, S. Leva, M. Mussetta, E. Ogliari, "The Optimum PV Plant for a Given Solar DC/AC Converter," Energies, **8**(6), 4853–4870, 2015, doi:10.3390/en8064853.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1007-1012 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Control of Soft Robotic Artificial Muscle with Hand Gesture Using Leap Motion Sensor

Victoria Oguntosin*, Akindele Ayoola E

Department of Electrical & Information Engineering, Covenant University, Ota, Ogun State, 112233, Nigeria

ARTICLE INFO

Article history:
Received: 01 June, 2020
Accepted: 18 July, 2020
Online: 20 October, 2020

Keywords:
Artificial Muscles
Pneumatic Actuator
Soft robot
Gesture-based control
Leap motion sensor

ABSTRACT

We describe the control design strategy used to control a soft robotic artificial muscle composed of silicone rubber using hand gesture signals. This artificial muscle is actuated with pneumatics, and therefore, the control strategy employed is through the regulation of air pressure within the inner chambers. Using the hand gestures of bringing the hands apart and together, the artificial muscle can be made to expand and contract with the gesture interface from the leap motion sensor. The advantage of the employed hand gesture control compared to switch control is that it provides a more natural interface for the regulation of air pressure within the artificial muscle through the use of electronic and automatic control. Possible areas of application include the use of the soft muscles for rehabilitation purposes and the combined system for developing a physiotherapy gaming device to exercise the hands and fingers of individuals that need to strengthen the muscles of the hands and fingers.

1. Introduction

Soft robots are specialized sets of actuators that possess infinite degrees of freedom and, therefore, able to perform complex motions and manipulations, which would be difficult for rigid manipulators to perform. The control of such infinite degrees of freedom is, therefore, an integral part of the study and design of soft actuators. This paper aims to control a soft robotic artificial muscle using gesture-based control. This paper is an addition to the study initially presented in International Conference on Soft Computing & Machine Intelligence (ISCMI) [1], which employs hand gestures to control a soft rotary actuator. In this work, the same principle is employed to contract and expand a Soft Robotic Muscle using gesture signals.

The pneumatic air within pneumatically powered soft actuators can be controlled using manual or electronics control. Manual control involves the operation of the control hardware such as piston and cylinder via manual operation [2] while electronics operation makes use of electronic devices such as air pump and ON/OFF solenoid valves for the control hardware interfaces [3][4]. While electronics control has the superior advantage of automatic control compared with manual operation of pistons and cylinders, an electronics-free control system is sought after. This is because soft robots are inherently compliant, and as such, can be used in environments where Electromagnetic Interferences (EMI) needs to be minimized. An electronics-free control system is required because the air pump and solenoid valves used for electronics control generate EMI waves and, as such will limit the use of soft robots in such controlled environments. Electronics-free control makes use of fluidic valves, logic circuits, and fluidic processors. The switching valves and pumps are composed entirely of silicone rubber [5].

Soft robot control can be over a long-distance communication channel [6]. The transfer of control commands from the master device to a slave device via a remote communication channel results in delays. For instance, a teleoperated soft actuator controlled on the master side using hand gestures has been studied for the measured delays. With the Message Queuing Telemetry Transport (MQTT) server utilized to convey packets of data over the internet to the soft actuator, delays, which include communication delays, elastic saturation delays occurring as a result of non-linear effects of pneumatic action, delays due to noise and hardware imperfections were observable. A Kalman filter [7] has been used to effectively minimize the delay due to elastic saturation.

Unlike silicone-based soft roots that move via material stretching, fabric-based eversion robots are made from non-extensible nylon fabric. Therefore, more complex control such as

*Victoria Oguntosin, Department of Electrical & Information Engineering, Covenant University, KM 10, Idiroko Road, Ota, Ogun State, Zip Code: 112233, Nigeria, Contact No: +2349018107052
Email: victoria.oguntosin@covenantuniversity.edu.ng

extension, steering, tip position, and orientation, and variable stiffness control is possible using air pressure control [8].

As a result of the multiple degrees of freedom exhibited by soft robots, it implies that multiple sensors are required to measure these degrees of freedom for feedback control. Embedding a large number of sensors within a soft robot is impractical because embedded sensors can limit movement [9] and can give rise to increased material and computational cost. Therefore, an observer or estimator is necessary to estimate the states of the soft robot based on the available measurements from the sensors available. Before observer design can be done, a model of the soft robot is first required. A soft robot has been modeled using a network of wavelet and a Kalman filter used to predict output measurements [10].

In this paper, a soft robotic muscle described in [11] is controlled for contraction and expansion using a gesture device. The choice of an artificial muscle based soft device compared to other forms of the soft actuator is due to its ability to contract and expand in such a manner that can follow natural motions of the hands. Soft robotic muscles made of silicone have the advantage of high contractibility [12] and high force compared to conventional McKibben muscles [13]. Soft robotic muscle actuators make use of vacuum pressure to contract via buckling of air channels [14] and make use of air pressure to expand via the opening of air channels [11]. This is In contrast to McKibben's artificial type muscles that shorten when pressurized air enters into the inner air cavities [15]. The novelty of the proposed solution is the control of artificial muscles for expansion and contraction purposes via hand gestures. Possible areas of application include the use of the soft muscles for rehabilitation purposes and the combined system for game purposes – a physiotherapy gaming device can be developed to exercise the hands and fingers of individuals that need to strengthen the muscles of the hands and fingers.

Hand gesture recognition refers to the combination of specialized sensors with mathematical algorithms to recognize the movement of the hands with the intention of controlling devices without touch. The applications of hand-gesture systems are in the areas of robotics, home automation, automatic sign language recognition, phone, and computer security and gaming devices. The sensors used for gesture recognition include magnetic sensors, inertial measurement units, single cameras, depth cameras, stereo cameras, and gesture-based gaming controllers. Hand gesture sensors can make use of 3D model-based algorithms, Skeletal-based algorithms (such as the Kinect and Leap Motion sensors), Appearance-based models (such as cameras) or Electromyography-based models (such as EMG electrodes) to track hand movements and recognize the gestures.

An accelerometer system for drawing English letters and Arabic numerals as well as recognizing directional movement for gaming, was implemented [16] as a hand recognition device for recognizing the handwritten gestures. Hand movement is captured with the use of an accelerometer while the DTW algorithm processes it and sent to a mobile device using the BlueTooth interface. The gesture recognition and motion detection algorithm were validated by collecting hand gesture and directional motion trace for three participants. Being an accelerometer based system, the device performed best with directional movement detection compared to shape and digits recognition.

In order to miniaturize hand gesture devices, a microsystem consisting of ultrasonic traducers was proposed [17]. The hand gestures implemented include hand tap, hand click, hand rotate, hand press, and hand rub, which are precisely the way people use the fingers in actual scenarios. Ultrasonic sensors were used to distinguish the movement of the hands, and RF signal processing algorithms were used to excerpt features for the machine learning classifiers. Accuracies of conventional classification methods with Radio Frequency (RF) and end-to-end architecture for 9 subjects were analyzed with accuracy greater than 80% for all the classifiers.

A segmentation method that uses both depth and RGB color data from a scene using a Kinect sensor to obtain the hand position and convex hull detection algorithm for identification of the fingertips position was presented [18]. The hand gesture image is first extracted, after which contour information is removed to give a sketch of the hand with palm and finger location highlighted. From the presented results, the proposed method was found to accurately determine hand gestures in realtime with distances of 0.6m, 1.0m, 1.5m, and 2.0m from the sensor. The gesture recognition rate was between 96.67% - 98.88% for the six gestures.

A multimodal hand gesture recognition system comprising of the combined use of a single Inertial Measurement Unit (IMU) and pressure-based transducers placed on the wrist has been evaluated [19]. The multimodal approach eliminates the problem of using a single IMU as a hand gesture detection system due to arm movements. The system comprised of an IMU unit (3-axis accelerometer, and 3-axis gyroscope) and acoustic signals (10 microphones) placed on the wrists of 10 subjects in order to detect and track thirteen common gestures for hand signaling.

A method for detecting gestures of the hand [20] using soft skin tactile sensors to measure wrist movements instead of using the more conventional methods involving IMUs, EMG, and FSR has been implemented. Eight specific gestures comprising the movement of the wrist could be distinguishable using the novel sensor. Filtering techniques were used to filter out pulse rate noise, which could be misinterpreted as a movement of the wrist. The verification of the performance of the proposed sensor involved comparison tests between the sensor with an EMG and FSR sensor was conducted. 97–99% accuracy of the stain sensor was obtained compared to the EMG and FSR sensors suggesting the superiority of the novel gesture recognition method with the tested conventional methods. Other conventional methods that have been used to control soft actuators other than hand gestures [21] include eye movement and tracking [22].

This section provided an introduction with background material on soft robotics control. It also describes hand gesture sensors and systems. Section two provides a description of the hand gesture control strategy employed to control the contraction and expansion of the soft robotic artificial muscle. Results and discussion are offered in Section three. The last section, Section four, provides a conclusion for the work.

2. Hand Gesture Control Design of Soft Robotic Muscle

In this work, a hand gesture system that controls a soft robotic muscle for contraction and expansion actuation is implemented.

The system design comprises of the hardware and software design. The hardware system consists of a leap motion sensor to detect the hands and the soft robotic muscle. The software system is made up of the control algorithms that detect the gestures responsible for contracting and expanding the soft muscles.

2.1. Leap Motion Sensor Interface

The Leap Motion device continually detects hand, finger, or tool-like objects inside its field of view when powered ON via connection to a computer. Three infrared LEDs and two infrared cameras within the device facilitate hand motion tracking. The device is able to track any object located with its 150^0 viewing area with a distance of 0.50m from it in the upward direction and 0.25m from it in other directions [23]. Hand and fingers can be detected by the Leap using the Software Development Kit (SDK) libraries. The leap sensor is displayed in Figure 1.

The leap motion sensor tracks the movement of the hands within its FOV by measuring distances in three directions, called the X, Y, and Z directions. The X distance is changed when a positioned hand moves left or right, the Y distance changes with a hand moving up and down from the sensor, while the Z distance changes with a positioned hand moving forward or backward.

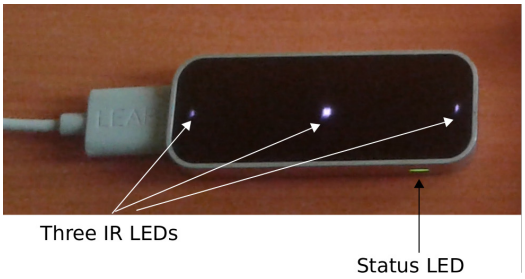


Figure 1: The Leap Motion sensor powered via USB connection to a computer; the green status LED is ON and the three infrared LEDs come ON as shown.

2.2. Soft Robotic Muscle Interface

The Soft Robotic Muscle (SRM) used is the same described in [24]. The detailed fabrication process, together with mathematical equations describing its model, is presented in [11]. The SRM muscle is comprised of soft modular units arranged in series to improve the rate of shortening. The SRM muscle contracts upon the application of vacuum pressure and expands to its original length when positive pressure is applied to its inner chambers. The muscle is also able to bend bi-directionally by controlled variation of air and vacuum pressure within its inner walls. The SRM muscle is shown in Figure 2. The parameters of muscle include the wedge angle, $\theta = 60^0$, wedge length, l_l=20mm, number of modular units, n=9, air channel height, h_c=40mm, wall thickness, t=2.5mm, and muscle length, l=160mm. The soft robotic muscle can contract and expand partially. The zero position of the SRM is at its normal length of 160mm; its contacted length is 104mm while its expanded length is 170mm.

2.3. Integrated Hand Gesture and Soft Robotic Interface

The software aspect of the system involves designing the hand gesture system responsible for contracting and expanding the SRM muscle. This process first begins with detecting the two hands and then implementing the software that measures the distance between both hands. Two gestures which we term the "contracting" and "expanding" gestures are implemented. The two gestures are described as follows:

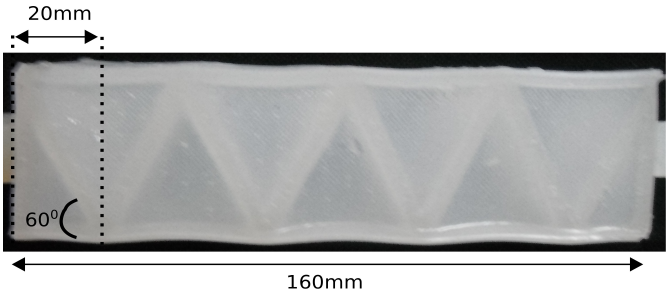


Figure 2: SRM Muscle having a total length of 160mm and defining angle of wedge as 60^0, wedge length is 20mm, number of wedges is 9, the height of air channel equals 40mm and wall thickness is 2.5mm

- **Contraction gesture**: This gesture is represented by bringing both hands together, i.e., decreasing the distance between the left and right hands.
- **Expansion gesture**: This gesture is represented by bringing both hands farther apart, i.e., increasing the distance between the left and right hands.

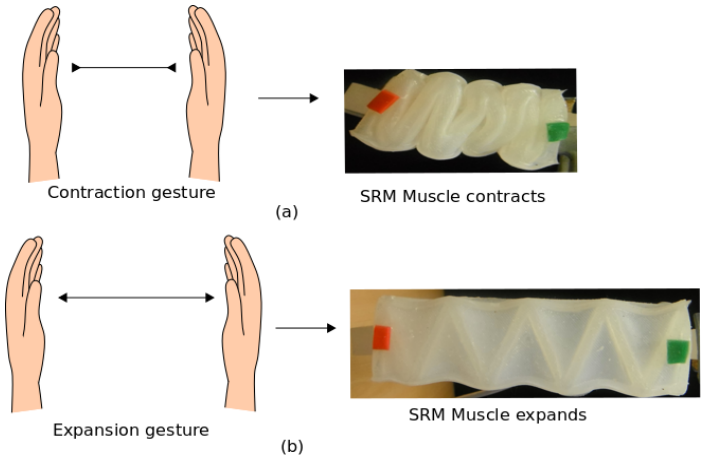


Figure 3: From (a), when the contraction gesture is executed, the SRM muscle contracts; From (b), when the expansion gesture is executed, the SRM muscle expands to its original length.

The two gestures were developed because they follow the natural motions of how a human will contract and expand the muscle as if holding it physically without an actuating source. Therefore, when the contraction gesture is executed, control commands are sent to the SRM muscle to contract it. Conversely, when the expansion gesture is performed, a control command from the leap sensor interface is sent to the electronics hardware [4] to cause the expansion of the soft robotic muscle. The rationale behind using a leap sensor is so that expansion and contraction control of the SRM muscle can be implemented over a remote distance. This gesture interface allows for a master-slave control where the master is the movement of the two hands, and the slave is the soft robotic muscle that follows the motions of the master. The pictorial representation of the system design is seen in Figure 3.

The software flowchart for implementing the integrated system comprising hand gesture detection and soft robotic muscle control is shown in Figure 4. As seen from the Figure, the distance between

the two hands is measured; if this distance is decreasing, i.e., its value is reducing, that means that a contraction gesture is being executed and a command is sent to the SRM muscle to contract it. If the measured distance is increasing, i.e., its value is getting larger from the previous timestamp, which means that an expansion gesture is being executed, and a command is sent to the SRM muscle to expand it to its normal resting length. If there is no change in the distance measured between the consecutive timestamps, then no gesture is being executed, and the SRM muscle does not change in length.

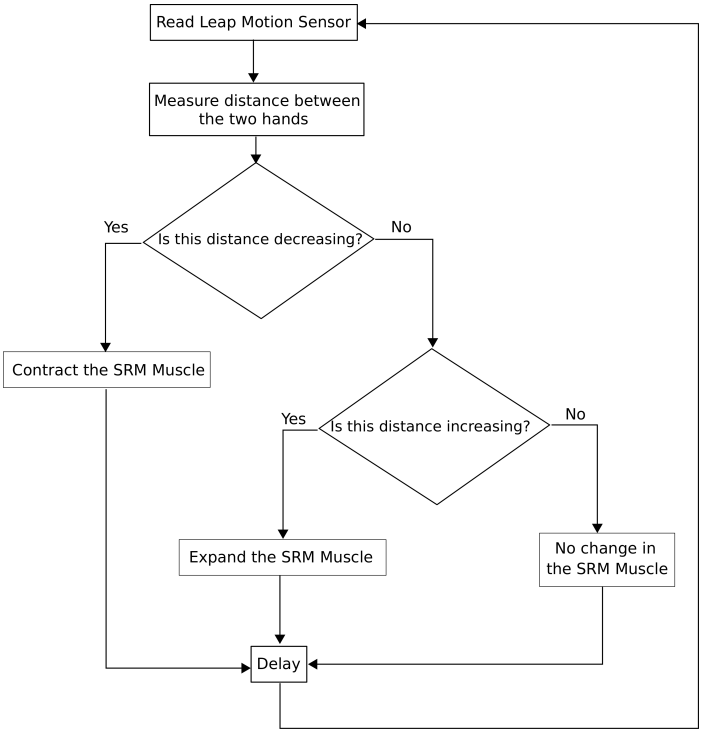


Figure 4: Flow chart for the control of soft robotic artificial muscle with hand gesture signals using the leap motion sensor

3. Results and Discussion

The controller hardware setup consisting of the electronic circuit, leap motion sensor, SRM Muscle, supply pumps for negative and positive pressurized air is shown in Figure 5. The detailed explanation of the experiment and description of the used circuit board and controller can be found in [3][4]. Through the USB connection of the Leap device to a computer, interpreted gestures commands are sent to the controller hardware with the microcontroller and leap connected to the same computer interface for the transfer control commands.

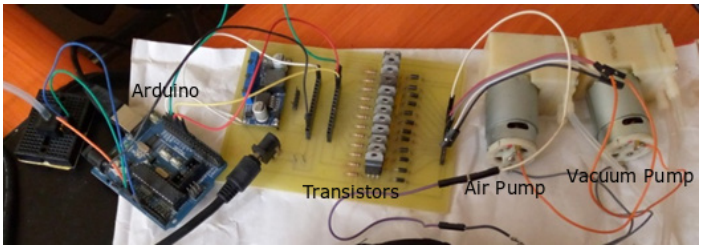


Figure 5: The hardware controller for the hand gesture system consisting of the controller circuitry, vacuum pump for contracting the muscle and air pump for expanding the SRM Muscle

Figure 6 shows absolute vacuum pressure and contraction ratio as a function of during contraction. A maximum absolute vacuum pressure of -20kPa is observed for a maximum contraction ratio of 0.35. The measured contraction time is 4 seconds

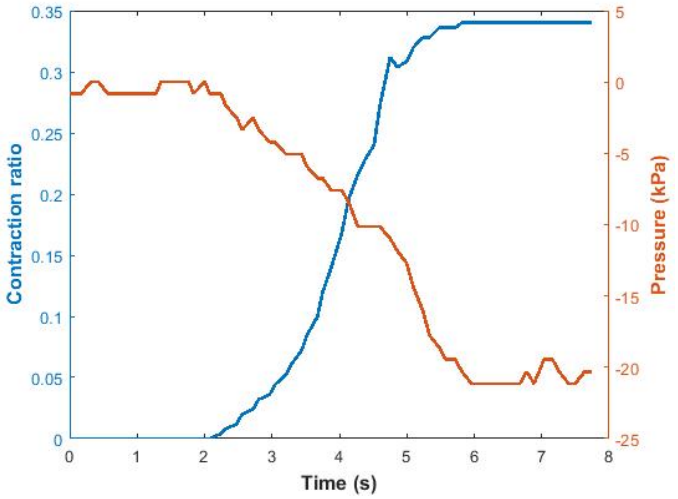


Figure 6: Contraction of the SRM Muscle, the contraction ratio moves from 0 to 0.35 of the muscle length in 4s, the vacuum pressure is shown to move from 0kPa to -20kPa during contraction

When the SRM Muscle is undergoing contraction, contraction ratio, and absolute vacuum pressure move from zero to the contraction ratio of 0.35 and vacuum pressure of -20kPa, respectively. Figure 7 shows absolute vacuum pressure and contraction ratio as a function of during expansion. The contraction ratio moves from 0.35, which is the contracted ratio to 0, which is the original muscle resting position in 6 seconds. The vacuum pressure is shown to move from -20kPa to -0kPa during expansion.

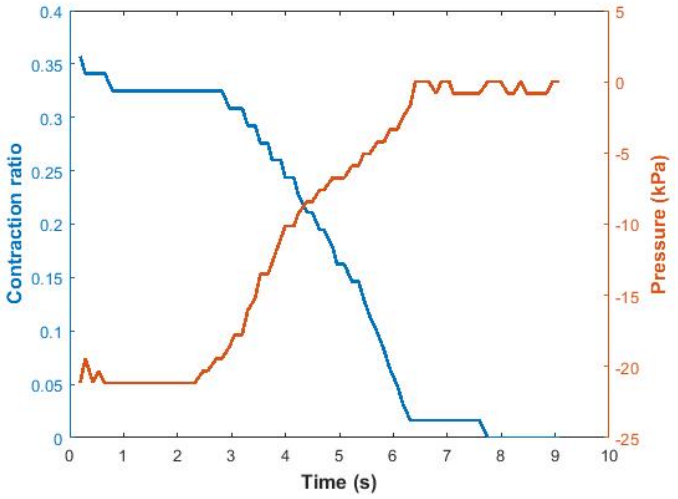


Figure 7: Expansion of the SRM Muscle, the contraction ratio moves from 0.35 which is the contracted form to 0 which is the original muscle resting position in 6s, the vacuum pressure is shown to move from -20kPa to -0kPa during expansion

When the SRM Muscle is undergoing expansion, contraction ratio, and absolute vacuum pressure move a contraction ratio of 0.35 and vacuum pressure of -20kPa to zero as shown. Through the use of the leap sensor to control the soft muscle, closed-loop

control operation of the SRM muscle opens the possibility to obtain a specified muscle contraction. To achieve the desired operation, visual information from the users' eyes measures the appropriate contraction or expansion length while control signals are sent to contract and expand the SRM muscle.

The novelty of the proposed solution is the control of artificial muscles for expansion and contraction purposes via hand gestures. Possible areas of application include the use of the soft muscles for rehabilitation purposes and the combined system for game purposes – a physiotherapy gaming device can be developed to exercise the hands and fingers of individuals that need to strengthen the muscles of the hands and fingers.

The demonstration of the control of the Soft Robotic Artificial Muscle with a hand gesture using the leap motion sensor is shown in Figure 8. A user performs the contraction and expansion gesture using the Leap motion sensor as a hand gesture interface to create contraction and contraction movements of the Soft Robotic Muscle. As shown, both hands must be within the 150^0 field of view of the leap sensor in order that the accurate distance between the hands is measured. Also, both hands should be within 0.50m from the sensor in the upward direction and 0.25m from the sensor in the x and y direction.



Figure 8: A user performs the contraction and expansion gestures using the Leap motion sensor as hand gesture interface to create contraction and contraction movements of the Soft Robotic Muscle

4. Conclusions

A control design strategy for the control of artificial muscles using hand gesture signals have been proposed and implemented through the regulation of air pressure within the inner chambers. The artificial muscle can be made to expand and contract as a result of interpreted gesture commands from the leap motion device. The advantage of the employed hand gesture control compared is that it provides a more natural interface for the regulation of air pressure within the artificial muscle through the use of electronic and automatic control. Applications of this work include teleoperated control of a soft robotic muscle using the natural motions of how the muscle will be contracted and expanded if it is physically held.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors acknowledge the management of Covenant University, Ota, Nigeria, for the publication support fund.

References

[1] V. Oguntosin, A. Akindele and E. Oladimeji, "Gesture-Based Control of Rotary Pneumatic Soft Robot Using Leap Motion Controller" in 6th International Conference on Soft Computing & Machine Intelligence(ISCMI), 2019, 169-174, 2019, DOI: 10.1109/ISCMI47871.2019.9004295

[2] F. Ilievski, A.D. Mazzeo, R.F. Shepherd, X. Chen and G.M. Whitesides, "Soft Robotics for Chemists" Angewandte Chemie,, **123**(8): 1930-1935. 2011, DOI: 10.1002/anie.201006464

[3] V. Oguntosin, S. Nasuto, and Y. Hayashi, "Embedded Fuzzy Logic Controller for Positive and Negative Pressure Control in Pneumatic Soft Robots" in UKSim-AMSS 19th International Conference on Computer Modelling & Simulation, 2017, 63–68. 2017, DOI: 10.1109/UKSim.2017.41

[4] V. Oguntosin, S. J. Nasuto and Y. Hayashi, "A Compact Low-Cost Electronic Hardware Design for Actuating Soft Robots", International Journal of Simulation Systems, Science & Technology (IJSSST), **16**(3), 1–11, 2016, DOI: 10.5013/IJSSST.a.16.03.06

[5] S. T. Mahon, A. Buchoux, M. E. Sayed, L. Teng and A. A. Stokes, "Soft Robots for Extreme Environments: Removing Electronic Control", IEEE International Conference on Soft Robotics, 782-787, 2019. DOI: 10.1109/ROBOSOFT.2019.8722755

[6] P..D..S..H. Gunawardane, R..E..A. Pallewela and Nimali T. Medagedara, "Tele-Operable Controlling System for Hand Gesture Controlled Soft Robot Actuator" in 2nd IEEE International Conference on Soft Robotics (RoboSoft), 2019, DOI: 10.1109/ROBOSOFT.2019.8722756

[7] P.D.S.H. Gunawardane, R.E.A Pallewela, A.P.T.D. Pathirana and N. T. Medagedara, "Kalman Filter Based Approach to Overcome Elastic Saturation and Pneumatic Disturbances in Tele-Operated Soft Actuators" in Advances in Manufacturing Technology XXXIII, 2019, DOI:10.3233/ATDE190023

[8] A. Ataka, T. Abrar, F. Putzu, H. Godaba, and K. Althoefer, "Model-based Pose Control of Inflatable Eversion Robot with Variable Stiffness", IEEE Robotics and Automation Letters, 2020, DOI: 10.1109/LRA.2020.2976326

[9] V. Oguntosin, A. Akindele, O. Alashiri,. "Vision Algorithms for Sensing Soft Robots", Journal of Physics: Conference Series, **1378**(3), 2019, DOI: 10.1088/1742-6596/1378/3/032102

[10] J. Y. Loo, K. C. Kong, C. P. Tan1 and S. G. Nurzaman "Non-linear System Identification and State Estimation in a Pneumatic Based Soft Continuum Robot", IEEE Conference on Control Technology Applications, 2019, DOI: 10.1109/CCTA.2019.8920693

[11] V. Oguntosin, A. Akindele, "Design and Characterization of Artificial Muscles from Wedge-Like Pneumatic Soft Modules." Sensors and Actuators A: Physical, 2019, DOI: 10.1016/j.sna.2019.07.047

[12] K. Han, N. Kim, and D. Shin, "A Novel Soft Pneumatic Artificial Muscle with High-Contraction Ratio", Soft Robotics, **00** (00), 2018, DOI: 10.1089/soro.2017.0114

[13] S. M. Mirvakili and I. W. Hunter, "Artificial Muscles: Mechanisms, Applications, and Challenges" Advanced Materials, 1704407, 1-28, 2017, DOI: 10.1002/adma.201704407

[14] D. Yang, M. S. Verma, J. So, B. Mosadegh, C. Keplinger, B. Lee, F. Khashai, E. Lossner, Z. Suo, and G. M. Whitesides, "Buckling Pneumatic Linear Actuators Inspired by Muscle", Advanced Materials Technologies, 1600055, 1-6, 2016, DOI: 10.1002/admt.201600055

[15] T. Abrar, F. Putzu, J. Konstantinova and K. Althoefer, "EPAM: Eversive Pneumatic Artificial Muscle" in 2nd IEEE International Conference on Soft Robotics (RoboSoft), 2019, DOI: 10.1109/ROBOSOFT.2019.8722787

[16] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. DOI: 10.1109/ICECCE47252.2019.8940756

[17] Y. Sang, L. Shi, and Y. Liu, "Micro Hand Gesture Recognition System Using Ultrasonic Active Sensing", IEEE Access, **6**, 49339-49347, 2018, DOI: 10.1109/ACCESS.2018.2868268

[18] X. Ma and J. Peng, "Kinect Sensor-Based Long-Distance Hand Gesture Recognition and Fingertip Detection with Depth Information", Hindawi Journal of Sensors, 5809769, 1-9, 2017. DOI: 10.1155/2018/5809769

[19] N. Siddiqui, R. H. M. Chan, "Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist" PLoS ONE **15**(1): e0227039. 2020, 1-12. DOI: 10.1371/journal.pone.0227039

[20] B. Sung-Woo and L. Seok-Pil, "Implementation of Hand Gesture Recognition Device Applicable to Smart Watch Based on Flexible Epidermal Tactile Sensor Array", MDPI micromachines, **10**, 692, 1-15, 2019. DOI:10.3390/mi10100692

[21] V. Oguntosin and A. Abdulkareem, "Design of a pneumatic soft actuator controlled via eye tracking and detection", Heliyon. **6**(7):e04388, 2020, DOI: 10.1016/j.heliyon.2020.e04388
[22] V. Oguntosin and A. Abdulkareem, "Hand gesture control and design of a rotary pneumatic soft actuator using leap motion sensor". Int J Intell Robot Appl, 2020, DOI: 10.1007/s41315-020-00140-5
[23] B. Sanders, Mastering Leap Motion. Packt Publishing, 2014
[24] V. Oguntosin, "Development of Soft Modular Robotics," Ph.D Thesis, University of Reading, 2017. ISNI: 0000 0004 6499 6177

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1020-1027 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Dissection of Quantitative Trait Loci (QTL), annotation of Single Nucleotide Polymorphism (SNP), and Identification of Candidate Genes for Grain Yield in *Triticum turgidum* L. var durum

Issame Farouk[*,1], Ahmad Alsaleh[2], Jihan Motowaj[3], Fatima Gaboun[4], Bouchra Belkadi[1], Abdelkarim Filali Maltouf[1], Zakaria Kehel[3], Ismahane Elouafi[5], Nasserelhaq Nsarellah[4], Dimah Habash[6], M. Miloudi Nachit[3]

[1]*Laboratory of Microbiology and Molecular Biology, Department of Biology, Faculty of Sciences, Med V University, Rabat, 10000, Morocco*

[2]*Dept. Science and Technology Bozok University, Yozgat, 66100, Turkey*

[3]*ICARDA. Av. Med Belarbi Alaoui, BP6299, Al Irfane, Rabat, 10000, Morocco*

[4]*INRA. Institut National de Recherche Agronomique Av. Med Belarbi Alaoui, Al Irfane, Rabat, 10000, Morocco*

[5]*ICBA. Dubai, 14660, United Arab Emirates*

[6]*Securewheat, AL3, St. Albans, UK*

ARTICLE INFO

Article history:
Received: 29 June, 2020
Accepted: 04 October, 2020
Online: 20 October, 2020

Keywords:
Durum wheat
SNP
Candidate genes
Grain yield
QTL
Linkage map

ABSTRACT

Durum wheat (Triticum turgidum L. var durum) is among the most important crops in the world. High and stable grain yield in diverse environments is the major objective in durum breeding programs. This trait is linked to the quantitative trait loci (QTL). For the detection of QTL linked to the grain yield, it is necessary to construct a high-density genetic linkage map. The aims of this study were to detect the candidate genes comprised in the QTLs on 2A chromosome linked to grain yield and annotate the single nucleotide polymorphisms (SNPs). The linkage map of Lahn/Cham1 population was used to identify QTLs. In multi-environmental analysis and employing bioinformatic approaches, 583 sequences corresponding to the SNPs markers selected from the detected QTL regions were analyzed, 122 SNP sequences were annotated of which 53% of the candidate genes were involved in stresses tolerance, 29.5% in plant development and growth, and 3.3% in cell transport. Moreover, 1.6% of candidate genes were retrotransposon and transposon 2.4% with unknown function. Further 9.8% were related in other cellular processes. The results also showed that 66.7% of the candidate genes harbored on 4B chromosome, were involved in stresses tolerance and 33.3% in plant development and growth. Additionally, in the specific and stressed environments analysis, the DNA sequences of the four QTL detected on 2A chromosome were used for homology search, 546 candidate genes were identified of which some were present in several QTL (F-box gene family, hydroxyproline-rich glycoprotein-like), retrotransposons and transposons and others. This study provided information on employing SNPs markers to detect candidate genes linked with grain yield trait in durum wheat in contrasting environments (dry, cold, hot).

1. Introduction

Durum wheat (*Triticum turgidum* L. var durum) is a tetraploid wheat, its genome is constituted by seven homoeologous groups of chromosomes A and B. Its grain use is reported in [1, 2]. The nutritional values for durum grain were published in earlier research works [1-3]. Grain yield is related to the QTL additive effects and to their interactions with the environment. Several wheat genetic linkage maps were constructed with various molecular markers, such as Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length

[*]Corresponding Author: Issame Farouk, Email: issame.farouk@gmail.com

Polymorphism (AFLP), Simple Sequence Repeats (SSR) and other markers [4-13]. The construction of saturated maps with large number of molecular markers of Single Nucleotide Polymorphism (SNP) [14-16] and Diversity Array Technology (DArT) [10, 17-23] are fundamental to detect candidate genes for desirable traits in crops, such as durum wheat. Both markers (SNP and DArT) are small sequence of DNA. These sequences can be compared with the public databases [URGI (Unité de Recherche Génomique Info) Blast of the Institut National de la Recherche Agronomique (INRA, France) and NCBI (National Center for Biotechnology Information, the USA)] to identify the candidate genes involved in a QTL controlling a specific trait. Recently the markers (SNP and DArT) were used for QTLs regions detection and for the identification of the candidate genes involved in these QTLs of wheat [24-29]. The high and stable grain yield across diverse environments constitutes the principal aim for durum breeding programs.

The aims of this work were to: 1) identify the SNPs mapped into QTL regions; 2) extend the SNPs sequence analysis in terms of functional and comparative analysis; and 3) identify the candidate genes constituting the QTLs harbored by 2A chromosome and linked to grain yield.

2. Materials and methods

2.1. *Plant materials*

The genetic population used in this study (Lahn/Cham1) was published in [8, 30] and presented in the conference [31]. The population was phenotyped for grain yield in 11 contrasting environments (cold / drought / heat). The characteristics of the parents: Lahn and Cham1 were described in earlier paper [8, 30].

The Lahn/Cham1 genetic map that consisted of 1425 SNPs, 216 SSR and 31 other molecular markers was constructed using IciMapping program v 4.1.0.0 was presented in [31].

The QTLs linked to grain yield detected using this map were used for candidate genes identification and SNPs annotation.

2.2. *SNPs annotation and candidate genes identification*

The methodology used for the SNPs annotation and the candidate genes identification is as flowing:

For the QTLs detection in specific and stressed environments, 9 QTLs were detected, 4 of which were localized on 2A, due to its strong contribution to grain yield [32, 33], the full DNA sequence of 2A chromosome were downloaded from URGI and the 4 DNA sequences corresponding to the 4 QTL were isolated based on markers flanking of these QTL regions [24, 25, 29] using the GeneStudio software. More informations about the environments, DNA sequences size, environmental conditions, QTL intervals and flanking markers are summarized in Table 1. Thereafter, the sequences isolated were served as queries for Blast search in NCBI against Rice and *Brachypodium* genomes to identify the candidate genes composing these QTLs detected in the four environments (EP, INC, RF, and RQ), see Figure 1

In multi- environmental analysis, the 583 SNPs selected from the QTLs regions and harbored by durum chromosomes [31] were analyzed through Blast (Basic Local Alignment Search Tool) search homology using the URGI Blast of (INRA), France (https://urgi.versailles.inra.fr/blast/blast.php) with the aim to expand the SNP sequences in terms of length [34] (Figure 2).

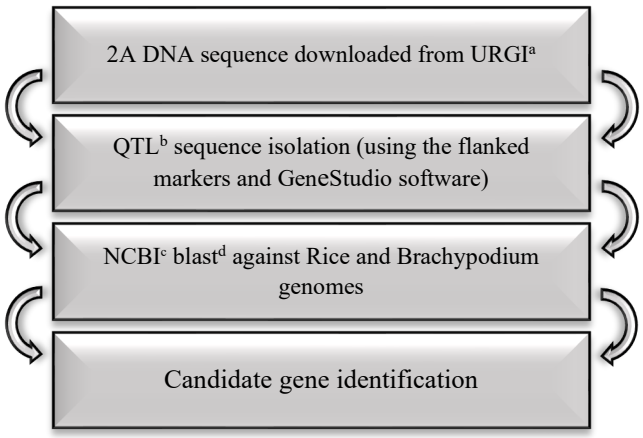


Figure 1: Candidate genes identification in specific and stressed environments

Subsequently, the derived genomic sequences (280bp) were served as query to identify the candidate genes using Blast search homology in NCBI (National Center for Biotechnology Information, the USA), see Figure 2. The Venn diagram was used to present the results.

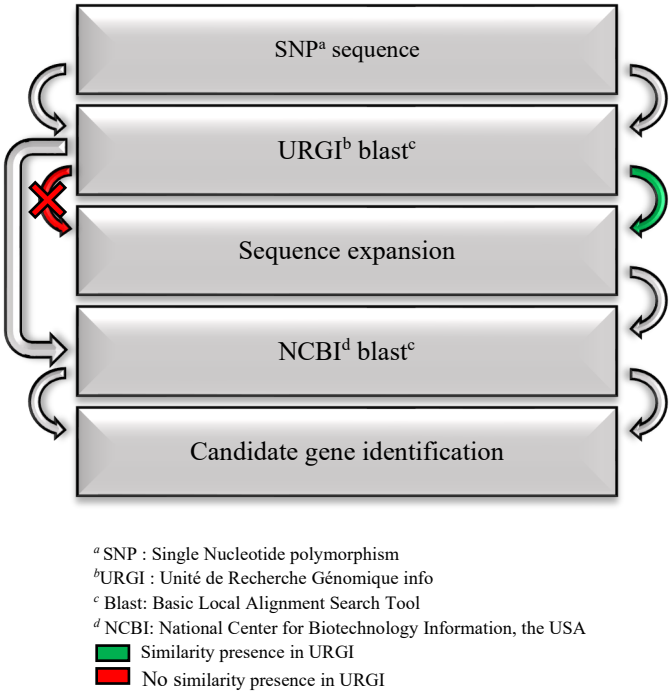


Figure 2: Candidate genes identification in multi-environmental analysis

More details about the candidate genes identified in combined and in specific environments QTLs analysis are shown in supplementary materials S1 and S2, respectively.

Table 1: QTL detected on 2A chromosome with the environmental conditions, QTL peak positions in cM, QTL regions interval size in centimorgan (cM) and in Mega bases (Mb) and with QTLs region flanking markers.

Environment	Environmental Conditions	Peak position (cM)	QTL[a] interval size		Flanking markers	
			cM	Mb	Left marker	Right marker
EP[b]	Cold	142	5.02	1.33	wsnp[f]_2280363	wmc[g]177bp190
INC[c]	Drought (vetch incorporation)	046	5.90	2.64	wsnp[f]_998898	cnl[h]127bp435
RF[d]	Drought	270	3.59	4.57	wsnp[f]_1000185	gwm[i]312bp189
RQ[e]	Heat	063	6.49	0.10	wsnp[f]_2276833	gwm[i]636bp111

[a]QTL: Quantitative Trait Loci

[b]EP: Early planting at Tel Hadya / Syria (ICARDA main station), 36°0.1'N 36°56'E, 284m, the sowing is in mid-October, 330mm of rainfall and 70mm of irrigation at sowing, the temperatures were 2 °C at Tillering and 20 °C at flowering

[c]INC: Incorporated in field (vetch biomass incorporation in soil) at same station, the sowing date in Mid-November with 330mm as rainfall, the temperatures were 8 °C at tillering and 24 °C at flowering

dRF: Rainfed (vetch above biomass harvested, no incorporation in soil) at the same station, sowing in Mid-November, the rainfall temperatures are also the same

eRQ: Raqqa station (35°57' N 39°0'E, 295m) with chickpea rotation, sowing in Mid-November, the rainfall was 150mm and 450 mm of irrigation, were 10 °C at tillering and 30 °C at flowering

[f]SNP: Single Nucleotide Polymorphism

[g]wmc: Wheat Microsatellite Consortium (SSR)

[h]cnl: Cornell University (EST-SSR)

[i]gwm: Gatersleben Wheat Microsatellite (SSR)

3. Results

3.1. *Candidate genes detection in specific and stressed environments*

The search for candidate genes in the QTLs corresponding to the environments (dry / cold / heat) show the presence of 407 candidate genes in Rice as in *Brachypodium* databases. Moreover, in the Rice database 139 genes identified were retrotransposons and transposons that have showed no correspondence in *Brachypodium* database.

3.1.1. Dry environment

As for the dry environments (RF and INC), the QTLs detected in these environments (Table 2) spanned on 4.57 Mb for RF and 2.64 Mb for INC. In RF (dryland) were identified 147 candidate genes and in INC (dryland with incorporation of vetch crop) 376 candidate genes, including genes involved in stress tolerance (biotic and abiotic stress) and in other cellular processes, the retrotransposons and transposons were also present in the two QTLs with a high frequency especially in the RF environment (Supplementary: Table S1). Also, it was noticed that the incorporation of legume vetch in the soil increased the expression of candidate genes. This expression is higher in dryland with vetch incorporation in soil (INC) than in dryland without vetch incorporation in the soil (RF). This may be due to increased availability of nitrogen and water through the organic matter incorporation in the soil.

3.1.2. Cold environments

In the cold environment (EP) characterized by low temperature during the tillering developmental stage (Table 2), the QTL detected was spanned on 1.33 Mb, it contained 17 candidate genes among those involved in plant development and abiotic/biotic stress tolerance: as Germin-like protein 8-14, protein far1-related sequence 5, putative FBD-associated F-box protein, Photosystem II 44 kDa protein and others. The retrotransposons and transposons are also present in this QTL (Supplementary: Table S1).

Table 2: The environmental conditions, the number of candidate genes identified for the four QTL linked to grain yield of the four stressed environments (EP, INC, RF and RQ) with the QTL interval size, and the number of candidate genes identified for Rice and *Brachypodium*

Environment	**Environmental Conditions**	**QTL[a] Interval size (Mb)**	**Candidates genes number**	
			Rice	***Brachypodium***
EP[b]	Cold	1.33	017	012
INC[c]	Drought (vetch incorporation)	2.64	147	113
RF[d]	Drought	4.57	376	277
RQ[e]	Heat	0.10	006	005

[a]QTL: Quantitative Trait Loci

[b]EP: Early planting at Tel Hadya / Syria (ICARDA main station), 36°0.1'N 36°56'E, 284m, the sowing is in mid-October, 330mm of rainfall and 70mm of irrigation at sowing, the temperatures were 2 °C at Tillering and 20 °C at flowering

[c]INC: Incorporated in field (vetch biomass incorporation in soil) at same station, the sowing date in Mid-November with 330mm as rainfall, the temperatures were 8 °C at tillering and 24 °C at flowering

[d]RF: Rainfed (vetch above biomass harvested, no incorporation in soil) at the same station, sowing in Mid-November, the rainfall temperatures are also the same

[e]RQ: Raqqa station (35°57' N 39°0'E, 295m) with chickpea rotation, sowing in Mid-November, the rainfall was 150mm and 450 mm of irrigation, were 10 °C at tillering and 30 °C at flowering

Table 3: SNPs annotated and not annotated in A and B genomes

Genome	SNPs			
	Annotated	Not annotated	Total	Percentage of SNPs annotated (%)
1A	15	24	39	38.46
2A	17	63	80	21.25
3A	6	33	39	15.38
4A	8	31	39	20.51
5A	1	11	12	08.33
6A	5	14	19	26.31
7A	7	46	53	13.20
A genome	**59**	**222**	**281**	**20.99**
1B	17	35	52	32.69
2B	16	63	79	20.25
3B	2	10	12	16.66
4B	9	64	73	12.32
5B	5	28	33	15.15
6B	13	29	42	30.95
7B	1	10	11	09.10
B genome	**63**	**239**	**302**	**20.86**

3.1.3. Hot environments

Further, the QTL detected in the hot environment (RQ) which is characterized by high temperatures during the anthesis developmental stage compared with the other environments (Table 2), the QTL region was spanned on a small interval of 0.1Mb that contained 6 genes, the first one is an LRR receptor-like serine/threonine-protein kinase gene involved in plant resistance to pathogens and the others are a retrotransposons and transposons (Supplementary: Table S1).

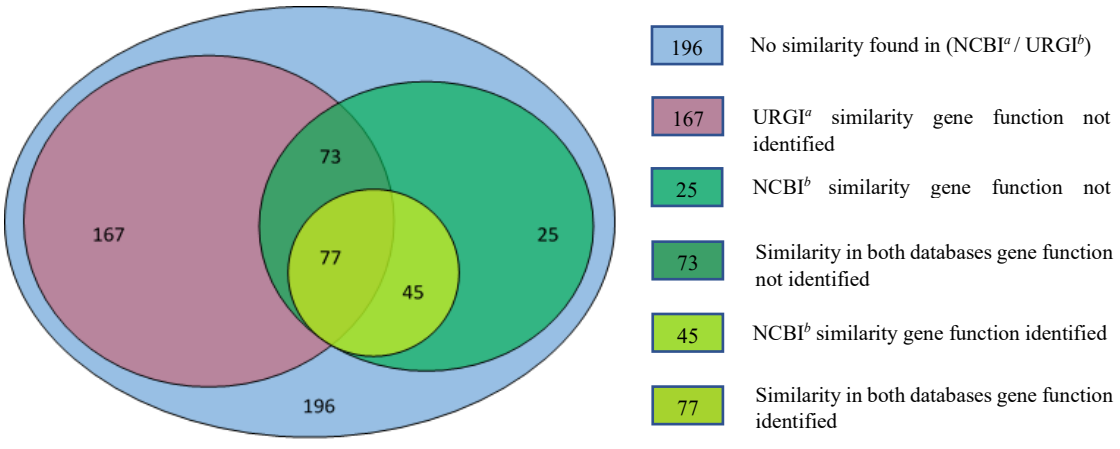
3.2. *Mapping of SNPs to identify candidate genes*

Of the mapped SNPs, 80 were mapped on 2A. Additional information for other chromosomes is shown in Table 3.

As for the candidate genes number, the 2A along 1B harbored the highest number (17), while in terms of annotation percentage for other chromosomes see Table 3.

The SNPs sequences availability provides the possibility of identifying gene loci contributing to grain yield.

Among the 583 SNP sequences selected, 122 sequences were annotated. the candidate genes involved in various growth, development processes, and stress tolerances have been identified, of which 77 sequences were enlarged with URGI database and annotated with NCBI. Whereas the remaining 45 SNPs sequences were annotated directly with NCBI, as no-similarity presence was available to enlarge in URGI database (Figure 3). Several sequences did not show any similarity in URGI and NCBI databases



[a]URGI : Unité de Recherche Génomique info
[c]NCBI: National Center for Biotechnology Information. The USA

Figure 3: Annotation and similarity study on Venn diagram of SNP.

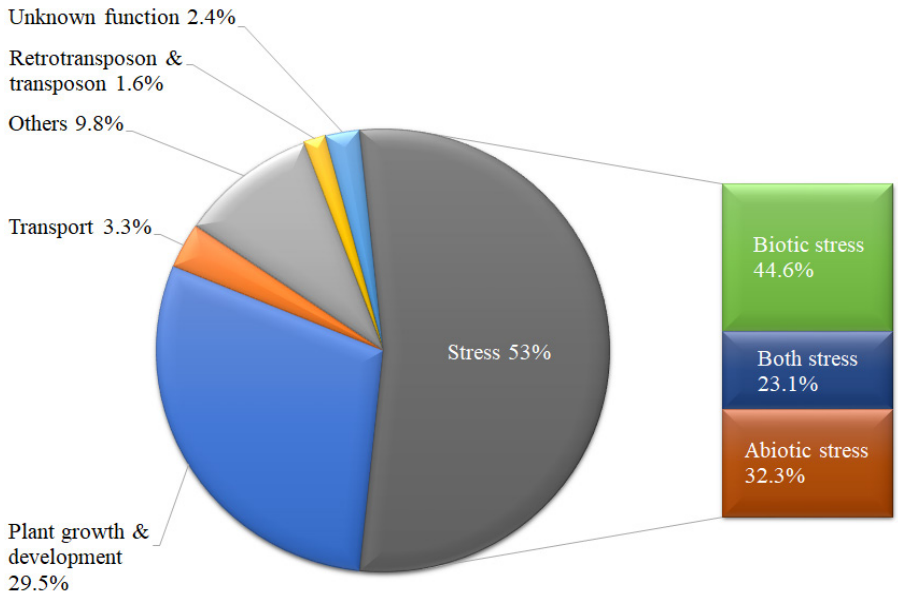


Figure 4: Functional classification of candidate genes co-located with QTLs linked to grain yield.

The 583 SNPs marked on Venn diagram (Figure 3) showed that 77 sequences were annotated (candidate genes identified) with similarity presence with both databases; while 45 sequences were annotated with similarity presence only in NCBI. Further, 73 were not annotated but had similarity in both databases. Also 167 SNP had similarity only in URGI and 25 only in NCBI, consequently both sets of SNPs were not annotated). The remaining196 sequences were not annotated, showing no similarity in the two databases.

The analysis with bioinformatic tools showed that 53% of candidate genes found were related to stresses tolerance, with 44.6% of them to biotic stress [powdery mildew, stem rust (sr32), leaf rust (lr10, lr34)], 32.3% to abiotic stress (cold heat, drought, salinity) and 23.1% to both stresses; 29.5% to plant development and growth; 3.3% to cell transport, and 1.6% to retrotransposon and transposon; and only 2.4% were with unknown function.

Further, 9.8% were involved in other cellular processes (secondary metabolism, pigments, herbicide detoxification, wall formation, etc.), see Figure 4. More informations about the SNPs annotation and candidate gene functions were shown in the supplementary table (S2).

Some candidate genes are involved in several processes, such as the basic helix-loop-helix (bHLH) proteins pertaining to the superfamily of transcription factors (TFs) and linked with wsnp_3026224. These proteins have been linked with the regulation of diverse biological processes as growth, development, and response to diverse stresses.

Moreover, in the 4BS, 57.1% of candidate genes found, were linked to stresses tolerance and 42.9% in plant development and growth.

4. Discussion

4.1. *Candidate genes detection in specific and stressed environments*

We found among 9 QTLs related to grain yield, four were located on 2A chromosome, these QTLs were detected in diverse stressed environments: EP characterized by cold, INC and RF by drought, and RQ by heat. These results are in accordance with several studies on grain yield QTLs located on 2A chromosome [24, 32, 33].

4.1.1. Dry environment

In the dry environment (INC), among the candidate genes were serpin-Z5, Probable glucuronosyltransferase Os04g0103100, Chalcone synthase 1, Fatty acid amide hydrolase isoform X1, Amino acid permease 3 (Supplementary: Table S1). This is in the agreement with the study of [30] where the transcriptomic pattern under progressive water stress of the two parents: Lahn and Cham1 (parents of the current studied population) and one recombined inbred line where the candidate genes cited above showed an overexpression. Such as the candidate gene coding for the mitochondrial metalloendopeptidase OMA1 identified in the QTL of the dry environment RF (Supplementary: Table S1). This was also confirmed by [30]. Similarly, the F-box and associated protein family were overexpressed in the dry environments RF and INC (Supplementary Table S1), but under cold stress was downexpressed [35]. Some genes of this large family under abiotic stress may be over or underexpressed depending on the stress [36]. As for the candidate gene coding for the Hydroxyproline-rich glycoprotein-like identified in the QTLs detected in the dry and cold environments: RF, INC, and EP (Supplementary: Table S1), this gene was associated with drought

tolerance in barley [37] and with the resistance to bacterial pathogen in rice [38].

Further, other candidate genes identified in the dry environment (RF): GDSL esterase/lipase EXL3; the cold environment (EP): UDP-glycosyltransferase 79-like isoform, and the glutathione S-transferase identified in both environments (Supplementary: Table S1). These genes play role in plant development and growth through the action on ABA homeostasis [39]; germination and growth [40]; detoxification and plant pathogen resistance [41].

4.1.2. Cold environment

In the cold environment (EP), among the candidate genes composing the QTL detected: 5'-3' exoribonuclease 3 isoform X1, NADH dehydrogenase; as for the QTL of dry environment (RF), it contains the genes: starch synthase3, chloroplastic/amyloplastic and a photosystem I P700 chlorophyll an apoprotein A1 candidate gene. While the gene coding for the Photosystem II 44 kDa protein were identified in the QTLs for the cold (EP) and dry (RF) environment, as for the ankyrin repeat family protein-like it was found in the INC environment (Supplementary Table S1). All these candidate genes expression was altered under water stress [30]. Some candidate genes identified in QTLs for INC and RF environments like zinc finger family genes play an important role in physiological traits of yield potential under heat stress [27]. Moreover, in the QTL for the RF environment, the gene coding for BTB/POZ and MATH domain-containing protein 2 was highly repeated in this QTL, this gene is involved in heat tolerance and in plant growth and development [42].

4.1.3. Hot environment

In the hot environment RQ, six candidate genes were identified, one coding for LRR receptor-like serine/threonine-protein kinase RCH1 involved in resistance to pathogens [43, 44]. The remaining candidate genes were belonging to the family of retrotransposon/transposase. This gene family was also present in the other QTLs especially in RF (dry environment) with a high frequency. The abundance of this gene family presence can be explained by their involvement in stress conditions, [30] reported that the activity of these genes was altered under water stress and by the fact that the retrotransposon/transposon may represents 80 to 90% of wheat genome [45].

4.2. *Detection of candidate genes in other chromosomes*

Additionally, the presence of the wsnp_993293 SNP marker in the chromosome 1A and its linkage with the candidate gene implicated in powdery mildew resistance, confirms earlier mention on one parent of the population (Lahn) that carries moderate resistance to powdery mildew. Additionally, the wsnp_2265986, located in the 1B chromosome and linked to 3-ketoacyl-CoA synthase 6-like, this candidate gene plays also a role in abiotic (drought, heat and cold) and in biotic stress resistance. This reflects the genes carried by the second parent Cham1 for good drought and heat tolerance, osmotic adjustment, and proline content.

The candidate genes covered the entire genome and were linked to different biotic (powdery mildew, rust, etc.) and abiotic (drought, cold, heat, etc.) stress resistance / tolerance (Fig 4).

In the current work, a large part of candidate genes harbored by the 2A, in addition to 4B chromosome, were related to stress tolerance. They were also involved in plant growth and development. The 4BS results corroborate earlier studies [25, 46-50] on the importance of the 4BS and its contribution to grain yield in stressed environments.

5. Conclusion

The candidate genes found were linked to yield performance under drought and temperature extreme conditions. Also, this study showed the importance of 2A chromosomes for contrasting environments (drought, cold, and heat).

In this study specific and general candidate genes were detected. For the specific environments, the candidate genes in cold environment (EP) involved the genes of 5'-3'exoribonuclease 3 isoform X1 and NADH dehydrogenase. As for the dry environment (RF) the genes included were starch synthase3, chloroplastic/amyloplastic, and a photosystem I P700 chlorophyll. Further, in the dry environment (RF), the gene coding for BTB/POZ and MATH domain-containing protein2 was highly repeated in this environment. This gene is also effectuates heat tolerance, plant development and growth. Similarly, the ankyrin repeat family protein-like was found in the dry environment (INC), it contains the genes synthase3, chloroplastic/amyloplastic and photosystem I P700 chlorophyll. For the general candidate genes involved in diverse stressed environments, the following genes were identified: The Photosystem II 44 kDa protein were identified in the cold (EP) and the dry (RF) environments. Also, some of the candidate genes are operating in all 3 stresses, such as the zinc finger family genes. This gene also plays a significant role in physiology of yield potential under heat stress frequently.

5.1.1. Conflict of Interest

The authors declare no conflict of interest.

5.1.2. Acknowledgement

We would like to thank the International Center for Agricultural Research in Dry Area (ICARDA) for the financial support of this study and the support staff of the CIMMYT/ICARDA durum-breeding program.

5.1.3. References

[1] I. Elouafi, MM. Nachit, "A genetic linkage map of the Durum x *Triticum dicoccoides* backcross population based on SSRs and AFLP markers, and QTL analysis for milling traits" Theor Appl Genet, **108**(3), 401-413, 2004. https://doi.org/10.1007/s00122-003-1440-8

[2] MM. Nachit, "Durum wheat breeding for Mediterranean drylands of North Africa and West Asia. In: Rajram S, Saari EE, Hetel GP (eds) Durum Wheats" Challenges and Opportunities. CIMMYT, Ciudad Obregon, Mexico, **14**, 14–27, 1992.

[3] El Yadini, I. Kahama, M. Labhilili, A. El Yadini, M. Azeqour, "La technique TILLING comme nouvelle biotechnologie pour la protection et l'amélioration du blé dur (TILLING : A new biotechnological method for protection and improvement of durum wheat)" J Mater Env Sci, **4**, 931-934, 2013.

[4] M. Maccaferri, M. Sanguineti, S. Corneti, J. Ortega, M. Salem, J. Bort, E. DeAmbrogio, L. del Moral, A. Demontis, A. El-Ahmed, "Quantitative trait loci for grain yield and adaptation of durum wheat (*Triticum durum* Desf.) across a wide range of water availability" Genetics, **178**(1), 489-511, 2008. https://doi.org/10.1534/genetics.107.077297

[5] B. Heidari, BE. Sayed-Tabatabaei, G. Saeidi, M. Kearsey, K. Suenaga, P. Gulick, "Mapping QTL for grain yield, yield components, and spike features in a doubled haploid population of bread wheat" Genome, **54**(6), 517-527, 2011. https://doi.org/10.1139/g11-017

[6] A. Blanco, M. Bellomo, A. Cenci, C. De Giovanni, R. D'ovidio, E. Iacono, B. Laddomada, M. Pagnotta, E. Porceddu, A. Sciancalepore, "A genetic linkage map of durum wheat" Theor App Genet **97**(5-6), 721-728, 1998. https://doi.org/10.1007/s001220050948

[7] SA. Dura, MA. Duwayri, MM. Nachit, "Detection of molecular markers associated with yield and yield components in durum wheat (*Triticum turgidum* L. var. *durum* Desf.) under drought conditions" Afr J Agr, **8**(19), 2118-2128, 2013. https://doi.org/10.5897/AJAR11.2031

[8] MM. Nachit, I. Elouafi, A. Pagnotta. A. El saleh, E. acono, M. Labhilili, A. Asbati, M. Azrak, H. azzam, D. Benscher, M. Khairallah, JM. Ribaut, OA. Tanzarella, E. Porceddu, ME. Sorrells, "Molecular linkage map for an intraspecific recombinant inbred population of durum wheat (*Triticum turgidum* L. var. *durum*)". Theor App Genet, **102**, 177-186, 2001. https://doi.org/10.1007/s001220051633

[9] A. Alsaleh, FS. Baloch, M. Derya, M. Azrak, B. Kilian, H Özkan, MM. Nachit, "Genetic Linkage Map of Anatolian Durum Wheat Derived from a Cross of Kunduru-1149× Cham1". Plant Mol Biol Rep, **33**, 209-220, 2014. https://doi.org/10.1007/s11105-014-0749-6

[10] Mantovani, M. Maccaferri, MC. Sanguineti, R. Tuberosa, I. Catizone, P. Wenzl, B. Thomson, J. Carling, E. Huttner, E. DeAmbrogio, "An integrated DArT-SSR linkage map of durum wheat" Mol Breed, **22**(4), 629-648, 2008. https://doi.org/10.1007/s11032-008-9205-3

[11] M. Nagel, S. Navakode, V. Scheibal, M. Baum, MM. Nachit, M. Röder, A. Börner, "The genetic basis of durum wheat germination and seedling growth under osmotic stress" Biol Plantarum, **58**, 681–688, 2014. https://doi.org/10.1007/s10535-014-0436-3

[12] KJ. Haile, MM. Nachit, H. Karl, B. Ayele, SM. Röder, "QTL mapping of resistance to race Ug99 of Puccinia graminis f. sp. tritici in durum wheat (*Triticum durum* Desf.)" Mol Breed, **30**, 1479–1493, 2012. https://doi.org/10.1007/s11032-012-9734-7

[13] S. Dura, M. Duwayri, MM. Nachit, F. Al Sheyab, "Detection of molecular markers associated with yield and yield components in durum wheat (*Triticum turgidum* L. var. *durum*) under saline conditions" Crop Past Sci, 64(10), 957-964, 2013. https://doi.org/10.5897/AJAR11.2031

[14] RM. Poecke, M. Maccaferri, J. Tang, HT. Truong, A. Janssen, NJ. Orsouw, S. Salvi, MC. Sanguineti, R. Tuberosa, EA Vossen, "Sequence based SNP genotyping in durum wheat" Plant Biotech J, **11**(7), 809-817, 2013. https://doi.org/10.1111/pbi.12072

[15] D. Trebbi, M. Maccaferri, P. de Heer, A. Sørensen, S. Giuliani, S. Salvi, M. Sanguineti, A. Massi, E. van der Vossen, R. Tuberosa, "High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.)" Theor App Genet, **123**(4), 555-569, 2011. https://doi.org/10.1007/s00122-011-1607-7

[16] W. Zhang, S. Chao, F. Manthey, O. Chicaiza, JC. Brevis, V. Echenique, J. Dubcovsky, "QTL analysis of pasta quality using a composite microsatellite and SNP map of durum wheat" Theor App Genet, **117**(8) 1361-1377, 2008. https://doi.org/10.1007/s00122-008-0869-1

[17] A. Blanco, P. Colasuonno, A. Gadaleta, G. Mangini, A. Schiavulli, R. Simeone, AM. Digesù, P. De Vita, AM. Mastrangelo, L. Cattivelli, "Quantitative trait loci for yellow pigment concentration and individual carotenoid compounds in durum wheat" J Cereal Sci, **54**(2), 255-264, 2011. https://doi.org/10.1016/j.jcs.2011.07.002

[18] A. Blanco, G. Mangini, A. Giancaspro, S. Giove, P. Colasuonno, R. Simeone, A. Signorile, P. Vita, AM. Mastrangelo, L Cattivelli, "Relationships between grain protein content and grain yield components through quantitative trait locus analyses in a recombinant inbred line population derived from two elite durum wheat cultivars" Mol Breed, **30**(1), 79-92, 2011. https://doi.org/10.1007/s11032-011-9600-z

[19] A. Gadaleta, A. Giancaspro, S. Giove, S. Zacheo, G. Mangini, R. Simeone, A. Signorile, A. Blanco, "Genetic and physical mapping of new EST-derived SSRs on the A and B genome chromosomes of wheat" Theor App Genet, **118**(5), 1015-1025, 2009. https://doi.org/10.1007/s00122-008-0958-1

[20] D. Marone, G. Laidò, A. Gadaleta, P. Colasuonno, D. Ficco, A. Giancaspro, S. Giove, G. Panio, M. Russo, P. De Vita, L. Cattivelli, R. Papa, A. Blanco, "A high-density consensus map of A and B wheat genomes" Theor App Genet, **125**(8), 1619-1638, 2012. https://doi.org/10.1007/s00122-012-1939-y

[21] MA. Russo, DBM. Ficco, D. Marone, P. De Vita, V. Vallega, C. Rubies-Autonell, C. Ratti, P. Ferragonio, V. Giovanniello, N. Pecchioni, "A major QTL for resistance to soil-borne cereal mosaic virus derived from an old Italian durum wheat cultivar" J Plant Interact, **7**(4), 290-300, 2012. https://doi.org/10.1080/17429145.2011.640437

[22] J. Ahmad et al., "Study of Physio-chemical Properties of POSS/Mineral Oil based Nanofluids," in 2018 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET), 1–5, 2018. DOI: 10.1109/PGSRET.2018.8685972

[23] Z. Peleg, Y. Saranga, T. Suprunova, Y. Ronin, MS. Röder, A. Kilian, AB. Korol, T. Fahima, "High-density genetic map of durum wheat x wild emmer wheat based on SSR and DArT markers" Theor App Genet, **117**, 103–115, 2008. https://doi.org/10.1007/s00122-008-0756-9

[24] H. Tura, J. Edwards, V, Gahlaut, M. Garcia, Sznajder, U. Baumann, F. Shahinnia, M. Reynolds, P. Langridge, HS. Balyan, "QTL analysis and fine mapping of a QTL for yield-related traits in wheat grown in dry and hot environments" Theor App Genet, 1-19, 2019. https://doi.org/10.1007/s00122-019-03454-6

[25] D. Xu, W. Wen, L. Fu, F. Li, J. Li, L. Xie, X. Xia, Z. Ni, Z. He, S. Cao, "Genetic dissection of a major QTL for kernel weight spanning the Rht-B1 locus in bread wheat" Theor App Genet, **132**(11), 3191-3200, 2019. https://doi.org/10.1007/s00122-019-03418-w

[26] R. Mérida-García, G. Liu, S. He, V. Gonzalez-Dugo, G. Dorado, S. Gálvez, I. Solís, P. Zarco-Tejada, J. Reif, P. Hernandez P, "Genetic dissection of agronomic and quality traits based on association mapping and genomic selection approaches in durum wheat grown in Southern Spain" PloS one **14**(2), e0211718, 1-24, 2019. https://doi.org/10.1371/journal.pone.0211718

[27] S. Pradhan, MA. Babar, G. Bai, J. Khan, D. Shahi, M. Avci, J. Guo, J. McBreen, S. Asseng, S. Gezan, et al, "Genetic dissection of heat-responsive physiological traits to improve adaptation and increase yield potential in soft winter wheat" BMC Genomics **21**, 315, 1-15, 2020. https://doi.org/10.1186/s12864-020-6717-7

[28] M.S. Rahman, K.J. Linsell, J.D. Taylor, M.J. Hayden, N.C. Collins, K.H. Oldach, "Fine mapping of root lesion nematode (*Pratylenchus thornei*) resistance loci on chromosomes 6D and 2B of wheat" Theor App Genet **133**, 635-652. 2020. https://doi.org/10.1007/s00122-019-03495-x

[29] J. Yu, Y. Miao, S. Yang, Z. Shi, N. Miao, M. Ding, H. Zhang, Y. Jiang, J. Rong, "Identification and mapping of a photoperiod response gene (QPpd.zafu-4A) on wild emmer wheat (*Triticum turgidum* L.) chromosome 4AL" Euphytica **215**, 146, 1-12, 2019. https://doi.org/10.1007/s10681-019-2469-3

[30] D. Habash, M. Baudo, M. Hindle, S. Powers, M. Defoin-Platel, R. Mitchell, M. Saqi, C. Rawlings, K. Latiri, J. Araus, et al, "Systems responses to progressive water stress in durum wheat" PloS one, **9**, e108431, 1-21, 2014. https://doi.org/10.1371/journal.pone.0108431

[31] I. Farouk, F. Gaboun, Z. Kehel, A. Alsaleh, B. Belkadi, I. Elouafi, J. Motowaj, A. Filali maltouf, MM. Nachit, "Dissection of QTL linked to grain yield and identification of candidate genes involved in grain yield formation using comparative SNP sequences analysis," in 2020 IEEE the first International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, pp. 1-6, 2020. https://doi.org/10.1109/IRASET48871.2020.9091989

[32] S. Li, J. Jia, X. Wei, X. Zhang, L. Li, H. Chen, Y. Fan, H. Sun, X. Zhao, T. Lei, et al, "A intervarietal genetic map and QTL analysis for yield traits in wheat" Mol Breeding **20**, 167-178, 2007. https://doi.org/10.1186/s12870-018-1308-3

[33] S. Sukumaran, M.P. Reynolds, C. Sansaloni, "Genome-Wide Association Analyses Identify QTL Hotspots for Yield and Component Traits in Durum Wheat Grown under Yield Potential, Drought, and Heat Stress Environments". Front Plant Sci **9**, 81, 1-16, 2018. https://doi.org/10.3389/fpls.2018.00081

[34] P. Colasuonno, MA. Maria, A. Blanco, Gadaleta, "Description of durum wheat linkage map and comparative sequence analysis of wheat mapped DArT markers with rice and *Brachypodium* genomes" BMC Genet, **14**, number (114), 1-9, 2013. https://doi.org/10.1186/1471-2156-14-114

[35] C. Xin, R. Hou, F. Wu, Y. Zhao, H. Xiao, W. Si, ME. Ali, L. Cai, J. Guo, "Analysis of cytosine methylation status in potato by methylation-sensitive amplified polymorphisms under low-temperature stress" J Plant Biol, **58**, 383-390, 2015. https://doi.org/10.1007/s12374-015-0316-1

[36] M. Jain, A. Nijhawan, R.Arora, P. Agarwal, S. Ray, P. Sharma, S. Kapoor, AK. Tyagi, JP. Khurana, "F-Box Proteins in Rice. Genome-Wide Analysis, Classification, Temporal and Spatial Gene Expression during Panicle and Seed Development, and Regulation by Light and Abiotic Stress" Plant Physiol, **143**, 1467-1483, 2007. https://doi.org/10.1104/pp.106.091900

[37] SF. Abou-Elwafa, "Identification of genes associated with drought tolerance in barley" Biol Plantarum **62**, 299-306. 2018. https://doi.org/10.1007/s10535-017-0765-0

[38] IS. Kumar, N. Zaharin, K. Nadarajah, "In silico identification of resistance and defense related genes for bacterial leaf blight (BLB) in rice" J Pure Appl Microbiol, **12**, 1867-1877, 2018. https://dx.doi.org/10.22207/JPAM.12.4.22

[39] Z. Liu, J-P. Yan, D-K. Li, Q. Luo, Q. Yan, Z-B. Liu, L-M. Ye, J-M. Wang, X-F. Li, Y. Yang, "UDP-Glucosyltransferase71C5, a Major Glucosyltransferase, Mediates Abscisic Acid Homeostasis in Arabidopsis" Plant Physiol, **167**, 1659-1670, 2015. https://doi.org/10.1104/pp.15.00053

[40] C-P. Lai, L-M. Huang, L-FQ. Chen, M-T. Chan, J-F. Shaw, "Genome-wide analysis of GDSL-type esterases/lipases in Arabidopsis" Plant Mol Biol, **95**, 181-197, 2017. https://doi.org/10.1007/s11103-017-0648-y

[41] G. Gullner, T. Komives, L. Király L, P. Schröder, "Glutathione S-Transferase Enzymes in Plant-Pathogen Interactions" Front Plant Sci **9**, 1836, 1-19, 2018. https://doi.org/10.3389/fpls.2018.01836

[42] K. Morimoto, N. Ohama, S. Kidokoro, J. Mizoi, F. Takahashi, D. Todaka, J. Mogami, H. Sato, F. Qin, J-S. Kim, et al, "BPM-CUL3 E3 ligase modulates thermotolerance by facilitating negative regulatory domain-mediated degradation of DREB2A in *Arabidopsis*" Proceedings of the National Academy of Sciences **114**(40), E8528-E8536, 2017. https://doi.org/10.1073/pnas.1704189114

[43] S. Padmarasu, DJ. Sargent, A. Patocchi, M. Troggio, P. Baldi, G. Linsmith, L. Poles, M. Jänsch, M. Kellerhals, S. Tartarini, R. Velasco, "Identification of a leucine-rich repeat receptor-like serine/threonine-protein kinase as a candidate gene for Rvi12 (Vb)-based apple scab resistance" Mol Breeding, **38** ,73, 1-14, 2018. https://doi.org/10.1007/s11032-018-0825-y

[44] D. Pradhan, D. Mathew, SK. Mathew, PA. Nazeem, "Identifying the markers and tagging a leucine-rich repeat receptor-like kinase gene for resistance to anthracnose disease in vegetable cowpea [*Vigna unguiculata* (L.) *Walp*.]" J Hortic Sci Biotech, **93**, 225-231, 2018. https://doi.org/10.1080/14620316.2017.1362962

[45] V. Jamilloux, J. Daron, F. Choulet, H. Quesneville, "De novo annotation of transposable elements: tackling the fat genome issue" Proceedings of the IEEE, **105**, 474-481, 2016. https://10.1109/JPROC.2016.2590833

[46] A.A. Diab, R. Kantety, N. Ozturk, D. Benscher, MM. Nachit, M. Sorrells, "Drought-inducible genes and differentially expressed sequence tags associated with components of drought tolerance in durum wheat" Sci Res Essay, **3**, 009-026, 2008.

[47] D. Habash, Z. Kehel, M. Nachit, "Genomic approaches for designing durum wheat ready for climate change with a focus on drought" J Exp Bot, **60**(10), 2805-2815, 2009. https://doi.org/10.1093/jxb/erp211

[48] MM. Nachit, "Association of grain yield in dryland and carbon isotope discrimination with molecular markers in durum (*Triticum turgidum* L. var. *durum*)" In Proceedings of 9th International Wheat Genetics Symposium, Saskatoon, ed. A.E. Slinkard, Canada, 218–223, 1998.

[49] M.M. Nachit, I.Elouafi, "Durum wheat adaptation in the Mediterranean dryland: Breeding, stress physiology, and molecular markers" In Challenges and strategies for dry-land agriculture, eds. S.C. Rao and J. Ryan, CSSA. Madison, WI: Crop Science of America and American Society of Agronomy, **32**, 203–218, 2004. https://doi.org/10.2135/cssaspecpub32.c13

[50] J. Li, S. Wen, C. Fan, M. Zhang, S. Tian, W. Kang, W. Zhao, C. Bi, Q. Wang, S. Lu, et al, "Characterization of a major quantitative trait locus on the short arm of chromosome 4B for spike number per unit area in common wheat (*Triticum aestivum* L.)" Theor Appl Genet, **133**, 2259–2269, 2020. https://doi.org/10.1007/s00122-020-03595-z

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 1030-1039 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

[1]ASTESJ

ISSN: 2415-6698

Simulated IoT Based Sustainable Power System for Smart Agriculture Environments

Shahenaz S. Abou Emira[1,2], Khaled Y. Youssef [3,*], Mohamed Abouelatta[1]

[1]*Faculty of Engineering, Electronics and Electrical Communication, Ain Shams University (ASU), Cairo, 39827, Egypt*

[2]*Faculty of Engineering, Electronics and Electrical Communications, October University for Modern Sciences and Arts (MSA), Cairo, 39827, Egypt*

[3]*Faculty of Navigation Science and Space Technology, Beni-Suef University (BSU), Beni-suef, 62511, Egypt*

ARTICLE INFO

Article history:
Received: 08 January, 2021
Accepted: 11 February, 2021
Online: 16 February, 2021

Keywords:
Agriculture
IoT
Load Scheduling
Solar power
Sustainability
State of charging

ABSTRACT

In vital energy applications especially the agricultural environments, the service of adaptive power utilization plays an essential role in facilitating the usage of Internet of Things systems. Such environments are distinguished by the large range of lands where most of the region lacks the commercial power lines. Reaching some high or deep sensing points is also difficult in such environments. The adaptive power system will pave the way to make a smart service for users to create a platform for real time interaction. It can enhance the reliability, stability and sustainability of power supply, and provide more humanized and various intelligent services for the users. The usage of adaptive power can be improved effectively using IoT technology with its strong data processing and reliable communication. In this paper, an algorithm is proposed to offer a sustainable power service for smart agriculture system to guarantee continuous system operation. It mainly rely on controlling the load demands and managing the renewable energy. A model is built on Matlab that governs the proposed algorithm and the results of the simulation are discussed.

1. Introduction

This paper is an extension of work originally presented in 15th International Computer Engineering Conference (ICENCO) [1]. The food demand increased due to the huge growth in population so it is essential to increase the food production and try to avoid any production losses. Agriculture has a major impact on food production so monitoring the plants' performance is very important. In order to avoid any crop losses, agro systems were designed. Remote sensors were used in these systems to measure light, humidity, soil moisture and temperature. The measured data will indicate whether the plants' performance is good or not and the data will be collected and sent to the user for further action[2, 3]. IoT is a new technology that is used recently in most of the smart systems. Using IoT technology in agro systems allow the plants to express its needs only when it is needed which will allow the efficient usage of resources. There are some challenges that appears in IoT systems, energy optimization is one of the main challenges that faces IoT. It is a major challenge due to the high number of devices that exist in the network which need high energy to keep it active for long time. It is essential to study this challenge as wireless sensors are commonly used in precision agriculture in remote areas and the agro systems depend mainly on sensors' power so battery drainage will cause the system to stop which isn't desired [4, 5].

The life time of sensor nodes needs to be long in order for the network to be active and to avoid battery drainage. In this paper, an energy management technique based on intelligent load scheduling is introduced for agro-applications. The battery state of charging and the solar power are the main parameters used in the proposed algorithm to activate or deactivate the loads. Based on the loads' priority and the availability of power provided by the battery or solar, loads are activated for specific time. The loads' priority is standard defined as the applications that need to be operating due to criticality are defined based on the standards of agro.

2. Literature Review

Researchers have implemented and introduced some systems using various methods and sensors to monitor the status of plants' health. It improved the agricultural production and helped in

*Corresponding Author: Khaled Y. Youssef, khalid_youssif@yahoo.com

preventing crop losses. In [6], the author illustrated the demonstration of a smart plant monitoring system. The system was enabled to discover any changes occurred in the measurements of light, moisture level and temperature. The plant received its required irrigation and illumination using a machine based curation. An Android device was used by the user in order to enable the user to override an operation which is a machine curated and in this case it is shifted to user based curation instead of machine based curation. In [7], the author explained a system that uses IoT technology in order to control and monitor an agricultural production. The sensors' data was gathered in the monitoring system from IoT devices and then stored at the database of the cloud in order to be accessible by the users. The users could control the actuators in the controlling system over the internet with the usage of IoT devices.

The main source of power for the systems presented previously is the sensors' batteries so systems that uses energy efficiently are studied. Energy plays a significant role in monitoring the environment of agriculture. Researchers presented various algorithms which assure the efficient usage of energy using energy harvesting, load scheduling and power reduction techniques. In [8], the author presented a method used in wireless sensor network for managing power efficiently. The system model consisted of three parts: sensing unit, transmitter unit , and power unit. Energy harvesting was used in the power unit as the power source used for charging the battery is the solar energy. Sending and collecting the data was the responsibility of the other two units. The microcontroller is utilized to monitor the sensed signal and if any disorder detected, it would be transmitted to the receiver.

In [9], the author concentrated on elongating the sensor nodes' life time by minimizing the consumption of energy. A solar cell was utilized to power the system to assure energy sustainability. The minimization of power consumption was achieved by proposing two methods of power reduction. Sleep/wake based on duty cycling was utilized and the second method was integrating the redundant data of the soil moisture with sleep/wake scheme. This research introduced the Sleep/wake on redundant data (SWORD).

In [10], the author presented an algorithm based on devices scheduling in order to optimize and reduce the electricity cost. Two methods were tested for load scheduling optimization either without the usage of renewable energy or with the usage of it. The results of the implementation showed that the cost is reduced up to 53% when combining the proposed algorithm with renewable energy and in case no renewable energy is combined with the algorithm, the cost is reduced up to 40%.

In [11], the author illustrated an algorithm that depends on electricity cost changes and renewable sources real time output in order to schedule the loads. The scheduling and management of energy was established according to the exchange of sensor data and control demands. It would update the real time output of the renewable sources and determine the devices' priority. After the sensed data was received, the energy management unit in the system updated the devices' priority status and the output of renewable sources. The energy management unit would transfer the devices to on or off status if required depending on the energy management and devices scheduling.

In [12], the author discussed various ways to utilize renewable energy in order to assure the efficient usage of energy, ,cost reduction, and handling loads. The scheduling of loads in the used algorithm was depending on time and overload management using multilevel priority. Based on the need of load scheduling during peak demand hour, a duration of time was determined. The loads with medium and high priorities would be activated during this time and the load with low priority would be deactivated. All load would be deactivated in case the limit was exceeded.

3. Proposed Architecture for Energy Efficient Agro Systems

As shown in Figure 1, the proposed architecture consists of four sensors: light sensor, color sensor, soil moisture sensor and temperature/humidity sensor. The proposed architecture is used for monitoring plants' performance in agro systems. The data measured by the soil moisture sensor and temperature/humidity sensor will be sent to Central Processing Server (CPS) to decide the proper time for irrigation after comparing it with the threshold values. Regarding the other two sensors, the user will take the action required based on the received data that were measured by these sensors in case there were any up normal values detected at the CPS. The goal is to ensure the continuous operation of this system by using energy efficiently to increase the life time of the system. Figure 2 shows a proposed architecture for sensor area network (SAN1) using energy efficient technique. It depends mainly on two parts:

3.1. Solar panel or battery used as a power source

- The source of power during day is the solar panel. It is used for supplying the loads during day as it is working and the power needed by the loads is available. Some loads will be switched off in case that the solar power isn't enough for suppling all loads and the rest will be on depending on the available solar power.
- The source of power during night is the battery. The active loads will be supplied by the battery and some loads will be off during night as these loads will not highly affect the system's operation.

3.2. The loads (sensors and actuators) which are divided into categories according to priority.

- Load 1 (High priority): the high priority loads highly affect the system's operation. This load must be on most of the time so the deactivation will be during night only in case that the state of charging of the battery is less than the threshold value.
- Load 2 (Medium priority): if the power is enough for supplying medium and high priority loads, this load will be active. It will be deactivated when the power isn't enough to supply all loads and this will occur mostly during night.
- Load 3 (Low priority): it doesn't have a huge impact on the system's operation so it can be deactivated during night. The user can activate this load upon his request if needed during night after checking the battery state of charging if it is within threshold values or not.

The soil moisture sensor is considered a high priority load due to its high criticality in precise agriculture applications as stated in

[13] .The soil moisture sensor is used in the proposed system to change the concept of scheduled irrigation to be on demand irrigation as its measured values help in deciding whether plants needs to be watered or not and based on these values irrigation takes place .The light sensor is considered a medium priority load in the proposed system due to its medium criticality as it needs to be active at specific time as it is used in the proposed algorithm to indicate whether the weather is sunny or cloudy and to indicate also whether it is day or night. Accordingly, the algorithm will switch to the power source that can be used in the current situation to supply the loads. The water quality sensor is considered a low priority load as irrigation is applied mostly during day so there is no need to schedule it to work during night but in the proposed algorithm it can be activated during night if needed upon the user request.

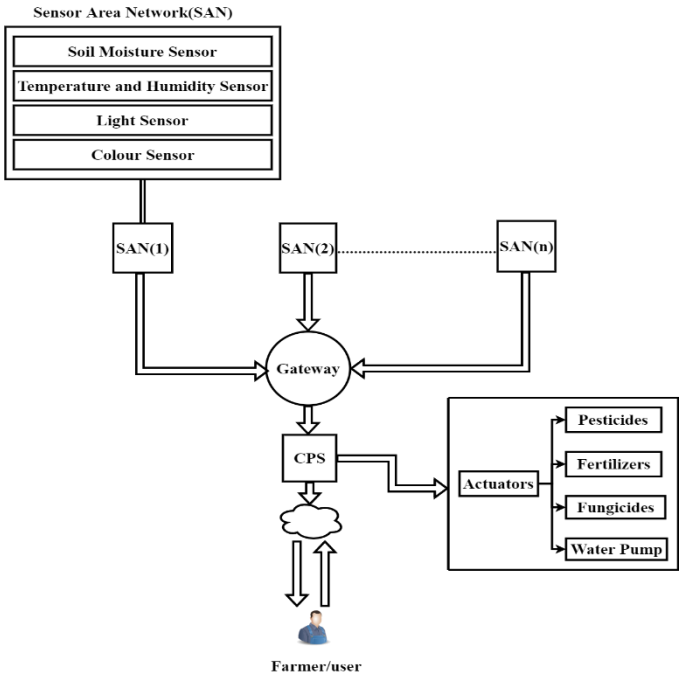


Figure 1: Architecture for Monitoring Agro Systems

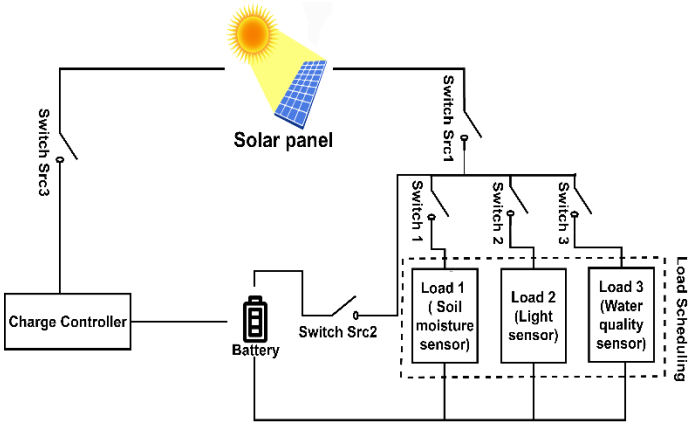


Figure 2: Representing Energy Efficient Algorithm in a Sensor Area Network (SAN 1)

4. Proposed Energy Efficient Algorithm

The proposed algorithm uses energy efficiently in agricultural applications to elongate the sensor nodes life time. The agro devices' dormant mode will be synchronized and controlled by the proposed algorithm based on the change of day/night. This algorithm doesn't depend on the marketable power supplies so it will guarantee a sustainable environment. The battery drainage issue will be solved to some extent which will enable a sustainable power system. The proposed energy efficient algorithm is dependent on the load scheduling technique and solar energy. There are two other methods for powering the system either by using battery only or by using battery along with load scheduling.

The proposed algorithm is compared with these two methods. In the first method as shown in Figure 3, there is no enough power for supplying the loads as they are active for only five consecutive hours. In the second method, load scheduling technique is used which saves more power therefore the loads are active for longer time as shown in Figure 3. However. The problem of battery drainage still exists and the battery needs to be replaced but it is not easy to replace it in agro-applications. Replacing the battery is not necessary in the proposed algorithm as the solar panel is used during day to recharge the battery. The loads are supplied by the battery only during night. The solar panel is used to supply the loads during day therefore the battery state of charging is maintained for longer time in the proposed algorithm as shown in Figure 3. As a result of using the proposed algorithm, the life time of the sensor nodes is increased due to the increase of the battery life time.

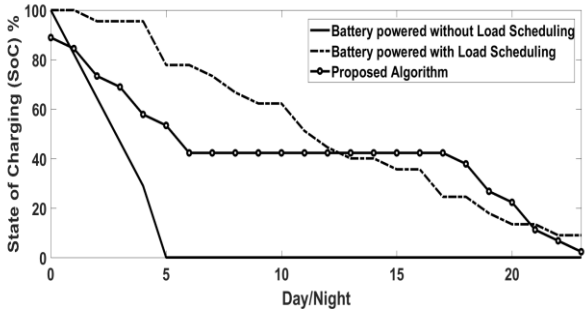


Figure 3: State of Charging versus Day/Night for The Two Power Methods Compared with The Proposed Algorithm.

The two main parameters used in the proposed algorithm are the availability of power and the criticality index of the load. The first parameter is represented as solar irradiance or State of Charging (SoC) of the battery. The second parameter indicates whether the load is critical or not to operate in the non-scheduled time. Based on these two main parameters, the agricultural ecosystems loads are transferred to on or off state. Intelligent load Scheduling and solar power are combined in this algorithm to efficiently implement the system using IoT.

Figure 4 represents a flow chart for a combined energy efficient technique using intelligent load scheduling and solar energy. The algorithm initially checks the capacity of solar power during day to decide whether to activate or deactivate the load. It is proven in the simulation that the idle duration varies based on the availability of solar power. For example, when the solar irradiance is reduced from 1000 W/m^2 to 500 W/m^2, the idle duration is increased by Δt. The battery is the source of power during night and it takes the lead in the absence of the sun to supply all the needed loads.

The loads are transferred to on or off state based on their criticality index as the loads are classified according to priority. The loads with high and medium priority are activated if SoC of battery is more than or equal to 75% and the unnecessary loads are deactivated. According to the availability of power, the loads' idle time is tuned and calculated. By time, the idle duration increases as a function of the battery drainage. If SoC is below 40%, the load with high priority is activated and the load with medium priority is deactivated and the idle duration increases based on the drainage rate of the battery. The inactive loads can be transferred to on state if needed during night based on user request.

The idle duration levels during day and night are shown in Table 1. The first level is selected during day as it is devoted for the dynamic solar irradiance scenario in which it varies during day hours from 100 W/m^2 to 1000 W/m^2. The other levels are devoted for the battery SoC variations. According to the availability of power, the idle duration for each load changes by Δt. Table 1 shows different scenarios where the idle duration level for each load changes by Δt and according to this change, the power gain is calculated for each load. The power gain shows the state of the consumed power at the current level compared to the previous level. As shown in Figure 5, the power gain increases as the Δt increases.

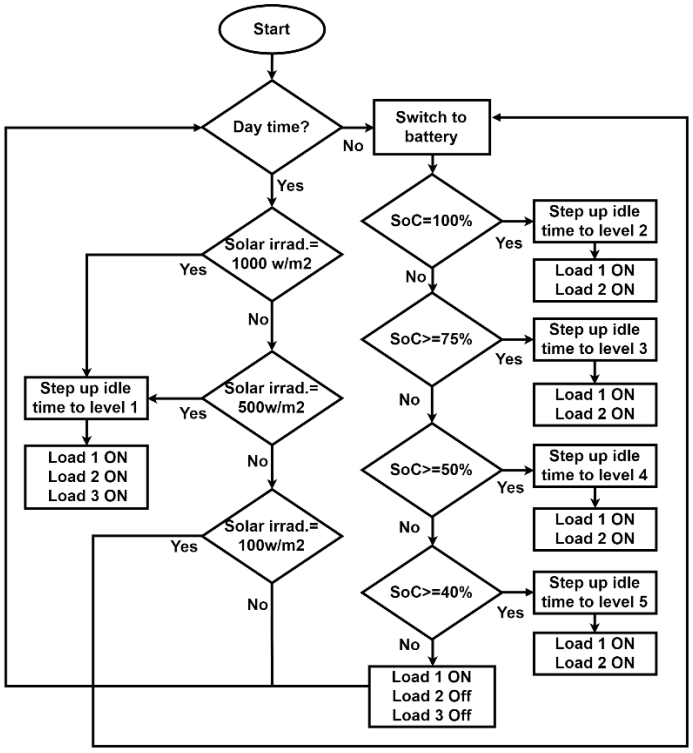


Figure 4: The Flow Chart Illustrates That the Decision of Activating or Deactivating the Loads is Taken After Checking State of Charging (SoC) of The Battery and the Availability of Light.

Table 1: Representing the Power Gain and The Changes in Idle Time for Different Scenarios

Level	Scenario	Load Priority	Active Hours	Δt (Hours)	Power Gain (dBm)
1	Dynamic Irradiance	High	11	10	46.48
		Medium	5	5	45.22
		Low	4	4	44.25
2	100%	High	12	1	36.47
		Medium	6	1	38.23
		Low	0	4	44.25
3	≥75%	High	10	2	39.48
		Medium	4	2	41.25
		Low	0	0	0
4	≥50%	High	6	4	42.49
		Medium	2	2	41.25
		Low	0	0	0
5	≥40 %	High	4	2	39.48
		Medium	1	1	38.23
		Low	0	0	0
6	<40 %	High	1	3	41.25
		Medium	0	1	38.23
		Low	0	0	0

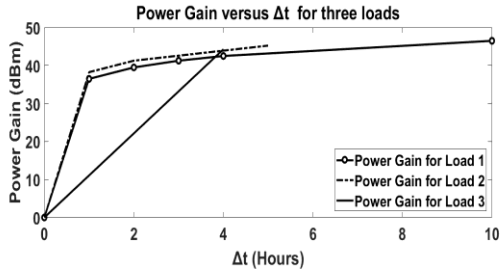


Figure 5: Power Gain versus The Change in Idle Time (Δt) for The Three Loads

It is very essential to choose the optimum sleep time (T_s) that will not affect the system operation. There are three significant parameters that must be taken into consideration while choosing T_s, State of Charging (SoC), Load current (I_L) , and Load index (η). The load index (η) ranges from 0.1 to 1. The value 1 indicates high priority, 0.5 indicates medium priority and 0.1 indicates low priority. The SoC is the difference between the battery rated capacity (C_{rated}) and the used capacity by loads (C_{used}) divided by the battery rated capacity. SoC can be calculated using (2):

$$T_s = fn(SoC, I_L, \eta) \quad (1)$$

$$SoC = \frac{C_{rated} - C_{used}}{C_{rated}} \times 100 \quad (2)$$

Figure 6 represents the charging stage of the battery which shows that the battery state of charging is increasing with time. As a result the sleep duration for the loads is decreasing due to the increase of the amount of power available in the battery. Therefore the sleep duration is inversely proportional to the SoC as shown in Figure 7. Figure 8 shows that the increase in load current results in a decrease in the battery state of charging.

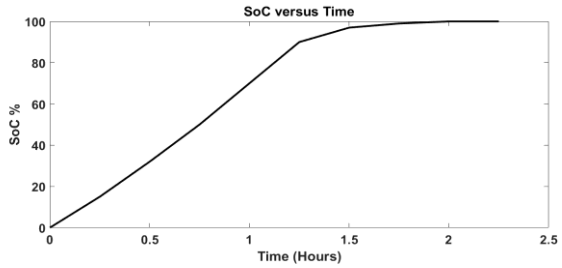


Figure 6: SoC Represented in a Charging Stage versus Time

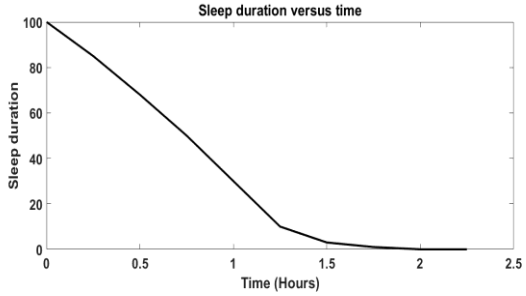


Figure 7: Sleep Duration versus Time

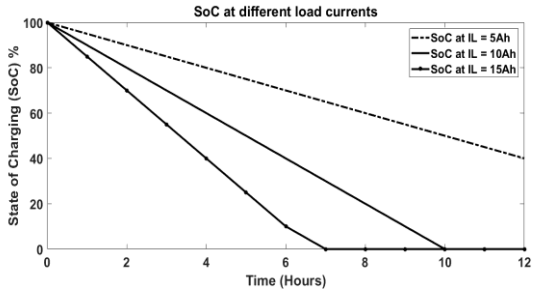


Figure 8: State of Charging at Different Load Currents versus Time

The active duration for each load depends on the maximum current consumed by the loads (I_{Lmax}), the current consumed by load 1 (I_{L1}) and the load index (η). Equation (3) is used to calculate the active duration for each load then calculate the sleep duration using (4). The active and sleep durations for the three loads are shown in Figure 9 and 10. Figure 11 represents the sleep duration for each load versus the battery SoC.

$$T_a = \frac{I_{Lmax} \times 60 \times \eta}{2 \times I_{L1}} \quad (3)$$

$$T_s = 60 - T_a \quad (4)$$

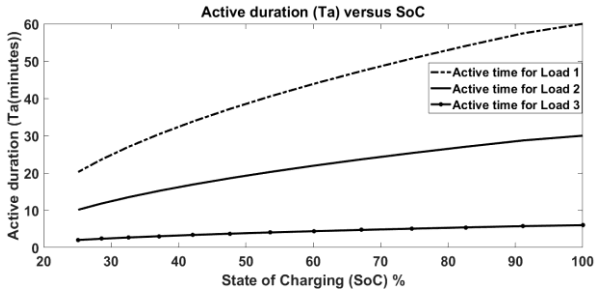


Figure 9: Active Duration for Three Loads versus State of Charging (SoC)

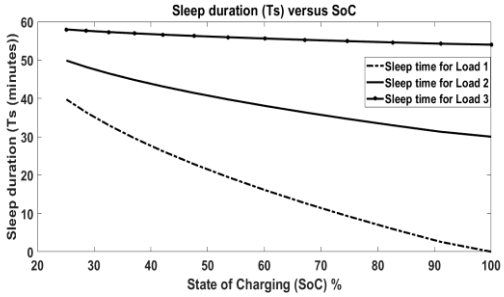


Figure 10: Sleep Duration for Three Loads versus State of Charging (SoC)

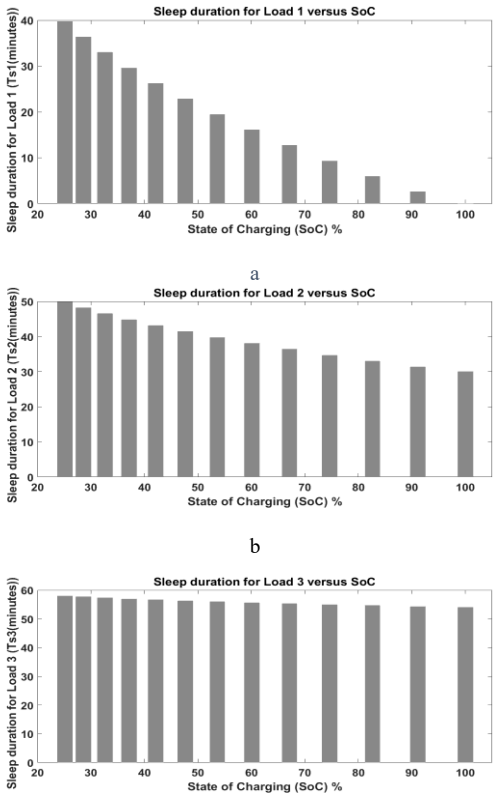


Figure 11: Sleep Duration for Each Load versus State of Charging (SoC)

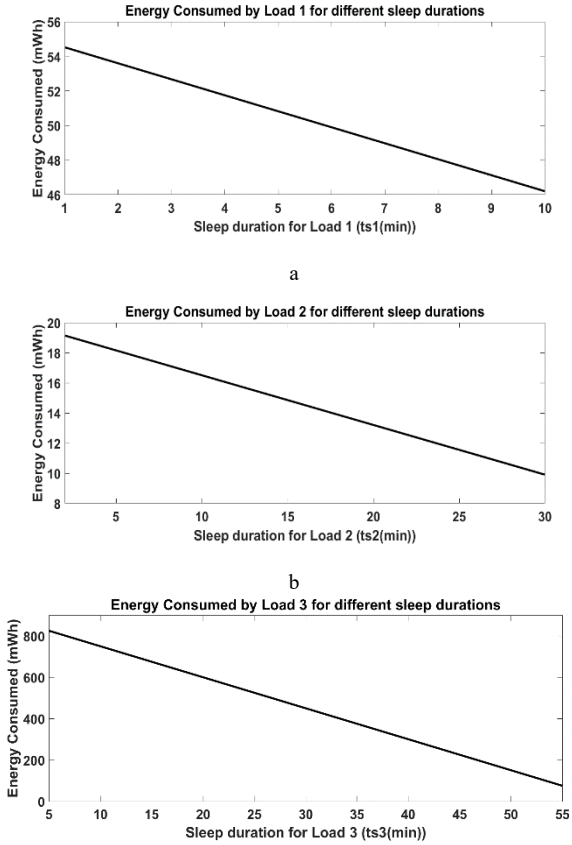


Figure 12: Energy Consumed by Each Load versus Sleep duration for Each Load

The energy consumed by each load for a uniform sleep duration depends on the power consumed by the load (P), the operating time (T_o), Sleep time for each load (T_s) and the number of sleep durations (n). The energy consumed by each load is calculated using (5). As shown in Figure 12, the energy consumed by each load decreases by increasing the sleep duration.

$$E = P \times T_o - P \times T_s \times n \quad (5)$$

5. Simulation Model Results

An agro environment is simulated on a Matlab Model that is designed using one node consisting of battery, three loads and solar panel. The designed model is used to test the algorithm by testing it at different scenarios that are shown in Tables 2 and 3. The solar irradiance and the state of charging must be checked first to enable the model to select the optimal scenario based on these two parameters. This scenario selection will guarantee the efficient use of energy which will increase the battery life time. The current and power of the solar panel is affected by solar irradiance so a decrease in solar irradiance causes a decrease in the current and power as shown in Figure 13. Low solar irradiance occurs as a result of the small amount of solar energy absorbed by the solar panel which results in a decrease in power and current. Due to the change in climatic conditions as shown in Table 2, solar irradiance changes. The solar power is affected by these changes and it will have an effect during day on deciding whether to switch loads on or off.

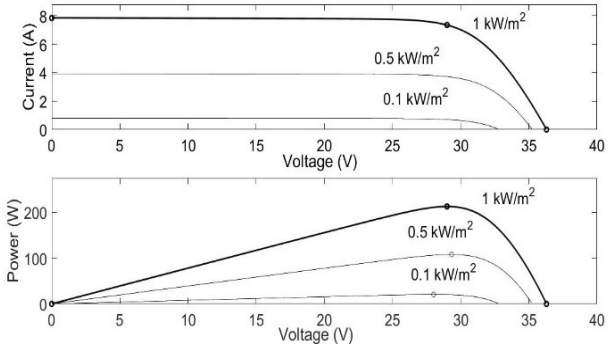


Figure 13: Power and Current at Different Solar Irradiances

Table 2 and 3 show the different scenarios that are designed depending on the battery state of charging and the changes in solar irradiance during day. The loads are activated or deactivated based on the availability of energy provided by battery or solar panel. High priority load is represented as load 1, medium priority load is represented as load 2 and low priority load is represented as load 3. Table 2 shows the four scenarios that are used usually during night. These scenarios are designed depending on the state of charging of the battery. During night, load 1 and load 2 are active for specific hours based on priority while load 3 is deactivated.

Table 2: Represents Different Scenarios of Load Scheduling During Night

Scenario #	Scenario Description	SoC (%)	hrs	Load Scheduling Scenarios		Load Active hours (hrs)
				Priority	Status	
1	Fully charged	100	12	High	On	12
				Medium	On	6
				Low	Off	0
2	Partially charged	75	12	High	On	10
				Medium	On	4
				Low	Off	0
3	Half charged	50	12	High	On	6
				Medium	On	2
				Low	Off	0
4	Battery draining	<40		High	On	1
				Medium	Off	0
				Low	Off	0

Figure 14 shows the battery state of charging as it decreases with time. It is essential to monitor the battery state in order to select the best scenario for supplying loads during night based on amount of power remained.

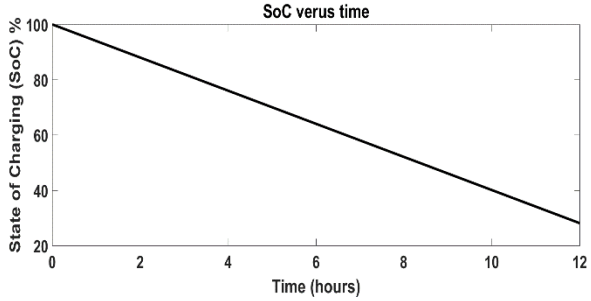


Figure 14: Battery State of Charging versus Time

The solar irradiance is divided into two types static and dynamic. Static solar irradiance is represented in the first three scenarios that are shown in Table 3. These scenarios are designed for various cases that might occur during day in case of static solar irradiance. The duration is constant for the first three scenarios as it is estimated to be 12 hours.

In case the solar irradiance is dynamic during day, this is represented in the fourth scenario shown in Table 3. Scenario 4 is designed due to the change of solar irradiance during day hours. The values of solar irradiance in the previous scenarios might happen in one scenario at different hours. The first scenario in Table 2 and the fourth scenario in Table 3 are simulated and shown in Figure 16 and 17.

Figure 15 shows the changes in solar irradiance during day while during night the loads are supplied by the battery. These variations have an effect on solar power provided to the loads which will decide the status of the load.

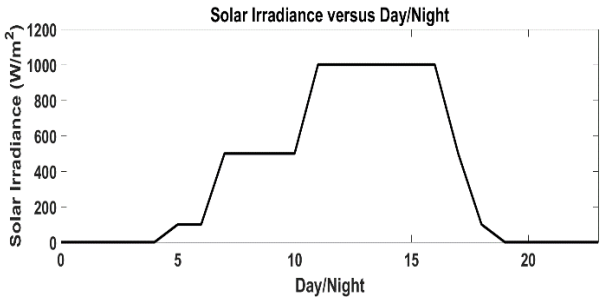


Figure 15: Solar Irradiance versus Day /Night

Table 3: Represents Different Scenarios of Load Scheduling During Day

Scenario #	Scenario Description	Climatic Conditions	Solar Irradiance	hrs	Load Scheduling Scenarios		Load Active hours (hrs)
					Priority	Status	
1	Static Irradiance	Direct sunlight	1000 W/m2	12	High	On	12
					Medium	On	10
					Low	On	8
2	Static Irradiance	Cloudy sunlight	500 W/m2	12	High	On	6
					Medium	On	4
					Low	On	2
3	Static Irradiance	Very cloudy	100 W/m2	12	Powered by Battery Only		
4	Dynamic Irradiance	Clear direct sunlight	1000 W/m2	6	High	On	6
					Medium	On	3
					Low	On	3
		Cloudy sunlight	500 W/m2	5	High	On	5
					Medium	On	2
					Low	On	1
		Very cloudy	100 W/m2	3	Powered by Battery Only		

The loads' status shown in Figure 16 has an effect on its total consumed energy as it changes due to the change in loads' status during day and night. The compensation factor (CF) is shown in Figure 22 as it is calculated for the eight scenarios. It is defined as the ratio between the loads' consumed energy and the energy supplied. During day hours, the solar irradiance changes and at noon, it reaches the maximum point. The solar power at the maximum point will be enough to supply all the loads as shown in Table 3 for specific hours.

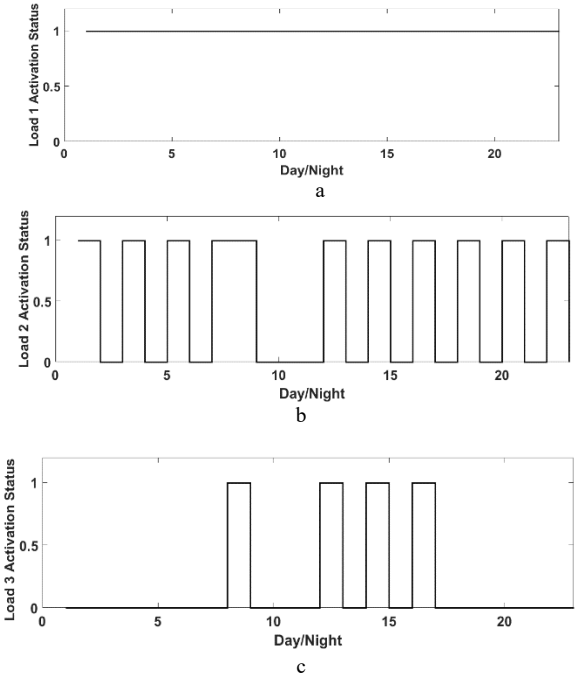


Figure 16: Loads (On/Off Status) versus Day/Night (Scenario 1 in Table 2 and Scenario 4 in Table 3)

Figure 16(a) shows the status of load 1 during day and night. During day hours, load 1 is supplied by solar and in this case the load is active. While during night, it is supplied by the battery.

Figure 16(b) shows the status of load 2, it is activated during day for only 5 hours. Load 2 is activated during night for 6 hours. Figure 16(c) shows the status of load 3, it is activated during day for only 4 hours and it is deactivated at night.

The calculation of load's consumed energy is according to the status of loads shown in Figure 16. The supplied energy either by solar or battery is efficiently consumed using the load scheduling scenarios to provide a sustainable system as shown in Figure 17.

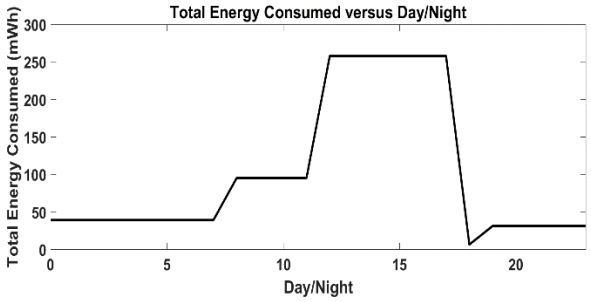


Figure 17: Total Energy Consumed versus Day/Night (Scenario 1 in Table 2 and Scenario 4 in Table 3)

Scenario 2 in Table 2 along with scenario 4 in Table 3 are tested and shown in Figure 18 and 19. Figure 18 shows the loads on/off status during day and night. The total consumed energy by the loads is calculated based on the status of loads as shown in Figure 19.

Figure 18(a) shows that during day hours, load 1 is activated as the power is provided by solar in this case while at night, the load

is still activated and supplied by the battery. Figure 18(b) shows that during day, load 2 is activated for only 5 hours and it is activated for 4 hours at night. Figure 18(c) shows that during day, load 3 is activated for only 4 hours but at night it is deactivated.

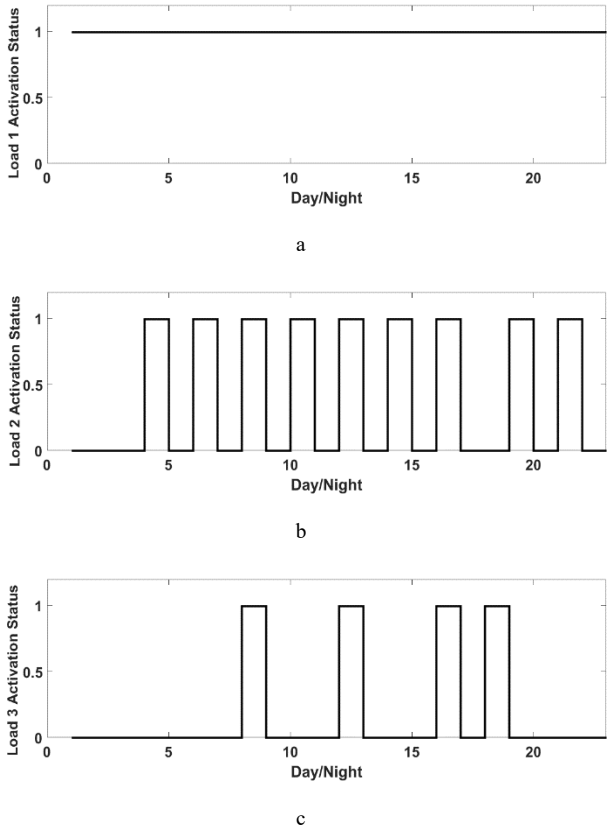


Figure 18: Loads (On/Off Status) versus Day/Night (Scenario 2 in Table 2 and Scenario 4 in Table 3)

The efficient usage of energy during day and night will provide a sustainable system. The total consumed energy by the loads during day and night in accordance to the loads' status is shown in Figure 19.

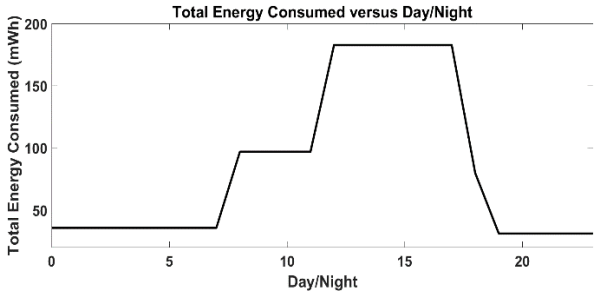


Figure 19: Total Energy Consumed versus Day/Night (Scenario 2 in Table 2 and Scenario 4 in Table 3)

Scenario 3 in Table 2 along with scenario 4 in Table 3 are simulated and represented in Figure 20 and 21. Figure 20 shows the status of the loads which indicates whether it is on or off during day and night. Figure 21 shows the total consumed energy by the loads during day and night based on the status of the loads.

Figure 20(a) shows the status of load 1 during day and night. During day, load 1 is activated by solar power for only 10 hours. During night, it is activated for only 6 hours as it is supplied by the battery. Figure 20(b) shows the status of load 2 during day and night. During day, load 2 is activated for only 5 hours and at night it is inactive. Figure 20(c) shows the status of load 3 during day and night. During day, load 3 is activated for only 4 hours and at night it is inactive. The supplied energy is consumed efficiently by the loads using the designed scenarios of load scheduling. The total consumed energy by the loads during day and night is shown in Figure 21.

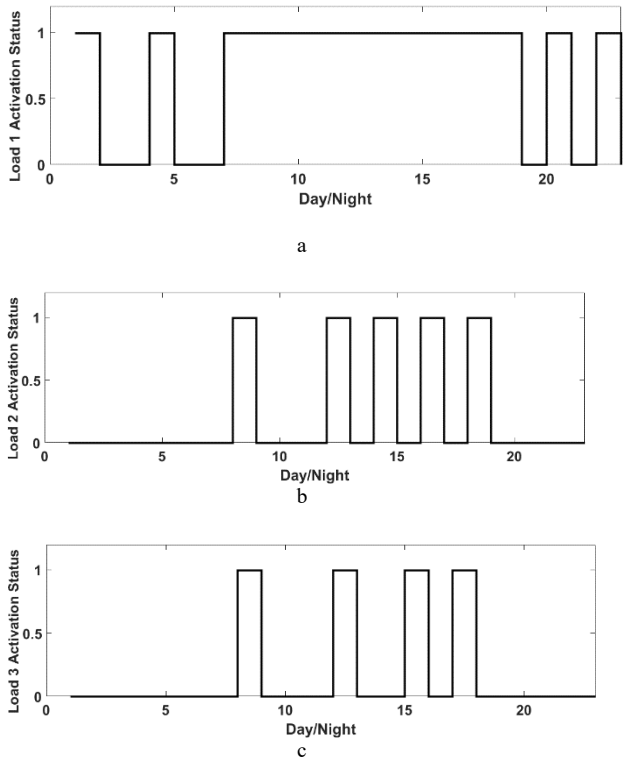


Figure 20: Loads (On/Off Status) versus Day/Night (Scenario 3 in Table 2 and Scenario 4 in Table 3)

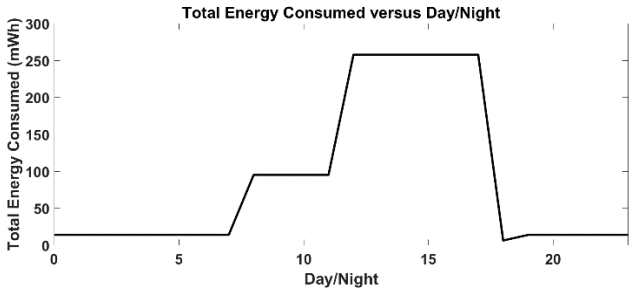


Figure 21: Total Energy Consumed versus Day/Night (Scenario 3 in Table 2 and Scenario 4 in Table 3)

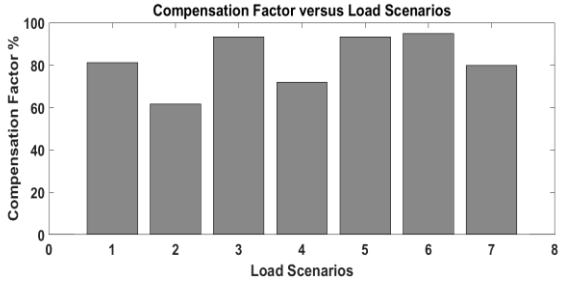


Figure 22: Compensation Factor (CF) versus Load Scenarios

Figure 22 represents the compensation factor (CF) which is calculated for all the designed scenarios. It is the consumed energy by the loads relative to the energy supplied by solar or battery. The compensation factor (CF) is used to show whether there will be an amount of remaining energy that can be used for battery charging or all the supplied energy is fully consumed by the loads.

6. Discussion

The proposed algorithm was simulated and its results are shown in Table 4. The main source that was used for supplying power to the loads during day is the solar panel. During day, there are four designed scenarios for load scheduling. In scenario 1, the three loads will be activated based on priority for a certain time and the energy consumed by the loads is 671.94 mWh at 1000 W/m^2. When the solar irradiance changes to 500 W/m^2, the solar power in this case isn't enough for supplying all loads so scenario 1 isn't suitable for this case as the loads activation period must be reduced.

In this case, scenario 2 is chosen as load 1 (highest priority) is activated for 6 hours while load 2 (medium priority) is activated for 4 hours and load 3 (lowest priority) is activated for 2 hours. In this scenario, the energy consumed by the loads is 184.32 mWh, the battery can be recharged during day using the remaining solar power. At 100 W/m^2, scenario 3 is selected as the loads are supplied by the battery because the solar power is not enough for supplying the loads in this case. The three solar irradiances 1000, 500 and 100 W/m^2 can vary during day so they are combined and represented in scenario 4. In scenario 4, the solar irradiance is 100 W/m^2 for 3 hours, 500 W/m^2 for 5 hours, and 1000 W/m^2 for 6 hours.

Table 4: Simulation Results for Energy Efficient Algorithm

Power Source	Scenario #	Power (mw)			Energy (mwh) for different scenarios	CF %
		Load 1	Load 2	Load 3		
Solar	1	4.62	1.65	75	671.94	81.26
	2	4.62	1.65	75	184.32	61.64
	3	4.62	1.65	75		
	4	4.62	1.65	75	359.07	71.77
		4.62	1.65	75		
		4.62	1.65	75		
Battery	1	4.62	1.65	75	65.34	93.33
	2	4.62	1.65	75	52.8	94.73
	3	4.62	1.65	75	31.02	79.93
	4	4.62	1.65	75		

During night, the algorithm was simulated by using the battery as the main source of power for supplying the loads. There are also four designed scenarios for load scheduling at night. In scenario 1, the loads are supplied by a fully charged battery so load 1 is activated for 12 hours while load 2 is activated for 6 hours as it is active for 1 hour every 2 hours and load 3 is inactive. In scenario 2, the battery SoC is 75 % so load 1 is active for 12 hours, load 2 is active for 4 hours as it is active for 1 hour every 3 hours and load 3 is inactive. Load 3 can be activated based on the user demand or it remains inactive till the availability of light. In scenario 3, the battery SoC is 50% so load 1 is active for 6 hours as it is active for 1 hour every 2 hours while load 2 is active for 2 hours as it is active for 1 hour every 6 hours and load 3 is inactive. In scenario 4, the battery SoC is below 40% so load 1 is active for 1 hour while load 2 and 3 are switched off till the availability of light in order for the loads to be powered by the solar and for the battery to be recharged. The proposed algorithm guarantees the efficient usage of energy by using the designed scenarios for load scheduling during day and night. It provides a sustainable system as it avoids the drainage of the battery which will prolong the system life time.

7. Conclusion

As a conclusion, it is essential to monitor plants' performance as it helps in increasing and saving the agricultural production. Monitoring the performance of plants is facilitated using IoT technology as it uses wireless sensors that enables the plants to talk and express its requirements. Energy optimization is a major challenge that faces IoT. In agriculture, wireless sensors are utilized in remote areas therefore they consume high energy for continuous operation. An energy efficient technique was proposed in this paper to avoid battery drainage in order to guarantee system sustainability and continuous operation. Solar power and intelligent load scheduling are combined in the proposed algorithm. The results showed that the system operation adapts to the availability of power represented as solar irradiance or battery SoC. It operates using solar panels during day based on solar irradiance but during night, it depends on the battery SoC. During day and night, intelligent load scheduling based on priority is applied to ensure continuous operation. The proposed algorithm elongates the sensors' batteries life time as it efficiently uses energy to ensure the system sustainability. The proposed algorithm enhanced the battery life time by 79.2 %.

References

[1] S.S. Abou Emira, K.Y. Youssef, M. Abouelatta, "Adaptive power system for IoT-based smart agriculture applications," in ICENCO 2019 - 2019 15th International Computer Engineering Conference: Utilizing Machine Intelligence for a Better World, IEEE: 126–131, 2019, doi:10.1109/ICENCO48310.2019.9027393.

[2] S.S. Sarmila, S.R. Ishwarya, N.B. Harshini, C.R. Arati, "Smart farming: Sensing technologies," in Proceedings of the 2nd International Conference on Computing Methodologies and Communication, ICCMC 2018, 149–155, 2018, doi:10.1109/ICCMC.2018.8487571.

[3] S. Rawal, "IOT based Smart Irrigation System," International Journal of Computer Applications, **159**(8), 7–11, 2017, doi:10.5120/ijca2017913001.

[4] S. Naveen, S. Hegde, "Study of IoT: Understanding IoT Architecture, Applications, Issues and Challenges," in 1st International Conference on Innovations in Computing & Networking (ICICN16), CSE, RRCE, 477–482, 2016.

[5] D.A. Vyas, D. Bhatt, D. Jha, "IoT : Trends , Challenges and Future Scope," International Journal of Computer Science & Communication, **7**(1), 186–197, 2016, doi:10.090592/IJCSC.2016.028.

[6] S. Sadasivam, V. Vadhri, Smart Plant Monitoring System, Informatik, Technische Universitat Darmstadt, Darmastatd, 2015.

[7] D. Markovic, R. Koprivica, U. Pesovic, S. Randic, "Application of IoT in monitoring and controlling agricultural production," Acta Agriculturae Serbica, **20**(40), 145–153, 2015, doi:10.5937/aaser1540145m.

[8] K.Y. Chan, H.J. Phoon, C.P. Ooi, W.L. Pang, S.K. Wong, "Power management of a wireless sensor node with solar energy harvesting technology," Microelectronics International, **29**(2), 76–82, 2012, doi:10.1108/13565361211237662.

[9] H.M. Jawad, R. Nordin, S.K. Gharghan, A.M. Jawad, M. Ismail, M.J. Abu-Alshaeer, "Power reduction with sleep/wake on redundant data (SWORD) in a wireless sensor network for energy-efficient precision agriculture," Sensors (Switzerland), **18**(10), 3450–3475, 2018, doi:10.3390/s18103450.

[10] H. Swalehe, P.V. Chombo, B. Marungsri, "Appliance scheduling for optimal load management in smart home integrated with renewable energy by using whale optimization algorithm," GMSARN International Journal, **12**(2), 65–75, 2018.

[11] X. Liu, L. Ivanescu, R. Kang, M. Maier, "Real-time household load priority scheduling algorithm based on prediction of renewable source availability," IEEE Transactions on Consumer Electronics, **58**(2), 318–326, 2012, doi:10.1109/TCE.2012.6227429.

[12] S. Trivedi, K. Parkh, K. Agrawal, "Load Scheduling for Smart Energy Mangement in Buildings with Renewable Power Generation," International Journal of Engineering Research And, **6**(03), 322–326, 2017, doi:10.17577/ijertv6is030349.

[13] M.R. Mohd Kassim, I. Mat, A.N. Harun, "Wireless sensor network in precision agriculture application," in 2014 International Conference on Computer, Information and Telecommunication Systems, CITS 2014, 1–5, 2014, doi:10.1109/CITS.2014.6878963.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1050-1055 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

VLSI Architecture for OMP to Reconstruct Compressive Sensing Image

Santosh Bujari[1,*], Saroja V Siddamal[2]

[1]Department of Electronics and Communication, SKSVMACET, Laxmeshwar, 5812116, India

[2]School of Electronics and Communication, KLE Technological University, Hubli, 580021, India

ARTICLE INFO

Article history:
Received: 03 August, 2020
Accepted: 28 September, 2020
Online: 20 October, 2020

Keywords:
Nyquist Criteria
Matching Pursuit
Sparsity
Threshold
FHT

ABSTRACT

A real-time embedded system requires plenty of measurements to fallow the Nyquist criteria. The hardware built for such a large number of measurements, is facing the challenges like storage and transmission rate. Practically it is very much complex to build such costly hardware. Compressive Sensing (CS) will be a future alternate technique for the Nyquist rate, specific to some applications where sparsity property plays a major role. Software implementation of Compressive Sensing takes more time to reconstruct a signal from CS measurements using the Matching Pursuit (MP) algorithm because of fetching, decoding, and execution policy. It is necessary to build hardware in CS. The author proposes one such VLSI Architecture (Hardware) for $256\ X\ 256$ *and* $512\ X\ 512$ *image. Various random matrices like Bernoulli, Partial Hadmard, Uniform Spherical, and Random Matrix are used to build hardware. FHT (Foreward Transform) with* $\pm 2\ to\ 6$ *threshold is applied to get CS measurements. The reconstruction time, Signal to Noise ratio (SNR), and Mean Square Error (MSE) are measured. Multiple time experiments are carried out and results show that for an image of size* $256\ x\ 256$*, SNR is* $25\ dB$ *and MSE is 166. For the image of size* $512\ x\ 512$*, the values are* $27dB$ *and 182. However, both the input images are resized to* $256\ X\ 256$ *so the reconstruction time is* $2.62\mu sec$ *which is less is compared to software implementation.*

1. Introduction

The digital sensing and processing based on the Nyquist criteria made a tremendous impact on real-time applications. The scale of data accumulated by digital conversion has increased from a few to exponential growth. In many emerging applications it will be very complex to store and transmit over the communication channel. Therefore, it is very much essential to find out an alternative for the Nyquist criteria which can reconstruct the signals with few numbers of samples than the Nyquist criteria. The Compressive Sensing theory (CS) [1-7], which is an asymmetric concept [8] with the simple decoder and dumb encoder, can reconstruct the sparse signals when the samples are acquired other than the Nyquist criteria.

$$Z = \Psi A \tag{1}$$

The signal acquisition in CS theory is the product of random measurement matrix (Ψ) and analog signal (A) leads to CS measurements (Z) [9, 10] defined by equation (1).

Z Ψ = I. N X N Identity matrix A

$$\text{N X 1}\begin{bmatrix} Z1 \\ Z2 \\ Z3 \\ Z4 \\ \vdots \\ \vdots \\ ZN \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots & 1 \end{bmatrix} * \begin{bmatrix} A1 \\ A2 \\ A3 \\ A4 \\ \vdots \\ \vdots \\ AN \end{bmatrix} \text{N X 1}$$

(2)

The Nyquist criterion as shown in equation (2), is extended to illustrate compressive theory, where A is an analog signal, Ψ is unity (identity) matrix of size N X N and Z is the Nyquist samples. In CS unity matrix is replaced with a random matrix such as Bernoulli Matrix [11], Uniform Spherical, Partial Hadmard, etc. Samples that are obtained by multiplying the analog signal with the random matrix are called CS measurements (Z) as shown in equation (3).

Z Ψ $m\ X\ N$ *matrix* A

*Corresponding Author: Santosh Bujari, santoshbujari@gmail.com

$$\begin{bmatrix} Z1 \\ Z2 \\ \vdots \\ Zm \end{bmatrix} = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} & \Psi_{14} & \cdots & \cdots & \Psi_{1N} \\ \Psi_{21} & \Psi_{22} & \Psi_{23} & \Psi_{24} & \cdots & \cdots & \Psi_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Psi_{m1} & \Psi_{m2} & \Psi_{m3} & \Psi_{m4} & \cdots & \cdots & \Psi_{mN} \end{bmatrix} \begin{bmatrix} A1 \\ A2 \\ A3 \\ A4 \\ \vdots \\ \vdots \\ AN \end{bmatrix} \quad (3)$$

These CS measurements are related to only sparse (big amplitude) signal, that's why m is less than N. Random matrix is modified as shown in equation (4) by including only those columns related to sparsity. Remaining signals (N − K) are not processed which may be located anywhere in the signal. Therefore, CS is an asymmetric theory.

$$Z \qquad \Psi \quad m\,X\,K \;\; matrix \qquad A$$

$$\begin{bmatrix} Z1 \\ Z2 \\ \vdots \\ Zm \end{bmatrix} = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} & \Psi_{14} & \cdots & \cdots & \Psi_{1K} \\ \Psi_{21} & \Psi_{22} & \Psi_{23} & \Psi_{24} & \cdots & \cdots & \Psi_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Psi_{m1} & \Psi_{m2} & \Psi_{m3} & \Psi_{m4} & \cdots & \cdots & \Psi_{mK} \end{bmatrix} \begin{bmatrix} A1 \\ A2 \\ \vdots \\ AK \end{bmatrix} \quad (4)$$

Various Matching Pursuit algorithms [12] are used to solve the linear equation (4). Orthogonal Matching Pursuit (OMP) is compatible for sparse recovery [13-15]. This OMP will search the sparse related columns iteratively which appears in the random matrix. These are the best approximate-correlated columns that correspond to the sparse signal. One best-correlated column is searched in the iterations and removed from the random matrix. Like this finally updated matrix (C) of dimension K x K is obtained. The sparse reconstructed signal is obtained by solving the Least Square Problem (LSP). The complete OMP procedure is bellowed:

1. **Declaration:** Declare residual Resi = Y, matrices $Q = zeros\,(m,K)$, $R = zeros\,(K,K)$, $R_{inv} = zeros\,(K,K)$ and the index set $C = zeros(m,k)$. Initialize iteration i to 1.
2. **Search Index:** Search the most correlated index by solving the optimization problems and generate Q and R matrices.

$$\mathrm{KK_i} = \arg\max_{j=1..N} \left| < \mathrm{Resi_{i-1}}, \Psi_j \right|$$

$$\Psi_j = Q_j R_j$$

3. **Residual Status Update:** Revise the index set c and estimate the residual.

$$C_i = \mathrm{C_{i-1}} \cup \{\mathrm{KK_i}\}$$

$$\mathrm{R_{inv}} = [\mathrm{R_{inv}}\ \Psi_{\mathrm{c_i}}]$$

$$Resi_i = Resi_{\mathrm{i-1}} - (Q_{\mathrm{i-1}} * \mathrm{Q_{i-1}}')Resi_{\mathrm{i-1}}$$

4. **Iterations:** $\boldsymbol{i = i + 1}$ and go back to search index step suppose $i < K$.
5. **LSP: Least Square problem,** Find the inverse corresponding $\boldsymbol{c}$ to get reconstructed sparse signal

$$A_R = (\mathrm{R_{inv_j}} * \mathrm{Q_j}' * \mathrm{Z_j})_{\mathrm{j=1.....K}}$$

Further write up of the paper is as follows. 2 briefs design methodology. In 3 VLSI architecture is proposed. The results of the work are shown in 4. Finally, work is concluded with future work in 5.

2. Design Methodology

2.1. OMP (Reconstruction Algorithm) Blocks

OMP is a sequential algorithm (as shown in Figure1) where the Optimization block is formed by combining the Search Index and Residual update. CS measurements (Z) which are in HEX format (256 X 256), Random Matrix (Ψ) of dimensions 128 x 256 and sparsity count (K = 64) are applied as inputs. Correlated columns (Ψ = 64 X 64) related sparse signals are identified. From these correlated columns, reconstructed signal ($\mathrm{C_K}$) is calculated by taking the inverse of the matrix as shown in equation (5) and (6) in LSP.

$$\mathbf{Z_K = C_K \Psi_K} \quad (5)$$

$$\mathbf{C_K = \Psi_K^{-1} Z_K} \quad (6)$$

$$\mathbf{A_R = IFHT\,(C_K)} \quad (7)$$

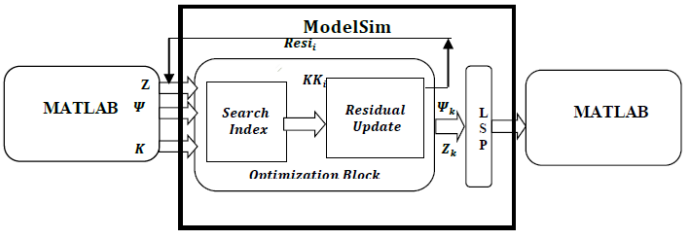


Figure 1: OMP Blocks

2.2. Input for OMP Architecture

MATLAB tool is used to generate input for OMP Architecture. The unsigned gray image of any dimensions is resized into to 256 X 256. Double Precision format for processing purposes and integer format for display purpose is followed. Since the image does not exhibit sparsity property in the normal domain so it is converted into FHT (Foreward Hadmard Domain). Less magnitude coefficient values are treated as zero by selecting a proper threshold value. This is asymmetric contrast to the Nyquist criteria. These zero value signals may be located anywhere in the image. These are converted to in HEX number and stored as a text file which is used as input for OMP VLSI Architecture. The complete procedure of creating a text file is shown in Figure2.

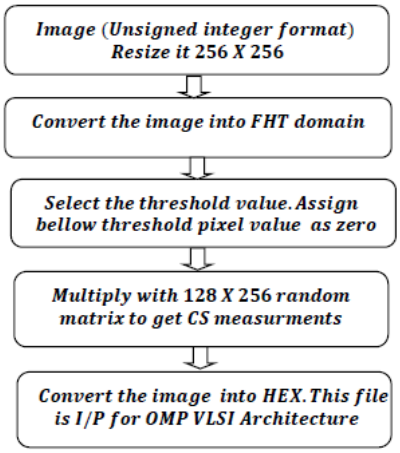


Figure 2: Text File Creation Flow

3. Proposed VLSI Architecture

The text file which is obtained from the previous step and the random matrix are stored in different RAM in the proposed OMP architecture as shown in Figure3. The tree structure (Wallace tree) as shown in Figure4 is used to obtain the inner product vector. The CS measurements in the form of 3 x 3 dimensions are applied to the Wallace tree along with random variables.

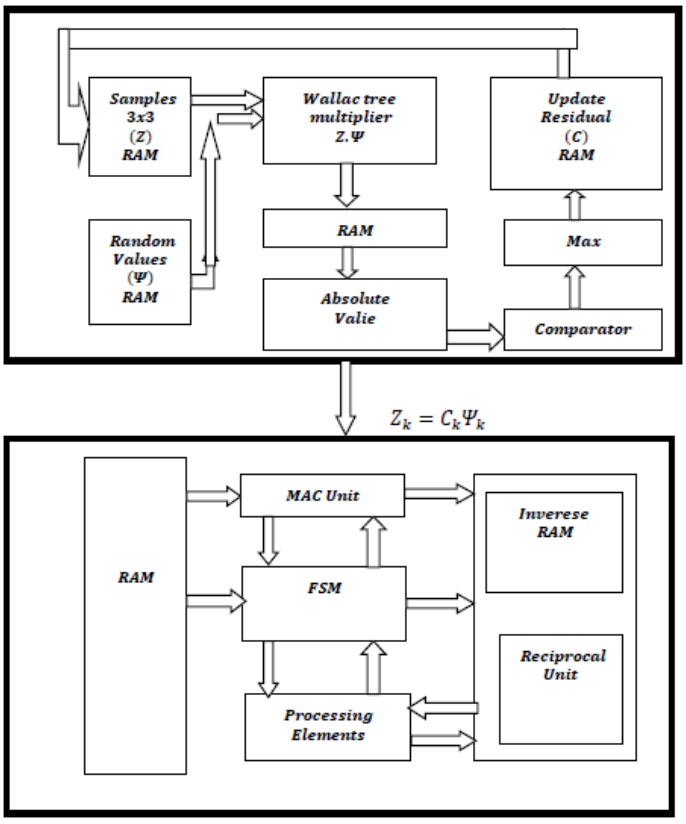


Figure 3: Proposed OMP Architecture

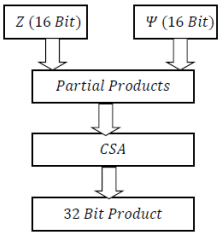


Figure 4: Multiplier (Wallace Tree)

In this proposed architecture, a carry-save adder (CSA) as shown in Figure5 is used to perform the fast addition of partial products. At the beginning of the calculations, the accumulator register is reset to zero. Then, the first partial product, P, is selected by the multiplexer and applied to the 16-bit adder. The other input of the adder comes from the 16-bit register. Hence, at this stage, the output of the 16-bit adder will be Acc=P + 0=P. With the upcoming clock tick, this will be stored in the accumulator register. Next, the second partial product, Q, will be chosen by the multiplexer. This will be added to the current value of the 16-bit register which gives Acc = P + Q. This procedure will be repeated for the remaining partial products. With each clock tick, a new partial product will be added and the result will be stored in the 16-bit register. The control unit is used to generate an appropriate signal for the select input of the multiplexer. Speed up of the 16-bit adder is achieved by implanting a carry-lookahead structure for a 16-bit adder.

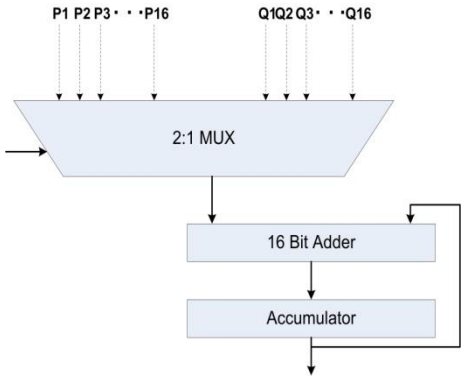


Figure 5: Carry save Adder

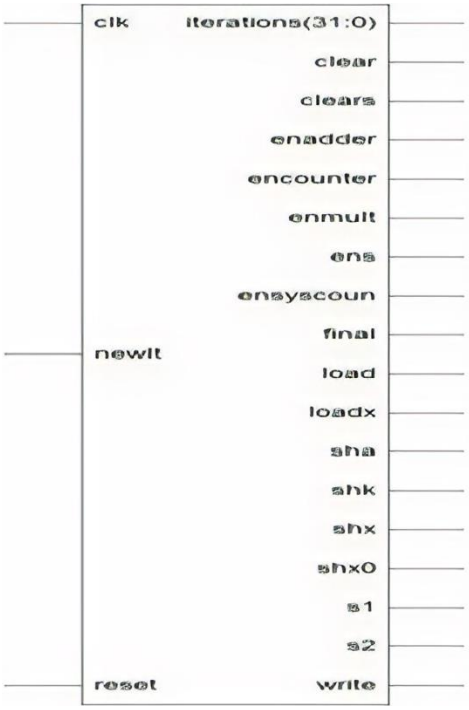


Figure 6: Control Path Block

In a single clock cycle, it cannot be completed since the carry bit has to propagate. CSA is required in different stages of the OMP algorithm. This adder architecture is critical as it is used for many numbers of operations. The complexity of the adder depends on the operation like subtraction, multiplication, and division. In general carry-save adders connected in a tree-like structure can reduce the number of operations from m to $m-1$ compare to parallel adders. This in turn contributes to reconstruction time. Within the tree structure, an adder is designed using $2:1$ multiplexers which can reduce the delay of higher multiplexers and it address reconstruction time. Once the product vector is obtained from the tree structure, the maximum value, and its corresponding column numbers are calculated using a comparator, and values are

stored in RAM represented as update residual in architecture. Later these values are removed from text files and like this for every 3x3 sample process repeats iteratively. This process repeats until K(Sparsity Count) values are searched. Finally, $Z_k = C_k \Psi_k$ values are obtained from the optimization block and later applied to the LSP block. As the matrix size increases, the number of computations like multiplication, addition/subtraction, and division is increasing. That's why the image is divided into a 3x3 matrix. All submatrix are processed iteratively which contributes towards reconstruction time. From the Optimization block, all the sparse signal information is obtained and it is stored in RAM of LSP architecture. The arithmetic right shifts are used in LSP to perform division and reciprocal operations. Shifting one bit right is dividing by 2 in binary. Compared to a division operator, a shift operator will execute in less number of cycles in all most all processor. This contributes to the reduction of reconstruction time. Moreover, it can be reused for shift operations wherever it is necessary. The entire LSP operation is controlled by the control path block as shown in Figure6, which is implemented in the form of a state diagram as shown in Figure7.

Depending upon the state diagram and data the control signals are generated, and the data out (reconstruction signal) will be calculated. Finally, these reconstructed signals are stored in the output text file.

3.1. State diagram

- Reset S_0: At the beginning, LSP architecture is initialized by storing all the signals (Z_k) obtained from the Optimization block. This state is considered a reset state. If control signal reset $= 0$, LSP remains in the same state otherwise new state will be initialized.
- New System S_1: Since LSP iterative in nature, the iteration counter is initialized to zero based on a clear signal in this state.
- New Iteration S_2: Values (Z_K, Ψ_k) is loaded in this state based on Load and loadx signal. When clears signal is in high, loaded values will be sent to the calculation state.
- Calcs S_3: Architecture in this state will start calculation (division) using processing elements and shift operator if sha and Shax signal is high. At the same time counter is incremented. It will be in the same state if the number of calculations is not equal to the number of elements considered. When the calculations become equal to number with counter then it will go to the next state called wait state.
- Wait S_4: Results of division obtained from the previous state are collected in this state.
- Calculation Reciprocal state S_5: (CalcR) In this state reciprocals are calculated iteratively on the outcomes obtained from the previous state. Encounter signals trigger the counter to keep track of the number of iterations. When the counter becomes equal to the number of iterations, state transition will occur based on shx.
- Calcx S_6: When the number of calculations is not equal to the number of reciprocal elements then S1 and S2 signals will be high which keeps reciprocal calculation in progress. If the number of calculations (iterations) is equal to K and then state changes to countsys state. If iteration is not equal to K then state transition switches to S_2 and the process repeats until all K blocks are completed.

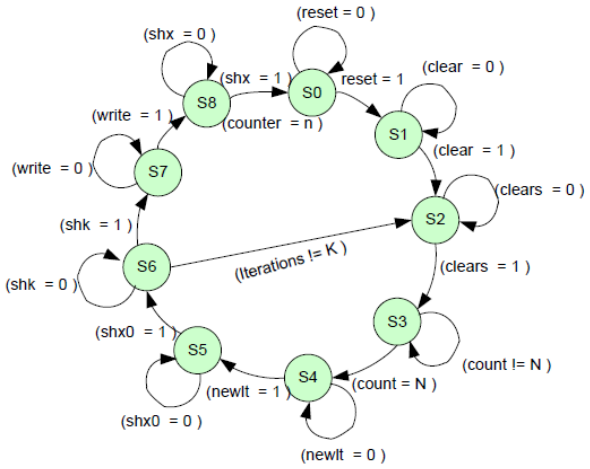


Figure 7: LSP State Diagram

- Countsys S_7: Once all K blocks are processed, memory is initialized in this state to store the results. State transition occurs when Write becomes high.
- Savex S_8: In this state, results are updated on the output text file.

All the control signals generated by the state diagram will be reflected in the LSP data path which is as shown in Figure 8.

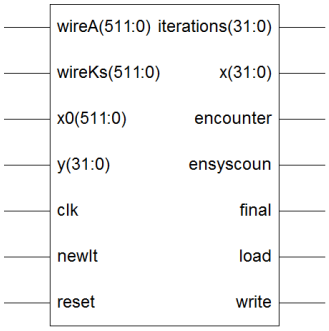


Figure 8: LSP Data Path

4. Results

On Intel Core with 1.2GH$_z$ processor and 3GB memory, several time experiments are carried out. Obtained results are tabulated to analyze the performance of our proposed architecture. Bellow steps are followed to get the results.

- Select any Image of dimensions more than or equal to 256 X 256.
- Create a text file in HEX.
- Use text file in Test-bench and reconstruct signal which will be stored in the output text file.
- Read the output text file and display it after processing

Steps 1, 2, and 4 are carried out in MATLAB 2014A. For the third step, the ModelSim PE Student Edition is used. Standard Lena image of dimension 256 x 256 and 512 X 512 with different image formats are used as input to the proposed architecture. Several times process is repeated and only the average value of selected parameters [16] is analyzed.

- MSE: Difference between original and reconstructed signal which is estimated using MSE defined by equation (8).

$$\text{MSE} = \sum_n \frac{(\text{A}-\text{Arec})^2}{\text{n}} \quad (8)$$

- Reconstruction time: Time taken by OMP VLSI architecture to generate output file which is measured in second.
- SNR: The amount of noise present in the signal reconstructed using OMP VLSI architecture which is defined by equation (9).

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_n \text{A}^n}{\sum_n(\text{A}-\text{Arec})^2}\right) \quad (9)$$

4.1. Analysis of Image Reconstruction Quality

- Standard Lena Image ($\text{N} = 256$ and 512).
- Measurement Matrix: $(128 \text{ X } 256)$
- Number of CS measurements = 128X 1.
- FHT transform coder.
- Threshold value ($\pm T_r$)= 3



Figure 9: Reconstructed 256 *X* 256 and 512 *X* 512 Image

Figure 9 is the reconstructed image using OMP VLSI Architecture indicates that the quality of the image degraded which will be one of the future challenges of OMP VLSI architecture. Simulation results of this work compared with MATLAB implementation [17] and Virtex 5 implantations [18, 19] as shown in Table I. Compared with other simulation work, the proposed design takes less reconstruction time. The reduction in reconstruction time is mainly due to less processing elements as input 3x 3 is fed into the system iteratively. To process 256x256 at a time require more hardware and more time to reconstruct

Table 1: Comparison of Proposed Design with Other Similar Work

Platform	**[17]**		**[18]**	**[19]**	**Proposed**
Tools Used	***MATLAB***		***Xilinx***		***Xilinx***
Target Device	***NA***		***Virtex − 5***		***Virtex − 4***
Signal Size (N)	256	512	256	256	256
Measurements (m)	128	128	64	64	128
Sparsity Count (K)	64	64	8	8	64
Reconstruction Time	***9sec***	***15sec***	***17μS***	***22μS***	***2.62μS***
MSE	550	580	-	-	166.60
SNR (dB)	23	26	-	-	25.91

4.2. Varying T_r with FHT Domain

Normally large amplitude signals hold most of the important information about the input. Even by neglecting lesser value amplitude signals, information can be extracted. To remove small amplitude signal threshold value from ± 2 to $\pm 6$is used in the FHT domain. To minimize the number of measurements larger threshold value should be selected. However, it affects the quality of the reconstructed signal. Therefore trade-off between the quality of the reconstructed signal and reconstruction time will be the parameter for selecting proper threshold value. Experiment analysis shows that for $\text{N} = 256$ or 512, the FHT threshold value should not be more than ± 7. SNR for different threshold is almost the same as shown in table 2.

Table 2: Threshold and SNR Analysis

Image Size	**Threshold ($\pm T_r$)**	**SNR (dB)**
256 X256	**2**	**25**
	4	**25**
	6	**25**

4.3. Device Utilities

Table 3 shows the device utilization report using the Xilinx ISE 8.1i version tool. Verilog code implemented on target device namely xc4vlx15-12sf363 (Vertex 4 FPGA). However, our work [1] and [17] are continued in this paper and the main objective of reduction of reconstruction time is achieved. The analysis and optimization of area, power consumption, and other VLSI parameters are considered as future work for this paper.

Table 3: Device Utilization Summary

Logic Utilization	**Used**	**Availability**	**Utilization**
Slices	2116	6144	34%
Slice F/F	2257	12288	18%
4 Input LUTs	3523	12288	28%
Bonded IOBs	212	240	88%
GCLKs	1	32	3%

5. Conclusion and Future Work

This paper presents semi-VLSI architecture which contributes to developing hardware, especially for the Compressive Sensing image. The implemented architecture has equal reconstruction for both 256x256 and 512x512 image i.e 2.62μS since it is resized to256x256. This is far better than software implementation and some hardware implementation. For FHT (Foreward Transform) ± 2 to 6 (less than 7) is to be used as a threshold for proper reconstruction. 25db and 27dB are the SNR for an image of size 256 x 256 and 512x512 respectively. Mean Square Error (MSE) is 166 which is less as observed in software. Finally, our work concludes that one of the objectives (reconstruction time) is achieved. Analysis and optimization of some objectives like area, power consumption, and other VLSI parameters are considered as future work of this paper

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Santosh, V. Siddamal, "A Survey and Theoretical View on Compressive Sensing and Reconstruction", I.J.Image, Graphics and Signal Processing, 2016, DOI: 10.5815/ijigsp.2016.04.01 MECS Journal

[2] D. L. Donoho, "Compressed sensing," in IEEE Transactions on Information Theory, **52**(4), 1289-1306, April 2006. DOI: 10.1109/MSP.2007.914728

[3] E. J. Candes and M. B. Wakin, "An Introduction To Compressive Sampling," in IEEE Signal Processing Magazine, **25**(2), 21-30, March 2008. DOI: 10.1109/MSP.2007.914731

[4] Shreyas Joshi, Saroja V. Siddamal "Performance Analysis of Compressive Sensing Reconstruction" IEEE sponsored 2nd International Conference on Electronics and Communication Systems (ICECS 2015), pp-724-729, 26th to 27th February KNU Coimbatore 2015, 978-1-4788-7225-8/15/$31.00 ©2015 IEEE.

[5] J. Romberg, "Imaging via Compressive Sampling," in IEEE Signal Processing Magazine, **25**(2), 14-20, March 2008. DOI: 10.1109/MSP.2007.914729

[6] K. V. Siddamal, Shobha P Bhat, Saroja V. S "A survey on compressive sensing" IEEE sponsored 2nd International Conference on Electronics and Communication Systems (ICECS 2015), 639-643, 26th to 27th February KNU Combaiture 2015, 978-1-4788-7225-8/15/$31.00 ©2015 IEEE.

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. "A simple proof of the restricted isometric property for random matrices". Const. Approx., **28**(3):2538211; 263, 2008.

[8] M. Davenport, J. Laska, P. Boufouons, and R. Baraniuk. "A simple proof that random matrices are democratic". Technical report TREE 0906, Rice Univ., ECE Dept., Nov. 2009.

[9] M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk. "Signal processing with compressive measurements". IEEE J. Select. Top. Signal Processing, **4**(2):4458211; 460, 2010.

[10] M. Davenport, J. Laska, P. Boufouons, and R. Baraniuk. "A simple proof that random matrices are democratic". Technical report TREE 0906, Rice Univ., ECE Dept., Nov. 2009.

[11] U. Dias and M. E. Rane, "Comparative analysis of sensing matrices for compressed sensed thermal images," 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), Kottayam, 2013, 265-270.

[12] Meenakshi and Sumit Budhiraja, "A survey of Compressive Sensing Based Greedy Pursuit Reconstruction Algorithms", IJIGSP, **7**(10), 1-10, Pub. Date: 2015-9-8

[13] J. A. Tropp and A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," in IEEE Transactions on Information Theory, **53**(12), 4655-4666, Dec. 2007.

[14] Lin Bai, Patrick Maechler, Michel and Hubert Kaeslin, "High Speed Compressed Sensing Reconstruction on FPGA using OMP and AMP", IEEE 2012.

[15] Pierre B, Hasan R and Abbes A, "High Level Prototyping and FPGA Implementation of The OMP Algorithm", in Information Sciences, Signal Processing and their Application Conference, 11th IEEE International Symposium on, 2012.

[16] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. DOI: 10.1109/ICECCE47252.2019.8940756

[17] S. S. Bujari and S. V. Siddamal, "Image Reconstruction Using Compressive Sensing Technique for Hardware Implementation," 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Msyuru, India, 1042-1046, 2018. doi: 10.1109/ICEECCOT43722.2018.9001308.

[18] J. L. V. M. Stanislaus and T. Mohsenin, "Low-complexity FPGA implementation of compressive sensing reconstruction," in Proc. Int. Conf. Comput. Netw. Commun. (ICNC), 671–675, 2013.

[19] J. L. V. M. Stanislaus and T. Mohsenin, "High performance compressive sensing reconstruction hardware with QRD process," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May, 29–32, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1056-1063 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Essential Features/Issues of a Multi-Phase Switching Synchronous Buck Regulator

Hani Ahmad-Assi*, Nour Sultan Gammoh, Mariana Awni Al Bader

Electrical Engineering Department, King Abdullah II School of Engineering, Princess Sumaya University for Technology, Amman, 11941, Jordan

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 28 September, 2020
Online: 20 October, 2020

Keywords:
Multi-phase buck converter
Phase shedding
Fast-transient
Current sensing

ABSTRACT

This paper addresses essential features/issues and proposes solutions that would improve the overall performance of a multi-phase buck converter. Low efficiency at light load is addressed with phase shedding, load fast transient and regulated output voltage spiking/dipping is addressed with novel helper technique at the point of load (output node). An Integrated current sensing is utilized to implement over-current-protection (OCP), in addition to its inherent function in current mode control. The phase shedding developed technique was used to enhance the efficiency of the converter. The number of phases rather increases or decreases, depending on the desired load. The proposed fast transient helper circuit is tested by inserting a 500mA transient current step in 100µs. Worst case spike of a 79.1mV was achieved at the output node; which is a reduction of 49.4% of the original response (160mV without the helper circuit). Worst case of a 35mV of output voltage dip was achieved; which is a reduced by 45.4% compared to the original response (77.1mV without the helper circuit). An integrated current sensing technique using current mirroring to equalize the drain voltages of main and replica (sense) PMOS devices was utilized. With this technique, the current in the replica (sense) PMOS device is a scaled down version of the current in the main PMOS device. The sensed currents from all three phases are added up and converted to a voltage. This voltage is compared to a reference voltage that represents the limit for over-current. This reference voltage is set to be 20% higher than the average total currents in the three phases combined.

1. Background

This paper is an extension of work originally presented in Power Electronics and Drive Systems (2019 IEEE 13th International Conference on PEDS). Unlike single phase converters, multiphase buck converters are considered one of the state-of-art regulator topologies that provide high current capability (example, some CPUs requires 10A-15A or even more at their peak performances). Compared to a single-phase converter, multi-phase converters were proven to have many benefits, such as higher efficiency, lower component sizes, and reduced output current and voltage ripple [1].

Different features and techniques are addressed in this paper to study the behaviour of the multiphase buck converter. In [2], phase shedding (PS) is defined as a technique used to improve the converter efficiency by disconnecting some phases at light load. When the load decreases, there is no need for using all the phases in the circuit, however, it is possible to reduce the number of phases as the output current is being distributed between those reduced phases. In [3], fast transient is presented as a technique that will guarantee a low transient spike/dip and preserve a stable supply voltage. The two main objectives for using the fast transient are having the shortest recovery time and having the smallest output voltage spike/dip. In [4], integrated current sensing was presented as a technique to implement over-current protection (OCP), and current-mode feedback control. Many different schemes have been executed to utilize the inductor current sensing techniques. Most of these current sensing schemes have limiting constraints such as high-power dissipation, difficult implementation and control, and process-dependence..

2. Features/Issues

In [1], the single-phase buck converter was used as a first step to develop a functional multi-phase buck converter. In this paper, it was also used to test the features added to the final multi-phase design (consists of power and control stages). The buck converter

*Hani Ahmad, Princess Sumaya University for Technology, Al-Jubaiha, Amman, Jordan;ZIP code 11941; PO BOX 1438, 00962795930140, h.ahmad@psut.edu.jo

(the power stage) was designed by sweeping the width of both NMOS and PMOS transistors so that it would handle the current passing through them. The width of the NMOS was chosen to be 23m and the width of the PMOS which is usually 2-3 times larger than the width of the NMOS, was chosen to be 53.2m. These sizes were selected based on RdSon which is designed to be within the typical industrial standard range of (50-100) mΩ.

This paper discusses three features/issues that will be added to the power stage of the design, starting from single phase to the three-phase final design. This is done to observe the impact of the features on each phase. The three features/issues are Phase Shedding, Fast Transient, and Integrated current sensing which will be discussed in the sub-sections A, B, and C.

The schematic design and simulation results of the proposed circuits were done in LTspice using 180nm standard CMOS technology [5].

2.1. Phase Shedding

Phase Shedding technique is defined as the disconnection of phases in multi-phase buck converters when using different loads. This technique is a requirement to achieve the highest efficiency possible. To implement this feature, an NMOS transistor is utilized to realize an ON/OFF switch function as shown in

Figure 1. This NMOS is added to the single-phase buck converter and placed at the output node as shown in Figure 2. An NMOS was chosen due to its smaller size and faster response when it is compared to a PMOS device. To make an NMOS transistor turns on, a voltage source (Vsource) is connected to its gate and to make it turns off, its gate is grounded.

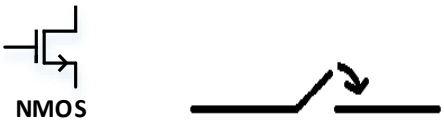


Figure 1: NMOS is used to implement an ON/OFF switch

Vsawtooth in the control stage is adjusted depending on the number of phases in the circuit. In multiphase converter, phase difference is applied to specify the angle that will be used depending on the number of phases in the circuit. Equation 1 was used to define the angle for each implementation. For single phase, the angle is at 0°. For two-phases, the angles are at 0° and 180°. For three-phases, the angles are at 0°, 120°, and 240°. As the number of phases increases, the ripple at the output is reduced.

$$360°/n \text{ ; where n is the number of phases} \qquad 1$$

Single-phase buck converter is utilized to test this phase shedding technique as shown in Figure 2. In this case, Vsawtooth was considered to be at 0°. The parameters used in the single-phase design are listed in Table *1*: Parameters for the single-phase buck converterTable 1 with Vin = 3.3V, Vref = 1.2V, and Vsupply = 2.19V.

Table 1: Parameters for the single-phase buck converter

Power stage	Control stage
Inductor (L1) = 2.17µH	Resistor (R2) = 13.3kΩ
Capacitor (C1) = 15µF	Resistor (R3) = 12.2kΩ
Resistor (R1) = 2.5Ω	Resistor (R4) = 53.7kΩ
	Capacitor (C2) = 86.07µF
	Resistor (R5) = 100kΩ
	Capacitor (C3) = 0.318nF

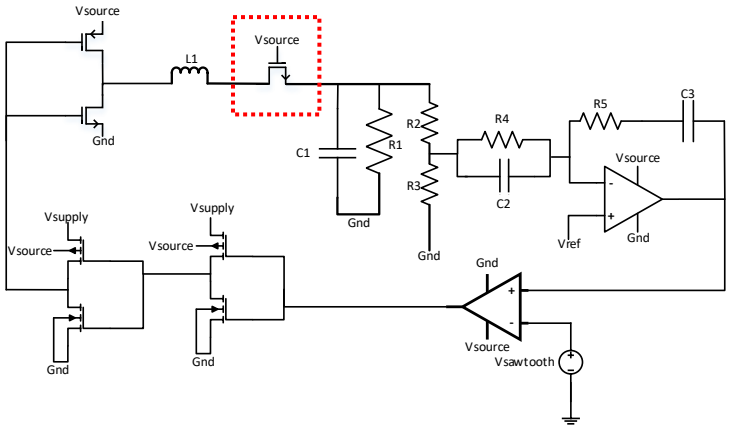


Figure 2: Single-phase Buck converter with NMOS switch used for Phase Shedding

When the NMOS is ON, the output voltage is 2.5V and the output current is 1A as seen in Figure 3.

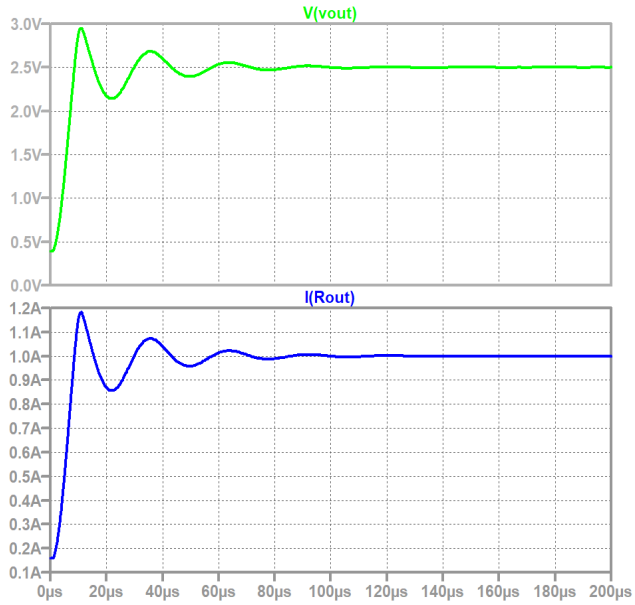


Figure 3: Output voltage and load current when switch is ON

Since the single-phase converter has the minimum number of phases, there is no need to turn off the NMOS switch. Therefore, a two-phase converter was used next to implement the effect of phase shedding as seen in Figure 4. The resistor R1 is a simple representation of the load impedance. Based on Ohm's law, $R = V/I$ and since the voltage is fixed (regulated), the current is changing (due to adding a phase). Since the current is changing, R1 is changed to fit the number of active phases (each phase adds same amount of current in this scenario). For two phase case, R1

is 1.25Ω and Vsawtooth was considered to be the summation of two signals at 0° and 180°.

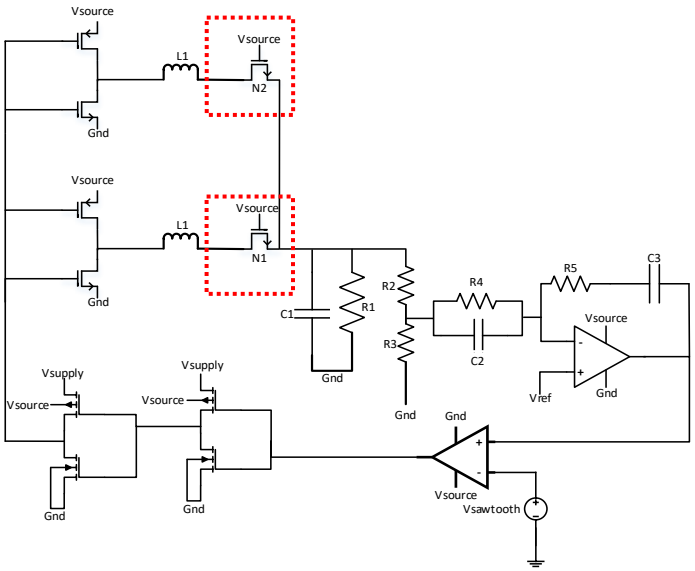


Figure 4: Two-phase Buck converter with NMOS switches used for Phase Shedding

When both N1 and N2 are ON, this means that both phases are being active. In this case, the output voltage is 2.5V and the output current is 2A, having 1A distributed in each phase as demonstrated in Figure 5. When N1 is ON and N2 is OFF, or vice versa, R1 is changed to 2.5Ω due to the number of active phases being decreased. This is donebased on equation 1 above and it consequently maintains output voltage at 2.5V. The output current is now 1A as observed in Figure 6. The results are summarized in Table 2.

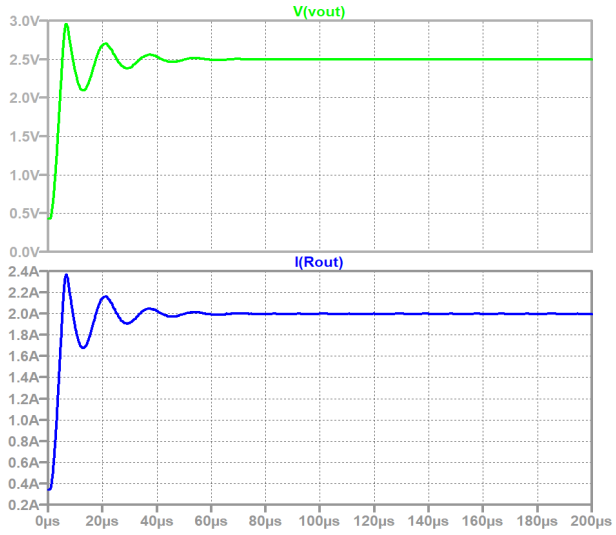


Figure 5: Output voltage and load current when both switches are ON

Table 2: Phase shedding results for two switches

N1	N2	Voltage / V	Current / A
ON	ON	2.5	2
ON	OFF	2.5	1

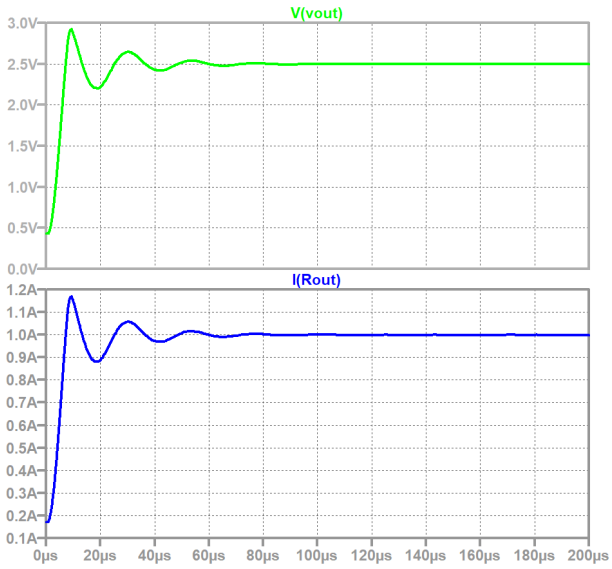


Figure 6: Output voltage and load current when one of the switches is ON

Applying this technique to the three-phase buck converter means having different load and different Vsawtooth. In three-phase scenario, R1 is decreased to 0.833333Ω and the summation of three sawtooth signals is used at 0°, 120°, and 240°. Figure 7 depicts this implementation.

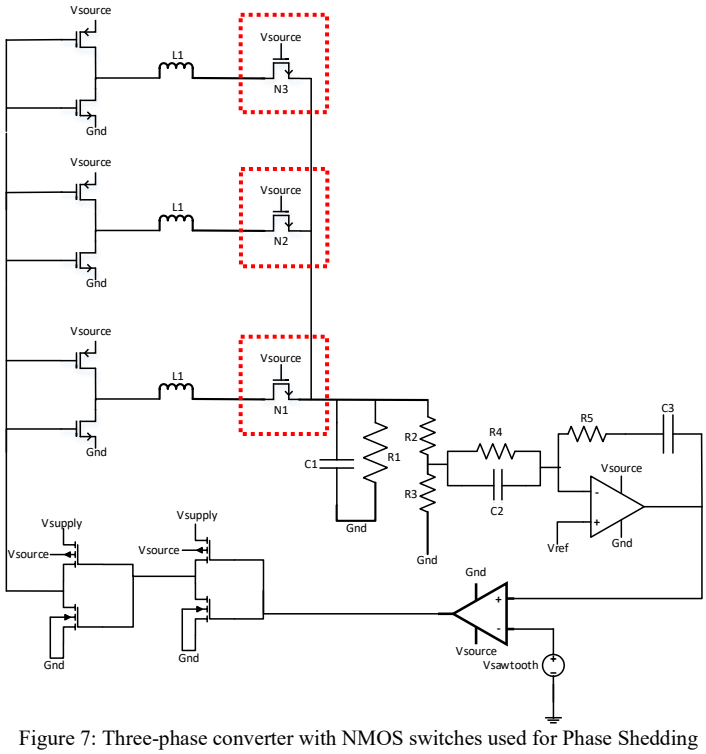


Figure 7: Three-phase converter with NMOS switches used for Phase Shedding

When N1, N2, and N3 are all ON, all phases are used, resulting in having an output voltage of 2.5V and an output current of 3A, having 1A distributed in each phase as shown in Figure 8. When N1 and N2 are ON and N3 is OFF (one phase is OFF), the used R1 is 1.25Ω as it is similar to the two-phase buck converter. The output voltage remains 2.5V and the output current is 2A (1A in each phase) as observed in Figure 9. When N1 is ON and N2

and N3 are OFF (one phase is ON), the simulation results are similar to the single-phase buck converter having 2.5V as an output voltage and 1A as an output current as depicted in Figure 10. The results are summarized in Table 3.

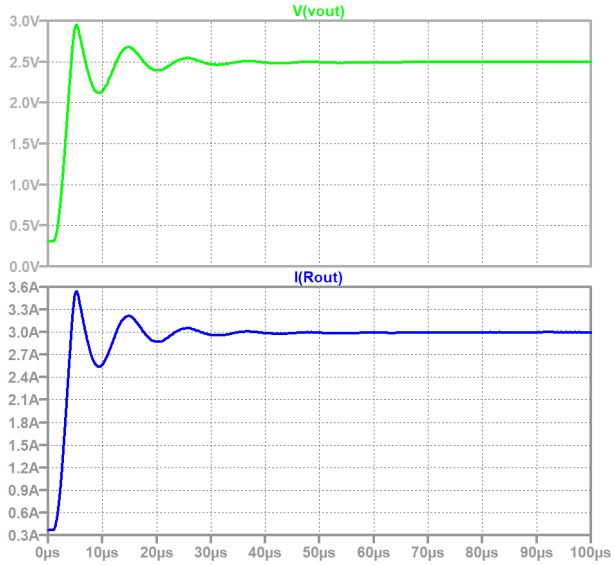


Figure 8: Output voltage and load current when all switches are ON

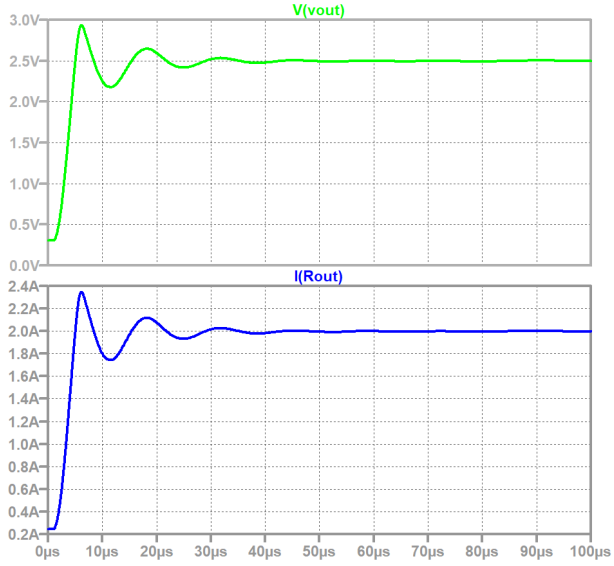


Figure 9: Output voltage and load current when one of the switches is OFF

Table 3. Phase shedding results for three switches

N1	N2	N3	Voltage / V	Current / A
ON	ON	ON	2.5	3
ON	ON	OFF	2.5	2
ON	OFF	OFF	2.5	1

2.2. Fast-Transient

As previously discussed, a current source can be used to force a step in the load current which would cause a spike or a dip in the output voltage [1]. The importance of the fast-transient feature is that the current profile of modern high-performance devices (such as a microprocessor) is very dynamic in nature which means a change from few microamps to full peak value within very small interval of time (say microseconds or even less) is expected. This definitely causes a spike/dip in the regulated output voltage. The dip below 10% of the regulated voltage may cause malfunctioning of the devices in the circuit. A spike above 10% of the regulated output voltage may cause damage to some devices (exceeds breakdown voltage).

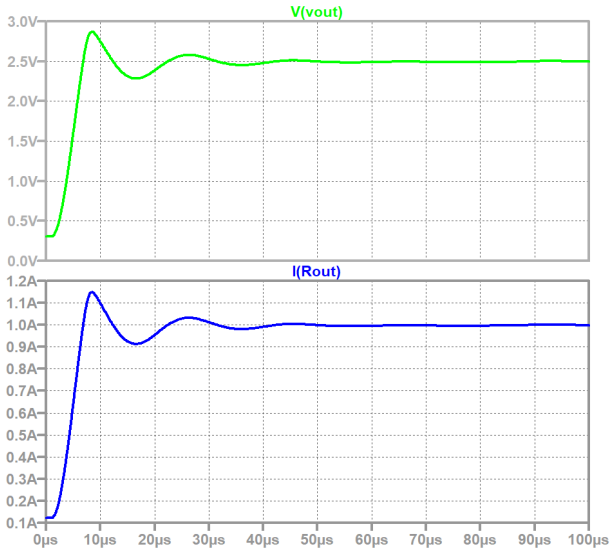


Figure 10: Output voltage and load current when one of the switches is ON

A transient current was inserted to test the response time that would take the controller to pull back the output to its regulated state and to measure the amount of worst spike/dip of the output voltage. In this paper, the current source with pulse signal of 500mA step (in both directions) was added parallel to the load to force the output voltage spike or dip. To mitigate/prevent this kind of problem, a proposed two-part circuitry was developed to assist in minimizing the voltage spike and dip to stay within the tolerable range. The developed solution is shown in Figure 11. The main job of the resistor voltage divider in helper circuit is to put the initial value for the gate voltage such that the transistor is off in the normal operating conditions (no load transient).

When spike occurs at the Vout node, the capacitor (C1) acts as a short circuit (due to high frequency signal of the fast-current step) which causes the gate voltage to increase. Therefore, NMOS transistor immediately starts to turn on. To overcome the spike, NMOS transistor starts pulling current from Vout node and this brings the spike back toward the regulated value. Results are shown in Figure 12 to Figure 15.

Table 4: Parameters of the helper circuits

Spike	Dip
Resistor (R1) = 30kΩ	Resistor (R3) = 32kΩ
Resistor (R2) = 1kΩ	Resistor (R4) = 1kΩ
Capacitor (C1) = 10µF	Capacitor (C2) = 1µF
NMOS size= 70mm	PMOS size= 200mm

When dip occurs at the Vout node, the capacitor (C2) acts as a short circuit which causes the gate voltage to decrease. Hence, PMOS transistor immediately starts to turn on. To overcome the dip, PMOS transistor starts pushing current to Vout node and this

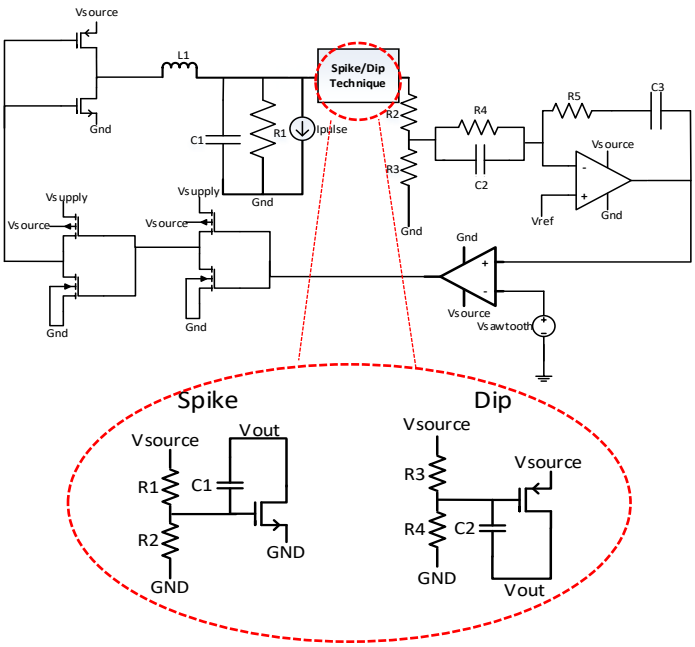


Figure 11: The helper circuits at the point of load



Figure 12: Output voltage spike without the helper circuit

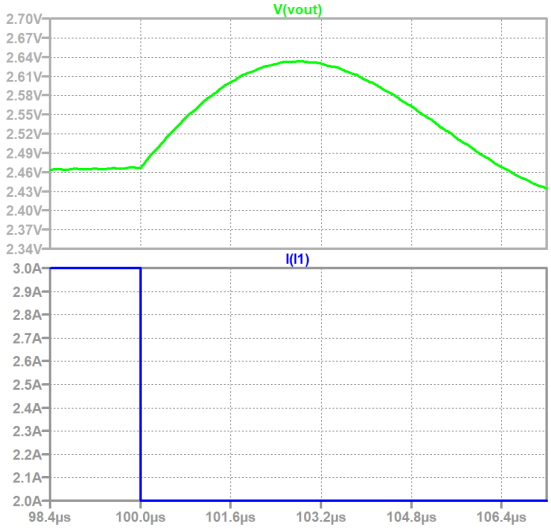


Figure 13: Zoomed-in graph when spike occurs without helper circuit

brings the output voltage back towards its regulated value. Results are shown in Figure 16 to Figure 19. The values of the helper circuits are shown in Table 4.

This fast-transient technique is introduced to the three-phase buck converter. One of the benefits is that the voltage ripple decreases as the number of phases in the circuit increases [1].

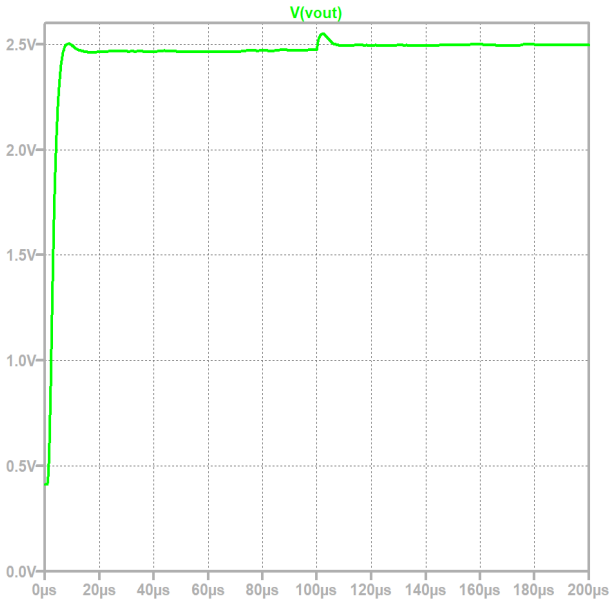


Figure 14: Output voltage spike with the helper circuit

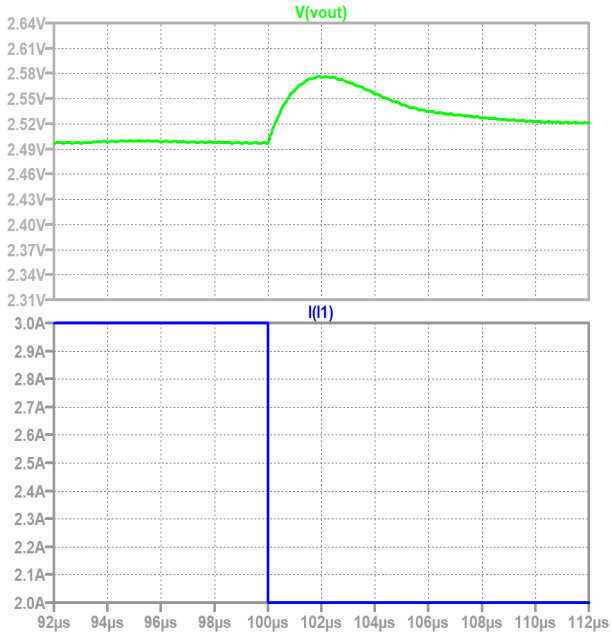


Figure 15: Zoomed-in graph when spike occurs with helper circuit

Table 5: Fast Transient results

Fast transient	Before the helper circuit	After the helper circuit
Spike	160mV	79.1mV
Dip	77.1mV	35mV

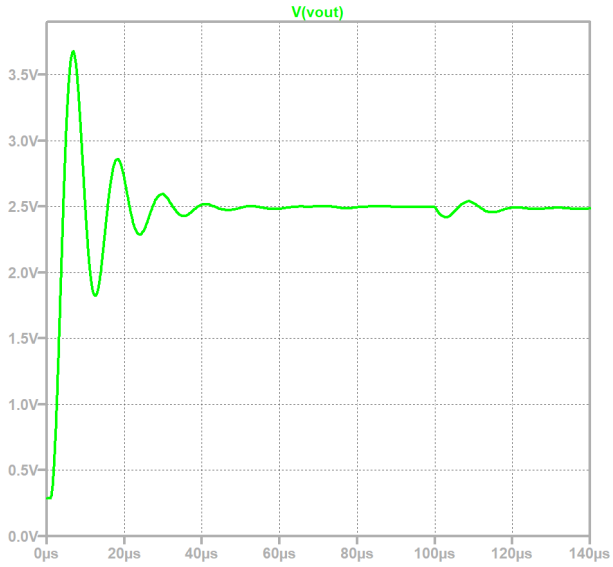


Figure 16: Output Voltage Dip without the helper circuit

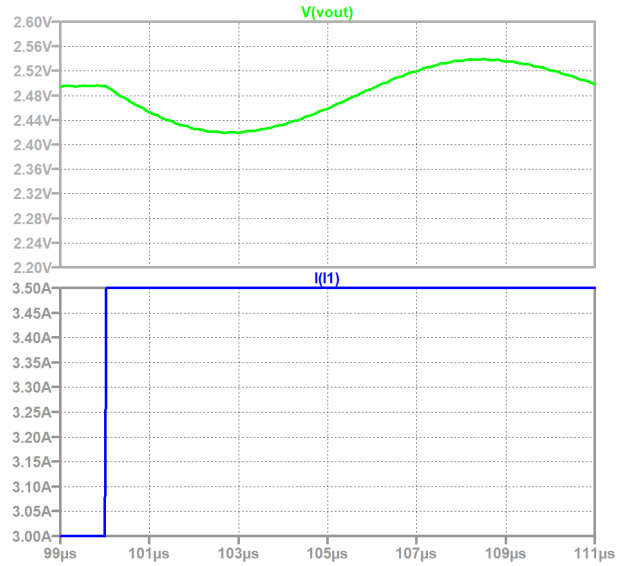


Figure 17: Zoomed-in graph when dip occurs without helper circuit

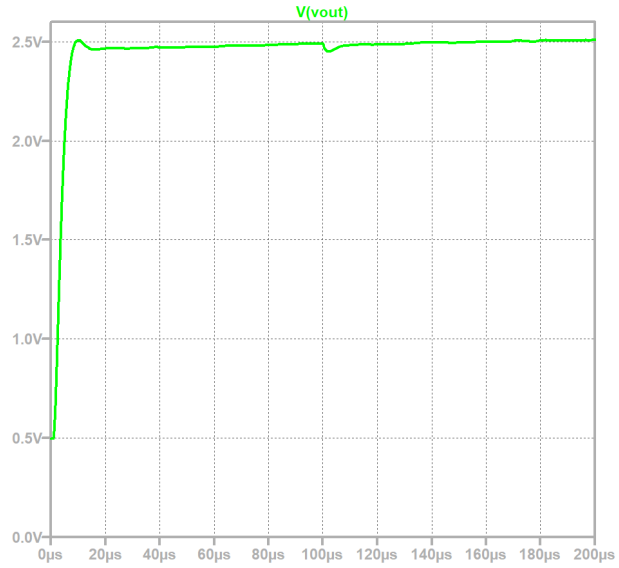


Figure 18: Output voltage Dip with the helper circuit

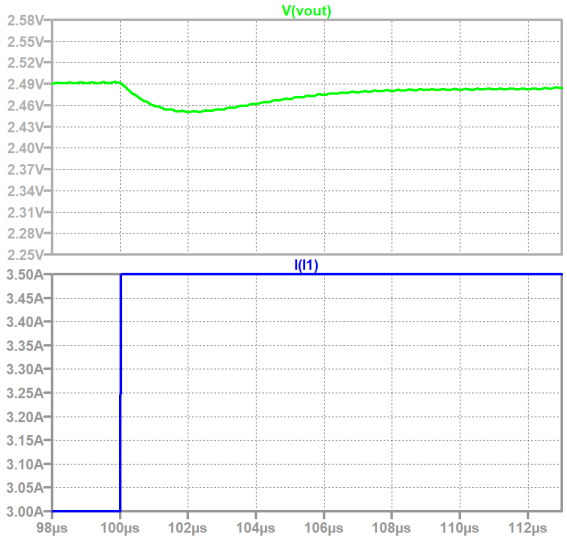


Figure 19: Zoomed-in graph when dip occurs with helper circuit

2.3. Integrated Current Sensing using Current Mirroring

Current sensing can be achieved using many conventional techniques such as Series-Sense Resistor [6]. This a simple technique that uses a resistor in series with an inductor is depicted in Figure 20. When knowing the value of the resistor, the voltage sensed across it will determine the current flowing through the inductor. This method allows higher power loss therefore, it has negative impact on the efficiency of the converter. The series resistor method can also be used by placing the sense resistor in series with the PMOS or NMOS.

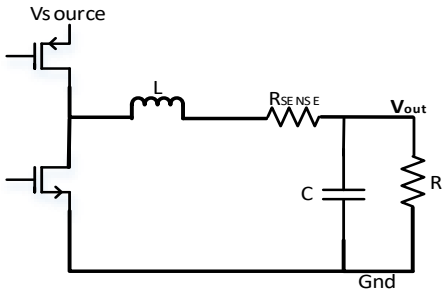


Figure 20: An example of series resistor used in current sensing

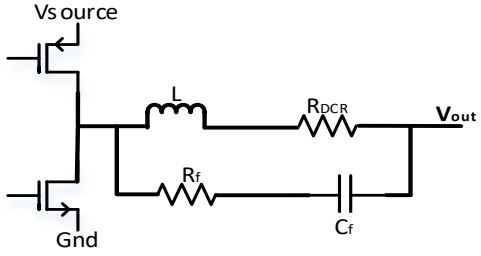


Figure 21: Integrated current sensing by adding a parallel capacitor Cf to the convertor's inductor L

DC resistance (DCR) of switching inductors with a low-pass RC circuit connected in parallel with the inductor is considered to be another method to sense the current as presented in Figure 21 [4]. The disadvantage of using this method is that the RC circuit

has to be chosen based on the inductor and resistor values which is not an easy task as some components are integrated, and others are not.

An Op-Amp can also be used as an additional method for a current sensing technique as shown in Figure 22 [6]. The Op-Amp is used to force the drain voltages of the transistors (power and sense) in the DC-DC converters to be equal. However, if the power and sense transistors width ratio increase (for efficiency reasons), the accuracy of the circuit degrades. The Op-Amp needs to have a high bandwidth and a high gain to have a very accurate current sensing method. Practically, there is not an Op-Amp that complies perfectly to the previous requirements.

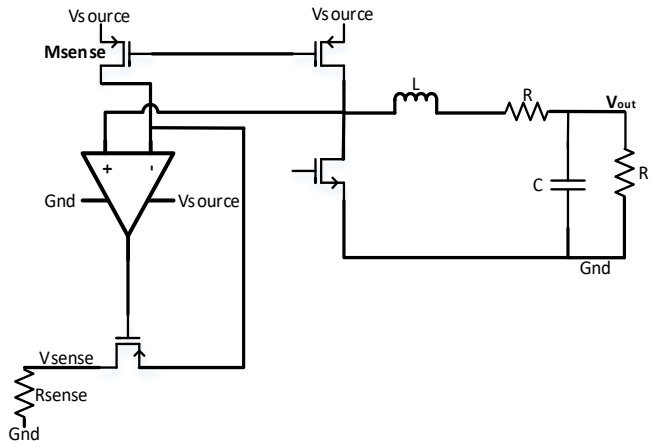


Figure 22: Integrated current sensing using op-amp and sense FET

The selected method for current sensing is the current mirroring circuit [7]. This method is selected due to the higher accuracy in addition to higher efficiency it provides compared to the previous mentioned methods. As shown in Figure 23, MP2 was connected to be a replica (sense) device for MP1. The two sources and gates are physically connected. The drains of these two devices are forced to be equal via the current mirror. Of course, each terminal of the main device and the corresponding sense device need to be equal in value since the both main and sense devices act as switches. The relationship between MP1 and MP2 sizes is presented as a ratio of 1: K. The drain currents flowing through MP1 and MP2 then will have a value of Ip1 and Ip2 which equals to Ip1/K.

In this current-sensing circuit, to have accurate current sensing, the voltages at NB and NC need to be almost equal (drain voltages for main and sense devices). As MP1 and MP2 are turned on, then MN1 is turned off, therefore the voltage at NB and NC is high because of currents Ip2 and Isense are equal and vice versa. This method can force the voltages at nodes NB and NC to be equal as noticed in Figure 24. The width of the transistor MP1 was chosen to have a value of W1 and transistor MP2 was chosen to have a width of W2 which equals to W1/K.

For the current mirror transistors, size sweeping/optimization was conducted to equalize both nodes NB and NC.

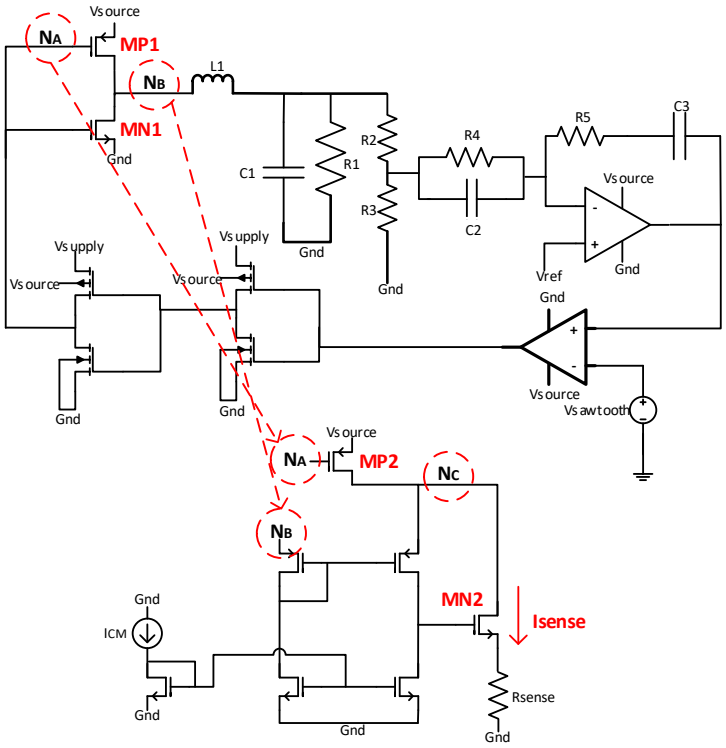


Figure 23: Integrated current sensing based on current mirror

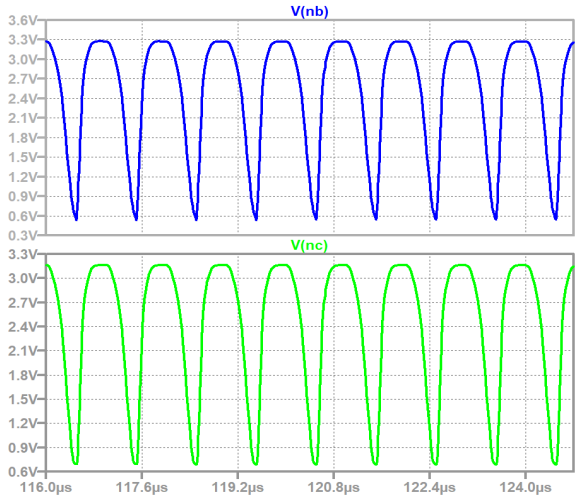


Figure 24: Voltages at Nodes NB and NC

Figure 25 shows the currents Ip1 and Ip2. It can be observed that Ip1 is identical to Ip2 except for the scaling factor of 1: K (Ip2=Ip1/K).

The sensed voltage across Rsense represents the output current of the converter. The output current can be sensed using two different options. The first option is by assuming that all phases are balanced and carry the same current which is 1A per phase. In this case, a single-phase converter can be used to represent whether an over-current condition has occurred or not. Therefore, a single-phase is enough to represent the total current of multi-phase converter. If Isense for the single phase is exceeded by 20% of a reference value, an over-current has occurred and switching will be halted.

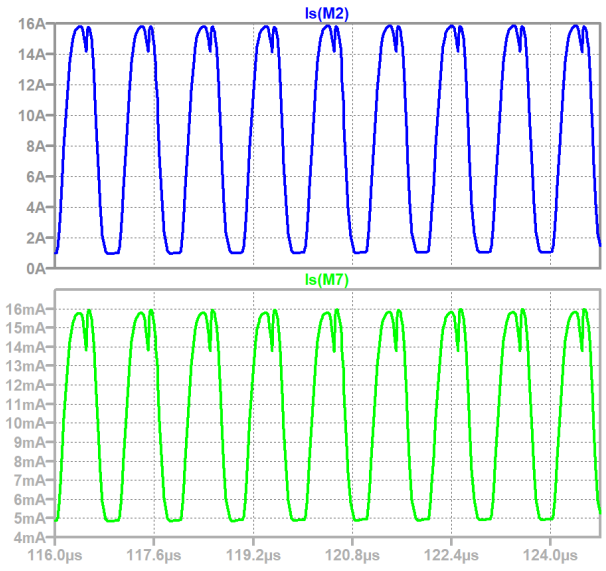


Figure 25: Ip1 and Ip2 of main and sense PMOSs

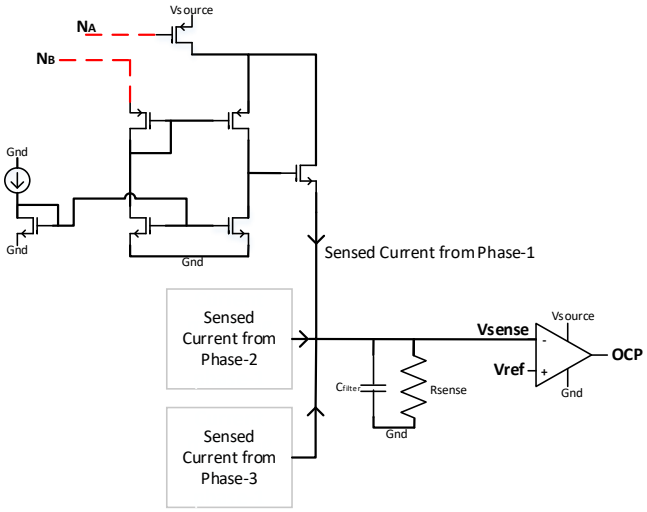


Figure 26: Option two combined sensed current for Three-Phase

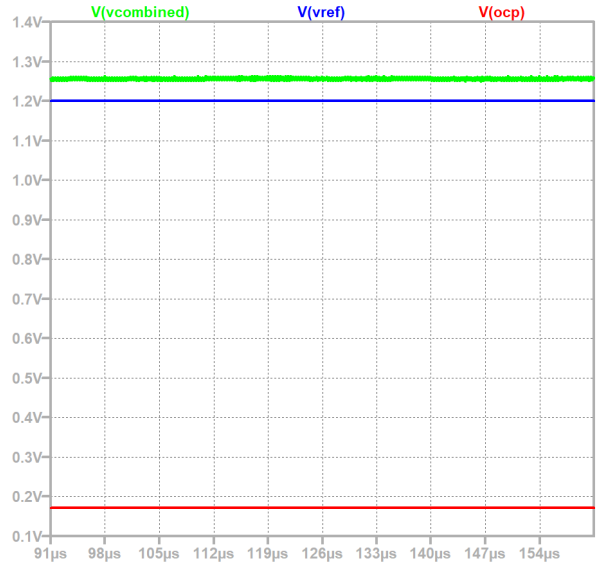


Figure 27: Three-Phase OCP indicator

Another option is possible by combining sensed currents of all phases together regardless of having the current being balanced between phases or not. In this case, the combined current, which is the total current of the converter, is 3A. A reference voltage with a value of 1.2V is presented to assist in detecting an over-current. The sensed voltage is compared to the reference voltage, therefore, if the sensed voltage value is higher than the reference voltage, then no over-current has occurred and vice versa. Of course, this 1.2V is a reference voltage that represent 20% higher than typical 3A. When over-current has occurred then the switching must be halted to protect the circuit. The second option was selected in this paper due to its higher practical value with the assumption of no current balancing. As a result, no over-current was detected in the final design as seen in Figure 26 and Figure 27.

3. Conclusion

It is well known that a multi-phase buck converter has higher capabilities, and efficiency compared to the single-phase buck converter. The phase shedding technique was implemented in the three-phase buck converter utilizing NMOS switches as they are used to disconnect some phases at light load. When the load decreases, there is no need for using all the phases in the circuit. It was concluded that by increasing the number of phases, the efficiency would be improved.

The proposed fast transient helper circuit is tested by inserting a 500mA transient current step in 100µs. Worst case spike of 79.1mV was achieved; which is a reduction of 49.4% of the original response (160mV) without the helper circuit. Worst case of 35mV of a voltage dip was achieved; which is a reduced by 45.4% compared to the original response (77.1mV) without the helper circuit. Moreover, the integrated current-sensing technique using current mirroring was used in the three-phase buck converter to detect the occurrence of an over-current, in order to protect the components in the circuit. A reference voltage with 20% higher than typical value is used to indicate an occurrence of an over current.

References

[1] N. S. Gammoh, M. A. Al-Bader and H. Ahmed-Assi, "Design and Simulation of a Three-Phase Switching Synchronous Buck Regulator," 2019.
[2] A. Costabeber, P. Mattavelli and S. Saggini, "Digital Time-Optimal Phase Shedding in Multiphase Buck Converters," 2010.
[3] Y. Zeng and H.-z. Tan, "Fast-transient DC-DC converter using an amplitude-limited error amplifier with a rapid error-signal control," 2019.
[4] Y.-S. Roh, Y.-J. Moon, J. Park, M.-G. Jeong and C. Yoo, "A Multiphase Synchronous Buck Converter With a Fully Integrated Current Balancing Scheme," 2015.
[5] 6 January 2012. [Online]. Available: http://ptm.asu.edu/modelcard/180nm_bulk.txt.
[6] H. P. Forghani-zadeh and G. A. Rincon-Mora, "Current-Sensing Techniques for DC-DC Converters," 2002.
[7] C.-H. Chang and R. C. Chang, "A Novel Current Sensing Circuit for a Current-Mode Control CMOS DC-DC Buck Converter," 2005.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1071-1081 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Review of Different Methods for Optimal Placement of Phasor Measurement Unit on the Power System Network

Ademola Abdulkareem, Divine Ogbe, Tobiloba Somefun*

Department of Electrical and Information Engineering, Covenant University, Ota, 112107, Nigeria

ARTICLE INFO

Article history:
Received: 06 August, 2020
Accepted: 11 November, 2020
Online: 14 December, 2020

Keywords:
Phasor Measurement Unit
Voltage Phasor
Current Phasor
Global Positioning System
Optimal PMU Placement
Power System Observability

ABSTRACT

Phasor Measurement Unit (PMU) is an integral device for tracking, protection, and regulation of the power network. PMU gives synchronised calculations of actual-time data for voltage phasor, current phasor, and the frequency. Placing PMU in every node to observe the power network is not realistic from an economic standpoint and even for big data management. Thus, raising the number of PMUs in the node network to achieve optimum monitoring of the power network is a mandatory issue. In tracking the problem of optimal PMU placement, various methods have been proposed by various researchers in recent times. To facilitate understanding among the various approaches, this study classified the existing methods into two headings; heuristic and mathematical(conventional)methods. The study discusses the different existing optimisation approaches used to solve the optimal PMU placement problems. The benefits, as well as the drawbacks of each approach, is thoroughly examined and wind up the discussion that the heuristic approach is a fast knowledge-based methodology, which was used as the optimization solver for the IEEE 14, 30 and Nigeria national grid 28 and 52 bus system.

1. Introduction

Phasor measurement unit is an electrical instrument that measures the phase angle and magnitude of current and voltage in the power network system, using a synchronised time source provided by global positioning system, which can provide accuracy up to one microsecond [1]. To decide the condition of a power network, the estimations of the state variables (phase angles and voltage magnitude), of all system nodes should be known. Realising these values helps to calculate the flow of real and reactive power in the power network. It will be very strenuous to consider system nodes one of after the other most especial a large system, this is one area in which intelligent devices find application [2] in this case PMU estimations solve the problem effortlessly.

At a particular period, the voltage phase angle and magnitude and current sinusoidal waveforms are expressed as real and imaginary numbers that are known as a phasor. The value of the magnitude is dependent on the amplitude of the sinusoidal waveform, while the phase angle is gotten relying on the time position as presented in the figures. Figure 1a shows the sinusoidal waveform is behind as for the cosine reference, as it is behind, the phase angle will be negative, though, in Figure 1b, the sinusoidal waveform is ahead as for the cosine reference. Subsequently, the phase angle is positive. Since the phasor measurement unit is provided with the global positioning system, every phasor estimation is time-tagged, consequently, permitting phasor estimations gotten from the phasor measurement units in different positions in a system to synchronise.

PMU firstly receives signals in analogue form. The signal being received contains estimated voltages and currents from current and voltage transformer. The anti-aliasing filter then used in eliminating high-frequency factors surpassing the sampling limit of Nyquist. It will, therefore, suppress the frequency factors that exceed the Nyquist sampling rate. The phase-locked oscillator divides the global positioning system one pulse each second into the needed number of pulses as regards the sampling of the waveform. Through the use of a 16-piece accuracy analogue to a digital converter, the input signals will be digitised at sampling moments when the time signals are sampled from the phase-locked oscillator before feeding it to the phasor microprocessor. The microprocessor resamples the signals from the digitised sample information and estimates the positive succession. Then the

*Corresponding Author: Tobiloba Somefun, km 10, Idiroko Road, Canaanland Covenant University, +2348037632337, Email: tobi.shomefun@covenantuniversity.edu.ng

approximate phasor is time-stepped before being moved using modems to a communication medium [3].

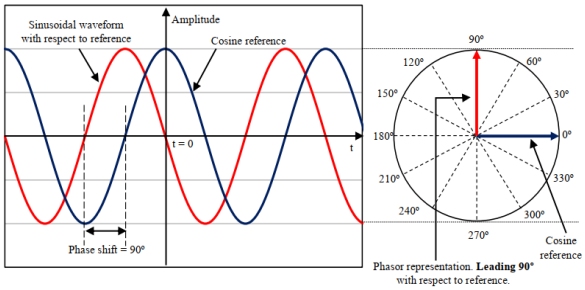


Figure 1: (a) phasor representation. -90º regarding cosine reference [3]

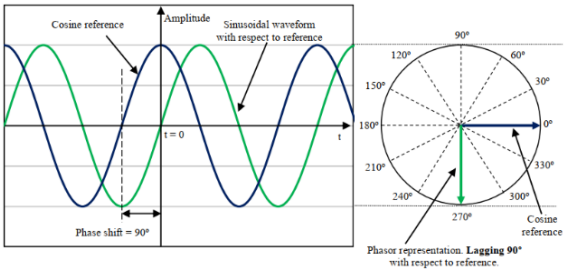


Figure 1: (b) phasor representation. 90º regarding cosine reference [3]

2. Optimal PMU Placement Problem Formulation

There are two main performances used in the analysis for observing power system; they are numerical observability and topological observability [4]. Numerical observability suffers from a lot of matrix computations; thus, it is not advisable to employ them for analysing power systems, especially large ones. Topological observability of a system is achieved when a full rank of traversing the tree is obtained. Some simplified and improved effective rules when topologically analysing a power system are illustrated below:

Rule 1: For a PMU-installed node, voltage phasor and the connecting branches current phasor are known as demonstrated in Figure 2. This measurement is known as direct measurement. In Figure 2, PMU is installed in node 1; consequently, the voltage phasor of the node is known, and the current phasor of the connecting line is also known following the first rule.

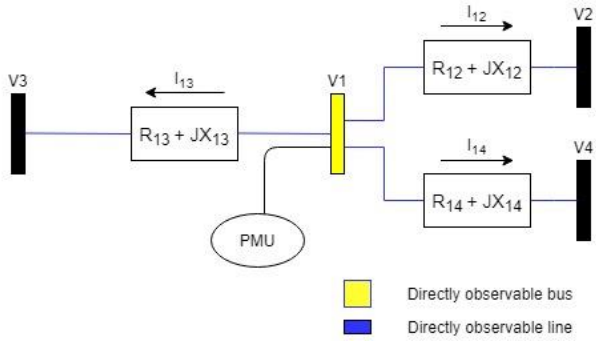


Figure 2: illustrating Direct measurement Rule 1

Rule 2: When the voltage phasor is known, and the current phasor at one end of the branch is also known, it guarantees that the other end is observable by determining the voltage phasor, as illustrated in Figure 3. This type of measurement is known as pseudo-measurement.

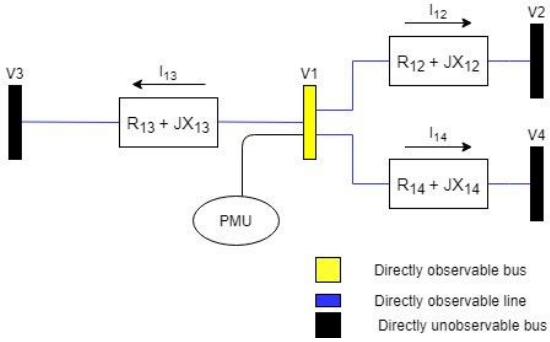


Figure 3: illustrating pseudo-measurement Rule 2

Since the value of the current phasors of the lines connected to node 1 is known, the value of the other nodes can be solved for by utilising Ohms law.

$$I_{12} = \frac{V_1 - V_2}{R_{12} + jX_{12}} \tag{1}$$

$$V_2 = V_1 - I_{12}(R_{12} + jX_{12}) \tag{2}$$

$$I_{13} = \frac{V_1 - V_3}{R_{13} + jX_{13}} \tag{3}$$

$$V_3 = V_1 - I_{13}(R_{13} + jX_{13}) \tag{4}$$

$$I_{41} = \frac{V_4 - V_1}{R_{14} + jX_{14}} \tag{5}$$

$$V_4 = V_1 + I_{41}(R_{41} + jX_{41}) \tag{6}$$

Rule 3: In a case where the voltage phasor for both ends is known, the current phasor of the connecting branch can be gotten as shown in Figure 4. Knowing the voltage phasor of node-1 and node-2, the current phasor of the connecting line can be calculated using Ohm's law.

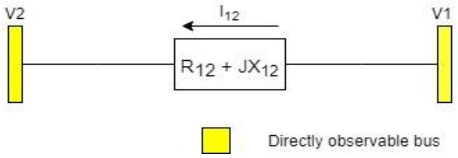


Figure 4: Illustrating observability of the branch current for rule 3

$$I_{12} = \frac{V_1 - V_2}{R_{12} + jX_{12}} \tag{7}$$

Rule 4: For a zero-injection node, knowing all the current of all connected branches, except one, calculating the current for the unknown branch can be achieved using Kirchhoff current law (KCL) this rule is demonstrated in Figure 5. node-3 is the zero-injection node, and current phasor of every branch is known except the current for the branch between node-3 and node-4.

Rule 5: For a zero-injection node where the voltage phasor is not known, if all the neighbouring nodes are known, the zero-injection node can be known through calculations, as shown in Figure 6. In the diagram, node-3 is the zero-injection node and the

voltages for the neighbouring nodes is known. By using node calculations, the voltage for node-3 is known.

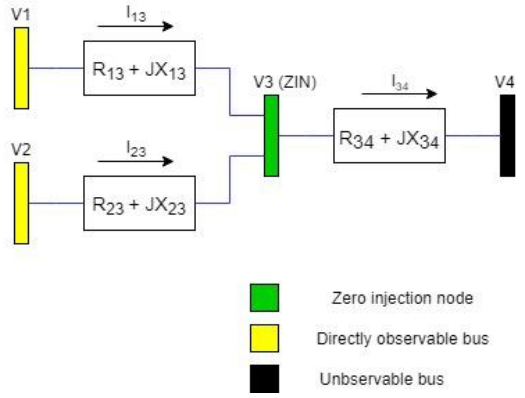


Figure 5: Model observability of zero-injection node for rule 4

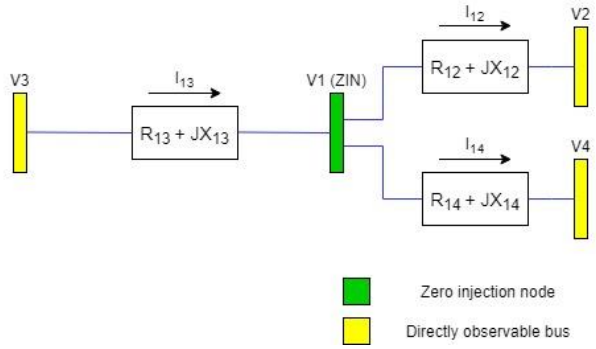


Figure 6 Modelling unobservable ZIN rule 5

$$V_3 = V_1 - I_{13}(R_{13} + jX_{13}) \quad (8)$$

$$V_3 = V_2 + I_{32}(R_{32} + jX_{32} \quad (9)$$

$$V_3 = V_4 + I_{34}(R_{34} + jX_{34}) \quad (10)$$

$$0 = I_{12} - I_{32} - I_{24} \quad (11)$$

Rule 6: For a group of neighboring zero-injection nodes that are not observable because the voltages are not known, but the neighboring nodes connected to the group of zero injection nodes are known. The unknown voltage phasors of the zero-injection nodes can be made known by using both Kirchhoff's current and voltage laws as illustrated in Figure 7. In this figure, the voltage of node-3 and node-4, but the voltages of the nodes connected to them are known.

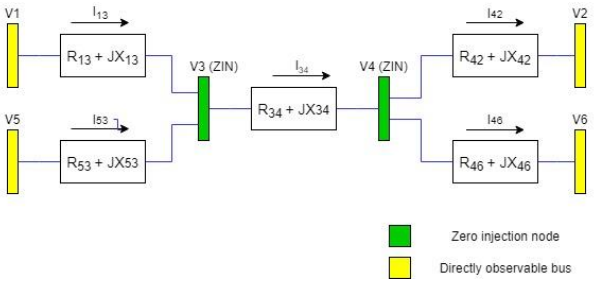


Figure 7 Modelling unobservable ZINs rule 6

Minimum number of PMU: One of the purposes of the optimal PMU placement problem is to resolve the minimum number of PMUs needed for a power system and their right allocation to make sure that the system is totally observable. This objective can be represented mathematically as shown:

$$\sum_{i}^{\text{N nodes}} A_{ij} \quad (12)$$

$$s.t\, B * A \geq 1 \quad (13)$$

$$I = [111 \ldots \ldots 1]^{T}_{N\,X\,1} \quad (14)$$

$$A(i) = \begin{cases} 1 & if\ PMU\ is\ installed\ on\ node\ i \\ 0 & if\ PMU\ is\ not\ installed\ on\ node\ i \end{cases} \quad (15)$$

$$B(i,j) = \begin{cases} 1 & if\ i = j \\ 1 & if\ nodes\ i\ and\ j\ are\ joined \\ 0 & if\ nodes\ i\ and\ j\ are\ not\ joined \end{cases} \quad (16)$$

3. Placement Methods

Several methods have been deployed for optimal location on power system network in different regards such as distributed generators and PMU [5,6]; in this section, several of the methods used to optimally site PMU are extensively reviewed.

3.1. Conventional Method

Two of the conventional methods that have been reported in previous studies are discussed under this subsection.

3.1.1. Integer Linear Programming (ILP)

Integer Linear Programming methods, also recognised as binary integer programming, considering both the system's injection and power measurement as well as PMU error calculation by state estimation. Integer linear programming algorithm is based on the individual vectors derived from the adjacency matrix of the transverse tree [7]. Upon decomposition, the PMUs are positioned ideally in the networks utilising the integer linear programming principle to reduce the deployment rate. Two types of ILP models are mixed integer linear programming (MILP) and binary integer linear programming (BILP) [8].

3.1.2. Integer Quadratic Programming (IQP)

Integer quadratic programming utilises the connectivity matrix that reflects the topology of the system to define the optimal PMU placement (OPP) problem. The quadratic function's objectives were structured, taking into account linear constraints and the vector of integer values. The method will decrease the number of PMUs by providing full network observability in standard operations as well as in outages [8].

In [9], the authors discussed factors that affect calculation precision, including measurements inconsistencies, number of PMU measurements, and quasi-power measurements implementation. Quasi-power uncertainty is determined by the propagation concept of uncertainty, and the optimum location of PMU is calculated by integer quadratic linear programming. Various types of state variables as well exhibited different sensitivities to different measurements [10].

3.1.3. Drawbacks of Conventional Methods

Usually, the conventional technique is utilised to model a network when the network is not so complex. Also, conventional

technique takes more computational time when compared with Heuristic techniques to solve a complicated network problem.

3.2. Heuristic Methods

Most of the Heuristic Methods that have been reported in previous studies are discussed under this subsection.

3.2.1. Genetic Algorithm

The foundation of genetic algorithm is to model natural selection which has no requirement for secondary functions like derivatives calculation. Some of the advantages of genetic algorithm which makes it more applicable to OPP problems are: the likelihood of the local minimum being trapped is reduced, Calculations are decreased from one state to the other, and the fitness assessment of each sequence directs the search [11].

In [12], they investigated optimally placing PMUs and communication links for distributed state estimation in distribution networks. They combined integer linear programming and genetic algorithm to find the least number of PMUs in a distribution network. A distributed state estimation system segmenting technique is created to minimise the burden it takes to compute and communicate [13]. The aim of the partitions is to balance the number of buses in subareas. By turning the network segmenting into the problem of assigning a centralised bus of each subarea, the scale-balanced partitioning scheme between subareas is obtained by applying genetic algorithms.

3.2.2. Particle Swarm Optimisation

Particle swarm optimisation is a similar technique to genetic algorithm, where population solutions are randomly assigned to a system firstly [14]. In [15], the authors described particles as entities that are hovering through multidimensions in space. For every particle, the best location is determined by the fittest position faced by that particle and its neighbouring particles. The process of the particle swarm optimisation begins with a primary point and velocity for every particle, where the velocity of each particle is restricted by reasons of not hovering in unusable space as well as overflowing forbiddance [16].

3.2.3. Tabu Search

The Tabu search algorithm is a combinatorial method, and it combines techniques such as conventional method and heuristic method. This method is majorly used in solving combinatorial optimisation problems in covering and scheduling. Tabu lists, which is a major element of tabu search, consists of numbers of states visited in recent times and unwanted states. Another major element of tabu search is the state definition and the encompassing area. The tabu search resets when it does not converge [17].

In [18], the authors solved the optimal PMU placement problem and achieved maximum observability of a system and maximum redundancy by utilising tabu search-based linear state estimation model [19]. This quick technique of topological observability analysis required a loss simulation function relying on the matrix of incidence to solve the optimal PMU placement problem. This technique also concerns the convenience and high speed to monitorable power network by manipulating integer numbers [20].

3.2.4. Differential Evolution

The method of differential evolution (DE) uses elements such as N-dimensional vectors to reduce progressing space functions. Mutation, fusion and choice are the key functions used in the global optimising process. Differential evolution approach is utilised extensively in various problems of cost function including multimodal, non-differentiable, and nonlinear functions. Certain advantages of this method include simultaneous computations, simple use, and strong synchronisation properties [21].

The researchers introduced multi-objective optimal PMU placement in [22] utilising a non-dominating filtering DE technique, that is an organic synthesis of the nondominated filtering procedure and DE technique Pareto. This theory put into consideration total measurement reliability and voluntary PMU failure to achieve complete observability of the network in addition to finding a solution to the optimal PMU placement problem [3] Using the differential evolution technique obtained from genetic algorithm contributed to the suggested method for the nondominated sorting differential evolution algorithm. It was stated that obtaining a particular and absolute Pareto front and having several Pareto ideal solutions was the improvement of this method.

3.2.5. Simulated Annealing

The simulated annealing method solves complex combinatorial optimisation where the existing solution is changed arbitrarily. The new changed solution is the worst change with a decreased probability as computing continues. The ideal outcome for a major problem of combinatorial optimisation requires an effective disturbance framework, cost function, space for the solution, and cooling plan for the simulated annealing to solve. Simulated annealing can be considered adequate by looking for a large-scale network and achieving good speed in seeking an ideal or near-ideal solution [23]

Authors in [24], investigated effective use of different PMU estimated values and exposure to extremely critical network data were put into consideration for the ideal deployment of PMUs to make the power system fully observable. A topological observability formulation based on the incidence matrix evaluated the original PMU arrangement to have a network with complete observability.

3.2.6. Ant Colony Optimization

A further concept used to present an optimisation problem solution is ant colony optimisation, which originally utilises the ant population. The function of the ant colony was to travel across neighbouring states of the problem by implementing an ideal controller for the stochastic local resolution, resulting in solving optimising problem. Evaporation of the pheromone trace and action of the daemon are procedures in ant colony optimisation. The use of ant colony optimisation to determine good paths across graphs will decrease computational issues [25].

Authors in [26], investigated on an enhanced ant colony optimization which was used to solve the ideal PMU placement problem for achieving a totally visible power network with the least number of PMUs and putting into consideration total measurement redundancy. The depth-first search as a conceptual

graph approach was used to create an estimation tree to evaluate if the power system was observable [27].

3.2.7. Iterated Local Search

The key viewpoint of iterated local search is to search for a subspace characterised by local optimal solutions rather than the entire solution space. By using an integrated heuristic, there is a series of solutions where the best option is achieved if repeated arbitrary tests of the heuristic were to be used [28].

The optimal PMU placement method introduced in [29] involves two phases which included an initial PMU dispensation to reach a visible network by using an iterated local search to decide the least number of PMUs required to make a system fully visible. Page rank placement formulation is utilised in this approach to determine the significance of every bus.

3.2.8. Mutual Information

Taking into account not only access to a system with maximum observability but also analysing irregularities in the network states that utilised mutual information between PMUs measurements values and system states, and information theoretical method to solve optimal PMU placement problems was addressed in [30].

3.2.9. Matrix Reduction

For the problems of placing PMUs, a distribution matrix occurs, whereby the graph indicates the distribution scope whenever a station is placed in different locations. The problem level is determined by the range of the distribution. The approach of matrix reduction aims to establish ideal placement by increasing the matrix of incidence. Using a formulation based on removing digital data, the matrix reduction technique was proposed to achieve the least number of PMU to reach observability of the network completely and reduce the computation process [31].

3.2.10. Imperialistic Competition Algorithm

Imperialistic competition technique is a new technique formulated to solve optimal placement problems. Just like other heuristic techniques, imperialistic competition algorithm begins with a first population named country which is a colony of two types. Competitiveness of the countries brings about the minimal of the problem. The effectiveness of the Imperialistic competition technique in wide area optimisation was evaluated on test buses and as confirmed [32].

3.2.11. Immune Genetic Algorithm

The Immune Genetic Algorithm was utilised to proffer solution to placing the PMU optimally utilising three impact vaccinations to create a topologically measurable power system. The two measures that occur in an immune genetic algorithm approach to defend against viruses and bacteria are vaccines and immunity alternatives. The vaccination basis is the combination of technical knowledge and previous data about the optimal PMU placement problem. Two operators, namely, crossover and mutation, were considered by the immune genetic algorithm, which was utilised to maximise the tests [33].

3.2.12. Biogeography Based Optimisation

The biogeography-based technique is a mathematical model for the movement of species from one region, which is known as a habitat to another region, rising and extermination conditions of species. This technique is used to provide solutions to the problem of optimisation through complex performance environments and to improve the efficiency of species migration and integration into the habitat [34].

3.2.13. Chemical Reaction Optimization

To find a solution to the OPP problems, a newly formed heuristic approach named population-based chemical reaction technique was implemented. Through modelling the behaviour and movement of gases in a chemical reaction, reaching a low energy steady state was the key goal of the chemical reaction techniques, which sought to achieve the least free energy level. High performance was demonstrated by the application of the chemical reaction technique to metrics and practical problems [35].

Using a new heuristic approach called chemical reaction technique and simplified chemical reaction technique model, optimal PMU placement problem solved to achieve a completely measurable power network considering the system with and without zero-injection nodes [36].

3.2.14. Artificial Bee Colony (ABC)

Artificial bee colony technique's key motivating criteria are food resources path, food resources range, and food resources performance information [37]. The developed formulation could be used in solving unimodal and multi-modal mathematical optimization problems. Also, a multiple objective optimal PMU placement algorithm, called binary-coded was implemented in solving the optimal PMU placement problem and obtain the least quantity of PMU and highest possible network redundancy. Consideration was given to eventuality constraint conditions like a one-line outage, and arrangements of the least quantity of needed PMUs were determined [37].

3.2.15. Bacteria Foraging Algorithm

Biological evolution propensity towards animal exclusion with weak feeding strategies that work to find, handle, and consume food and spread the genes of animal performance in feeding approaches as they are more probable to achieve productive results. Weak foraging approaches after several years are either discarded or revamped [4]. Four operators utilised in the developed bacterial foraging algorithm were chemotaxis, swarming, reproduction, and removal and distribution [38].

3.2.16. Cellular Learning Automata

Cellular learning automata technique is cantered on the use of training automaton to change cellular automaton state conversion likelihood. This approach begins by determining each cell's inner condition. Instead, on the basis of cellular learning automata law, the strengthening signal for every training automata is calculated. Maintaining this operation will achieve the intended state [39].

In [29], a CLA approach was used to examine the integration of two opposing objective functions like the least number of PMUs and total measurement reliability to achieve a completely measurable power network. This research put into consideration

the drawbacks of flexibility, including PMU/branch loss and traditional measurements and zero-injection nodes.

3.2.17. Hybrid Methods

Utilising a formulated hybrid genetic algorithm and simulated annealing, a solution was introduced in [38] to ideally place PMUs and remote terminal units for a large network of interconnected grids determined by a multiarea system state estimate. Using remote terminal units and conventional measurements, a PMU was incorporated into a power network to ensure the calculated state more precisely and reduce the expense of conventional measurement and cost of the remote terminal unit. Utility centre of one multiarea area to reach system measurements required a single PMU placement, as PMU measurement calculated the network node voltage. A criterion for the PMU was the node with the highest connecting line. The identification of faulty data was achieved by taking into account the vital measurement of each region.

A synthesis of least traverse tree technique with the optimised genetic algorithm is a hybrid method named MST-GA presented in [40] to achieve the least number of PMUs required to make a system fully observable and to consider optimising redundancy. This approach enhanced mutation operation taking into account grid topology information. As the key and side steps, fusion and mutation were used as an activity to create new entities, the outcome of the new analysis of this approach was a decrease in the number of PMUs required and a variety of solutions [41].

To easily capture the objectives and contributions of the previous studies with respect to methods/techniques deployed Table 1 is developed while Table 2 gives the comparison of conventional and heuristic methods with respect to four factors namely; the size of the problem, data requirement, accuracy, and effectiveness.

3.2.18. Drawbacks of Heuristic Methods

To efficiently implement the heuristic technique, it takes a specific level of expertise and experience [4,8]. Often these functionality specialists are difficult and costly to employ, particularly if they need domain knowledge. If the analysts are not included in the development, they may not be informed of the technical design constraints or why such design choices have been made [8]. Heuristic tests are poorly organized, and thus risk identifying one-time, low-priority issues. Lastly, heuristic assessments do not allow for an assessment of the performance of re-designs [3].

Table 1: Objectives and contributions of investigated studies

Author(s)/Reference Number	Objective	Technique	Considerations	Gaps	Benefits
B. Mallikarjuna, P. Gopakumar, M. J. B. Reddy and D. K. Mohanta [23]	Ideal positioning of PMUs to allow monitoring of incorrect data. simulated annealing technique is implemented with the stochastic new approach	Simulated Annealing	detection of critical measurements	Zero-injection nodes (ZIN)& Measurement Redundancy was not considered	Observability of the power network with critical measurements
C. Margarida, K. Xenia and V. Ana, [42]	Limiting the quantity of PMU allocation utilising stochastic simulated annealing algorithm	Stochastic Simulated Annealing	Detection of critical measurements is introduced as a consequence function	The algorithm was not Line loss, PMU the outage	Detection of critical measurements
Z. Zhida, Y. Hao, L. Peng, L. Peng, K. Xiangyu, W. Jianzhong and W. Chengshan [12]	Reducing the quantity of PMUs, and determining their geographical allocation, achieving a total network of measurable power. A genetic algorithm-based procedure is used, and PMUs are installed with current phasor measurements as the maximum quantity of simultaneous lines are highlighted in all network nodes	Genetic Algorithm	Connection of PMUs with the number of current phasors to be calculated	Quality of the result were not documented.	Needed channels for current in the optimisation problem
C. Peng, H. Sun and J. Guo. [43]	Limiting the quantity of PMUs and increasing the redundant measurements. In place of one optimal solution, Pareto-ideal solutions are presented	Non-Dominated SortingGenetic Technique	Zero-injection nodes & Measurement Redundancy	It was assumed that PMUs capacity is limitless. PMU channels limit was not considered	Offering Pareto-ideal facade forcontradictory objectives, solution repair
S. M. Mazhari, H. Monsef, H. Lesani and A. Fereidunian [39]	A multi-objective ideal allocation to minimise the quantity of PMUs and increase the accuracy of measurements. PMU allocation is also analysed by the implementation of a standardised observability function based on traditional non-synchronous	Cellular Learning Automata	Traditional measurements, PMU failure, Branch loss	This technique might not give the least quantity of PMUs to make the network observable	High performanceinelectrical net works on a large scale
A. A. Hany, Y. A. Almoataz and V. Mukherjee [25]	The ABC definition is used to attain the least quantity of PMUs to achieve a total measurable power network, fulfilling measurement redundancy	Artificial Bee Colony	One branch loss, zero-injection nodes	The pheromone variation span is restricted to interval [0, -1],	System viability and efficiency is shown by the comparison of the simulation results with the previous works

N. C. Koutsoukis, N. M. Manousakis, P. S. Georgilakis and G. N. Korres [19]	Two contending purposes involving the least quantity of PMU and maximum redundancy on the network	Tabu Search	Zero-injection nodes & Measurement Redundancy	Placement of PMU in ZINwas not allowed in the initialization.	Optimal problems solved with high precision and reduced calculation
R. Ramachandran and S. Karthick [44]	Limiting the quantity of PMU placement and developing a parallel tabu search formulation	Parallel Tabu Search	Communication restraint, Zero-injection nodes, and matrix condition of the state estimations	Traditional measurements, PMU failure, Branch loss was not considered	Reduced calculation period
M. Nazari-Heris and B. Mohammadi-Ivatloo [4]	Two contending purposes involving the least quantity of PMU and maximum redundancy	Clonal Algorithm	Total redundancy	ZIN enhancement of secondary voltage control operation was not considered	Maximum velocity of the procedure, achieving practical schemes
M. Zhou, V. Centeno, A. Phadke, Y. Hu, D. Novosel and H. Volskis [31]	Reducing the quantity of PMUs. Utilising pre-processing techniques and providing solutions using mathematical based methods	Matrix Reduction	Computer-generated data removal, Pre-processing technique and matrix reduction procedure, utilising Lagrangian relaxation	Quality assessment was not recorded because there was no other tested optimal placement set for the Brazilian network	Minimising the size of allocation model and the calculation exertion, employed in a large-scale network
S. Md, I. H. Md, M. Abido, A.-F. Taher and M. A.H [15]	Limiting the quantity of PMUs so as the power network can be fully observed. Developing a hybrid technique based on binary particle swarm optimisation	Binary Particle Swarm Optimisation	Increasing redundancy in the power network, one PMU and Multi-PMU failure	loss of PMU and branch loss were ignored	Maximum velocity of the procedure and reduced calculation period
A. A. Hany, Y. A. Almoataz and V. Mukherjee. [25]	Reducing the quantity of PMU in the system. Developing an improved ant colony technique	Ant Colony Optimisation	High redundant measurements in the system	The pheromone variation span is restricted to interval [0, -1],	Avoiding inactivity behaviour and elevated velocity of the procedure, employing a graph theoretic technique
S. S. Noureen, V. Roy and S. B. Bayne [8]	Reducing the quantity of PMU and remote terminal units with critical measurements free	Hybrid Genetic Algorithm and Simulated Annealing	Traditional measurement and RTU faulty data recognition, one-line outage	Fault monitoring, ZIN was ignored	Relevant to existing power networks observed utilising remote terminal units
S. Mahapatra, S. Pandab and S. C. Swaina [33]	Limiting the quantity of PMU placement and increasing measurements redundancy, taking considerations of traditional measurements	Bacterial Foraging Algorithm	Total redundancy & zero-injection nodes	The linear model could not correctly depict complex dynamics of the network, particularly in disturbances.	Appropriate for existing power networks due to modelling standard measurement
M. N. Seyyed and M. Javad [21]	Limiting means square error by achieving the least quantity of PMUs, with or without the presence of standard measurement	Differential Evolution	Normal measurement least square error of state estimate	Fault monitoring, Zero-injection nodes was ignored	The precise, fast and straightforward procedure, ability to employ in multiple objectives problem
B. Rohit and B. Biplab [45]	Limiting the cost of installation of Mismodelling dissimilar cost of PMUs for various nodes	Particle Swarm Optimisation	Dissimilar cost of PMU allocations	It was assumed that PMUs capacity is limitless. PMU channels limit was not considered	Taking consideration of real-time equipping cost of PMUs, limiting entire cost in place of the quantity of PMUs
B. Rohit and B. Biplab [46]	Reducing the number of meters and PMUs Taking consideration of one line/meter loss and one line/PMU loss	Biogeography Based Optimisation	Zero-injection nodes, line loss, SCADA meter loss, PMU the outage, PMU/branch loss	Traditional measurements, PMU failure, Branch loss was not considered	using simulated node reduction method for decreasing the size of the network
M. Hurtgen and J.-C. Maun. [28]	Reducing the overall quantity of PMUs utilising PageRank placement technique and iterated local search	Iterated Local Search	PMU outage	Zero-injection nodes (ZIN)& Measurement Redundancy was not considered	Taking consideration of eventualities, simple comprehension and implementation
S. S. Noureen, V. Roy and S. B. Bayne. [8]	Limiting the quantity of PMUs. Observing guide nodes needed for secondary voltage control	Branch and Bound	Zero-injection nodes enhancement of secondary voltage control operation	PMU outage, branch loss, measurements redundancy, considered	Observing guide nodes for boosting speed of the voltage control system
Y. Zhao, P. Yuan, Q. Ai and T. Lv [22]	Reducing the quantity of PMUs to monitor both the overall network and fault in the power network	Differential evolution	Fault monitoring, Zero-injection nodes	The constraints utilized ignored some optimal solutions	Utilising PMUs for power network monitoring and fault detection

S. S. Noureen, V. Roy and S. B. Bayne [8]	Limiting means square error by achieving the least quantity of PMUs	Differential evolution	Taking consideration of incessant alterations in the power network's topology	Fault monitoring, ZIN was ignored	Integrating PMU allocation problem into state estimate.
A. Ahmadi, Y. Alinejad-Beromi and M. Moradi [16]	Limiting the quantity of PMUs, recommending new the rule for the network to be observed topologically	Modified binary particle swarm optimisation	Zero-injection nodes, PMU/Branch loss	Measurement redundancy was not studied, and the problem of local least affected the results	Establishing new rules of topologically Observable valuation for limiting the quantity of the needed PMUs
H. Mohammadi, G. Khademi, D. Simon and M. Dehghani [34]	Two contending purposes involving the least quantity of PMU and maximum redundancy	Multi-objective biogeography-based optimisation	measurements redundancies, zero-injection nodes.	The large amount of cost evaluations utilized made a little difference in the solution	The proposed technique generates well-circulated Pareto-optimal solutions
M. Nazari-Heris and B. Mohammadi-Ivatloo [4]	Multi-objective model, reducing the quantity of PMUs and increasing measurements	Genetic algorithm	Considering single branch/single PMU loss considered	ZIN enhancement of secondary voltage control operation was not ignored	Very little population and iterations needed are less
L. Qiao, C. Tao, W. Yang, N. Rohit, F. Franz and M. D. Ilić [30]	Reducing the quantity of PMUs utilising anInfo-theoretic approach	Mutual information	PMU outage, standard measurement	ZIN was not considered	Modelling the improbabilities in the network states
A. Bashian, M. Assili, A. Anvari-Moghaddam and O. R. Marouzi [11]	Reducing the quantity of PMUs needed for a total observable network and increasing measurements redundancy. A topologically observable rule of zero-injection nodes is also presented	Binary imperialistic completion algorithm	PMU outage, branch loss, measurements redundancy,	Cost considering ZIN was ignored	Quick integration, less deviance, competence of locating global optimal and zero standard deviation
S. Li and Z. Meng [37]	Artificial bee colony technique is implemented to attain least Quantity of PMUs to achieve total observability of the power network, a fitting measurement redundancy	Artificial bee colony	One branch loss and Zero-injection nodes	The technique in this study was only tested on small scale network	Viability and performance of the technique validated by comparison of the simulation outcomes with the related previous research

Table 2: Comparison of Conventional and Heuristic Technique

S/N	Properties	Conventional Technique	Heuristic Technique
1	Size of the Problem	Usually, the conventional technique is utilised to model a network when the network is not so complex.	Heuristic technique is a form of computation which tries to mimic human or animal capability in a very simplified manner in order to proffer solution to the problem of PMU allocation
2	Data requirement	The conventional method utilises physical principles to obtain network relationships, which are typically rightfully generalised with several assumptions and need a previous understanding of the nature of the data relationships.	The heuristic technique focuses on data based on the data only to decide the framework and variables to solve a network problem, with fewer assumptions about the network's nature.
3	Accuracy	As compared to Heuristic technique, because of several factors implemented, the conventional technique does not have a better degree of correctness.	Owing to very little possibilities of failure, Heuristic technique has higher precision.
4	Effectiveness	The conventional technique takes more computational time when compared with Heuristic techniques to solve a complicated network problem.	The heuristic technique can be used effectively in performing routine and tasks that consume time.

4. Results

According to the comparison in Table 2, heuristic technique did better in terms of complexity of the network, data requirement, and level of correctness and effectiveness as regards to time. For this research genetic algorithm is selected the optimization solver for optimal placement of PMUs on IEEE14, 30 and Nigerian 330KV (28 and 52 bus system) networks. It was selected in place of other heuristic algorithms because it is one of the pioneer evolutionary algorithms, simplicity and ease of implementation, and logical reasoning behind the use of operators like selection, crossover and mutation. The mutation operator is executed bit by bit after offspring is produced, for every node in the chromosome expression, an evenly appropriated random number is created between zero and one. If the number produced is lower than the likelihood of the mutation, then the corresponding bit will be

modified from zero to one or one to zero, thereby putting or eliminating the PMU on the appropriate node. Once all the offspring have been generated, the current population and the created offspring are categorized based on the decreasing order of the objective function provided, and a tournament selection is carried out to determine the new generation. Again, a difference is made based on the size on the problem, the top 10 chromosomes are chosen to utilize the tournament selection scheme and inserted directly into the current generation. The suggested approach is implemented to find a solution to the problem by taking consideration of these cases: Normal operation; and Case considering Zero Injection (ZIN)

4.1. Normal operation

In this case, the solution for the optimal PMU placement problem is obtained by ignoring ZIN.

Table 3: Normal operation

	IEEE14	**IEEE 30**	**NG 28**	**NG 52**
Number of PMU	4	10	11	17
PMU location Bus number	2, 6,7, 9	1, 5, 8, 11, 12, 19, 23, 26, 29	1, 3, 4, 12, 13, 17, 20, 21, 23, 24, 27	3, 4, 6, 8, 10, 16, 19, 24, 26, 27, 32, 35, 39, 41, 42, 46, 51

Table 3 shows PMU placement result for the IEEE 14, IEEE 30, Nigeria bus 28 and 52 considering normal arrangement of the bus.

4.2. Case considering ZIN

In this case, the solution for the optimal PMU placement problem is obtained by considering ZIN.

Table 4: Locations of the ZINs

Bus	**Zero Injection location**
IEEE-14	7
IEEE-30	6, 9, 22, 25, 27,28
NG-28	11
NG-52	-

Table 4 shows the location of the zero-injection node for each bus considered.

Table 5: Considering Zero Injection Node

	IEEE14	**IEEE 30**	**NG 28**	**NG 52**
Number of PMU	3	7	10	N/A
Bus PMU location	2, 6, 9	1, 2, 10, 12, 19, 23, 27	1, 3, 4,12, 17, 20, 21, 23, 24, 27	N/A

Table 5 shows PMU placement results for the IEEE 14, IEEE 30, Nigeria bus 28, and 52 considering zero injection node. When a zero-injection node is considered, the number of PMUs needed to observe a system is reduced because that particular bus is merged with a connecting node.

Table 6: System of Redundancy Index (SORI)

Bus	GA	ILP [54]	IQP [55]
IEEE14	NPMU: 4 SORI: 19	NPMU: 4 SORI: 19	NPMU: 4 SORI: 19
IEEE 30	NPMU: 10 SORI: 52	NPMU: 10 SORI: 50	NPMU: 10 SORI: 50
NG 28	NPMU: 11 SORI: 46	N/A	N/A
NG 52	NPMU: 17 SORI: 71	N/A	N/A

Table 6 shows a benchmark with previous studies that utilized a mathematical method considering a normal operation. The result is favourably compared with the previous work carried out on IEEE 14 and IEEE 30 bus system. However, the result obtained in this study is superior to the mathematical methods since it gives the same number of PMU and a higher quality of SORI. Based on this fact, Genetic Algorithm is recommended and is adequately fit heuristic technique for PMU placement which improved resilience of the system and enhanced its reliability.

5. Conclusion

The problem of placing PMU optimally was solved using GA heuristic technique and the quality of the result was compared to results of previous research, from the comparison as shown in the result, it can be observed that the system of redundancy index for the heuristic algorithm used did better than the mathematical programming method when implemented on the IEEE 30 bus system, and achieved the same quality for the IEEE 14 bus system. Methods such as conventional and heuristic techniques, exist to solve the optimal placement problems; this study presented a thorough review study on both techniques. Stabilisation and control of Optimal PMU Placement (OPP) have become a power network necessity. We covered a magnificent set of techniques in this study. This would give researchers a structure for classifying these methods and choosing the appropriate solution prior to applying for PMU placement. Analysis of optimisation techniques in this research shows that minimising the OPP would be a framework and help identify optimisation conditions for projections in time to come.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research work and publication charge is fully funded by Covenant University Centre for Research, Innovation and Discovery (CUCRID).

References

[1] A.G. Phadke, "Synchronized Phase Measurement in Power System," IEEE Computer Applications in Power, **6**(2), 10–15, 1993, doi:10.1109/67.207465.

[2] P.O. Oluseyi, T.E. Somefun, O.M. Babatunde, T.O. Akinbulire, O.O. Babayomi, S.A. Isaac, D.E. Babatunde, "Evaluation of energy-efficiency in lighting systems for public buildings," International Journal of Energy Economics and Policy, **10**(6), 2020, doi:10.32479/ijeep.9905.

[3] N. Rahman, A.F. Zobaa, "Optimal PMU placement using topology transformation method in power systems," Journal of Advanced Research, **7**(5), 625–634, 2016, doi:10.1016/j.jare.2016.06.003.

[4] M. Nazari-Heris, B. Mohammadi-Ivatloo, Application of heuristic algorithms to optimal PMU placement in electric power systems: An updated review, Renewable and Sustainable Energy Reviews, 2015, doi:10.1016/j.rser.2015.04.152.

[5] T.E. Somefun, C.O.A. Awosope, A. Abdulkareem, A.S. Alayande, "Deployment of power network structural topology to optimally position distributed generator within distribution system," Journal of Engineering Science and Technology Review, **13**(1), 2020, doi:10.25103/jestr.131.2.

[6] T.S. Shomefun, A. Ademola, C.O.A. Awosope, A.I. Adekitan, "Critical review of different methods for siting and sizing distributed-generators," Telkomnika (Telecommunication Computing Electronics and Control), **16**(5), 2018, doi:10.12928/TELKOMNIKA.v16i5.9693.

[7] S.P. Singh, A.K. Thakur, S.P. Singh, "PMU Placement for Maximum Observability of Power System under Different Contingencies," in Energy Procedia, 2017, doi:10.1016/j.egypro.2017.05.208.

[8] S.S. Noureen, V. Roy, S.B. Bayne, "Phasor measurement unit integration: A review on optimal PMU placement methods in power system," in 5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017, 2018, doi:10.1109/R10-HTC.2017.8288967.

[9] Y. Xueliang, Z. Hengxu, J. Cao, "Analysis of influencing factors in hybrid state estimation," in Asia-Pacific Power and Energy Engineering Conference, APPEEC, 2016, doi:10.1109/APPEEC.2016.7779892.

[10] N. Xie, F. Torelli, E. Bompard, A. Vaccaro, "A graph theory based methodology for optimal PMUs placement and multiarea power system state estimation," Electric Power Systems Research, 2015, doi:10.1016/j.epsr.2014.08.023.

[11] A. Bashian, M. Assili, A. Anvari-Moghaddam, O.R. Marouzi, "Co-optimal PMU and communication system placement using hybrid wireless sensors," Sustainable Energy, Grids and Networks, 2019, doi:10.1016/j.segan.2019.100238.

[12] Z. Zhao, H. Yu, P. Li, X. Kong, J. Wu, C. Wang, "Optimal placement of PMUs and communication links for distributed state estimation in distribution networks," Applied Energy, 2019, doi:10.1016/j.apenergy.2019.113963.

[13] P.P. Bedekar, S.R. Bhide, V.S. Kale, "Optimum PMU placement considering one line/one PMU outage and maximum redundancy using genetic algorithm," in ECTI-CON 2011 - 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011, 2011, doi:10.1109/ECTICON.2011.5947933.

[14] R. Babu, B. Bhattacharyya, "Optimal allocation of phasor measurement unit for full observability of the connected power network," International Journal of Electrical Power and Energy Systems, 2016, doi:10.1016/j.ijepes.2016.01.011.

[15] M. Shafiullah, M.I. Hossain, M.A. Abido, T. Abdel-Fattah, A.H. Mantawy, "A modified optimal PMU placement problem formulation considering channel limits under various contingencies," Measurement: Journal of the International Measurement Confederation, 2019, doi:10.1016/j.measurement.2018.12.039.

[16] A. Ahmadi, Y. Alinejad-Beromi, M. Moradi, "Optimal PMU placement for power system observability using binary particle swarm optimization and considering measurement redundancy," Expert Systems with Applications, 2011, doi:10.1016/j.eswa.2010.12.025.

[17] C. Lu, Z. Wang, Y. Yu, "Optimal PMU placement for pessimistic dynamic vulnerability assessment," IET Generation, Transmission and Distribution, 2018, doi:10.1049/iet-gtd.2017.1261.

[18] Q. Xu, L. Yu, Y. Wang, X. Kong, X. Yuan, "Optimal PMU Placement for the Distribution System Based on Genetic-Tabu Search Algorithm," in 2019 IEEE PES Innovative Smart Grid Technologies Asia, ISGT 2019, 2019, doi:10.1109/ISGT-Asia.2019.8881465.

[19] N.C. Koutsoukis, N.M. Manousakis, P.S. Georgilakis, G.N. Korres, "Numerical observability method for optimal phasor measurement units placement using recursive tabu search method," IET Generation, Transmission and Distribution, 2013, doi:10.1049/iet-gtd.2012.0377.

[20] A. Yda Amira, G. Fathi, A. Amira, G.F.-M.-C. on Systems, & S., U. 2011, "Optimal PMU placement for full network observability case of the tunisian network," Ieeexplore.Ieee.Org, 1–5, 2011, doi:10.1109/ssd.2011.5986789.

[21] S.S.M. Nosratabadi, J. Modarresi, J.M.-2017 S.G. Conference, undefined 2017, "Phasor measurement units placement considering double contingency by differential evolution algorithm based on pareto method," Ieeexplore.Ieee.Org, **2018-Janua**, 1–6, 2018, doi:10.1109/SGC.2017.8308885.

[22] Y. Zhao, P. Yuan, Q. Ai, T. Lv, & T.L.-I.J. of E.P., U. 2015, "Optimal PMU placement considering topology constraints," Elsevier, **73**, 240–248, 2015, doi:10.1016/j.ijepes.2015.04.022.

[23] B. Mallikarjuna, P. Gopakumar, M.J.B. Reddy, D.K. Mohanta, "A case study on optimal phasor measurement unit placement for emerging Indian national smart grid," in International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings, Institute of Electrical and Electronics Engineers Inc.: 1956–1960, 2017, doi:10.1109/SCOPES.2016.7955789.

[24] S. Akhlaghi, "Optimal PMU placement considering contingency-constraints for power system observability and measurement redundancy," in IEEE Power and Energy Conference at Illinois (PECI), Urbana, IL, USA, IEEE:, 2016, doi:10.1109/PECI.2016.7459251.

[25] A.Y. Abdelaziz, H.A. Abdelsalam, V. Mukherjee, "Optimal PMU Placement in a Distribution Network Considering Network Reconfiguration," Ieeexplore.Ieee.Org, 2015, doi:10.1109/ICCPCT.2014.7054997.

[26] B.K.S. Roy, A.K. Sinha, A.K. Pradhan, "Optimal phasor measurement unit placement for power system observability - A heuristic approach," in IEEE SSCI 2011 - Symposium Series on Computational Intelligence - CIASG 2011: 2011 IEEE Symposium on Computational Intelligence Applications in Smart Grid, 77–82, 2011, doi:10.1109/CIASG.2011.5953335.

[27] R. Salem, A.Y. Abdelaziz, A.M. Ibrahim, R.H. Salem, Optimal PMU Placement for Complete Observability Using Heuristic Methods, 2012.

[28] M. Hurtgen, J.C. Maun, "Optimal PMU placement using Iterated Local Search," International Journal of Electrical Power and Energy Systems, **32**(8), 857–860, 2010, doi:10.1016/j.ijepes.2010.01.028.

[29] S. Chakrabarti, E. Kyriakides, "Optimal Placement of Phasor Measurement Units for Power System Observability," IEEE TRANSACTIONS ON POWER SYSTEMS, **23**(3), 1433, 2008, doi:10.1109/TPWRS.2008.922621.

[30] Q. Li, T. Cui, Y. Weng, R. Negi, F. Franchetti, M.D. Ilic, "An information-theoretic approach to PMU placement in electric power systems," IEEE Transactions on Smart Grid, **4**(1), 446–456, 2013, doi:10.1109/TSG.2012.2228242.

[31] M. Zhou, V.A. Centeno, A.G. Phadke, Y. Hu, D. Novosel, H.A.R. Volskis, "A preprocessing method for effective PMU placement studies," in 3rd International Conference on Deregulation and Restructuring and Power Technologies, DRPT 2008, 2862–2867, 2008, doi:10.1109/DRPT.2008.4523897.

[32] A. Ghasemkhani, A. Anvari-Moghaddam, J.M. Guerrero, B. Bak-Jensen, "An Efficient Multi-objective Approach for Designing of Communication Interfaces in Smart Grids," Ieeexplore.Ieee.Org, Nov. 2020, doi:10.1109/ISGTEurope.2016.7856286.

[33] S. Mahapatra, S. Panda, S.C. Swain, "A hybrid firefly algorithm and pattern search technique for SSSC based power oscillation damping controller design," Ain Shams Engineering Journal, **5**(4), 1177–1188, 2014, doi:10.1016/j.asej.2014.07.002.

[34] H. Mohammadi, G. Khademi, D. Simon, M. Dehghani, Multi-Objective Optimization of Decision Trees for Power System Voltage Security Assessment, Nov. 2020, doi:10.1109/SYSCON.2016.7490524.

[35] M.H.F. Wen, J. Xu, V.O.K. Li, "Optimal Multistage PMU Placement for Wide-Area Monitoring," IEEE TRANSACTIONS ON POWER SYSTEMS, **28**(4), 2013, doi:10.1109/TPWRS.2013.2277741.

[36] J. Xu, M.H.F. Wen, V.O.K. Li, K.C. Leung, "Optimal PMU placement for wide-area monitoring using chemical reaction optimization," in 2013 IEEE PES Innovative Smart Grid Technologies Conference, ISGT 2013, 2013, doi:10.1109/ISGT.2013.6497845.

[37] S. Li, Z. Meng, "Optimal PMU placement based on improved binary artificial bee colony algorithm," in 2017 IEEE Transportation Electrification Conference and Expo, Asia-Pacific, ITEC Asia-Pacific 2017, Institute of Electrical and Electronics Engineers Inc., 2017, doi:10.1109/ITEC-AP.2017.8081008.

[38] P.K. Ghosh, "Complete and incomplete PMU observablity for hybrid state estimation," in 6th International Conference on Computation of Power, Energy, Information and Communication, ICCPEIC 2017, Institute of Electrical and Electronics Engineers Inc.: 867–870, 2018, doi:10.1109/ICCPEIC.2017.8290487.

[39] S. Mahdi Mazhari, S. Member, H. Monsef, H. Lesani, A. Fereidunian, "A Multi-Objective PMU Placement Method Considering Measurement Redundancy and Observability Value Under Contingencies," Ieeexplore.Ieee.Org, 2013, doi:10.1109/TPWRS.2012.2234147.

[40] B. Ozsoy, M. Gol, "A Hybrid State Estimation Strategy with Optimal Use of Pseudo-Measurements," in Proceedings - 2018 IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT-Europe 2018, Institute of Electrical and Electronics Engineers Inc., 2018, doi:10.1109/ISGTEurope.2018.8571513.

[41] M.M. Rezaoui, D. Ridha, A.A. Laouid, A. Kouzou, R.D. Mohammedi, "Optimal PMUs Placement Using Hybrid PSO-GSA Algorithm," Ieeexplore.Ieee.Org, 25–27, 2019, doi:10.1109/ICPEA1.2019.8911178.

[42] M. Carvalho, X. Klimentova, A. Viana, "Observability of power systems with optimal PMU placement," Computers and Operations Research, **96**, 330–349, 2018, doi:10.1016/j.cor.2017.10.012.

[43] C. Peng, H. Sun, J. Guo, "Multi-objective optimal PMU placement using a non-dominated sorting differential evolution algorithm," International Journal of Electrical Power and Energy Systems, **32**(8), 886–892, 2010, doi:10.1016/j.ijepes.2010.01.024.

[44] R. Ramachandran, S. Karthick, "Optimal PMU Placement for Tamil Nadu Grid Under Controlled Islanding Environment," Procedia Technology, **21**, 240–247, 2015, doi:10.1016/j.protcy.2015.10.021.

[45] B. Rohit, B. Biplab, "Optimal Placement of Phasor Measurement Unit using Binary Particle Swarm Optimization in Connected Power Network," in 2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON), IEEE, Allahabad, India: 1–5, 2015, doi:10.1109/UPCON.2015.7456695.

[46] K. Jamuna, Y.L. Selva, G. Gnanasekaran, "Placement of PMUs for power system observability using biogeography based optimization," in Proceedings - 2013 3rd International Conference on Advances in Computing and Communications, ICACC 2013, 355–358, 2013, doi:10.1109/ICACC.2013.76.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 6, No. 2, 1074-1083 (2021)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Synthesis of CVs Using a Context-free Grammar

Darren Tafadzwa Semusemu [*], Abejide Ade-Ibijola

Formal Structures, Algorithms and Industrial Applications Research Cluster, Department of Applied Information Systems, School of Consumer Intelligence and Information Systems, University of Johannesburg, Bunting Road Campus, Johannesburg, 1021, South Africa

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 09 December, 2020
Online: 28 April, 2021

Keywords:
Context-free grammar
Synthesis of things
CV presentation

ABSTRACT

To get hired, an aspiring candidate needs a good CV/résumé. This is not an easy task as it must be readable, well structured, concise with no grammatical errors, and containing all relevant information needed by employers. In this paper, we present a newly designed context-free grammar for the dynamic synthesis of a CV. The grammar rules were implemented in a software tool called Flex-CV. Flex-CV makes use of user input and selected CV templates to produce many CV instances in LaTeX. Examples of CVs generated with Flex-CV can be found at `https://tinyurl.com/cv-instances`*. We then evaluated this tool based on its perceived usefulness among job seekers. Evaluation results indicated that this tool will be useful to those aiming to improve their employment prospects through better CV presentation.*

1 Introduction

Curriculum Vitaé, often abbreviated as "CV", roughly translates in Latin as "life's course" [1]. It is important to note that in some countries CV and résumé are not synonymous. In commonwealth countries, a CV is merely a short document — detailing professional and academic careers — serving as a summary of the job seekers personal information, employment, education, qualification, and other information relating to a job applicant [2]. While in other countries such as USA, Canada, Australia, India, etc; a CV is a comprehensive document mainly used in academic [3]–[5] and medical fields [6, 7]; such that, a résumé — meaning short summary — is used for industry, non-profit, and public sector job applications [8]. This paper refers to CV and résumé synonymously as a means towards job application.

A CV is usually provided to a potential employer when seeking employment; this serves as a means to screen applicants before the interview process is conducted [9]. In search for employment, CV presentation may greatly influence the likelihood of a job seeker being a potential job candidate [10]. Previous studies give an insight into how relevant information and personal characteristics from a CV are often omitted or incorrectly stated in such a way employers are not able to extract positive attributes [11].

With successful job search comes employment such that we can easily account for our most needed basic needs — including living expenses such basic housing, food, and water among other needs. Apart from also providing much needed economic growth [12]. Employment also contributes to our mental health and physical health [13]–[15] — affecting confidence and self-efficacy [16]. As a result, endeavours undertaken to improve CV writing significantly increase the chances of obtaining employment [9].

In computing, string generation and manipulation has become important in many different fields [17]–[19]. In this paper, we focus on providing a theoretical and practical solution to the problem of CV presentation. This is achieved by designing a new context-free grammar (CFG) for the synthesis of a user's CV, which we implemented in a tool that aids with the job search process. The use of grammars is critical as we can specify input by the user, and get a pseudorandomly generated number of CVs, each having shuffled sections and subsections, with minor errors pre-checked. A CFG consists of several rules which are simply a model of a language. One rule which does not depend on context can be applied in order to generate a sentence [20].

Many tools exist as solutions to CV problems but many are lacking in features and fail to deliver on what makes a good CV, i.e. readable, well structured, concise with no grammatical errors, and containing all relevant information. However, there exist many job search intervention programs and they seek to further develop skills such as communication. It is known that a tool that helps with CV presentation can improve an applicant's CV for a better job search process. In order to address issues related to CVs a means

[*]Corresponding Author: Darren Tafadzwa Semusemu, darrensemusemu@gmail.com

for job applications, Figure 1 illustrates the process this research undertakes for the synthesis of a CV. This process requires that a user completes a form with the details that should go into their CV, then the CV synthesiser normalises, standardises and synthesises a CV based on existing templates underlined by CFG rules. The result is an instance of a CV (and in some cases, many instances of a CV for the candidate to choose from).

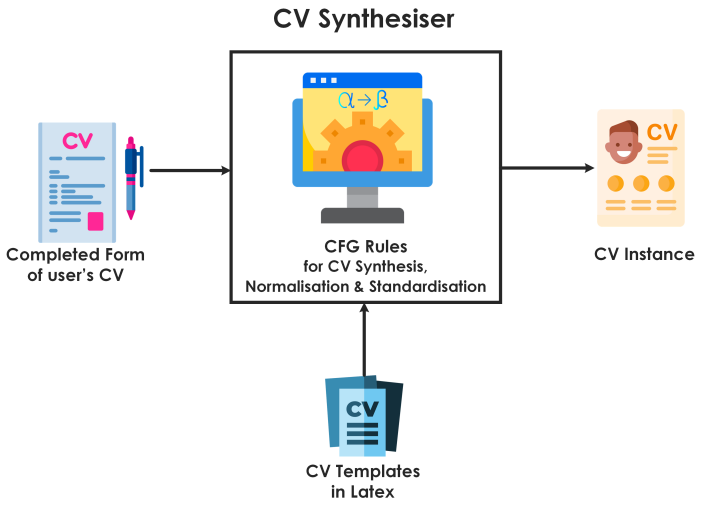


Figure 1: The Process of CV Synthesis

Recently there has been many works on the synthesis of things using CFGs [21]–[24]. This paper makes the following contributions to these trend of works:

1. designed a new CFG for the dynamic synthesis of a CV, and
2. presented an online tool that implements the newly designed CFG rules.

This paper is structured as follows. Section 2 present a background to this paper, Section 3 discussed related works. In Section 4, we presented the design of a new CFG for the synthesis of a CV. Section 5 showcases the implementation of a tool (named Flex-CV) that implements the CFG rules and sample results of synthesised CVs. In Section 7 the conclusion and future work is presented.

2 Background

In this section, we first give an in-depth analysis on the use of a CV as a means towards employment and as a self presentation tool. We consider factors during initial job fit valuations and common sections found in CVs.

2.1 A Means to Employment: Major Theories

In a study aimed at understanding how CVs are handled by recruiters, in [2], the author noted there exist defining factors influencing hiring managers decisions during the hiring process; such that, these factors need to be well understood. Although dependent on the communication channel, these factors help filter out candidates in the recruitment process. Some of these factors include and are not limited to: conducted interviews [25], CV presentation [26], related work experience and skills [27], personal or behavioural traits [28], any form of formal assessments [29], gender [12], social economic status [30], and the recruitment channel used [31]. For an individual, employment and its perceived benefits are important.

In [32], the author stated that employment offers benefits associated with our psychological well-being (latent benefits) and benefits associated with remuneration (manifest benefits). Individuals only realise manifest benefits when engaged in paid work, while employment then provides latent benefits such as time structure, social interaction, shared goals, status, and activity [33, 34]. However, unemployed individuals lack these benefits. When compared to their employed counterparts, they then experience lower psychological health [34]. In [35], the author note that psychological health — our emotional and mental well-being — was significantly lower in studies with unemployed long-term workers. Further supported in [15] and [13], the study notes that for unemployed individuals, physical health greatly suffers as the transition to unemployment starts; as for mental health, the decline starts even before the result of unemployment. However, job search interventions can be set as a means to help with employment.

In an aim to better understand the effectiveness of job search interventions, in [9], the author identified four theoretical mechanisms which may be used as a guide for planning job search interventions. These four mechanisms are as follows [9]:

Behavioural learning theory: result of behaviour modification [36]. As opposed to learning from direct instruction, this learning theory suggests it may be more effective to change behaviours when the desired behaviour is constantly reinforced. As such, job searching consists of many sets of behaviours, some of which include networking and CV writing.

Theory of planned behaviour: result of predicting one's behaviours [37]. In job search, this theory aims to predict behavioural performance, as such, predicting the effort put in by the job seekers and the likelihood of success.

Social cognitive theory: result from the analysis of social behaviours associated with physical functioning [16]. This theory notes the importance of goal setting, specifically, the relationship that exists between a job seeker's present reality and their expected outcome.

Coping theory: results from managing stress over time [38]. In the context of job search, unemployment harms the psychological and physical well-being. This has serious outcomes when reconsidering employment.

Encompassing some of the major theories discussed, in [9], the author identified a self-regulated framework of critical components most seen in job search programs. The focus being on skills development and motivation enhancement. Skills development was identified as being made up of: (1) a self-presentation component, where training and helping an individual's skills (e.g. CV/resume writing, interview preparations, application forms) is important; and (2) a teaching job skills component; whereby job seekers are taught to find relevant jobs based on skills such as how to network in order

to obtain possible job leads over others. In the next subsection we discuss the use of a CV as a self-presentation component.

2.2 Curriculum Vitae: A Self Presentation Component

Self-presentation makes use of the following major theories: Behavioural learning and Social cognitive theory [9]. As previously stated, improving self-presentation provides assistance related to CV writing. In line with interview question conducted by recruiters, CV presentation is one of the main factors affecting the application success even before the screening process is conducted by recruiters [12]. In order to screen applicants, [2] acknowledge how recruiters use different tools, including CVs, as one of the more formal recruitment techniques towards employment. In [10], the author argue that the way job seekers present themselves on CVs, as well as during interviews, greatly influences their chance of employment. However, most job seekers are not aware of what makes an overall good CV.

Initial valuations of job fit are made by examining an applicant's CV, thereafter, comparing knowledge, skills, and abilities of the individuals with stated job requirements [10]. In [11], the author indicated how during the job search process, individuals ignore and dismiss information that enables an employer to discern important attributes and traits. As with job search interventions perceived success [9, 39], aiding job seekers by highlighting their skills through self-presentation strategies, such as CVs, greatly increases the chances of employment.

In [2], the author note of various changes in the Human Resources domain, one being of how there is a shift towards more autonomy resulting in less human oversight during the recruitment process. As it is common with the use of CV screening applications, missing key information (e.g. a key word) results in several applicants being left out. It is up to these applicants to have to be willing to adjust to these screening processes in hopes of employment. In an attempt to increase the likelihood of employment, [12] indicate that correctly highlighting knowledge, skills, and attributes in CVs is a step towards closer step to obtaining employment.

There are key problems in current literature. In [40], the author note of how in literature, more endeavours are undertaken to understand and further describe the realities of unemployment; as a result, work aimed at assisting individuals has attracted less interest. In an effort to better understand the success associated with job search interventions, the likelihood of obtaining employment is three times higher when job seekers are participants in job search intervention programs [40]. Contrary, there are many other claims of that show no inherent benefits associated with job search programs [9].

In [41], the author argue on the convergence of artificial intelligence (AI) and its perceived economic reasoning. One of the primary goals of AI is on the development of computational methods for natural language understanding [42]. Recently there has been rise on works in natural language processing (NLP). In NLP, it is important to look at a target's language grammatical structure; thereafter, through syntax we are then able to apply the rules of a certain language's grammar [43]. In [17], the author notes of how artefact generation has advanced in many notable different fields of work, some of which include the synthesis of things and procedural content generation.

3 Related Work and Gap

In this section, we discuss related works with regards to synthesis of things, procedural generation and similar tools which exists for CV synthesis. We compare these tool by discussing what they offer and what they lack. We also note the gap and motivation of this study.

3.1 Synthesis of Things

Several papers have been published on the synthesis of things in different environments. Some related work in this domain are as follows:

Program Synthesis using Natural Language In [44], the author presented a framework for taking natural language inputs and producing a domain specific language as a result.

Synthesis of Integration Problems and Solutions: In [17], the author presented grammar rules, along with a tool, for the synthesis of integration problems along with the solutions.

Generation of Practice Programs in Python Using CFGs, in [22] the author presented automatic generation of procedural python programs.

SQLizer: Query Synthesis from Natural Language In [23], the author presented an end-to-end system whereby from natural language, SQL queries can be synthesised.

Interactive Synthesis using Free-Form Queries In [24], the author present a tool for code assistance where the system uses free from queries as input and outputs Java code adhering to the syntactic rules of the language.

Synthesis of Social Media Profiles In [45], the author presented a Probabilistic CFG which was then implemented in a synthesiser tool for social media profiles.

3.2 Procedural Content Generation (PCG)

Widely used in game development, procedural generation makes use of human generated assets which are used to algorithmically produce computer content. In [46], the author noted that PCG is not random but based on a systematic approach to content generation, and that simple PCG may not be adaptive unless an adaptive approach is taken. Game content usually generated includes game structures, maps, levels, characters designs, and rules [47]. Some of the work related to PCG includes the following:

The Level Generation Competitions: competitors submit level generators based on the Mario Bros game series [46, 48].

No Man's Sky: a game where players are able to explore 18 quintillion unique planets and moons [49].

SpeedTree: a middleware used for vegetation generation in game development [50].

Borderlands: a game where the weapon system generation was done algorithmically [51].

Based on the input obtained through web forms, we can then use PCG, with CV templates as assets, to layout and present in LaTeX the results of a synthesised CV.

3.3 Existing Tools for CV Synthesis

There exist many tools for the synthesis of a CV. However, these tools are limited in use leaving much to be desired for a better CV. Table 1 lists some examples of these tools, gives a description on the uses these tools, and compares with the tool (Flex-CV) we have developed.

3.4 Gap and Motivation

3.4.1 Gap

Prior to this work, there has not previously been a CFG for the synthesis of a CV. This work, in doing so, is a first to help with CV presentation problems with the use of CFGs. Current tools which exist make use of template based CV generation, whereas Flex-CV is dynamic in design, based on its use of grammar rules. As a result, Flex-CV can dynamically generate many CV instances compared to the static templates used by existing tools.

3.4.2 Motivation

Many tools exist which can help with constructing a CV, however, there are limitations which affect the rules of what makes a good CV. Given our aim is to produce a tool that will help with the job search process. We consider CV normalisation and standardisation towards a comprehensive tool for CV generation. This allows us to address common errors, focusing on capitalisation mistakes. We can then address one of the main factors affecting job application success.

3.5 Why Use Context-free Grammars?

CFGs are used to specify context-free languages. We have used CFGs for the synthesis of a CV based on the following reasons:

Production of Pseudorandom CV Sections One reason why we have chosen to use a CFG for the synthesis of CVs is the ability to write rules that can be compiled into modules that generate pseudorandom (or shuffled) permutations of CV sections. This is shown in Section 4, where rules allow the generation of CV sections that have different orderings, i.e. `Education` may come before `Work Experience` in one derivation, or it may be ordered differently in another derivation sequence.

Not a "Machine Learning" Task Another reason why CFG is appropriate for this task is that statistical approaches have not performed "spectacularly enough" to *fool humans*, in the synthesis of natural language texts [59, 60]. This is because Machine Learning models perform well with estimated guesses and not exactness — whereas, in the composition of a text like CV, results have to be 100% exact. Hence, CFG becomes appropriate for this task, as it can be proven that correct rules, will always generate correct strings (or texts).

Sharing Grammar Rules Using CFGs allows others to easily expand on the grammar rules and to synthesis other patterns of CVs (or other forms of text) based on the re-use of substructures within the CFG rules. This lays a scientific foundation to assist those who undertake such research in the future.

Applications of FLAT Formal Languages and Automata Theory (FLAT) is a theoretical aspect of Computer Science; mostly not attributed to many real-life applications (few exceptions are in the design of compilers, and language recognisers). Using a CFG, an aspect of FLAT, demonstrates that FLAT can find applications in topics such as PCG and social sciences application.

4 Design: CFG For CV Sythensis

In this section, we present a newly designed CFG for the synthesis of a CV. A CFG consists of grammar rules, whereby, the rules are a finite set. We define a four tuple CFG, $G = (N, \Sigma, P, S_{\text{start}})$ where N consists of non-terminal symbols as a placeholders, set Σ of terminal symbols, set P of production rules, and S_{start} as our start symbol. In this section, we describe a CFG for the synthesis of a CVs.

4.1 Grammar Rules

Here we define a rule for the start symbol of this grammar, denoted as S_{start}. In Rule 1, we define S_{start} as consisting of compulsory and optional sections that appear in a CV.

$$S_{\text{start}}\texttt{<Comp\_Sec>}^{+}\texttt{<Opt\_Sec>}^{*} \tag{1}$$

In Rule 2, we proceed to define the compulsory sections as consisting of biographical information, work experience, and education/qualifications obtained. We ensure that the compulsory part of the synthesised CV always starts with the biographical information and contains either an educational section, experience section, or both sections ordered differently.

$$\begin{aligned}\texttt{<Comp\_Sec>} \longrightarrow\ & \texttt{<Bio\_Sec>(<Edu\_Sec>} \mid \texttt{<Exp\_Sec>} \mid \\ & \texttt{<Edu\_Sec><Exp\_Sec>} \mid \texttt{<Exp\_Sec><Edu\_Sec>} \mid \\ & \lambda) \end{aligned} \tag{2}$$

In Rule 3, we proceed to define the optional sections as consisting of either relevant skills, activities, awards, publications, or projects. We ensure that any optional section in the CV appears only once and any in order possible .

$$\begin{aligned}\texttt{<Opt\_Sec>} \longrightarrow\ & x \subseteq \{\texttt{<Skills\_Sec>}, \texttt{<Projs\_Sec>}, \\ & \texttt{<Publ\_Sec>}, \texttt{<Awrds\_Sec>}, \texttt{<Acts\_Sec>}\} \\ & \ni : \mid x \mid \neq 0 \end{aligned} \tag{3}$$

Table 1: Existing Tools for CV Synthesis

Tool	Description	Limitations	Comparison with Flex-CV
Microsoft Word's CV Assistant.	Powered by LinkedIn, this tool assists individuals to create a CV. The main use of this tool is showing professional CVs of notable individuals in specific industries or according to job specifications as examples [52].	Microsoft Word's CV Assistant helps by showing attributes based on related roles, and/or industry. However, no emphasis is put on an individuals own content.	Emphasises on an individuals content, giving a more comprehensive CV as a results.
LinkedIn Resume Builder	Leverages the details (experience, education, skills, etc.) a user has on the LinkedIn social platform and builds a CV based on that available information [53].	The rules of a good CV and a good LinkedIn profile are distinct and there exist a problem if a profile is incorrectly setup up as a CV based on a user profile.	Focuses on the rules of what makes a good CV rather than what makes a good LinkedIn profile.
Vmock	A resume reviewing platform that allows students to upload resume and then receive feed. This platform works by bench-marking and scoring the uploaded towards a specific targeted audience [54].	This tool is great at formatting content, however, limited at analysing content.	Focuses not only on analysing content through data normalisation and standardisation.
Ceev	A google chrome extension that transforms a LinkedIn profile page into a CV ready for print. In addition, this tool allows for the customisation of a template [55].	This tool relies on the use of LinkedIn and the same problems with LinkedIn résumé builder apply to this tool. A good LinkedIn profile does not necessary result in a good CV.	Focuses on the rules of what makes a good CV rather than what makes a good LinkedIn profile.
Resume.io	An online résumé builder that enable users to easily create a résumé. This online tool includes automatic spell-checker, summary generator, and a cover letter builder [56].	This tool is not freely available as pricing options are in place. This is not ideal for individuals looking for employment with no financial means to support themselves.	Does not limit usability in hopes for monetary gain but seeks to helps individuals better their livelihoods.
Resumonk	An online tool that by default, you can add contact information, summary, experience, education and skills. This tool also allows for the importing of a LinkedIn generated CV [57].	This tool is not freely available as pricing options are in place. This is not ideal for individuals looking for employment with no financial means to support themselves.	Does not limit usability in hopes for monetary gain but seeks to helps individuals better their livelihood.
Zety CV Builder	Allows individuals to re-create their CV in a better looking way. It allows a user to pick a template, fill in relevant details and customise the document for a unique look [58].	This tool is not freely available as pricing options are in place. This is not ideal for individuals looking for employment with no financial means to support themselves.	Does not limit usability in hopes for monetary gain but seeks to helps individuals better their livelihood.

In Rule 4, we define personal information as having an a name, and the other details (i.e. arrangements of id, email, contact number, and gender) related to the individual. In Rule 5, ${}^4P_4$ is a k-permutation of x without repetitions. We ensure those arrangements are dynamic as the order may differ based on a chosen template or formatting.

$$\texttt{<Bio\_Sec>} \longrightarrow (\texttt{<Title>} \mid \lambda)\texttt{<Fst\_Nm>}(\texttt{<Mid\_Nm>} \mid \lambda)\ \texttt{<Lst\_Nm>}(\texttt{<Other\_Dtls>} \mid \lambda) \tag{4}$$

$$\texttt{<Other\_Dtls>} \longrightarrow x \in \{\texttt{<Id>}, \texttt{<Email>}, \texttt{<Phone>}, \texttt{<Gender>}\}^4 \ni: x \in {}^4P_4 \tag{5}$$

In Rule 6, we define the education section as having one or more instances of qualifications obtained. Each entry consisting of the following: qualification, date obtained, institution, location, and the description. Here ${}^4P_4$ is a k-permutation of x (i.e. arrangements of qual, date, inst, and place without repetitions). We ensure those arrangements, with the excepting of the description, are dynamic as

the order may differ based on a chosen template or formatting.

$$\begin{aligned} \texttt{<Edu\_Sec>} &\longrightarrow \texttt{<Edu\_Ent>}^{+} \\ \texttt{<Edu\_Ent>} &\longrightarrow \texttt{<Edu\_Field><Desc>} \\ \texttt{<Edu\_Field>} &\longrightarrow x \in \{\texttt{<Date>}, \texttt{<Inst>}, \texttt{<Place>}, \texttt{<Qual>}\}^{4} \\ &\ni: x \in {}^{4}P_{4} \end{aligned} \tag{6}$$

In Rule 7, we define the experience section as having one or more instances of relevant experience. Each entry consisting of the following: the date, organisation, location, position, and description. Here ${}^{4}P_{4}$ is a k-permutation of x (i.e. arrangements of date, org, place, and pos without repetitions). We ensure those arrangements, with the excepting of the description, are dynamic as the order may differ based on a chosen template or formatting.

$$\begin{aligned} \texttt{<Exp\_Sec>} &\longrightarrow \texttt{<Exp\_Ent>}^{+} \\ \texttt{<Exp\_Ent>} &\longrightarrow \texttt{<Exp\_Field><Desc>} \\ \texttt{<Exp\_Field>} &\longrightarrow x \in \{\texttt{<Date>}, \texttt{<Org>}, \texttt{<Place>}, \texttt{<Pos>}\}^{4} \\ &\ni: x \in {}^{4}P_{4} \end{aligned} \tag{7}$$

In Rule 8, we define the skills sections as having transferable skills (i.e. soft skills) and technical skills (i.e. hard skills). We ensure that either hard skills, soft skills, or both or none can appear in this section.

$$\begin{aligned} \texttt{<Skills\_Sec>} &\longrightarrow \texttt{<Hard\_Skills>} \mid \texttt{<Soft\_Skills>} \mid \\ &\longrightarrow \texttt{<Hard\_Skills><Soft\_Skills>} \mid \\ &\longrightarrow \texttt{<Soft\_Skills><Hard\_Skills>} \end{aligned} \tag{8}$$

In Rule 9, we define the projects section as having one or more instances of relevant projects. Each entry consisting of the following: the date, title, organisation, and description of the project. Here ${}^{3}P_{3}$ is a k-permutation of x (i.e. arrangements of date, title, and org without repetitions). We ensure those arrangements, with the excepting of the description, are dynamic as the order may differ based on a chosen template or formatting.

$$\begin{aligned} \texttt{<Projs\_Sec>} &\longrightarrow \texttt{<Projs\_Ent>}^{+} \\ \texttt{<Projs\_Ent>} &\longrightarrow \texttt{<Projs\_Field><Desc>} \\ \texttt{<Projs\_Field>} &\longrightarrow x \in \{\texttt{<Title>}, \texttt{<Date>}, \texttt{<Org>}\}^{3} \\ &\ni: x \in {}^{3}P_{3} \end{aligned} \tag{9}$$

In Rule 10, we define the publications section as having one or more instances of published works. Each entry consisting of the following: the year, title, publisher, and journal or conference name. Here ${}^{3}P_{3}$ is a k-permutation of x (i.e. arrangements of year, title, publisher, and journal/conference without repetitions). We ensure those arrangements are dynamic as the order may differ based on a chosen template or formatting.

$$\begin{aligned} \texttt{<Publ\_Sec>} &\longrightarrow \texttt{<Publ\_Ent>}^{+} \\ \texttt{<Publ\_Ent>} &\longrightarrow x \in \{\texttt{<Title>}, \texttt{<Year>}, \texttt{<Publisher>}, \\ &\quad \texttt{<Journ\_Conf>}\}^{4} \ni: x \in {}^{4}P_{4} \end{aligned} \tag{10}$$

In Rule 11, we define the awards section as one or more instances of awards received. Each entry consisting of the following: the year, and a description of the award.

$$\texttt{<Awrds\_Sec>} \longrightarrow (\texttt{<Year><Desc>})^{+} \tag{11}$$

In Rule 12, we define the activities section as one or more instances of involved activities. Each entry consisting of the following: the year, and a description of the activity.

$$\texttt{<Acts\_Sec>} \longrightarrow (\texttt{<Year><Desc>})^{+} \tag{12}$$

4.2 CV Normalisation and Standardisation

In this section we present algorithms for CV normalisation and standardisation. Firstly, we present new definitions of standardisation and normalisation of CVs in our context. This is given in Definitions 1 and 2.

Definition 1 (CV Normalisation) *We define the normalisation of a CV as a way to ensure no capitalisation errors and mistakes are present within the CV.*

Definition 2 (CV Standardisation) *We define the standardisation of a CV as the use of CFG rules to enforce the automatic organisation of CV sections and flow of CV contents.*

Algorithm 1: Normalisation

```
Function NormaliseText(text):
    isNewSentence ← true;
    new_text ← null;
    foreach chr ∈ text do
        new_char ← null;
        if isNewSentence = true then
            new_char ← CapitaliseChar(char);
            isNewSentence ← false;
        else if isEndingPunctation(chr) then
            new_char ← char;
            isNewSentence ← true;
        new_text ← new_text + new_chr;
    end
    return new_text;
End Function
```

Algorithm 1 shows that once submitted, a CV synthesiser then receives the data which is then normalised to ensure no capitalisation mistakes are found.

Algorithm 2: Standardisation

```
Function Standardisation(cv_data, section_order):
    template ← null;
    foreach section ∈ section_order do
        section_data ← CFGRules(cv_data, section);
        template ←
          template + TemplateCode(section_data);
    end
    return template;
End Function
```

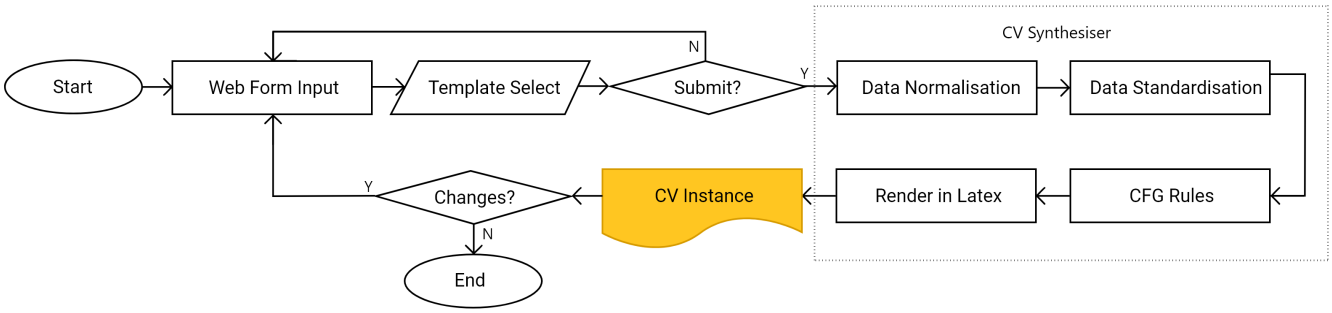


Figure 2: Flow diagram of Flex-CV Web Implementation

The synthesiser proceeds to standardise the data to a common format. Indicated in Algorithm 2, the synthesiser applies CFG rules on a section by section basis and prepares it be rendered as a PDF in LaTeX based on the selected template.

5 Implementation and Results

The rules defined in Section 4.1 can be used to dynamically synthesis an instance of a CV, whereby an individual's information is provided. In this section, we present implementation details of the designed grammar rules and the results obtained.

5.1 Implementation Details

The CFG presented in Section 4.1 was implemented in a web application (Flex-CV) using the ASP.NET Framework. Web forms were used to collect user data regarding content of the CV. In order to carry out this study, the CV sections used were those outlined in our grammar rules. Figure 2 illustrates the implementation of Flex-CV as a flow diagram.

As opposed to mobile applications, web applications provide certain benefits to our research. Essentially, mobile applications live and run on a device itself while web applications can seamlessly be assessed through any web browser. Web applications need not be downloaded and can be used on any modern smart phone. Hence, web forms were used to collect user data regarding content of the CV, i.e. biographical information, education, experience, skills, projects, publications, activities, and awards. The front-end (user interaction side) of the application can be viewed at: `https://cv-synthesis.web.app`. This then communicates with a back-end (server side) which is responsible for receiving data, implementation of our algorithms on the data, and sending results back to the front-end.

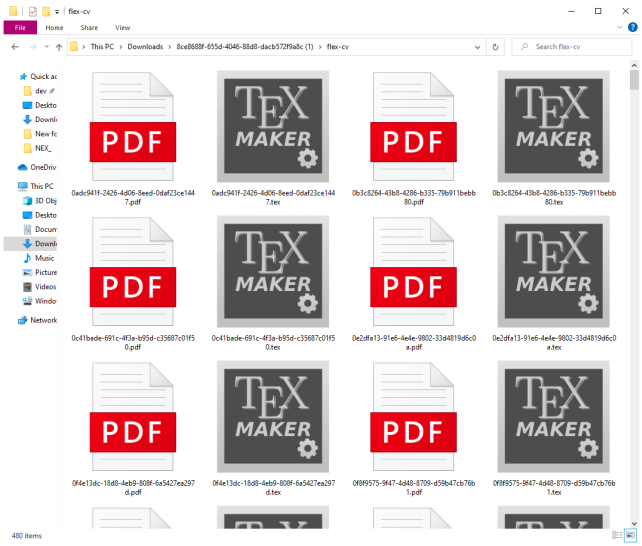


Figure 3: Output CVs from Flex-CV

5.2 Results

Upon submission, the user may generate a single CV or permutation of possible instances (*see* Figure 3). As for the option to generate a single CV instance, a PDF is rendered in LaTeX and returned along with an editable TEX file for future reference. As for the option to generate multiple instances, all possible instances with regards to the section ordering are returned to the user. Rendering these CV documents in LaTeX then provides high quality and consistent output using the TEX typesetting. As for using and providing an editable TEX file a user can then seamlessly use that to make changes, or use Flex-CV again for the benefits it offers.

5.3 More Results

Using the designed grammar rules, more results of synthesised CVs and the TEX files can be found at: tinyurl.com/cv-instances.

6 Evaluation of Flex-CV

In this section, we present the results of the survey that evaluates the impact of Flex-CV. This survey was conducted with the aim of understanding Flex-CV and its usefulness. This survey was conducted

online and respondents were individuals looking to improve their CVs. The survey was distributed online in a manner that allowed random respondents to complete the survey.

Results in Figure 4 indicate that 55.6% of respondents often struggle with writing a good CV, while none indicated as having not struggled. As seen in Figure 5, only 5.6% indicated as never having discovered grammatical errors in their CV. These results are quite problematic, however with Flex-CV we are able to help with CV presentation as means towards potential employment for individual. With 83.3% having tried Flex-CV, majority agree (Figure 6) with Flex-CV potentially providing candidates with lesser or no errors. Job search intervention can be setup to assist in job search programs, this can be seen in Figure 7 as 61.1% had never used a CV/résumé assistant tool before, however as indicated in Figure 8, an overwhelming majority later then stated as likely to use a tool such a tool in the future.

Correctly highlighting knowledge, skills, and attributes is a step closer to obtaining employment. Considering whether individuals struggle with writing a rather good CV, we indicate this as a problematic issue given how a good CV/resume is a gateway to employment and benefits that come with it [33]. Considering whether respondents had ever discovered grammatical errors in their CVs, we note that a way to lessen error occurrences is needed as they are found. Considering all the results, we therefore state that job search intervention can assist in the job search process. As such, Rayman and Atanasoff [40] note that the likelihood of obtaining employment is three times higher when job seekers are participants in job search intervention programs. This further indicates how tools, such as Flex-CV, are essential as job search intervention assistants.

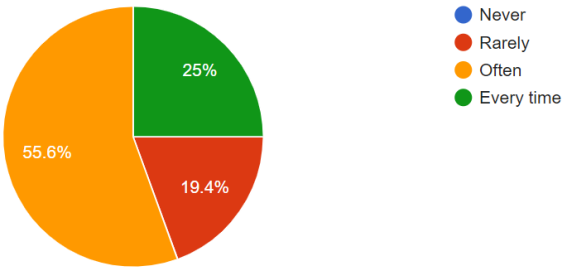


Figure 4: Struggle With Writing A Good CV

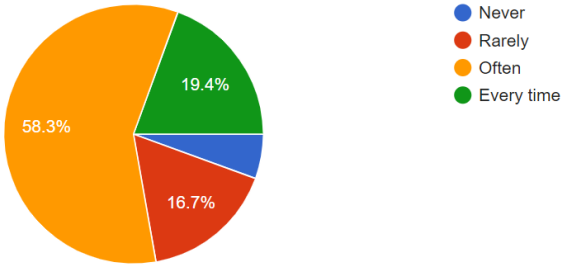


Figure 5: Discovered Grammatical Errors In CV

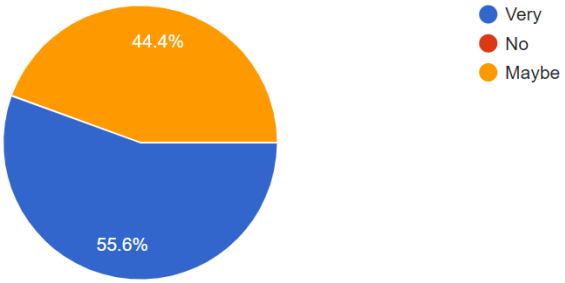


Figure 6: Flex-CV Can Give Candidates Good CVs With Lesser Or No Errors

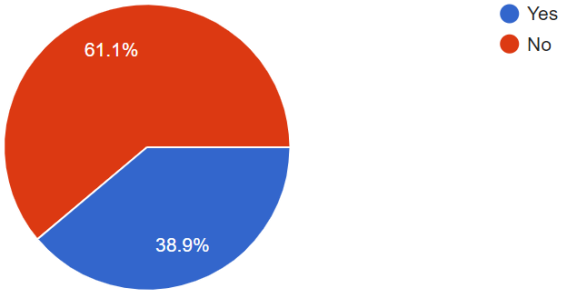


Figure 7: Used A CV Assistant Tool

7 Conclusion and Future Work

7.1 Conclusion

In this paper, we presented a CFG for the dynamic synthesis of a CV. The use of formal grammars in synthesising things has recently had many works done and in this paper, grammar rules were implemented in a software tool that produces CV instances to help with CV presentation problems. We presented a survey to many individuals who found the tool to be useful as a means towards better employment prospects through better CV presentation. The impact of job search interventions, Flex-CV included, are beneficial as a way to help with employment in many communities.

7.2 Future Work

From here, we will expand the grammar rules to be more complex and include the following: more sections, usually those not common in many CVs; and more templates.

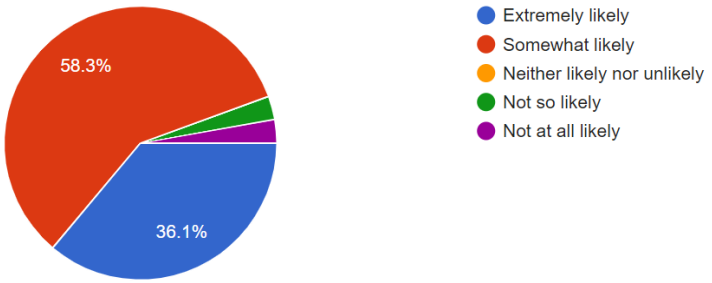


Figure 8: Likely To Use Such A Tool In The Future

References

[1] E. Forsberg, "Curriculum Vitae – The Course of Life," Nordic Journal of Studies in Educational Policy, **2016**(2-3), 33742, 2016, doi:10.3402/nstep.v2.33742.

[2] D. R. Dickson, K. Nusair, "An HR perspective: the global hunt for talent in the digital age," Worldwide Hospitality and Tourism Themes, **2**(1), 86–93, 2010, doi:10.1108/17554211011012612.

[3] C. Cañibano, B. Bozeman, "Curriculum vitae method in science policy and research evaluation: the state-of-the-art," Research Evaluation, **18**(2), 86–94, 2009, doi:10.3152/095820209X441754.

[4] R. Woolley, T. Turpin, "CV analysis as a complementary methodological approach: investigating the mobility of Australian scientists," Research Evaluation, **18**(2), 143–151, 2009, doi:10.3152/095820209X441808.

[5] M. G. D'Onofrio, "The public CV database of Argentine researchers and the 'CV-minimum' Latin-American model of standardization of CV information for R&D evaluation and policy-making," Research Evaluation, **18**(2), 95–103, 2009, doi:10.3152/095820209X441763.

[6] M. S. Kramer, "Medical Research: A Prescriptive View," Pediatrics, **95**(1), 82–84, 1995.

[7] U. Sandström, "Combining curriculum vitae and bibliometric analysis: mobility, gender and research performance," Research Evaluation, **18**(2), 135–142, 2009, doi:10.3152/095820209X441790.

[8] R. Popken, "The Pedagogical Dissemination of a Genre: The Resume in American Business Discourse Textbooks, 1914-1939," JAC, **19**(1), 91–116, 1999.

[9] S. Liu, J. L. Huang, M. Wang, "Effectiveness of job search interventions: A meta-analytic review." Psychological bulletin, **140**(4), 1009, 2014.

[10] C. A. Higgins, T. A. Judge, "The effect of applicant influence tactics on recruiter perceptions of fit and hiring recommendations: a field study." Journal of Applied Psychology, **89**(4), 622, 2004.

[11] N. Azrin, T. Flores, S. Kaplan, "Job-finding club: A group-assisted program for obtaining employment," Behaviour Research and Therapy, **13**(1), 17 – 27, 1975, doi:https://doi.org/10.1016/0005-7967(75)90048-0.

[12] S. Klasen, F. Lamanna, "The impact of gender inequality in education and employment on economic growth: new evidence for a panel of countries," Feminist economics, **15**(3), 91–132, 2009.

[13] F. McKee-Ryan, Z. Song, C. R. Wanberg, A. J. Kinicki, "Psychological and physical well-being during unemployment: a meta-analytic study." Journal of applied psychology, **90**(1), 53, 2005.

[14] M. Modini, S. Joyce, A. Mykletun, H. Christensen, R. A. Bryant, P. B. Mitchell, S. B. Harvey, "The mental health benefits of employment: Results of a systematic meta-review," Australasian Psychiatry, **24**(4), 331–336, 2016.

[15] J. Stauder, "Unemployment, unemployment duration, and health: selection or causation?" The European Journal of Health Economics, **20**(1), 59–73, 2019.

[16] S. Liu, M. Wang, H. Liao, J. Shi, "Self-regulation during job search: The opposing effects of employment self-efficacy and job search behavior self-efficacy." Journal of Applied Psychology, **99**(6), 1159, 2014, doi:10.1037/a0036692.

[17] A. Ade-Ibijola, "Synthesis of Integration Problems and Solutions," in 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), 185–190, IEEE, 2019, doi:10.1109/ISCMI47871.2019.9004291.

[18] D. Pawade, A. Sakhapara, M. Jain, N. Jain, K. Gada, "Story Scrambler–Automatic Text Generation Using Word Level RNN-LSTM," International Journal of Information Technology and Computer Science (IJITCS), **10**(6), 44–53, 2018.

[19] J. Hirschberg, C. D. Manning, "Advances in natural language processing," Science, **349**(6245), 261–266, 2015.

[20] J. Jeuring, "Applications of Grammars," Citeseer, 2006.

[21] A. Ade-Ibijola, "Synthesis of regular expression problems and solutions," International Journal of Computers and Applications, **42**(8), 748–764, 2020, doi:10.1080/1206212X.2018.1482398.

[22] A. Ade-Ibijola, "Syntactic Generation of Practice Novice Programs in Python," in S. Kabanda, H. Suleman, S. Gruner, editors, ICT Education, 158–172, Springer International Publishing, Cham, 2019.

[23] N. Yaghmazadeh, Y. Wang, I. Dillig, T. Dillig, "SQLizer: query synthesis from natural language," Proceedings of the ACM on Programming Languages, **1**(OOPSLA), 1–26, 2017.

[24] T. Gvero, V. Kuncak, "Interactive synthesis using free-form queries," in 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, volume 2, 689–692, IEEE, 2015.

[25] S. P. Golen, L. P. Grasso, C. L. Moeckel, "Barriers to communication during interviews for accounting jobs," Journal of Education for Business, **70**(5), 272–277, 1995.

[26] S. B. Knouse, "Impressions of the resume: The effects of applicant education, experience, and impression management," Journal of Business and Psychology, **9**(1), 33–45, 1994.

[27] E. Flouri, A. Buchanan, "The role of work-related skills and career role models in adolescent career maturity," The Career Development Quarterly, **51**(1), 36–43, 2002.

[28] C. L. Hulin, T. A. Judge, "Job attitudes," Handbook of psychology, 255–276, 2003.

[29] S. B. Morris, R. L. Daisley, M. Wheeler, P. Boyer, "A meta-analysis of the relationship between individual assessments and job performance." Journal of Applied Psychology, **100**(1), 5, 2015.

[30] V. K. Borooah, A. Dubey, S. Iyer, "The effectiveness of jobs reservation: caste, religion and economic status in India," Development and change, **38**(3), 423–445, 2007.

[31] C. Gorter, J. v. Ommeren, "Sequencing, timing and filling rates of recruitment channels," Applied Economics, **31**(10), 1149–1160, 1999.

[32] M. Jahoda, Employment and unemployment: A social-psychological analysis, volume 1, CUP Archive, 1982.

[33] D. Bartrum, P. A. Creed, et al., "Explanations for deteriorating wellbeing in unemployed people: Specific unemployment theories and beyond," Unemployment and health: International and interdisciplinary perspectives, 1, 2006.

[34] C. R. Wanberg, "The individual experience of unemployment," Annual review of psychology, **63**, 369–396, 2012.

[35] K. I. Paul, K. Moser, "Unemployment impairs mental health: Meta-analyses," Journal of Vocational behavior, **74**(3), 264–282, 2009.

[36] M. L. Rothschild, W. C. Gaidis, "Behavioral learning theory: Its relevance to marketing and promotions," Journal of marketing, **45**(2), 70–78, 1981.

[37] I. Ajzen, "The theory of planned behaviour: Reactions and reflections," Psychology & Health, **26**(9), 1113–1127, 2011, doi:10.1080/08870446.2011.613995, pMID: 21929476.

[38] R. S. Lazarus, "Coping theory and research: Past, present, and future," Fifty years of the research and theory of RS Lazarus: An analysis of historical and perennial issues, 366–388, 1993.

[39] J. Li, M. Zhu, W. Lu, G. Zhou, "Improving semantic parsing with enriched synchronous context-free grammar," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1455–1465, 2015.

[40] J. Rayman, L. Atanasoff, "Holland's Theory and Career Intervention: The Power of the Hexagon," Journal of Vocational Behavior, **55**(1), 114 – 126, 1999, doi:https://doi.org/10.1006/jvbe.1999.1701.

[41] D. C. Parkes, M. P. Wellman, "Economic reasoning and artificial intelligence," Science, **349**(6245), 267–272, 2015.

[42] E. Brill, R. J. Mooney, "An overview of empirical natural language processing," AI magazine, **18**(4), 13–13, 1997.

[43] A. Reshamwala, D. Mishra, P. Pawar, "Review on natural language processing," IRACST Engineering Science and Technology: An International Journal (ESTIJ), **3**(1), 113–116, 2013.

[44] A. Desai, S. Gulwani, V. Hingorani, N. Jain, A. Karkare, M. Marron, S. Roy, "Program synthesis using natural language," in Proceedings of the 38th International Conference on Software Engineering, 345–356, 2016.

[45] A. Ade-Ibijola, "Synthesis of social media profiles using a probabilistic context-free grammar," in 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), 104–109, IEEE, 2017, doi: 10.1109/RoboMech.2017.8261131.

[46] J. Togelius, E. Kastbjerg, D. Schedl, G. N. Yannakakis, "What is Procedural Content Generation? Mario on the Borderline," in Proceedings of the 2nd International Workshop on Procedural Content Generation in Games, PCGames '11, Association for Computing Machinery, New York, NY, USA, 2011, doi: 10.1145/2000919.2000922.

[47] J. Togelius, A. J. Champandard, P. L. Lanzi, M. Mateas, A. Paiva, M. Preuss, K. O. Stanley, "Procedural Content Generation: Goals, Challenges and Actionable Steps," in S. M. Lucas, M. Mateas, M. Preuss, P. Spronck, J. Togelius, editors, Artificial and Computational Intelligence in Games, volume 6 of *Dagstuhl Follow-Ups*, 61–75, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2013, doi:10.4230/DFU.Vol6.12191.61.

[48] N. Shaker, J. Togelius, G. N. Yannakakis, B. Weber, T. Shimizu, T. Hashiyama, N. Sorenson, P. Pasquier, P. Mawhorter, G. Takahashi, et al., "The 2010 Mario AI championship: Level generation track," IEEE Transactions on Computational Intelligence and AI in Games, **3**(4), 332–347, 2011, doi: 10.1109/TCIAIG.2011.2166267.

[49] R. E. Cardona-Rivera, "Cognitively-grounded procedural content generation," in Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[50] R. Van Der Linden, R. Lopes, R. Bidarra, "Procedural generation of dungeons," IEEE Transactions on Computational Intelligence and AI in Games, **6**(1), 78–89, 2013, doi:10.1109/TCIAIG.2013.2290371.

[51] J. Togelius, G. N. Yannakakis, K. O. Stanley, C. Browne, "Search-Based Procedural Content Generation," in C. Di Chio, S. Cagnoni, C. Cotta, M. Ebner, A. Ekárt, A. I. Esparcia-Alcazar, C.-K. Goh, J. J. Merelo, F. Neri, M. Preuß, J. Togelius, G. N. Yannakakis, editors, Applications of Evolutionary Computation, 141–150, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[52] Microsoft, "Use Resume Assistant and LinkedIn for great resumes," 2020, available at https://support.microsoft.com/en-us/office/use-resume-assistant-and-linkedin-for-great-resumes-444ff6f0-ef74-4a9c-9091-ffd7a9d1917a?ui=en-us.

[53] LinkedIn, "LinkedIn Resume Builder- Frequently Asked Questions — LinkedIn Help," 2020, available at https://www.linkedin.com/help/linkedin/answer/94034/linkedin-resume-builder-frequently-asked-questions?lang=en.

[54] Vmock, "Career Acceleration Platform," 2020, available at https://www.vmock.com/.

[55] M. Samuels, "design a resume, build a profile, find a job." 2020, available at https://ceev.io/.

[56] C. Y. J. W. R. Resume.io, "Create Your Job Winning Resume," 2020, https://resume.io/.

[57] Resumonk, "Resume Maker: Create a standout professional resume and CV," 2020, available at https://www.resumonk.com/.

[58] Zety, "Zety Online Resume Maker: Quick, Effective, Try for Free," 2020, available at https://zety.com/.

[59] H. Cho, "Investigation of error tolerant nature of machine learning algorithms," .

[60] I. W.-B. MIT IDE, "What Machine Learning Can and Cannot Do," 2018, available at https://medium.com/mit-initiative-on-the-digital-economy/what-machine-learning-can-and-cannot-do-6788a818776.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1075-1092 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Nature Inspired and Transform Based Image Encryption Techniques: A Comparative Study

Bhagyashri Pandurangi R[*,1], Chaitra Bhat[2], Meenakshi R. Patil[3]

[1]Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi, 590008, India

[2]Robert Bosch Engineering and Business Solutions, EEA, Bengaluru, 560 103, India

[3]Electronics and Communication Engineering, Jain AGM Institute of Technology, Jamakhandi, 587301, India

ARTICLE INFO

Article history:
Received: 01 September, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Bio inspired techniques
Chaotic maps
Fractional transforms
Radial Hilbert transforms
Transform domain encryption

ABSTRACT

In this paper, performances of two variations of chaos based algorithms are compared. First algorithm is a self-adaptive color image encryption algorithm is proposed based on Radial Hilbert Transform and chaos. This technique uses chaotic random phase masks operated on the transformed pixels to increase the randomness in confusion and diffusion operations. Also, a random jumbling process is used at the final stage to increase the randomness in the cipher image. Part of the plain image is used to generate the keys for encrypting another part. Second algorithm is inspired by the bio operations resembling confusion and diffusion. Use of a scrambler improves the performance of this algorithm. Proposed work elaborates the results of the suitability analysis conducted on various kinds of input images, namely, satellite images, face images and handwritten signature images. Performance parameters considered for the analysis include horizontal correlation, vertical correlation, diagonal correlation, and net changing pixel rate, unified average change in intensity, entropy and encryption time taken by the encryption techniques.

1. Introduction

With the increase in multidimensional applications of internet based multimedia content, secure transmission of multimedia data including images has become the need of the day. Biometric image based security is deployed in all the electronic gadgets. Electronic transactions are involved in day today life. Various multimedia applications include transmission variety of images though the computer networks. Image encryption algorithms vary in the way they are developed, the domain in which they encrypt and the inputs they accept. Since the algorithms need to satisfy a variety of constraints imposed by different applications, there is a need for adaptivity. Adaptive encryption algorithms meet the security standards without compromising the quality of the transmission. Self-adaptive encryption techniques improve the security of the encryption. Bringing adaptivity at various stages of encryption algorithm is a promising field of current era.

As image security is an elementary requirement in variety of applications ranging from IoT to aerospace communication, variety of constraints are imposed on the encryption methods.

Researchers have developed various algorithms to meet the specifications of the application intended. Based on implementation, image encryption techniques can be classified into cover a wide range of classes including:

- Nature Inspired Encryption
- Parallel Encryption
- Transform Domain Encryption
- Partial / Selective Encryption
- Self-Adaptive Encryption, etc

Many hybrid algorithms are developed recently with a proper blend of these classes, considering the relative advantages of various classes. For the comparative study proposed, two different algorithms are considered. First one is based on nature inspired encryption and the second is a blend of transform domain and self-adaptive encryption.

Majority of the nature inspired encryption algorithms are based on genetic algorithms (GA). Efficient image encryption algorithms exploit the basic techniques of the GAs which simulate processes in natural systems necessary for evolution.

2. Literature Survey

Genetic algorithms initialize a random population and define fitness of population (p). These algorithms randomly choose

*Corresponding Author: Bhagyashri Pandurangi R, brpandurangi@git.edu

parents from p. Crossover operation is carried out on these parents, resulting in the creation of new population (p+1). The new population is mutated and the fitness of population (p+1) is determined. The entire process is repeated until best individual is good enough. This concept can be effectively used in the accomplishment of confusion and diffusion of pixels required for encryption. Genetic Algorithms possess the following limitations:

- requires a greater number of iterations for optimum solution
- convergence is spread across many parameters [1].

To overcome these limitations, it is essential to find and to improve GA and to increase the speed of convergence. Various methods have been proposed to improvise the efficiency of GA [2-7]. In [8], a Chaos Genetic Algorithm (CGA) is proposed that employs logistic map to create the initial population. Limitation of this technique is that it still cannot generate diverse content in mutation for some complex circumstances. Work taken up in [9] is a feature selection method based on chaos GA using two different chaotic maps to keep up and improve the capability in global searching.

Various approaches to achieve confusion and diffusion based on genetic operations are presented in [10-14]. They utilize the fitness function for key generation and generation of cipher image. The pixels are permuted with DNA (Deoxyribo nucleic acid) based combinations to improve the encryption quality. A scheme for the security of medical images based on the features of genetic algorithms is discussed in [15]. Limitations of these algorithms are that few of the parameters are not tested for the cipher images. The algorithms are efficient but complex in nature. Work in [16] cryptanalyzes a DNA based color encryption technique. Also, there is a different approach using predicted key pair values to breach the security of these algorithms as demonstrated in [17].

Since the DNA based techniques do not change the pixel values non-linearly, they are prone to known plaintext attack. If one plain image is compromised, the enciphered image can be easily reconstructed without the help of keys. Optical encryption techniques target parallel processing of the image with multiple degree of freedom.

The image pixels are multiplied with random phase masks in the process of Double Random Phase Encryption (DRPE) in spatial and frequency domain. Optical encryption is performed with the keys generated in the transform domain. Work presented in [18-25] describe encryption techniques based on various transforms including fractional Fourier transform (FrFT), Radial Hilbert transform (RHT), Discrete Cosine Transform (DCT) and Hartley transform (HT). Optical encryption exploits several image parameters, including color, amplitude, phase, spatial frequency, polarization, etc., to arrive at a robust and secure encryption. Also, the complications can be regulated by proper selection of the number of rounds and the fractional order.

Fractional transforms play a vital role in image encryption. They exhibit many beneficial assets of general transforms, and have an additional independent component, its fraction. If this fraction equals to zero, the output image is a modulated version of the input image. When the fraction equals to one, fractional transform is equivalent to the original regular transform. With the fraction varying from zero to unity, different versions of the signals interpolating between modulated form and the regular transformed version can be generated. Thus, the free component serves as a key for encipherment.

The fractal compression technique is another promising method for image encryption. It is a lossy compression technique that suits for the images having high similarity at different parts. Natural and texture images are the best examples for this kind of images. A compression-encryption algorithm based on the fractal coding is discussed in [26]. Even though the compression ratio is high, the algorithm suffers at the quality of encryption. Balancing the compression ratio and the encryption quality is an interesting topic for current research.

Fractional Fourier transform has found effective in image encryption and noise removal. A multiple parameter FrFT based scheme is discussed in [27], in which, the order is considered as a vector. This scheme provides flexibility for the dimension of transform order and improvises the strength of encrypted image while retaining low computational complexity and hardware cost. Fractional Fourier transform with vector power multi-parameters is another general form of the FrFT which has great significance in information processing security. Performance of various transform based techniques is compared in [28]. Even though it claims that chaos based encryption with discrete wavelet transform gives better performance, the NPCR (number of changing pixel rate) and UACI values need to be improved. A properly designed scrambler can improve these parameters.

Remote sensing (RS) is a booming technology that provides the dynamic information of the earth surface remotely and without contact. Remote sensing technology has been developed with an exponential rate due the improved research on the computers and graphic and playing a vital role in the national, economic, construction and defense related issues [29].

Latest studies of primary focus are on the protection of digital contents communicated using RS technology. Satellite remote sensing utilizes image distribution and characteristics with various structures of remote sensing image [30]. Recently purposeful hack assaults have demonstrated that interruption into satellite information is not an unimaginable assignment. A group at the Embry Riddle Aeronautical University figured out how to get National Oceanic and Atmospheric Administration (NOAA) satellite symbolism with fundamental contraption worked as a component of an exploratory venture and by utilizing open sources accessible from the Internet [31].

Cryptanayzers have been successful in breaking few of these algorithms with known/chosen plaintext cryptanalysis. The reason is that the same set of encryption keys are used to encipher different input plain images. Unfortunately, all these cryptosystems use the same key for quite a long period. The key set can be easily compromised by encrypting some distinct images like a completely white image or completely black image, and then associating the result with the equivalent enciphered output. In order to make the algorithms robust against the kind of attacks, a plaintext-related diffusion and confusion needs to be incorporated. The key stream rudiments can be selected under the control of plain pixel. Secret keys must be associated not only to the control parameters but also to the parameters related to plain

image, in order to make the known/chosen plaintext attack harder to succeed.

Self-adaptive encryption is considered as a method that has the encryption considering input image details. The dependency on the plain image content is strongly desirable characteristics of a good encryption algorithm since two images differing with few pixels will get encrypted into totally dissimilar patterns. The proposed technique generates the encryption keys dependent on the plain image features. Also, the technique arrives at a second level of self adaptivity since one part of the image gets encrypted by the influence of the other part.

Optical systems deploy transform based algorithms, which include, Fourier transform (FT), the fractional Fourier transform (FrFT), radial Hilbert transform (RHT). To add to these, there are also discrete fractional Fourier transform (DfrFT), discrete fractional cosine transform (DFrCT), and discrete fractional Sine transform (DFrST). Haar wavelet transform de-correlates the image pixels by decomposing into averaging and differencing components [32].The Hilbert transform exhibits edge enhancing properties and hence finds its use in processing of images and phase observation [33]. Radial Hilbert transform can increases the strength with expanded key space [34].

Satellite images are widely used in various applications including remote sensing, weather forecasting and vegetation based analysis, mineral/rock differentiation, plant species identification, forest classification, coastal water analysis, soil/vegetation differentiation, etc. [35]. Satellite images can be classified as infrared images, water vapor and visible images. Some features of the satellite images are high correlation among the pixels, bulkiness and large size. Since most of them are confidential, the traditional encryption methods lack in efficiency for bulky images and have a poor speed. To encrypt bulky images, a secure and high speed encryption algorithm is necessary.

The paper compares the performance of chaos based bio-inspired image encryption [36] and chaos based self-adaptive encryption using radial Hilbert transform [37]. The test is performed with three sets of images, namely, satellite images, face images and signature images as input data. Self-adaptive technique divides the plain image into sub-images and adds chaos. The encrypted image after jumbling yields an unsystematic, disordered image [38]. In the second method, image encryption is accomplished using multi-point crossover and mutation serving as bio-operational tools for confusion and diffusion.

3. Description of proposed techniques

3.1. Chaos Based Self Adaptive Encryption Using Radial Hilbert Transform (Technique 1)

Proposed technique employs Radial Hilbert transform. Hilbert Transform in fractional domain has been explored by Lohmann et al. [39]. Hilbert Transform has a special feature that it selectively emphasizes the features of the given input image during the spatial filtering process. Generally Hilbert Transform applied to an image builds its edge-enhanced version. In support to this, the fractional HT modifies the nature of the edge enhancement. A Hilbert filter yields an edge-enhanced image in a single dimension. 2D edge enhancement for the image can be accomplished by merging the response of two Hilbert filters orthogonal to each other. Still, these transformed images preserve the basic symmetry in rows and columns. A uniformly edge-enhanced image is the result of its radial version of the Hilbert transform. Since the Hilbert transform is helpful in enhancing the edges of one the input image, it is commonly used in image processing and phase observation. Three dimensional radially symmetrical Hilbert transform is mathematically represented as

$$H(a, b, 3) = \exp(i * x * \theta(a, b, 3)) \quad (1)$$

where x is a user defined parameter.

Self-adaptive encryption employs the keys generated form one part of the image to encrypt the other part. Proposed algorithm has the following components.

3.1.1. Image Encryption

Flow diagram of algorithm is as shown in Fig.1.Plane images are initially re-sized to MxN and divided into 4 sub-images with size (M/2, N/2). It is divided into four sub-images (I_1, I_2, I_3 and I_4). First sub-image I_1 is encrypted using three input parameters: radial Hilbert transform, random chaos mask and a key x selected by the user. The Second sub-image I_2 is encrypted with a key calculated as the fraction of mean to standard deviation the previous sub-image I_1. This value also functions as an input to the radial Hilbert transform. Remaining sub-images are encrypted similarly. The encrypted sub-images are joined together and pixel positions are jumbled using a randomly generated matrix to get the final encrypted image. For decipherment, conjugate of the radial Hilbert transform and two random chaos masks are considered.

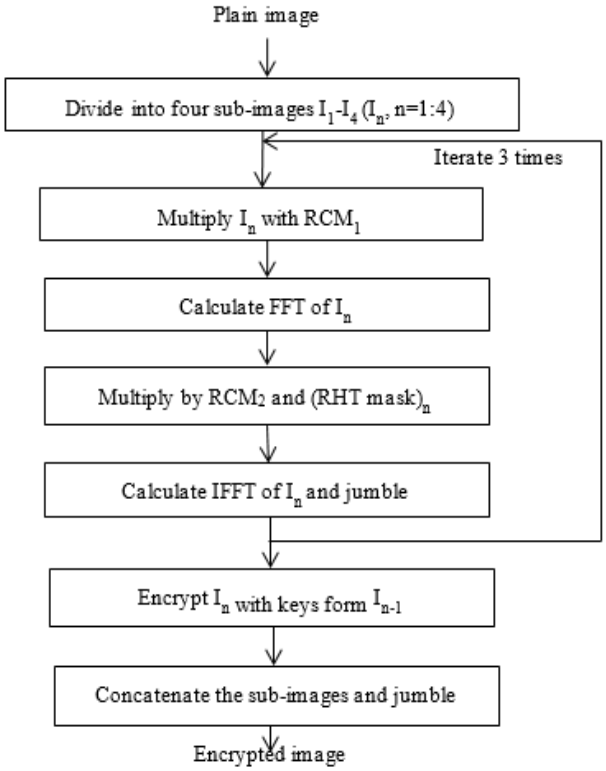


Figure 1: Radial Hilbert transform based image encryption process

3.1.2 Creation of Random Chaos Mask (RCM)

Two random three dimensional matrices having dimensions equal to that of the plain image 'i' are considered. These are considered as A_1 and A_2. Two random matrices are generated as

$$Ø_1(a, b, 3) = \exp(i * \theta_1(a, b, 3)) \quad (2)$$

$$Ø_2(a, b, 3) = \exp(i * \theta_2(a, b, 3)) \quad (3)$$

where $Ø(x, y, 3)$ is the random chaos mask generated, $\theta(x, y, 3)$ is the Fourier transform of randomly generated matrix. These values are utilized in the generation of chaotic images A1 and A2 using logistic map which serve as a random chaos masks. The process is depicted in Figure 2.

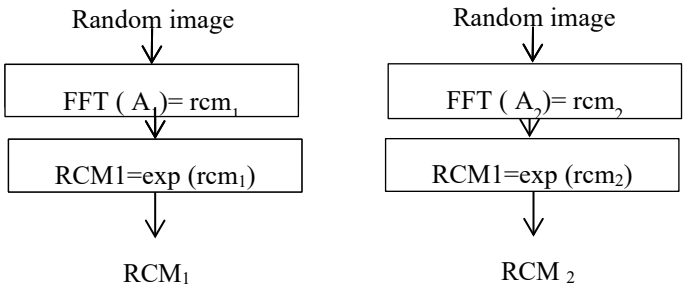


Figure 2: Generation of random phase mask (RCM)

3.1.3 Transformation with Radial Hilbert Transform

Each of sub-images undergoes a transformation as per equation (4), where x is a user defined value for sub- image I_1. For I_2-I_4, x is fraction of mean and standard deviation of previous sub-images. $\theta(a,b,3)$ is the imaginary part of the Fourier transform for the sub- image I_n and H(a,b,3) is the Hilbert transform for sub-image I_n. By introducing radially symmetrical Hilbert transform, we can use it in three dimensions. The equation for radial Hilbert transform is given by

$$H(m, n, 3) = \exp(i * x * \theta(m, n, 3)) \quad (4)$$

The conjugate of radial Hilbert transform parameter is expressed as

$$H_1(a, b, 3) = \exp(-i * x * \theta(a, b, 3)) \quad (5)$$

3.1.4 Results

The algorithm is applied on color images downloaded from Bhuvan, a Geo-portal of ISRO [40]. The images published on Bhuvan in 2D/3D domain are captured using many sensors and inform about the water, soil and land.

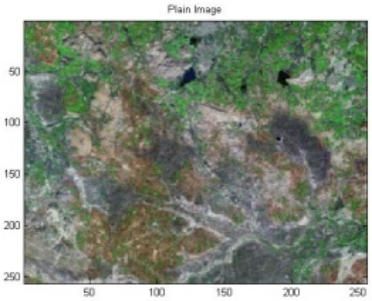


Figure 3: Input image for encryption

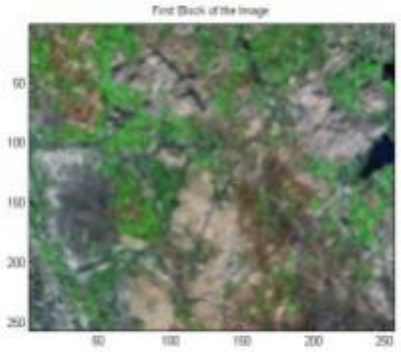


Figure 4: Sub-image I_1

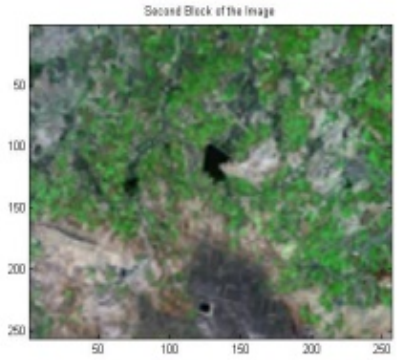


Figure 5: Sub-image I_2

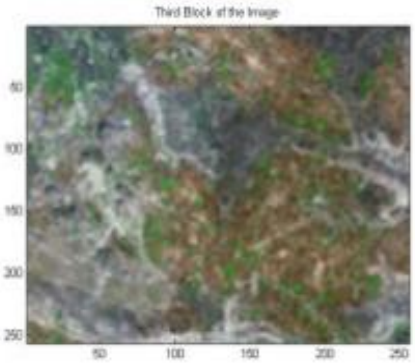


Figure 6: Sub-image I_3,

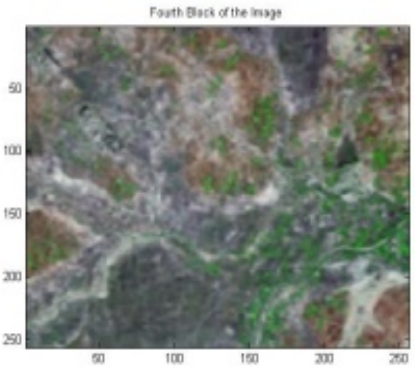


Figure 7: Sub-image I_4

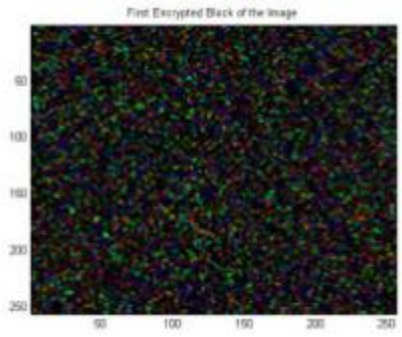


Figure 8: First encrypted block

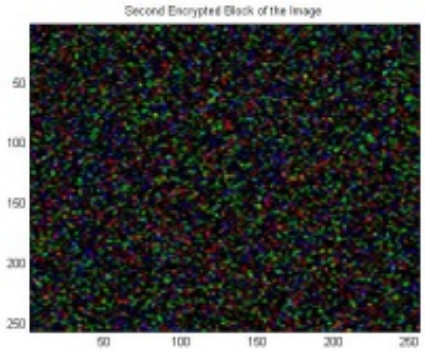


Figure 9: Second encrypted sub-block

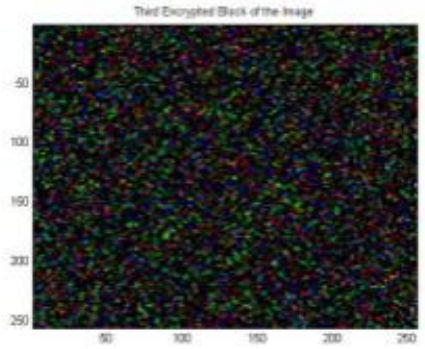


Figure 10: Third encrypted sub-block

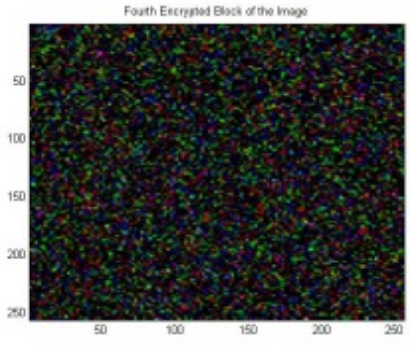


Figure 11: Fourth encrypted sub-block

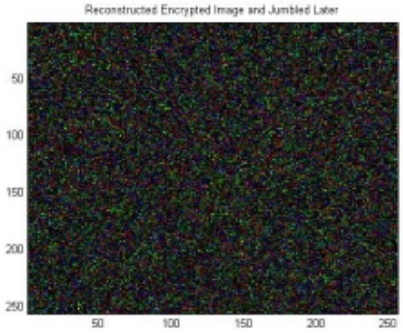


Figure 12: Complete encrypted image

Input satellite image shown in Figure 3 is divided into four smaller sub-images as displayed in Figure 4 -7. Each block is encrypted by the keys generated using the previous sub-image. The process is iterated three times for the entire image. Encrypted sub-images are displayed in Figure 8-11. Complete encrypted image is presented in Figure 12.

3.1.4.1 Histogram Analysis

Histogram analysis provides an overview of strength of confusion and diffusion for an image encryption algorithm, in presence of statistical attack. Histogram of plain image is obtained as shown in figure 13 (a) – (c). Histogram of encrypted image is shown by figure 14 (a) – (c). Histogram of the encrypted image gets exponentially distributed due to radial Hilbert transform.

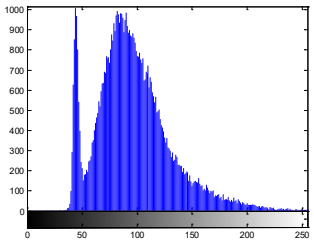


Figure 13(a): Histogram of plain image red channel

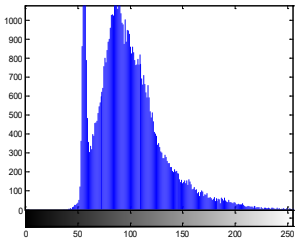


Figure 13(b): Histogram of plain image green channel

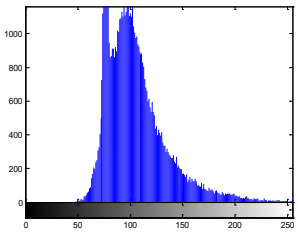


Figure 13(c): Histogram of plain image blue Channel

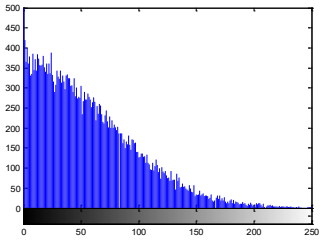


Figure 14(a): Histogram of encrypted image red channel

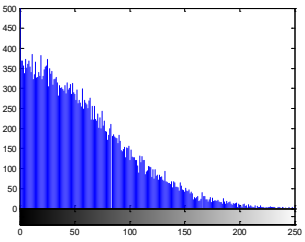


Figure 14 (b) Histogram of encrypted image green channel

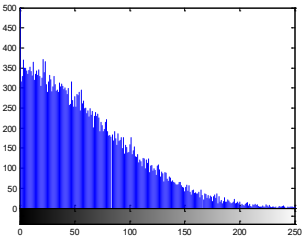


Figure 14(c): Histogram of encrypted image blue channel

3.1.4.2 Correlation of Two adjacent pixels

Correlation coefficients r(x,y) in horizontal, vertical and diagonal directions are calculated with the equation 6.

$$r(x,y) = \frac{COV(x,y)}{\sqrt{D(x)}\sqrt{D(y)}} \quad (6)$$

where x and y are values of two neighboring pixels in the image, D(x) equals to the variance of x and COV(x, y) indicates the covariance of x and y.

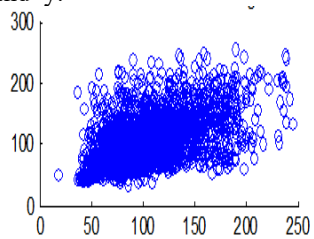


Figure 15(a): Horizontal pixel distribution of plain image for red plane

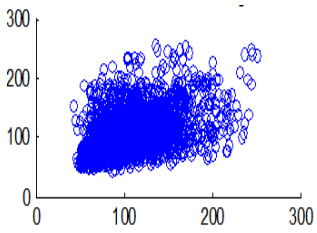


Figure 15(b): Horizontal pixel distribution of plain image for green plane

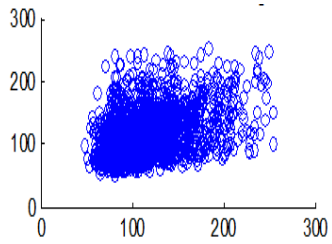


Figure 15(c): Horizontal pixel distribution of plain image for blue plane

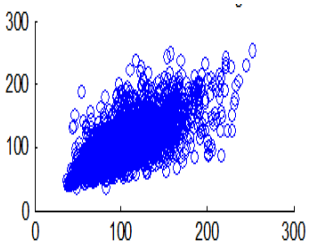


Figure 16(a): Vertical pixel distribution of plain image for red plane

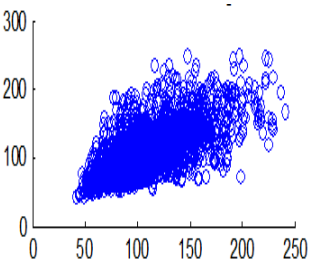


Figure 16(b): Vertical pixel distribution of plain image for green plane

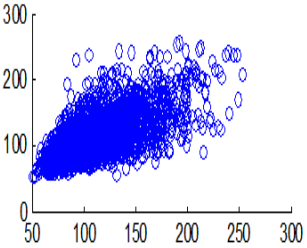


Figure 16(c): Vertical pixel distribution of plain image for blue pixel

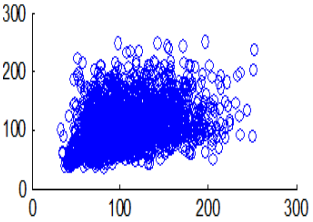


Figure 16(a): Diagonal pixel distribution of plain image for red plane

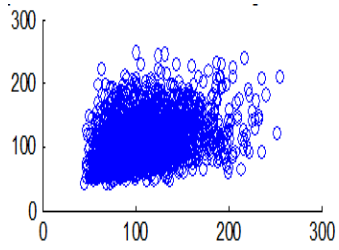


Figure 16(b): Diagonal pixel distribution of plain image for green plane

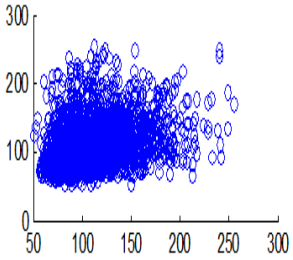


Figure 16(c): Diagonal pixel distribution of plain image for blue pixel

The experiment was conducted with 4096 pairs of neighboring pixels selected arbitrarily from the plain image and the enciphered image. Pixel distribution in Horizontal, Vertical and Diagonal directions corresponding to plain satellite image and encrypted image are plotted in figures 14 (a) –(c), 15 (a) – (c), 16 (a) –(c) and 17 (a) – (c), 18 (a) – (c), 19 (a) – (c) respectively. Encrypted image exhibits a uniform pixel distribution, which is a desirable property of a good encrypting algorithm to resist the statistical attacks. Correlation coefficients are given in table 1.

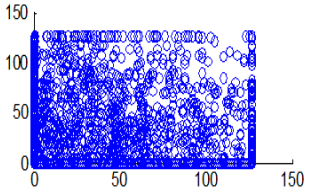


Figure 17(a): Horizontal pixel distribution of encrypted image for red plane

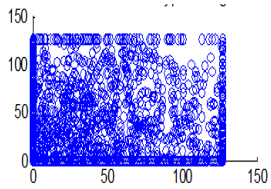


Figure 17(b): Horizontal pixel distribution of encrypted image for green plane

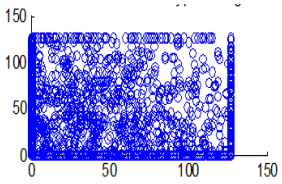


Figure 17(c): Horizontal pixel distribution of encrypted image for blue plane

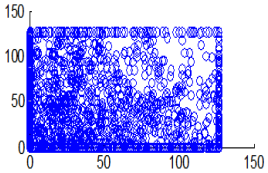


Figure 18(a): Vertical pixel distribution of encrypted image for red plane

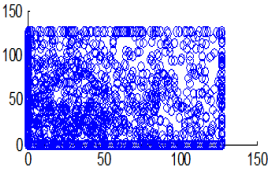


Figure 18(b): Vertical pixel distribution of encrypted image for green plane

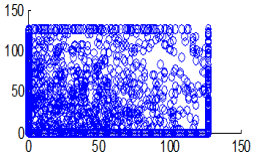


Figure 18(c): Vertical pixel distribution of encrypted image for blue plane

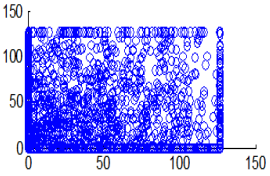


Figure 19(a): Diagonal pixel distribution of encrypted image for red plane

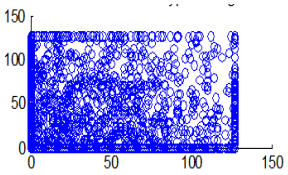


Figure 19(b): Diagonal pixel distribution of encrypted image for green plane

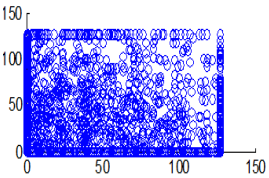


Figure 19(c): Diagonal pixel distribution of encrypted image for blue plane

Table 1: Correlation coefficients of two adjacent pixels in plain and encoded image

	Plain Image			Cipher Image		
Color	*R*	*G*	*B*	*R*	*G*	*B*
HC	0.6316	0.5629	0.5239	0.0057	-0.0013	0.0027
VC	0.7814	0.7525	0.7312	-0.0072	0.0393	-0.0021
DC	0.5247	0.5058	0.4179	-0.0056	-0.0189	-0.0170

3.1.4.3 Sensitivity Analysis

Since this algorithm encrypts one sub-image 3 times it is less prone to differential attacks. The initial key is obtained from the user and a special operation on the initially encrypted image acts as the input for the next part of the original image using radial Hilbert transform. This makes the algorithm more powerful as the complete image cannot be decrypted at once with a single key.

Four unique keys are required to decrypt the complete image. Input to radial Hilbert transform is image dependent. If the input given to radial Hilbert transform while decrypting is changed slightly, it leads to error propagation to the complete sub-image. The whole encrypted image changes as the sub images are interdependent. Thus the sensitivity is four times more than the algorithm with single key for the whole image. Figure 20 shows the deciphered image with a wrong key.

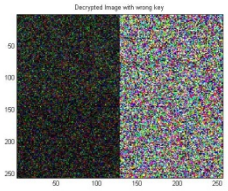


Figure 20: Deciphered image with a wrong key

To resist brute force attack, the algorithm must possess a very large key space. The technique proposed exhibits key sensitivity up-to 24 bits after the decimal point (10^{24}) and before the decimal point. There can be a range of 10^{2128} combinations of keys.

3.1.4.4 Differential analysis

3.1.4.5 Entropy Analysis

Information entropy of an image is a basic parameter to measure the randomness of pixels. A value of entropy near to 8 implies a random distribution.

Table 2: NPCR, UACI and entropy results

	Encrypted Image		
Color	*R*	*G*	*B*
NPCR (%)	99.6368	99.6017	99.6201
UACI (%)	33.5262	33.6828	33.5840
Entropy (bits)	7. 7499	7. 4897	7.3896

3.2. Nature Inspired Technique (Technique 2)

The general block diagram of the proposed encryption algorithm is shown in figure 21. The process includes a quantification unit, a logistic map generator, a mutation block, a crossover block and a scrambler. Chaotic sequence generator generates four chaotic sequences. The quantifier transforms these chaotic sequences into four streams of keys applied to crossover and mutation operations. Crossover unit alters the order of the image pixels along the horizontal and vertical direction (confusion). The mutation unit disguises the encryption image with a random mask image (diffusion). The scrambler alters the position of the image pixels using a random sequence. Functionality of the various units is described as follows:

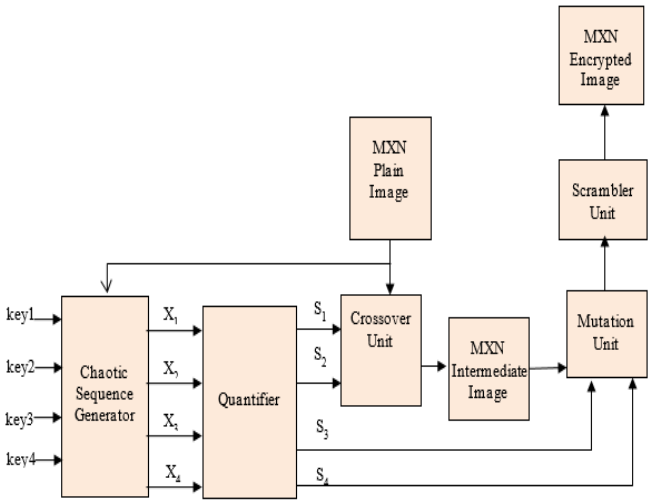


Figure 21: Architecture of the nature inspired image encryption algorithm

3.2.1 Chaotic Sequence Generator

Chaotic generator unit produces four random sequences with logistic map depending on the control parameters (μ1, μ2, μ3, μ4) and initial seed values (X_{10}, X_{20}, X_{30}, X_{40}). These keys form the set of shared secret values used in the cryptosystem. All the plain image pixels are added up and divided by (MxNx255). This fractional value is scaled down to 10% and added with the initial parameters. Thus the chaotic sequences generated are dependent on initial keys as well as the plain image. This makes the encryption algorithm stronger as the values X_1, X_2, X_3 and X_4 are not same with any two different images. Logistic map is used for the generation of chaotic sequences.

3.2.2 Quantifier

The quantifier receives four chaotic arrays X_1- X_4 produced by the chaotic map and reforms them to four sequences S_1- S_4. The length of the 1st and 3rd key streams is chosen equal to M, and the length of the 2nd and 4th key streams is made equal to N, where M × N is the dimension of the plain image.

3.2.3 Crossover Unit

Let K_i and K_i+1 denote two sequential elements present in the key stream. Rows/columns numbered Ki and K_i+1 are undergo crossover process. Cut points occupy the positions in accordance with the following equation:

$$r_1 = \max\{1, |K_i - K_{j\_} + 1|\text{modL}\} \quad (7)$$

$$r_2 = \max\{1, r_1 + |K_i - K_{j\_} + 1|\text{modL}\} \quad (8)$$

$$r_k = \max\{1, r_{k-1} + |K_M - K_N + 1|\text{modL}\} \quad (9)$$

where k refers to the number of cut points, L = M or N = the length of the row/column, and (r_1, r_2, …, r_K) are the localities of

cut points. Crossover operation at multiple points is executed by exchanging the even or odd numbered sectors of i and j.

3.2.4 Mutation Unit

Mutation unit disguises the transitional values rearranged after crossover by XORing with a random image. The random mask image is generated before the encryption process starts and shared between the two end users. Mutation unit changes each and every pixel in the intermediary image by XORing with an arbitrary pixel selected by the values of the S_3 and S_4 from the secrete mask image. Any $(i, j)^{th}$ pixel in the enciphered image is attained by XORing the corresponding pixel in the transitional image and $(p_i, q_j)^{th}$ pixel of the enciphered image, where $p_i \epsilon S_3$ and $q_i \epsilon S_4$.

3.2.5 Scrambler Unit

The scrambling is incorporated by altering the positions of the mutated image pixel values as per the elements in a random sequence generated by a random seed. The process for a sample sequence {2, 3, 1… N} is tabulated in table 3.

Table. 3: Scrambler unit design

	Random Sequence					
Random Sequence		2	3	1	…	N
	2	(2,2)	(2,3)	(2,1)	…	(2,N)
	3	.	.	.	.	.
	.	.	.	.	.	.
	..	.	.	.	.	.
	N	(N,2)	(N,3)	(N,1)	..	(N, N)

3.2.6 Experimental results

A sample satellite image shown in Figure 22 is enciphered using the proposed method. Key used is {3.7158, 0.11, 3.89858, 0.25, 3.76158, 0.35, 3.8458, 0.6520}. Figure 23 shows the encrypted image.



Figure 22: Original image,



Figure 23: Encrypted Image

3.2.6.1 Histogram Analysis

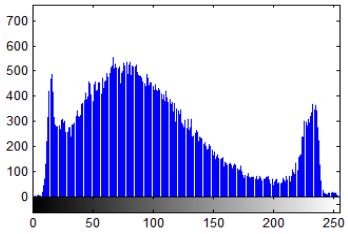


Figure 24(a): Histogram of plain image for red plane

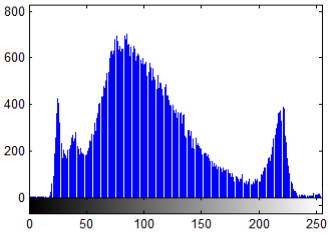


Figure 24(b): Histogram of plain image for green plane

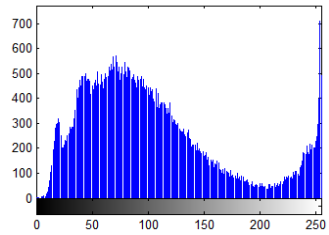


Figure 24(c): Histogram of plain image for blue plane

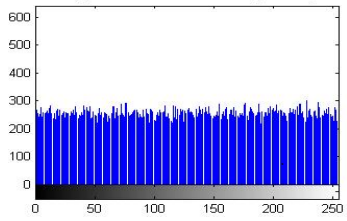


Figure 25(a): Histogram of cipher image for red plane

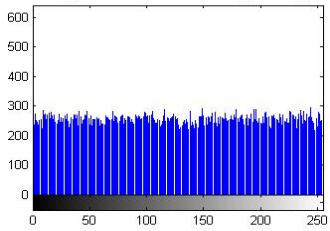


Figure 25(b): Histogram of cipher image for green plane

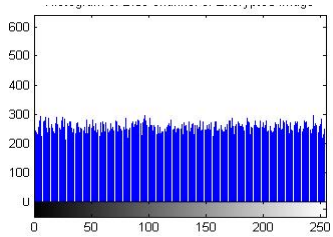


Figure 25 (c): Histogram of cipher image for blue plane

The histograms of plain image are presented in figure 24 (a)-(c). From figure 25(a)-(c), it is witnessed that the histogram of the encrypted image differs completely from the histogram of the plain image and has a uniform distribution, which is the desired property to make the algorithm hard for cryptanalysis.

3.2.6.2 Correlation Analysis

Correlation between two contiguous pixels corresponding to the plain image and cipher image is tested in this experiment. Horizontal, Vertical and Diagonal pixel distribution of plain image and cipher image for red, green and blue planes is shown in figures 26 (a) –(c), 27(a) –(c), 28(a)-(c) and 39(a)-(c), 40(a) –(c), 41(a)-(c) respectively. It is observed that the distribution is almost same and uniformly distributed for all the three-color planes in the cipher image.

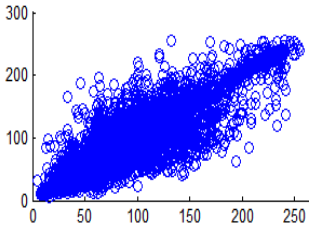


Figure 26(a): Horizontal pixel distribution of plain image for red plane

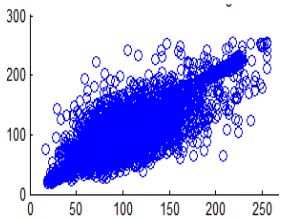


Figure 26(b): Horizontal pixel distribution of plain image for green plane

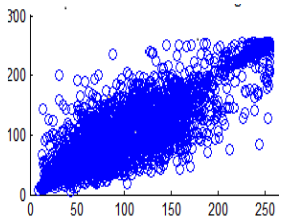


Figure 26(c): Horizontal pixel distribution of plain image for blue plane

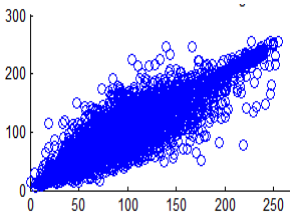


Figure 27(a): Vertical pixel distribution of plain image for red plane

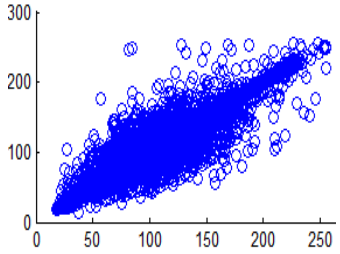


Figure 27(b): Vertical pixel distribution of plain image for green plane

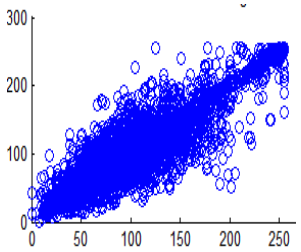


Figure 28(c): Vertical pixel distribution of plain image for blue plane

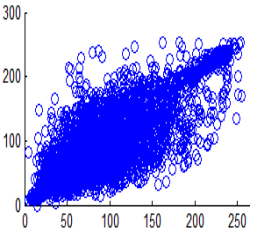


Figure 29(a): Diagonal pixel distribution of plain image for red plane

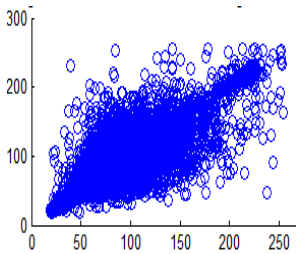


Figure 29(b): Diagonal pixel distribution of plain image for green plane

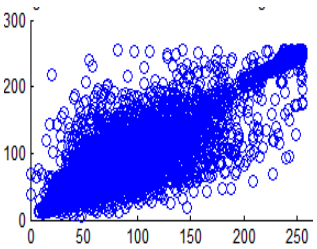


Figure 29(c): Diagonal pixel distribution of plain image for blue plane

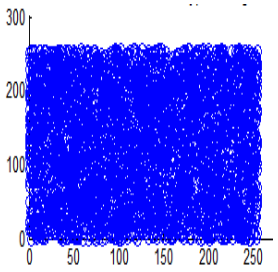


Figure 30(a): Horizontal pixel distribution of cipher image for red plane

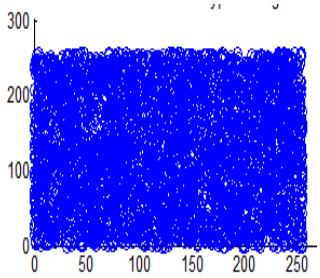


Figure 30(b): Horizontal pixel distribution of cipher image for green plane

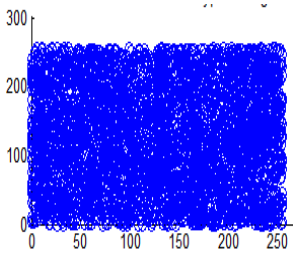


Figure 30(c): Horizontal pixel distribution of cipher image for blue plane

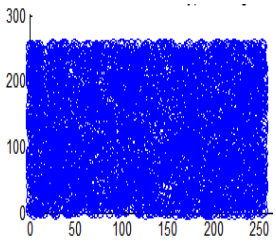


Figure 31(a): Vertical pixel distribution of cipher image for red plane

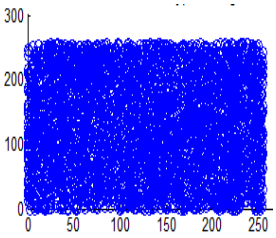


Figure 31(b): Vertical pixel distribution of cipher image for green plane

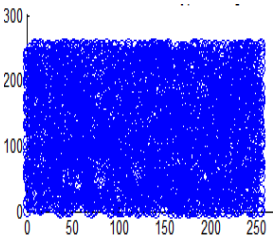


Figure 31(c): Vertical pixel distribution of cipher image for blue plane

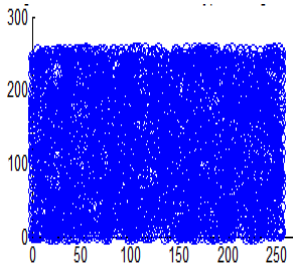


Figure 32(a): Diagonal pixel distribution of cipher image for red plane

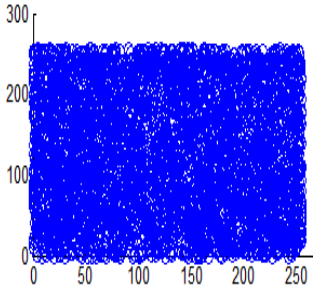


Figure 32(b) Diagonal pixel distribution of cipher image for green plane

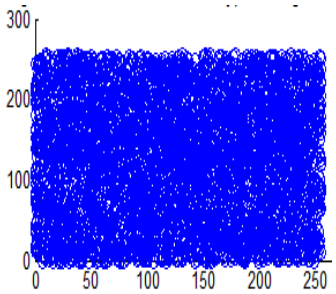


Figure 32(c): Diagonal pixel distribution of cipher image for blue plane

The following equations are used to obtain the correlation coefficients r(x,y) in horizontal (HC), vertical (VC) and diagonal (DC) directions:

$$r(x,y) = \frac{COV(p,q)}{\sqrt{D(p)}\sqrt{D(q)}} \quad (10)$$

where p and q represent gray scale values of two neighboring pixels in the image, D(p) denotes the variance of p and COV (p, q) indicates the covariance of p and q.

The experiment was executed by selecting 4096 random adjacent pixel sets from the plain image and the ciphered image. Correlation coefficients are calculated using equation 3.3 and documented in table 4. It can be observed that the cipher image correlation coefficients are uniformly distributed in the range of 0-255 for all the three planes. This shows the strength of the algorithm towards statistical attacks.

Table 4: Correlation coefficients of two neighboring pixels in input and enciphered image

	Plain Image			Cipher Image		
Color	*R*	*G*	*B*	*R*	*G*	*B*
HC	0.9545	0.9491	0.9223	-0.0005	-0.0008	0.0185
VC	0.9680	0.9558	0.9542	-0.0006	-0.0266	0.0052
DC	0.9269	0.9034	0.8875	-0.0242	-0.0109	-0.0051

3.2.6.3 Differential Analysis

The number of changing pixel rate and the unified averaged changed intensity are effective parameters used in the assessment of the robustness of an image encryption algorithm against differential attacks. Parameter NPCR implies the number of pixels

changed over the encrypted image with only one single pixel in the plain image being varied. The UACI index represents difference in the intensity between plain and encrypted images on an average. The equations 1.9 and 1.10 are used in the calculation of NPCR and UACI respectively. The results for input satellite image shown in figure 3.5(a) are given in table 5 that shows the resistance at differential attacks

Table 5: NPCR, UACI and entropy results

	Encrypted Image		
Color	*R*	*G*	*B*
NPCR (%)	99.6368	99.6017	99.6201
UACI (%)	33.5262	33.6828	33.5840
Entropy (bits)	7.9974	7.9974	7.9973

4. Comparative Study

A huge variety of chaos based image encryption algorithms have been modeled and developed in the past few years. Usually these algorithms are tested on a single image and the conclusion is drawn based on the performance parameters obtained. There arises a need to test the algorithms for a set of images with identical features and decide about the suitability of the algorithms for the input dataset. Proposed work attempts to analyze the results obtained with the algorithms tested using three datasets, namely, satellite images, face images and handwritten signature images. Sample images from each of these datasets are shown in Figure 33, 34 and 35.



Figure 33: Satellite image



Figure 34: Face image



Figure 35: Signature Image

The three datasets have a vast variation in their features. Satellite images have many small segments which show a high level of pixel variations. Also, the content is massive. Face images have few different segments with identical colors. Content is almost uniform in a segment. Signature images have only two colors, blue and white. Pixels are almost identical except the blue colored area. Fifty images belonging to each class are considered in the comparative study.

4.1. Statistical features

Statistical parameters for the three sample images are provided in table 2. Satellite images have almost equal variation in the three planes, with an average centered at the middle value in the range of 0-255. Face images have a bit higher deviation in the three planes and the average is smaller as compared to satellite images. Signature images are centered at near white pixel because of the large white background. Variation is quite different for the three planes. However, all the three sets follow an identical pattern with respect to the correlation coefficients.

Table 5: Statistical parameters for the three datasets

Parameter	Satellite image	Face image	Signature image
mean	102.0983	85.4429	212.9507
standard deviation (R plane)	45.7829	53.3737	91.6546
standard deviation (G plane)	40.5145	44.4940	95.4395
standard deviation (B plane)	50.0683	52.0977	10.6353
correlation coefficient (R plane)	1	1	1
correlation coefficient (G plane)	0.8618	0.9209	0.9999
correlation coefficient (B plane)	0.9715	0.9480	0.3898
Entropy	7.303112	7.307502	2.740406

Results obtained with fifty images from each dataset are compared in the following section to decide about the suitability of the algorithm for the particular dataset. Performance parameters considered for the analysis include horizontal correlation, vertical correlation, diagonal correlation, net changing pixel rate, unified average change in intensity, entropy and encryption time taken by the encryption technique.

4.2. Results for Satellite Images

Fifty satellite images are considered for testing. They are RGB images with sizes varying from 128x128 to 1024x1024 pixels. The images are resized to 256x256 to bring the uniformity in the result. Performance parameters included in the analysis are correlation coefficient, NPCR, UACI, entropy and time required for encryption. Figure 36 (a) to (f) shows the plot of the pixel correlations for fifty encrypted satellite images.

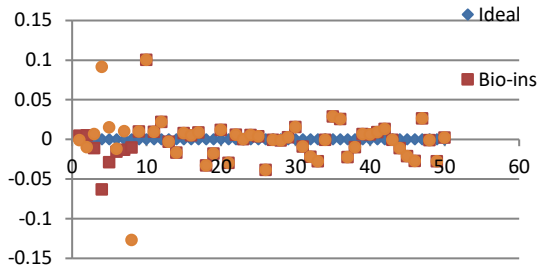


Figure 36(a): Horizontal correlation for 50 images

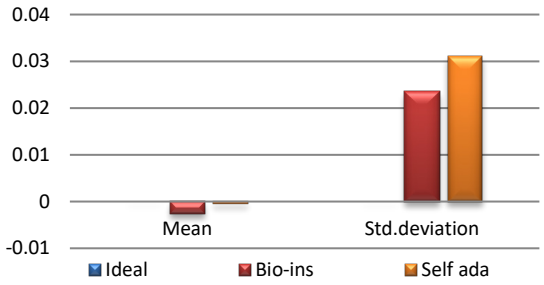


Figure 36(b): Mean Horizontal Correlation

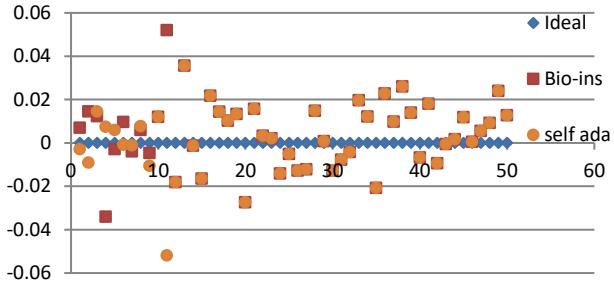


Figure 36 (c): Vertical Correlation for fifty images

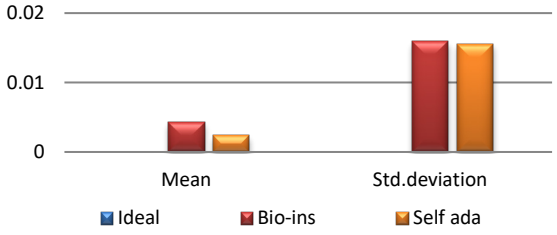


Figure 36 (d): Mean vertical correlation

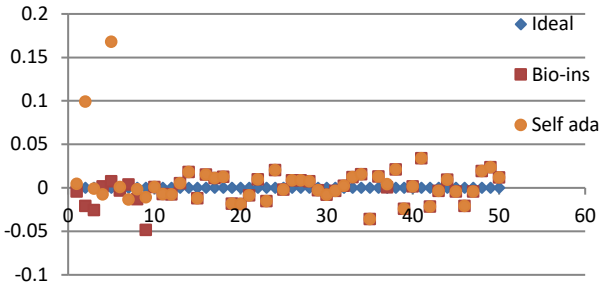


Figure 36(e): Diagonal Correlation for fifty images

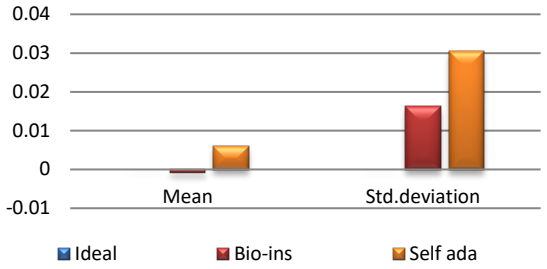


Figure 36(f): Mean Diagonal Correlation

Results of performance comparison for satellite images are displayed in Figure 37.

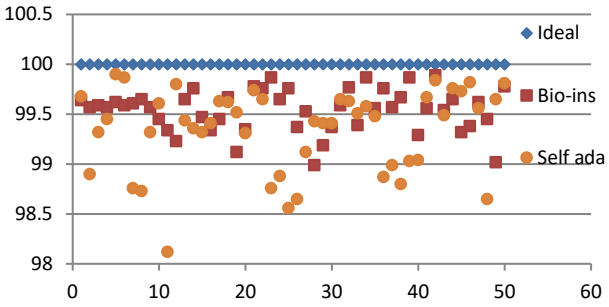


Figure 37(a): NPCR for fifty images

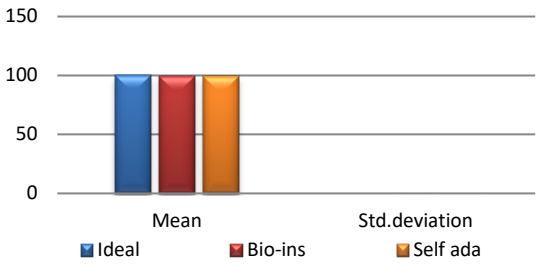


Figure 37 (b): Mean NPCR

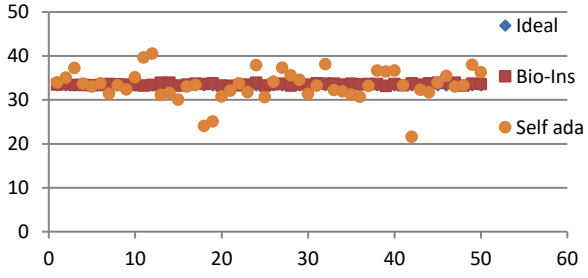


Figure 37(c): UACI for fifty images

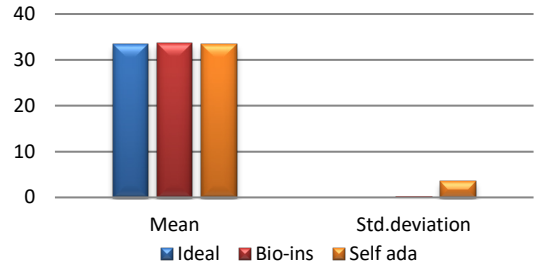


Figure 37(d): Mean UACI

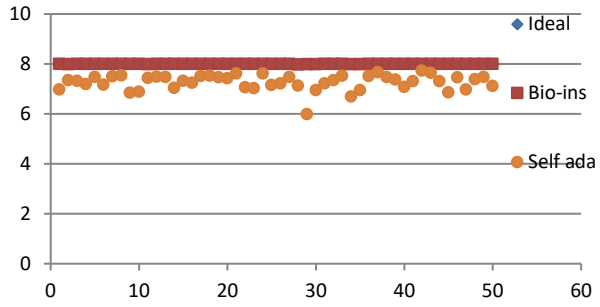


Figure 37(e): Entropy for fifty images

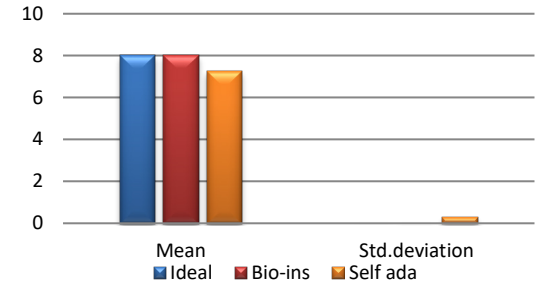


Figure 37(f): Mean Entropy

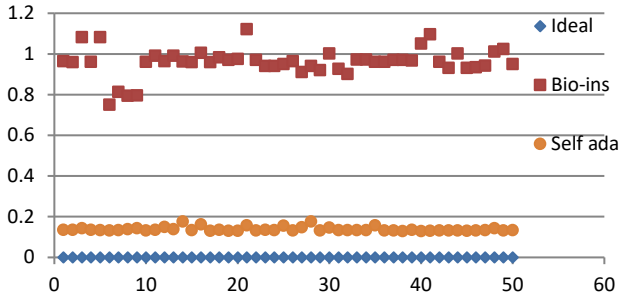


Figure 37(g): Encryption time for fifty images

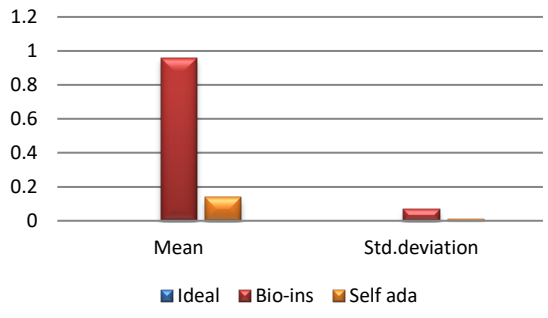


Figure 37(h): Mean encryption time

Referring to these results, Bio-inspired technique performs well with respect correlation coefficients, but requires more time for encryption. Self-adaptive method works well except entropy. Since the satellite images are generally bulky in size and sometimes need to be transmitted in real time, encryption time is also an important parameter to decide the capability of an algorithm. Self-adaptive encryption technique seems to be more time-efficient for the encryption of the satellite images.

4.3. Results for Face Images

Fifty RGB color images containing human face are provided as the inputs for testing. These images are obtained from an open source database [41] and have size of 180x200 pixels. Figure 38(a) to (f) displays the correlation parameters obtained for the fifty encrypted face images.

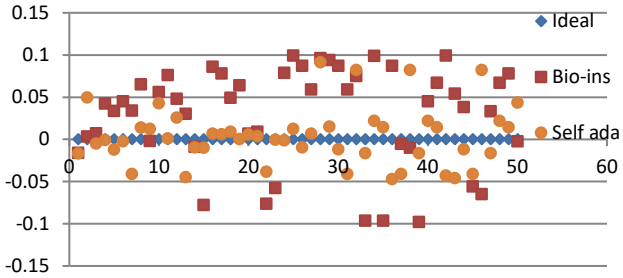


Figure 38(a): Horizontal Correlation for fifty images

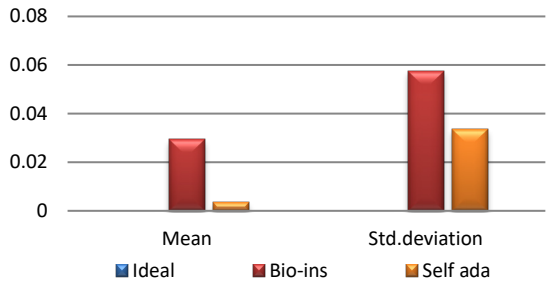


Figure 38(b): Mean Horizontal Correlation

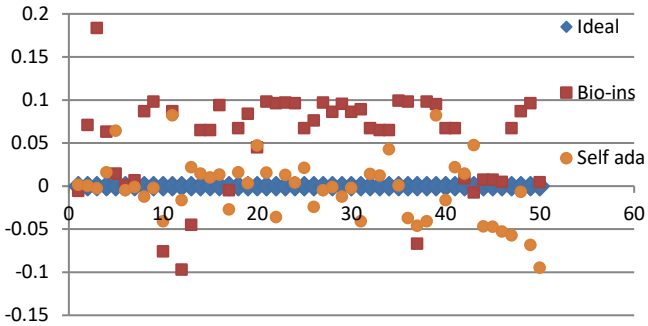


Figure 38(c): Vertical Correlation for fifty images

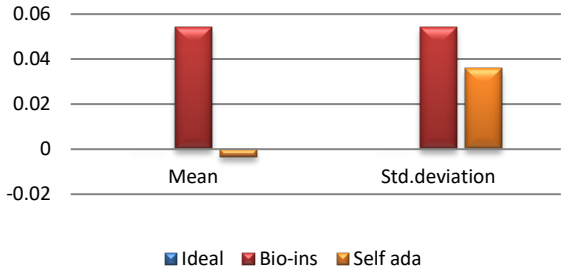


Figure 38(d): Mean Vertical correlation

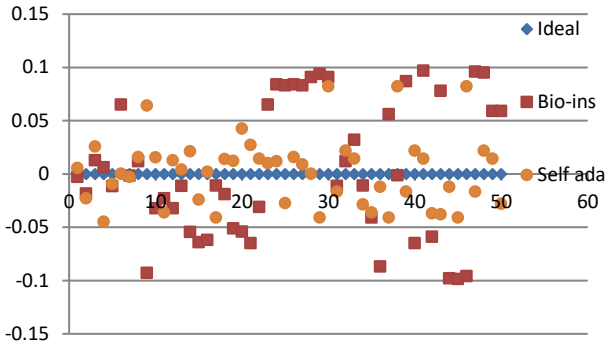


Figure 38(e): Diagonal correlation for fifty images

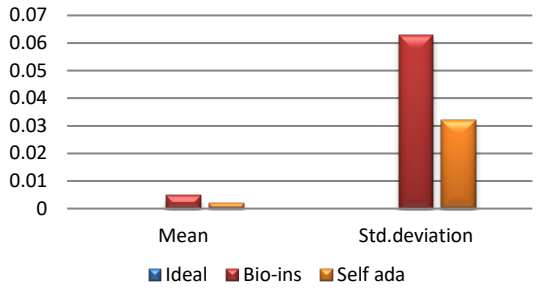


Figure 38(f): Mean Diagonal Correlation

Performance comparison for face images is visualized in Figure 39.

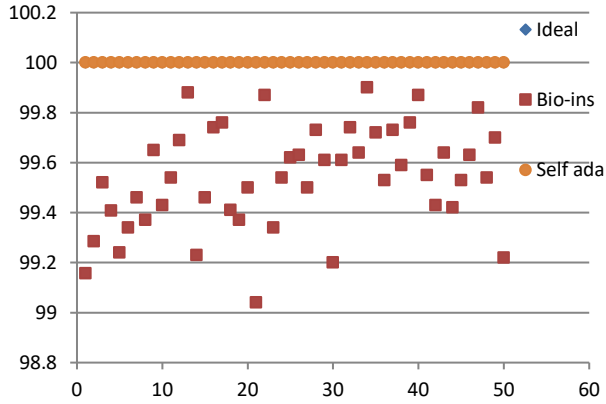


Figure 39 (a): NPCR for fifty images

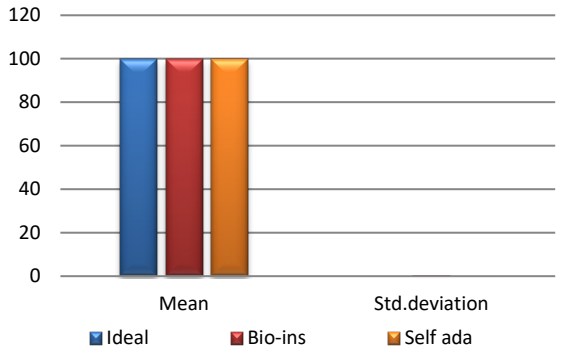


Figure 39(b): Mean NPCR

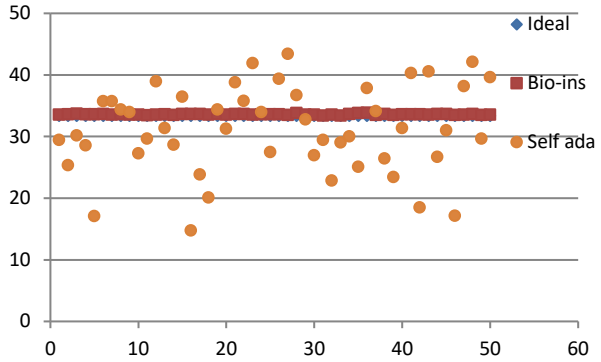


Figure 39(c): UACI for fifty images

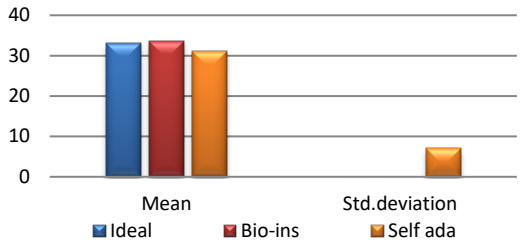


Figure 39(d): Mean UACI

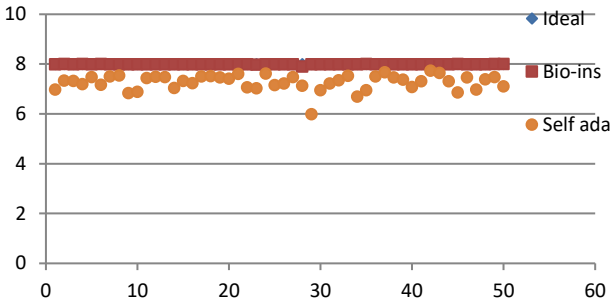


Figure 39(e): Entropy for fifty images

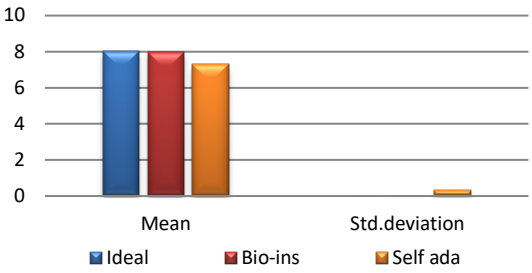


Figure 39(f): Mean Entropy

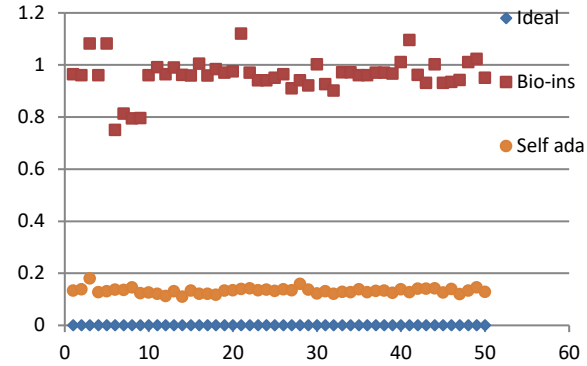


Figure 39(g): Encryption time for fifty images

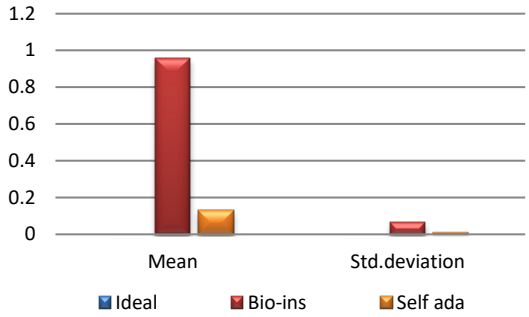


Figure 39(h): Mean Encryption time

Results show that bio-inspired algorithm performs low with correlation test and takes more time for encryption. Self-adaptive algorithm is the fastest but lacks in performance with respect to UACI and entropy values. Face has an important significance in biometric based identification and authentication applications. However, the size of the images does not vary much and the need for real time transmission has limited scope. Hence, any of these algorithms can be chosen for encryption of this kind of images.

4.4. Results for Signature images

Fifty sample signatures are collected from fifty different people with a mobile app. Signatures are RGB images with size varying from 320x320 pixels to 675x675 pixels with white background. Performance parameters obtained for the fifty signature images are documented in Figures. 40 (a) to (f) and 41(a) to (h)

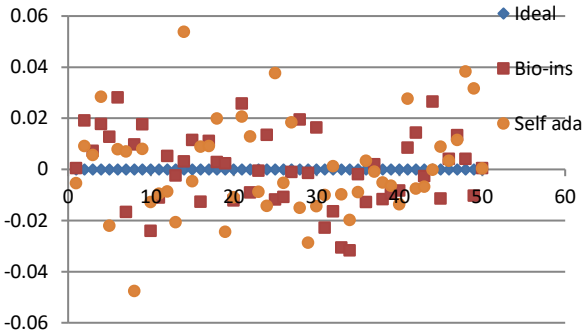


Figure 40(a): Horizontal correlation of fifty images

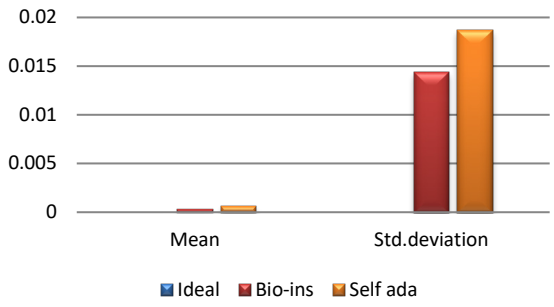


Figure 40(b): Mean Horizontal correlation

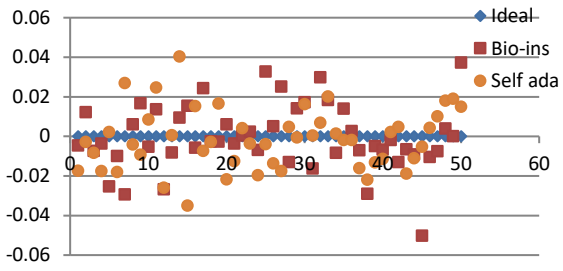


Figure 40(c): Vertical correlation for fifty images

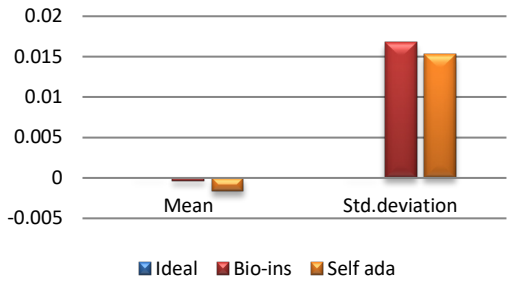


Figure 40(d): Mean vertical correlation

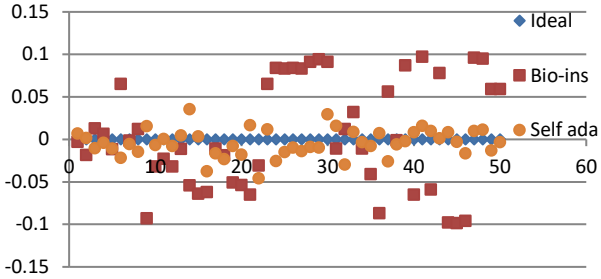


Figure 40(e): Diagonal correlation for fifty images

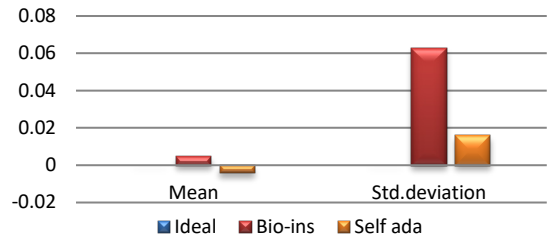


Figure 40(f): Mean Diagonal correlation

Below images Performance comparison for signature images from figure 41 (a) – (g)

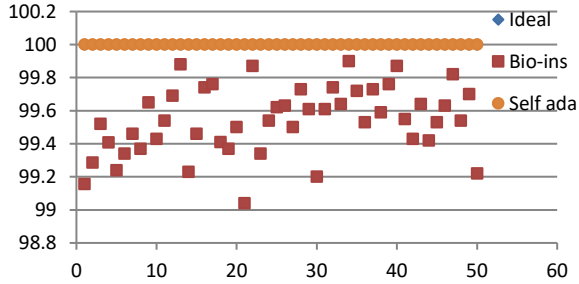


Figure 41(a): NPCR for fifty images

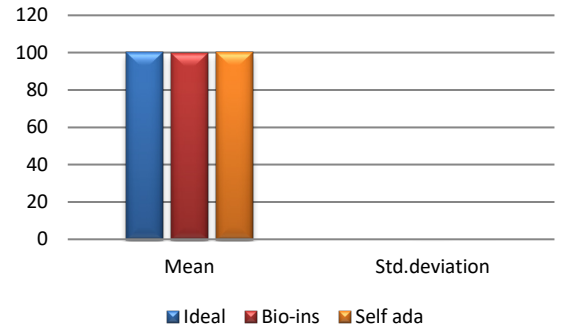


Figure 41(b): Mean NPCR

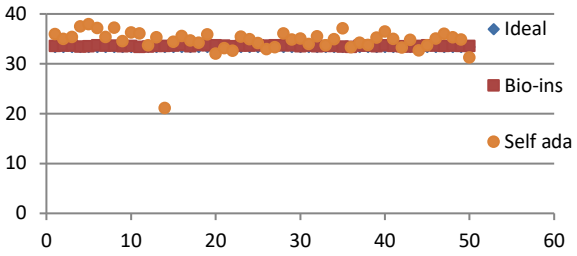


Figure 41(c): UACI for fifty images

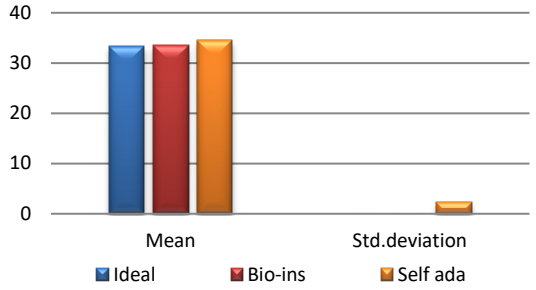


Figure 41(d): Mean UACI

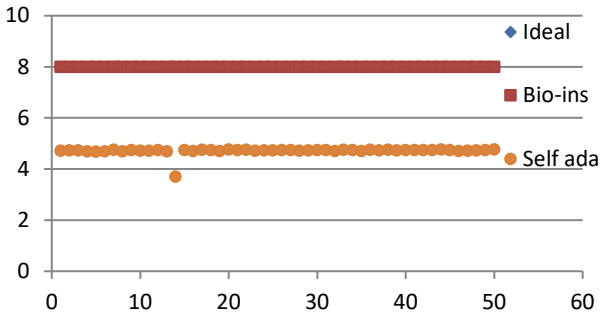


Figure 41(e): Entropy for fifty images

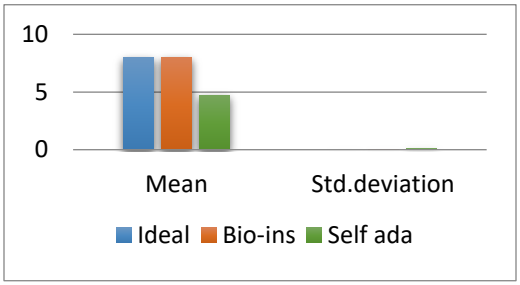


Figure 41(f): Mean Entropy

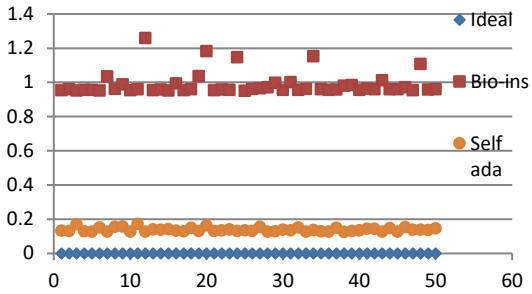


Figure 41(g): Encryption time for fifty images

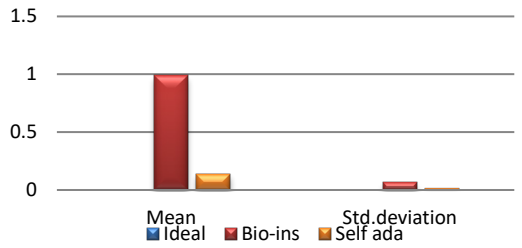


Figure 41(h): Mean encryption time

It can be observed that nature inspired encryption technique performs high for correlation, NPCR, entropy and UACI. Self-adaptive technique has high performance for correlation, NPCR and time, with entropy and UACI values deviating from the ideal values.

Signature images are mainly involved in online shopping and banking applications. Security of them is the primary factor rather than the time taken. Real time processing of these images has limited scope. Also, size of the signature images is not bulky and does not vary much. Therefore, nature inspired technique suits better for the signature image encryption in comparison with the self-adaptive technique.

5. Conclusion and future scope

Suitability analysis is performed on three datasets with different statistical features. Results obtained reveal that the bio-inspired technique provides better performance with respect to NPCR, UACI and entropy, but takes more time to encrypt. Self-adaptive encryption performs well in consideration with correlation and NPCR. Limitation of this method is its low performance in the UACI and entropy tests. Suitability analysis finds out that satellite images can be encrypted more efficiently using self-adaptive method. Performance is almost equal for face images. Signature image encryption can achieve better quality by using bio-inspired method and fast partial method.

References

[1] M. Javidi and R. Hosseinpourfard, "Chaos Genetic Algorithm Instead Genetic Algorithm", The International Arab Journal of Information Technology, . **12**(2), 163-168, March 2015. https://doi.org/10.1007/s11633-017-1107-6

[2] D. Haufu, L. Xiao-lu., L. Xue, "An Improved Genetic Algorithm for Combinatorial Optimization," in Proceedings of the IEEE International Conference on Computer Science and Automation Engineering, Shanghai) 58-61, 2011. https://doi.org/10.1109/CEC.2014.6900496

[3] K. Tang, "An Improved Genetic Algorithm based on A Novel Strategy for Nonlinear Programming Problems," Computers and Chemical Journal, **35**(3), 615-621, 2011. https://doi.org/10.1016/j.compchemeng.2010.06.014

[4] F. Ye F., Y. Haiyang, and J. Xueshou, "An Improved Constrained Optimization Genetic Algorithm," in Proceedings of IEEE International Conference on ICIS, Xiamen, China) 435-439, 2010. https://doi.org/10.1109/ICICISYS.2010.5658317

[5] K. Shankar, P Eswaran, "An Efficient Image Encryption Technique Based on Optimized Key Generation in ECC Using Genetic Algorithm", Artificial Intelligence and Evolutionary Computations in Engineering Systems, , **394**, 705-714. https://doi.org/10.1007/978-81-322-2656-7_64.

[6] H. Nematzadeh, RasulEnayatifar, bHomayunMotameni, Frederico GadelhaGuimarães, Vitor NazárioCoelho, "Medical image encryption using a hybrid model of modified genetic algorithm and coupled map lattices", **110**, 24-32, 2018. https://doi.org/10.1016/j.optlaseng.2018.05.009.

[7] X. Wang, Hui li Zhang, "A novel image encryption algorithm based on genetic recombination and hyper-chaotic systems", Nonlinear Dynamics, **83**, 333-346, 2016. https://doi.org/10.1007/s11071-015-2330-8.

[8] Cheng T., Wang C., Xu M., and Chau W., "Optimizing Hydropower Reservoir Operation using Hybrid Genetic Algorithm and Chaos," Water Resources Management, **22**(7), 895-909, 2008. https://doi.org/10.1007/s11269-007-9200-1

[9] Chao-Lin Kuo, et.al, "Image Encryption Based on Fuzzy Synchronization of Chaos Systems", IEEE 37th Annual Computer Software and Applications Conference)45 3-461, 2013. https://doi.org/10.1109/COMPSAC.2013.23

[10] Abdul Hanan Abdullaha, Rasul Enayatifara,, Malrey Lee, "A hybrid genetic algorithm and chaotic function model for image encryption", International Journal on Electronics and Communication (AEÜ), **66**, 806– 816,2012. https://doi.org/10.1016/j.aeue.2012.01.015

[11] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. doi: 10.1109/ICECCE47252.2019.8940756.

[12] V. Srikanth, et.al, "Bit-Level Encryption of Images using Genetic Algorithm", TECHNIA, International Journal of Computing Science and Communication Technologies, **3**(1), 2010. https://doi.org/10.1007/978-3-642-30111-7_75

[13] K.A Al-Utaibi., El-Alfy, "A bio-inspired image encryption algorithm based on chaotic maps", IEEE Congress on Evolutionary Computation (CEC)) 87-92, 2010. https://doi.org/10.1109/CEC.2010.5586463

[14] Hassan Al-Mahdi , Yaser Fouad, "Design and analysis of DNA Binary Cryptography Algorithm for Plaintext", International Journal of Engineering and Technology (IJET), 10(3) 699-706, 2018. https://doi.org/10.21817/ijet/2018/v10i3/181003055

[15] Narendra K. Pareek, Vinod Patidar, "Medical image protection using genetic algorithm operations", Soft Computing, **20**(2), 763–772, 2016. https://doi.org/10.1007/s00500-014-1539-7

[16] Yuansheng Liu,et.al, "Cryptanalyzing a RGB image encryption algorithm based on DNA encoding and chaos map", Optics & Laser Technology, 60(6)111–115, 2014. https://doi.org/10.1016/j.optlastec.2014.01.015

[17] H. Wen et.al, "Breaking an Image Encryption Algorithm Based on DNA Encoding and Spatiotemporal Chaos', Entropy, **21**(7), 246-263, 2019. https://doi.org/10.3390/e21030246

[18] Areeba Fatima, Isha Mehra and Naveen K Nischal , "Optical image encryption using equal modulus decomposition and multiple diffractive imaging", Journal of optics, **18**, 2016, 10.1088/2040-8978/18/8/085701.

[19] R. Kumar, Basanta Bhaduri, "Optical image encryption using Kronecker product and hybrid phase masks", Optics & Laser Technology, **95**, 51-55, 2017. https://doi.org/10.1016/j.optlastec.2017.03.041

[20] Ravi Kumar, Basanta Bhaduri, "Optical image encryption in Fresnel domain using spiral phase transform", Journal of optics, **19**, 2017. 10.1088/2040-8986/aa7cb1

[21] M. Joshi, Chandra Shakher, Kehar Singh, "Image encryption and decryption using fractional Fourier transform and radial Hilbert transform', Optics and Lasers in Engineering, **46**(1), 522– 526, 2008. https://doi.org/10.1016/j.optlaseng.2008.03.001

[22] S. Liu, "A review of optical image encryption techniques", Optics & Laser Technology, **57**(1), 327–342, 2014. https://doi.org/10.1016/j.optlastec.2013.05.023

[23] Z. Liu, et.al, "Color image encryption by using the rotation of color vector in Hartley transform domains", Optics and Lasers in Engineering, **48**, 800–805, 2010. https://doi.org/10.1016/j.optlaseng.2010.02.005

[24] N. Singh, A. Sinha, "Optical image encryption using fractional Fourier transform and chaos", Optics and Lasers in Engineering, **46**(2) 117–123, 2008. https://doi.org/10.1016/j.optlaseng.2007.09.001

[25] Y. Liang & G. Liu "Color image encryption combining a reality-preserving fractional DCT with chaotic mapping in HIS space". Multimedia Tools and Applications, **75**(11), 6605-6620, 2016. https://doi.org/10.1007/s11042-015-2592-7

[26] Dr. Emad S. Othman, Dr. Mohammed M. Sakre, "Compression and Encryption Algorithms for Image Satellite Communication", International Journal of Scientific & Engineering Research, **3**(9), 1-4, 2012.

[27] Qiwen Ran, et.al, "Vector power multiple-parameter fractional Fourier transform of image encryption algorithm", Optics and Lasers in Engineering, **62**, 80–86, 2014. https://doi.org/10.1016/j.optlaseng.2014.05.008

[28] Ensherah A. Naeem, et.al, "Efficient implementation of chaotic image encryption in transform domains", The Journal of Systems and Software, **97**(3), 118–127, 2014. https://doi.org/10.1016/j.jss.2014.07.026

[29] R. A. Schowengerdt, "Remote sensing, third edition: models and methods for image processing", Academic Press, 2006.

[30] Anil K. Jain, et.al, "Statistical Pattern Recognition: A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. https://doi.org/10.1109/34.824819

[31] Panigrahi, Sushant, and Toran Verma, "Texture image classification using neuro fuzzy approach." International Journal of Engineering and Computer Science, **2**(1), 2309-2313, 2013.

[32] Sura F. Yousif, ""Grayscale Image Confusion and Diffusion Based on Multiple Chaotic Maps", 1st International Scientific Conference of Engineering Sciences - 3rd Scientific Conference of Engineering Science (ISCES), 2018. https://doi.org/10.1109/ISCES.2018.8340538

[33] Jianhua Wu, Mengxia Zhang and Nanrun Zhou,"Image encryption scheme based on random fractional discrete cosine transform and dependent scrambling and diffusion", Journal of Modern optics, 2016. https://doi.org/10.1080/09500340.2016.1236990

[34] P. Maan, Hukum Singh, "Non-linear Cryptosystem for Image Encryption Using Radial Hilbert Mask in Fractional Fourier Transform Domain", 3D Research, **9**(53), 112-119, 2018. https://doi.org/10.1007/s13319-018-0205-8

[35] V. Srikanth, et.al, "Bit-Level Encryption of Images using Genetic Algorithm", TECHNIA, International Journal of Computing Science and Communication Technologies, 3, issue1, 2010.

[36] R. Bhagyashri. Pandurangi, Meenakshi R. Patil, "A Nature Inspired Color Image Encryption Technique to Protect the Satellite Images", International Journal of Current Engineering and Technology, **9**(3), May 2019. https://doi.org/10.14741/ijcet/v.9.3.4

[37] R. Bhagyashri. Pandurangi, Meenakshi R. Patil, Chaitra Bhat, "Comparison of Bio-inspired and Transform based Encryption Algorithms for Satellite Images", Third IEEE International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysore, India, December 2018. https://doi.org/10.1109/ICEECCOT43722.2018.9001344

[38] H. Gao, et.al, "A new chaotic algorithm for image encryption", Chaos, Solitons& Fractals, **2**(29), 393-399, July 2006. https://doi.org/10.1016/j.chaos.2005.08.110

[39] Adolf W. Lohmann, David Mendlovic, and Zeev Zalevsky, "Fractional Hilbert transform," Optics Letters, **21**(4), 281-283, 1996. https://doi.org/10.1364/OL.21.000281

[40] https://bhuvan-app1.nrsc.gov.in/imagegallery/bhuvan.html

[41] http://cmp.felk.cvut.cz/~spacelib/faces/faces94.html

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1082-1088 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Human Emotion Recognition Based on EEG Signal Using Fast Fourier Transform and K-Nearest Neighbor

Anton Yudhana[1,*], Akbar Muslim[1], Dewi Eko Wati[2], Intan Puspitasari[2], Ahmad Azhari[3], Murein Miksa Mardhia[3]

[1]*Faculty of Industrial Technology, Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, 55166, Indonesia*

[2]*Faculty of Teacher Training Education, Department of Teacher Training of Early Childhood, Universitas Ahmad Dahlan, Yogyakarta, 55166, Indonesia*

[3]*Faculty of Industrial Technology, Informatics Department, Universitas Ahmad Dahlan, Yogyakarta, 55166, Indonesia*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 28 November, 2020
Online: 14 December, 2020

Keywords:
Electroencephalography (EEG)
Brainwaves
FFT
KNN

ABSTRACT

Human emotional states can transform naturally and are recognizable through facial expressions, voices, or body movements, influenced by received stimuli. However, the articulation of emotions is not practicable by every individual, even when feelings of joy, sadness, or otherwise are experienced. Biomedically, emotions affect brain wave activities, as the continuously functioning brain cells communicate through electrical pulsations. Therefore, an electroencephalogram (EEG) is used to capture input from brain signals, study impulses, and determine the human mood. The examination generally included observing a person's frame of mind in response to a given stimulus where the immediate results were inconclusive. In this study, the associated classifications were normal, focused, sad, and shocked. The raw brainwave data from 50 subjects were recorded by employing a single-channel EEG called the Neurosky Mindwave. Meanwhile, the assessments were performed while the candidates' minds were stimulated by listening to music, watching videos, or reading books. The Fast Fourier Transform (FFT) method was utilized for feature extractions, along with the K-nearest neighbours (K-NN) for classifying brain impulses. The parameter k had a value of 15, and the average classification accuracy was 83.33%, while the highest accuracy for the focused emotional state was 93.33%. The Neurosky Mindwave in conjunction with the FFT and KNN techniques is potential analytical solutions to facilitate the enhanced identification of human emotional conditions.

1. Introduction

The signal of the brainwaves has specific traits and characteristics of the person. The brainwaves cannot be imitated or interpreted by individuals so that there is no chance of resemblance. Identity recognition is essential to identify the features of the person [1].

Identity development is done using biometric technology, which aims at personality-implantation through the use of parts of the body or social behavior—application of Identity Recognition using Biometrics Technology for self-recognition by using parts of the body or human behavior [2-6]. In [7], the author has presented machine-learning algorithms to categorize EEG dynamics according to subject self-reported emotional states during music listening. Support Vector Machine was employed to classify four emotional states and obtained an accuracy of 82.29% across 26 subjects. Vanitha [8] has used Hilbert-Huang Transform (HHT) to remove artifacts and relevant time-frequency features from the EEG signal, and the extracted features are manipulated to detect stress levels using Support Vector Machine (SVM) Classifier on real-time stress detection.

EEG is used as a diagnostic tool by medicine. In comparison, a broader application of the EEG is to analyze a person's mental state, thought patterns, or emotional state [9]. The EEG signal's visual observation is real, given the very low amplitude of the EEG signal, and the pattern is very complex. In addition, EEG signals are strongly influenced by various variables, including emotional conditions, health, respondent activity, environment, electrical disturbances from other organs, and external stimuli.

*Corresponding Author: Anton Yudhana, eyudhana@ee.uad.ac.id

[10]. Conducting the introduction of the EEG signal requires a pattern of brain activity that is prominent and constant. This pattern emerged due to specific stimulation from volunteers. EEG signals are shown to provide more insight into the subject's emotional processes and responses. The recorded EEG signals will get the waveform, the duration, the orientation of the signal, and the signal rhythm. [11, 12].

The EEG signal pattern data's shape and position are not yet known because the signal pattern and response speed response of each individual's brain to the received stimuli are different [13]. Therefore, we need a signal processing method EEG that can pattern specific brain activity [14].

In this study, the Fast Fourier Transform (FFT) based spectrum analysis method converts the EEG signal from the time domain to frequency domain. The FFT results were classified by K-Nearest Neighbor (KNN) as a training data set to characterize the brain wave output from the EEG as a material for analyzing differences in brain waves in each emotional state studied.

2. Data Acquisition

2.1. Neurosky Mindwave Mobile

Neurosky Mindwave Mobile is a brain-computer interface (BCI) device that can be used wirelessly and connected to several devices such as computers and smartphones via Bluetooth connection to collect input from the human brain wave activity [15].

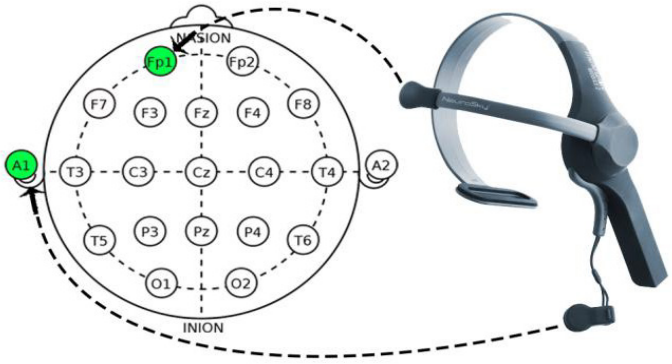


Figure 1: Neurosky Mindwave Mobile and electrode's position

Mindwave provided single-channel EEG in the form of an electrode and placed in the forehead or Fp1 position to measure the brain's electrical voltage and ear-clip on the left ear or A1 position as ground. With Neurosky's ThinkGear and eSense technologies, connected devices can receive a real-time visual representation of brain waves generated from the electrodes [16, 17].

2.2. Subjects

In this study, EEG data were recorded from 50 subjects aged around 20 - 22 years old who were in healthy condition. All of the subjects were undergraduate students and were informed about the purpose of this experiment.

2.3. Stimuli

The use of appropriate stimuli is needed to get better brain wave recording results where a psychologist has validated the stimulus used for this study. Respondents were given stimuli in the form of listening to music, watching videos, and mathematical calculations to influence brain wave activity in the subject to produce an emotional condition that was following the object to be studied [18, 19].

2.4. Procedure of Data Collecting

Before starting data collection, make sure that the respondent is in good health. The data collection process is carried out in a quiet and calm place to make subjects feel comfortable and reduce disturbances affecting data collection. We use the computer laboratory at Ahmad Dahlan University for this research.

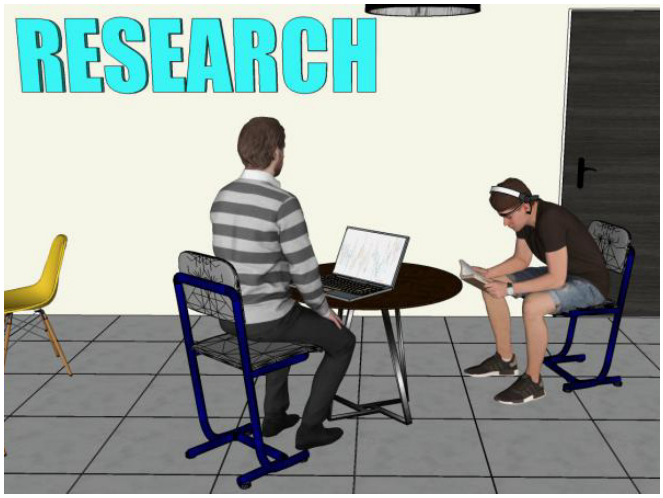


Figure 2: Illustration of the data collecting process

Data collection will be needed as much as 200 data with 50 respondents and carried out four times with a span of at least two days. The recording time for data collection is 7 minutes. When the recording begins, the respondent will be given a stimulus in instrumental music, video, and math questions [20].

3. Methods

3.1. Fast Fourier Transform (FFT)

FFT is applied in various fields,from digital signal processing and solving partial differential equations to algorithms to multiplicate integers in large numbers. FFT is an algorithm for calculating discrete Fourier transforms (DFT) quickly and efficiently. Because many signals in a communication system are continuous in the time domain, we use the Fourier transform to represent the frequency domain [21].

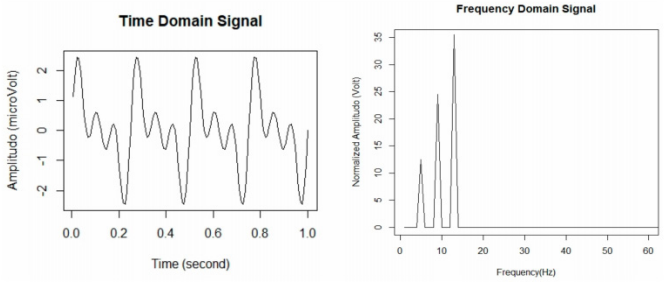


Figure 3: Time-domain signal and frequency domain of EEG signals

FFT is a reliable method for signal processing in the form of a sine wave as EEG signals, and it was proven in previous studies

related to real-time applications because of superior speed compared to other available methods. However, it is not suitable for the analysis of short EEG signals [22]. The FFT formula is defined in (1).

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \quad (1)$$

$e^{-j2\pi ft}$

FFT algorithm is used to reduce the complex transformations performed by DFT. For comparison, when using DFT, then the transformation complexity is N2. Meanwhile, with the FFT algorithm and faster transformation times, the complexity of the transformation is too decreased to (N) log (N).

3.2. K-nearest neighbor

K-Nearest Neighbor is a method for classifying objects based on training data closest to the object. The data is then presented with a distance measure to be processed into a mathematical calculation. In this classification, the calculation is done by using Euclidean Distance. The training data with the closest distance are said to be neighbors and then sorted from the closest distance to the farthest distance. Each neighbor is different from one another or the like. Similar neighbors with the largest number among *k* neighbors correspond to the object being classified [23]. To find the neighbor, it uses distance metrics like euclidean distance as given in equation (2) .

$$D(x, y) = \Sigma_{k=1}^{N} \sqrt{(x_k - y_k)^2} \quad (2)$$

where *D (x, y)* is the distance between *x* and *y* facilities, x_k-y_k is the *x* and y coordinates of variable *k* (*k* = 1,3,5,… n). The value of k must be an odd number, and the number of training data is even to avoid ambiguity. This value is necessary to facilitate the achievement of classification results from the number of closest neighbors. When there is a class with the most neighbors, the test data will get that class result [24].

3.3. Hardware Design

The block diagram above shows that Neurosky Mindwave Mobile is a major component in this hardware system. Neurosky captures input from brain waves via electrodes attached to the FP1 area of the respondent's head and sends an EEG signal via a Bluetooth connection, which acts as a communication protocol between Mindwave and PC / Laptop.



Figure 4: Hardware system diagram

3.4. Software Design

The block diagram above explains that when raw EEG data is obtained from recording brain waves using Neurosky Mindwave on respondents who have been given a stimulus. Raw data taken in the signal pre-processing stage is cut 2 seconds early to minimize noise at the beginning of the recording, due to external interference such as a Bluetooth connection between the recording device and the application, or due to the imperfect installation of the device on the respondent's head. The next stage of feature extraction is carried out using the FFT algorithm to convert the EEG signal in the time domain into the frequency domain and extract all important frequency components from the EEG signal such as Delta waves (0.1 - 3 Hz), Theta (4 - 7 Hz.), Alfa (8 - 12 Hz), Beta (12 - 30 Hz), and Gamma (30 - 100 Hz) [25]. From the results of feature extraction, a huge amount of data is obtained, and data simplification must be done by not including data that is below the average value so that the classification results are more accurate for each emotion. Furthermore, grouping training data and test data are carried out using the KNN classification method in mathematical calculations with the Euclidean formula to calculate the closest distance to the results of the FFT feature extraction. The final step is to determine the accuracy of the KNN classification results on waves using the Confusion Matrix.

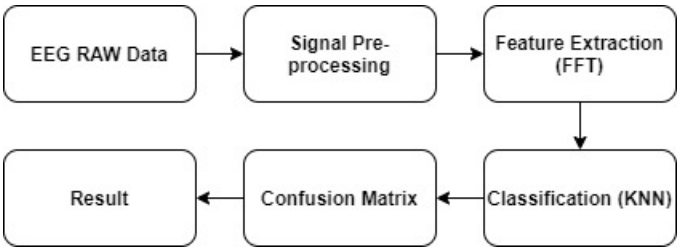


Figure 5: System design diagram

4. Result

In this research, several experimental stages were carried out. The first step is signal pre-processing to reduce noise in raw data, then feature extraction to change the form of RAW data from brain waves to the FFT spectrum. Second, monitoring the FFT results in each emotional state, followed by taking all the highest data from the FFT spectrum with a range of 0.2 Hz. Finally, KNN is used to classify emotions (normal, focused, sad, and shocked) and calculate accuracy using a confusion matrix.

4.1. Feature Extraction of Raw Data

As previously explained, FFT feature extraction with a sampling frequency of 512 Hz is used to convert the time domain signal from raw data to the frequency domain. We use MATLAB2016b software to do this process. The results of running the program can be seen in Figure 6 below.

The picture above shows the feature extraction program results, which produces two different forms of signal output with the same raw data input from subjects who have been given a stimulus in the form of listening to instrumental music. It can be seen that the signal in the time domain has data along with 2.13×10^5 ms with the highest amplitude of 631, which is in the range of 1.32×10^5 ms, while the signal in the frequency domain has data with a range of 0 - 50 Hz where the highest amplitude value is at a frequency of 2.26 Hz. These results will show a different value for each respondent due to different human brain wave activity against the given stimulus.

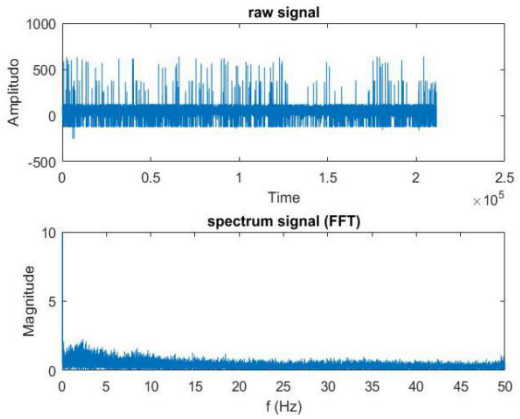


Figure 6: Feature extraction result of raw data

4.2. FFT Spectrum Comparison

In this stage, a comparison of four signal combinations resulted from FFT feature extraction that represented the subject's emotional condition. The results of running programs are shown in Figure 7.

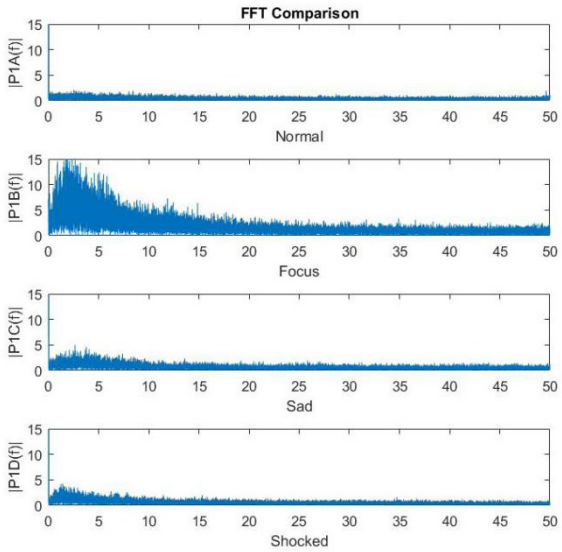


Figure 7: FFT spectrum of all emotional state

It can be seen that the spectrum of the FFT signal in four emotional states, namely Normal, Focus, Sad, and Shocked. When the spectrum is observed, there is a significant difference from the shape of the spectrum wherein the Focused and Normal emotional state, the difference in the form of the spectrum is very recognizable compared to the Sad and Shocked emotional states.

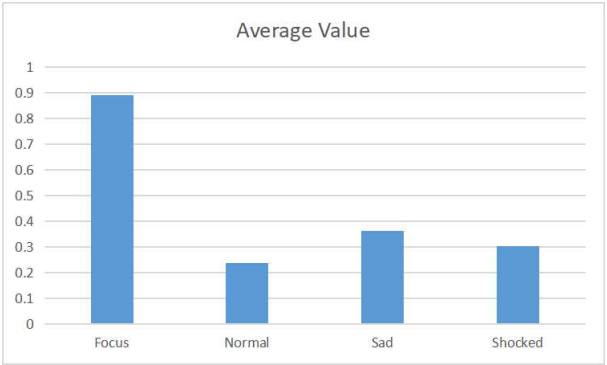


Figure 8: The average value of amplitude in the FFT spectrum

Then the amplitude value is compared to prove the truth of the observations. However, each spectrum with a frequency range of 0 - 50 Hz produces hundreds of thousands of data so that only the average value from data will be taken. Normal emotions get an average score of 0.2386, Focus 0.8904, Sad 0.3625, and Shocked 0.3044. Figure 8 shows the average value of the FFT spectrum.

From a comparison of the average values on each FFT spectrum, focus emotional state has the highest amplitude and lowest for Normal emotions, between Sad and Shocked emotions there is a small difference in the average results so that it can be proven the truth of observation result on the FFT spectrum.

4.3. Collecting Data Based on Spectrum Peaks

A reduction process is carried out at this stage at a frequency of 50 Hz to 10 Hz because the fluctuation of data values in the FFT spectrum always occurs at frequency range 0-10 Hz. Figures 9 to 12 show the results of the reduction process.

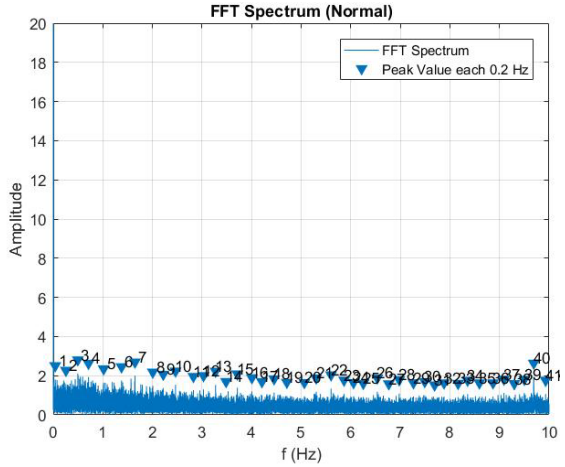


Figure 9: FFT Spectrum of Normal emotion

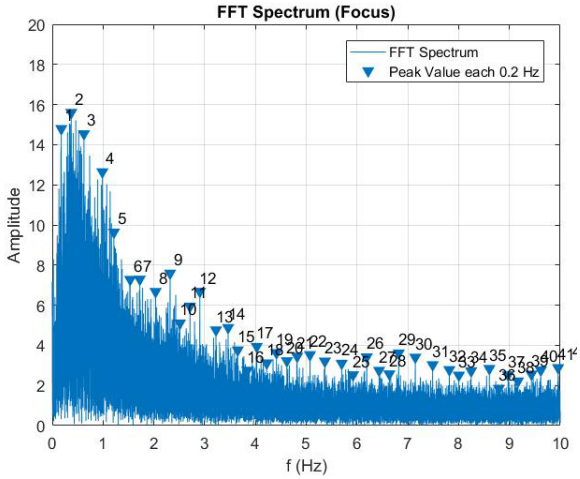


Figure 10: FFT spectrum of Focus emotion

The Values at the highest peaks of the marked frequency spectrum will be collected from the figure above. The frequency in the range 0 - 10 Hz produces 40 data by looking for the highest peak value every 0.2 Hz. Then, from 40 data, each 10 data's average value will be sought to produce four average values used

as material for the classification process. The results of the average value are shown in table 1.

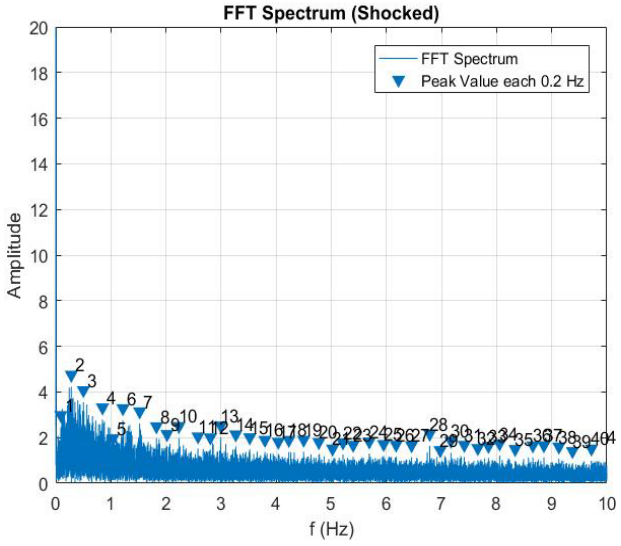


Figure 11: FFT spectrum of Shocked emotion

Table 1: Frequency average value

Focus	Normal	Sad	Shocked
8.891733187	1.908573527	4.199081623	3.059306615
3.371044125	1.2510798	2.052021058	1.633249441
2.851979474	1.124556395	1.583934658	1.250179144
2.597699756	1.202698894	1.554512291	1.136366638

If the reduction process is not carried out in the FFT spectrum, then based on the previous discussion (4.2), it is explained that the frequency in the range 50 Hz produces hundreds of thousands of EEG data and a minimum average value, which can complicate the classification process. We get between the results of previous observations and the current calculations to get very similar results from this process. It can be proven that the results of recording brain waves with the stimulus used to produce good data quality.

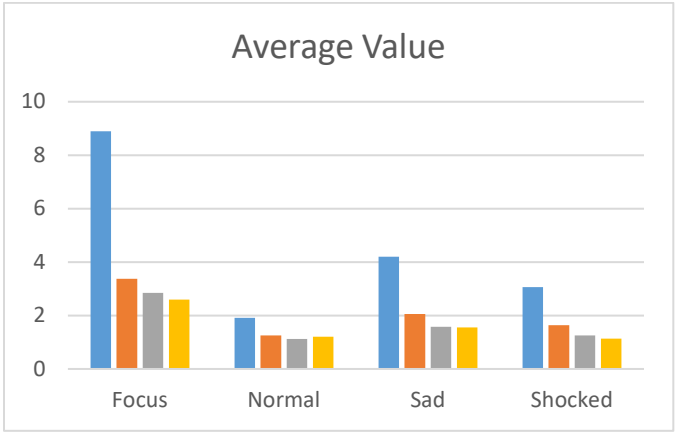


Figure 12: Average value of FFT based on spectrum peaks

4.4. KNN Classifier

In this process, 50 respondents studied and recorded brain waves four times obtained 200 data with 140 training data and 60 testing data. The KNN classification is then carried out to determine the results of the subject's actual emotional state from the 60-testing data.

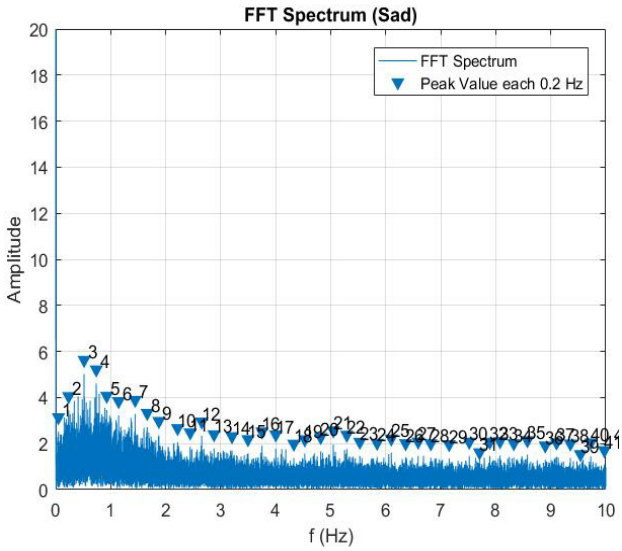


Figure 13: FFT spectrum of Sad emotion

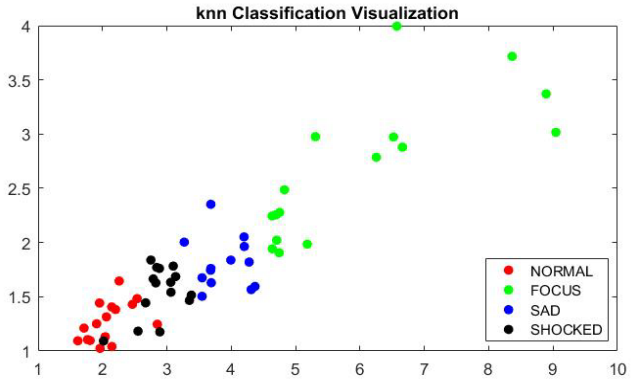


Figure 14: Visualization of KNN classification

The KNN classification results run through MATLAB2016b are displayed in a scatter chart and the coordinate points. It can be seen that each color at the coordinate point is a representation of the emotional condition of the subject, which forms a group based on the value obtained from 60 testing data. If observed, there are several color points whose coordinate values are closer to other neighboring groups. This indicates a change in the status of the color point. more clear results are shown in Table 2 below

Table 2: Result of KNN Classifier

	Match	Error	Match	Error	Match	Error
***k* Values**	**7**		**15**		**23**	
Normal	11	4	**13**	**2**	13	2
Focus	13	2	**14**	**1**	12	3
Sad	9	6	**12**	**3**	11	4
Shocked	12	3	**11**	**4**	12	3
Total	45	15	**50**	**10**	48	12

The table above shows that the KNN classification results using the k = 15 value are the best choice among k = 7 or k = 23. Sixty

test data produces 50 data that corresponds to the original emotional state status, and 10 data does not match. Then perform accuracy testing using the confusion matrix as in table 3 below.

Table 3: Accuracy result using confusion matrix

	Accuracy		
***k* Values**	**7**	**15**	**23**
Normal	73.33%	**86.66%**	86.66%
Focus	86.66%	**93.33%**	80.00%
Sad	60.00%	**80.00%**	73.33%
Shocked	80.00%	**73.33%**	80.00%
Average	74.99%	**83.33%**	79.99%

From the table above, the highest average accuracy is obtained from the results of the KNN classification using a value of k = 15 with a result of 83.33%, where the accuracy value of the focused emotional state gets the highest result of 93.33%, followed by a normal emotional state of 86.66%, 80% sad emotional state, and shocked get the lowest accuracy of 73.33%. Differences in accuracy results using the right stimulus dramatically affects the accuracy obtained. Regulations about sad emotions and focus are needed to help improve the accuracy of the results obtained.

5. Conclusion

This study aims to identify human emotions towards a given stimulus by applying the FFT method as feature extraction from the raw data generated by recording brain waves using the Neurosky Minwave mobile with the KNN method using the variable k = 15 to classify the FFT spectrum. Accuracy in each emotional state studied was performed using a confusion matrix where the best accuracy was obtained for Focus emotions as much as 93.33%, followed by Normal emotions 86.66%, Sad 80%, and Shocked getting the lowest accuracy results of 73.33%. The results of the average accuracy in the classification of brain waves were 83.33%. This research found that FFT and KNN can be used as alternatives in the EEG signal analysis method.

Acknowledgment

The research is funded by Ministry of Research and Technology / National Agency for Research and Innovation Fundamental Research with contract number: 1627.9/LL5/PG/2020

References

[1] A. Yudhana, S. Mukhopadhyay, I.R. Karas, A. Azhari, M.M. Mardhia, S.A. Akbar, A. Muslim, F.I. Ammatulloh, "Recognizing Human Emotion patterns by applying Fast Fourier Transform based on Brainwave Features," Proceedings - 1st International Conference on Informatics, Multimedia, Cyber and Information System, ICIMCIS 2019, 249–254, 2019, doi:10.1109/ICIMCIS48181.2019.8985227. This paper is an extension of work originally presented in International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS).

[2] I. Jayarathne, M. Cohen, S. Amarakeerthi, "BrainID: Development of an EEG-based biometric authentication system," 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016, (October), 2016, doi:10.1109/IEMCON.2016.7746325.

[3] Z. Pang, J. Li, H. Ji, M. Li, "A new approach for EEG feature extraction for detecting error-related potentials," 2016 Progress in Electromagnetics Research Symposium, PIERS 2016 - Proceedings, 3595–3597, 2016, doi:10.1109/PIERS.2016.7735380.

[4] B. Trowbridge, C. Rodriguez, J. Prine, M. Holzemer, J. McCormack, and R. Integlia, "Gaming, fitness, and relaxation," in Games Entertainment Media Conference (GEM), 2015 IEEE, 1–4, 2015.

[5] D. Bright, A. Nair, D. Salvekar, S. Bhisikar, "EEG-based brain controlled prosthetic arm," Conference on Advances in Signal Processing, CASP 2016, 479–483, 2016, doi:10.1109/CASP.2016.7746219.

[6] A. Yudhana, Sunardi, Priyatno, "Development of Door Safety Fingerprint Verification using Neural Network," Journal of Physics: Conference Series, **1373**(1), 2019, doi:10.1088/1742-6596/1373/1/012053.

[7] Y.P. Lin, C.H. Wang, T.P. Jung, T.L. Wu, S.K. Jeng, J.R. Duann, J.H. Chen, "EEG-based emotion recognition in music listening," IEEE Transactions on Biomedical Engineering, **57**(7), 1798–1806, 2010, doi:10.1109/TBME.2010.2048568.

[8] V. Vanitha, P. Krishnan, "Real time stress detection system based on EEG signals," Biomedical Research (India), 2016(Special Issue 2), S271–S275, 2016.

[9] S. Al Irfan, A. Yudhana, S.C. Mukhopadhyay, I.R. Karas, D.E. Wati, I. Puspitasari, "Wireless Communication System for Monitoring Heart Rate in the Detection and Intervention of Emotional Regulation," Proceedings - 1st International Conference on Informatics, Multimedia, Cyber and Information System, ICIMCIS 2019, 243–248, 2019, doi:10.1109/ICIMCIS48181.2019.8985210.

[10] P.D. Purnamasari, P. Yustiana, A.A. Putri Ratna, D. Sudiana, "Mobile EEG Based Drowsiness Detection using K-Nearest Neighbor," 2019 IEEE 10th International Conference on Awareness Science and Technology, ICAST 2019 - Proceedings, (June), 1–5, 2019, doi:10.1109/ICAwST.2019.8923161.

[11] M. Soleymani, S. Asghari-Esfeden, Y. Fu, M. Pantic, "Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection," IEEE Transactions on Affective Computing, **7**(1), 17–28, 2016, doi:10.1109/TAFFC.2015.2436926.

[12] V. Starcevic, D. Berle, K. Viswasam, A. Hannan, D. Milicevic, V. Brakoulias, E. Dale, "Specificity of the Relationships Between Dysphoria and Related Constructs in an Outpatient Sample," Psychiatric Quarterly, **86**(4), 459–469, 2015, doi:10.1007/s11126-015-9344-8.

[13] D. Nie, X.W. Wang, L.C. Shi, B.L. Lu, "EEG-based emotion recognition during watching movies," 2011 5th International IEEE/EMBS Conference on Neural Engineering, NER 2011, 667–670, 2011, doi:10.1109/NER.2011.5910636.

[14] R.A. Surya, A. Fadlil, A. Yudhana, "Identification of Pekalongan Batik Images Using Backpropagation Method," Journal of Physics: Conference Series, 1373(1), 2019, doi:10.1088/1742-6596/1373/1/012049.

[15] E.C. Djamal, M.Y. Abdullah, F. Renaldi, "Brain computer interface game controlling using fast fourier transform and learning vector quantization," Journal of Telecommunication, Electronic and Computer Engineering, **9**(2–5), 71–74, 2017.

[16] S. Sengupta, S. Biswas, S. Nag, S. Sanyal, A. Banerjee, R. Sengupta, D. Ghosh, "Emotion specification from musical stimuli: An EEG study with AFA and DFA," 2017 4th International Conference on Signal Processing and Integrated Networks, SPIN 2017, 596–600, 2017, doi:10.1109/SPIN.2017.8050019.

[17] A. Azhari, L. Hernandez, "Brainwaves feature classification by applying K-means clustering using single-sensor EEG," International Journal of Advances in Intelligent Informatics, **2**(3), 167–173, 2016, doi:10.26555/ijain.v2i3.86.

[18] R. Das, E. Maiorana, P. Campisi, "EEG Biometrics Using Visual Stimuli: A Longitudinal Study," IEEE Signal Processing Letters, **23**(3), 341–345, 2016, doi:10.1109/LSP.2016.2516043.

[19] N. Masood, H. Farooq, "Investigating EEG patterns for dual-stimuli induced human fear emotional state," Sensors (Switzerland), **19**(3), 1–23, 2019, doi:10.3390/s19030522.

[20] E.C. Djamal, D.P. Pangestu, D.A. Dewi, "EEG-Based Recognition of Attention State Using Wavelet and Support Vector Machine," 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA), 139–144, 2016, doi: 10.1109/ISITIA.2016.7828648.

[21] C.S. Huang, C.L. Lin, L.W. Ko, S.Y. Liu, T.P. Su, C.T. Lin, "Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels," Frontiers in Neuroscience, **8**(SEP), 1–12, 2014, doi:10.3389/fnins.2014.00263.

[22] R. Jaswal, "Brain Wave Classification and Feature Extraction of EEG Signal by Using FFT on Lab View," International Research Journal of Engineering and Technology, 1208–1212, 2016.

[23] A.E. Putra, C. Atmaji, F. Ghaleb, "EEG-Based Emotion Classification Using Wavelet Decomposition and K-Nearest Neighbor," Proceedings - 2018 4th International Conference on Science and Technology, ICST 2018, **1**, 1–4, 2018, doi:10.1109/ICSTC.2018.8528652.

[24] S. Shukla, R. Chaurasiya, Emotion Analysis Through EEG and Peripheral Physiological Signals Using KNN Classifier, Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering 2018 (ISMAC-CVB), 2019, doi:10.1007/978-3-030-00665-11

[25] B. Senzio-Savino, M.R. Alsharif, C.E. Gutierrez, K. Yamashita, "Synchronous emotion pattern recognition with a virtual training environment," Proceedings of the 2015 International Conference on Artificial Intelligence, ICAI 2015 – WORLDCOMP, 650–654, 2019.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1089-1095 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

5G, Vehicle to Everything Communication: Opportunities, Constraints and Future Directions

Boughanja Manale*, Tomader Mazri

Advanced Systems Engineering, Electrical Engineering, Networks and Telecommunications System, Ibn Tofail Science University, Kenitra, 14070, Morocco

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 14 November, 2020
Online: 14 December, 2020

Keywords:
5G
V2X
Security
Challenges

ABSTRACT

The 5G, as a new source of telecommunication infrastructure technology, has attracted many stakeholders to promote the progress of its standards and the development of its technology industry. The 5G reinforces new technologies and delivers vehicles for everything services (V2X) to drivers and passengers. It also offers several advantages. On the other hand, due to pervasive network connectivity, there are significant trust, security, and privacy issues for vehicles this affects the overall performance of 5G V2X. This paper provides a comprehensive review; and particularly; an overview of the mobile communication system and introduces the Next Generation (NG). additionally, the paper concentrates on the fifth-generation structure and its use cases. The 5G-V2X as a new field of communication requires more attention and focus, moreover, it aims to present the problems facing this new technology as well, and some future directions for research have been highlighted.

1. Introduction

Telecommunication changed radically in recent years, from traditional wired telephony oriented services to data-based services, from non-shrewd devices to smart handheld and versatile mobile computers.

Mobile ad hoc networks (MANET) formed a network without infrastructure because the mobile nodes in the network dynamically establish a path to transmit data packets. The strength of this network is its rapid development, robustness, flexibility, and mobility. With the rapid growth of this network and the massive focus on data creation through mobile phones and devices. Previous generations become unable to manipulate this data because they achieve their limits. For this purpose, a fifth-generation has been developed. The appearance of 5G has given several advantages to mobile involvement and some new technologies appear with it [1] to enhance communication between devices:

- Millimeter waves
- Small cells
- Massive multi-input multi-output (MIMO)
- Beamforming
- Full-duplex.

The 5G named also next-generation (NG), is emerging the network. Due to their capacity to handle large and diverse sets of devices, the NG is expected to bring a true wireless worldwide web (WWWW) which will have the ability to support massive data rates and a huge number of user applications and services. Moreover, it will reduce latency and will improve energy efficiency.

The 5G aims to create a networked society, where anyone can view and share information anytime from anywhere [2]. It will bring wireless connectivity for everything that is connected. Due to this fact, 5G faces some issues that can be summarized as follows:

- The number of connected devices has increased considerably;
- Traffic has increased significantly;
- Wide variety of applications with various characteristics and demands.

Our goal is to provide a detailed review where we will show the development brought to 5G and present some use cases and application of 5G, especially we will discuss the relation between 5G and V2X communication and application. This paper is an extension of an original work presented at the international conference of computer science and renewable energies (ICCSRE) [3]. In the previous article, we have shown that with the application of 5G the generation of data will increase. For that, it is essential to implement a dedicated storage method to manage this large

*Corresponding Author: Boughanja Manale, Salé,Morocco, 2126 68 81 32 32 & boughnja.manale@gmail.com

quantity of data. In this article, we will focus more on NG itself; explain the difference between 5G and prior generation, and some applications that allow generating this data. For this reason, our paper is organized as follows: In the first section, we will present an overview of the development of mobile networks. Then, we will introduce the 5G new radio access technologies in section two. We will provide some applications for 5G in section four. Further, in section five, we will introduce the 5G V2X application, and we will present the security issues that face the 5G V2X services. To address this security issue, we present some future directions for further studies to promote greater efforts on trust, confidence, assurance, and trustworthiness in 5G V2X services. Finally, we will conclude our paper.

2. Overview

So far, four generations of cellular communication systems have already been adopted [4]. Starting with the 1G that relies on an analog FM cellular system to 4G. With the huge interest in data creation and connected devices, the past generation achieved their limits. To deal with this problem the 5G has been created to solve the problem of the frequency spectrum cutting [5]. The 5G performances will develop later on with the progressive introduction of new technologies: use of millimeter waves, small cell, massive MIMO (or smart micro antennas), beamforming, and full-duplex [6], we will briefly explain those new advances.

- Massive MIMO: the concept was created in [7]. The purpose is to permit the system to achieve higher throughput. The massive MIMO can provide multiple services to various customers simultaneously in all directions.
- Millimetric waves: mobile wireless systems have been always in the band 300MHz - 3GHz. However, the spectrum becomes congested with the huge increase in mobile data traffic and the number of connected devices and the frequency band is coming to an end, creating a signal cutoff [8]. The proposed solution was to use millimeter waves. This latter refers to the wave between 30 and 300 GHz, and their wavelength varies between 1 and 10 mm [9].
- Small cells: are considered to be low-power radio switching equipment controlled by the administrator [10]. We can describe a small cell as a little power base station that will be very nearby to the usual towers and transmit the signal without having any break of connection or attenuation of the signal bypassing all obstacles.
- Beamforming: given the number of antennas used in 5G, the signal will be transmitted from the base station to the devices in an unmonitored manner [11]. This technique permits the orientating of the signal.
- Full duplex (FD): this technique brings to the device the capability to Send and get the data on the identical frequency band. FD improves the efficiency of the network.

The figure 1 summarizes those advances.

3. 5G new radio access technology

5G wireless is conceived to provide a connected environment, where information can be accessed and shared anywhere and anytime, by anyone and deliver a wireless connection for anything. However, with this requirement, the 5G will face several challenges like the volume of traffic the large variety of applications, and so on. To overcome this difficult task, 5G wireless access demands not just new functionality, but also a much more spectrum and larger frequency bands. The millimeter waves provide a huge spectral range, wide transmission bandwidth for the 5G [12]. The figure 2 shows the frequency spectrum.

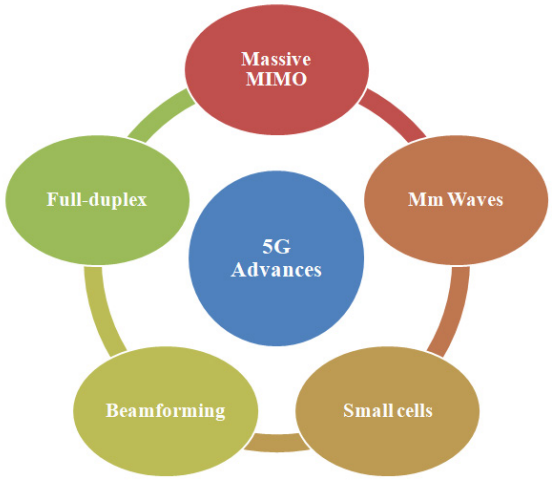


Figure 1: 5G Advances

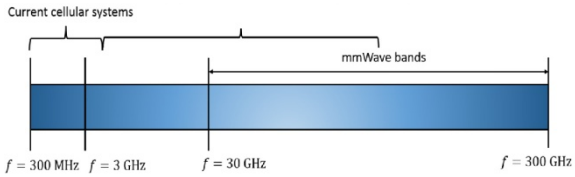


Figure 2: The frequency spectrum of the current and NG communication system

The past generations of cellular mobile systems were operating below 6 GHz frequencies, but with the implementation of the mm waves frequencies, much more spectrum is available and can deal with the huge amount of connected devices With the use of this principle in both 5G and Long Term Evolution (LTE), will have covered to have wireless access [13]. The appearance of 5G did not stop the evolution of LTE. LTE continues its evolution to meet the growing demand for 5G. This evolution is called the LTE evolution [14]. Figure 3 shows the LTE and 5G radio access.

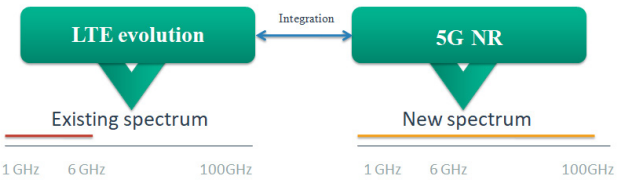


Figure 3: 5G radio access

3.1. Spectrum for 5G

Due to the lack of spectrum in the conventional mobile wireless band regulator started to envisage alternative approaches to manage the spectrum availability. For this reason, one of the major adjustments between the past generation of mobile communication and the fifth generation is the use of a higher frequency spectrum in the millimeter-wave interval [15]. It is about technological evolution and revolution, mixed to the spectrum. This approach is

made to make the spectrum more available with a higher frequency to exploit the frequency above 6 GHz and provide more bandwidth for mobile applications. The regulators have taken some measures for the proper functioning of the new generation among them:

- Spectrum sharing: technologies that will allow wireless broadband applications to operate bands assigned to multiple sites.
- Spectral efficiency: to minimize the bandwidth needed to supply a given quality of service level.
- Advanced MIMO: for higher modulation and lower signaling throughput.
- Densification: 5G communication system aims to get mobile customers closer to base stations, particularly by implementing a dense network of small cells.

All these developments reduce the amount of spectrum required for a single user, which indirectly leads to a greater spectrum availability to serve more users with an increased Quality of Services (QoS).

3.2. Use cases for 5G

5G will need to bring solutions for a variety of equipment running within a completely heterogeneous environment. The new human-centric applications include several fields: augmented reality, virtual reality, and gaming. This requires an extremely higher flow rate and reduced delay. In other words, we can differentiate two segments for machine-model communication:

- The Massive Internet of things (MIoT): which is categorized by a large number of low-cost device connections, in this case, a small amount of data supported by each device with extended battery lifetime and a wide range of coverage. This type is used more in the field of intelligent building, transport logistics, and agriculture.
- The critical internet of things (CIoT): shaped with highly dependable, extremely low-latency connectivity. For example, it can support autonomous vehicles, robotic surgery, and road safety.

Unlike previous products, 5G aims to bring improved support for a diversity of distinct services. 5G target three sets of use cases with very unique features:

- Enhanced Mobile Broadband (eMBB): This kind of use case is a direct extension of the conventional mobile broadband connection scenario based on the current mobile telecommunication standards. This set of boxes addresses broadband access everywhere, and in highly inhabited areas. This later provides the user with multiple services such as multimedia and 3D video [16].
- Large scale machine to machine communication (mMTC): with high latency and stability demands. This type of use case mainly relates machine to machine communication (MTC). It enables mobile communication with vehicles, this type requires simultaneous broadband and high mobility capacity. The problem with the mMTC that requires increased reliability and decreased latency [17].
- Ultra-Reliable Low Latency Communication (URLLC): the development of IoT is leading to a multiplication of wireless attached devices carrying TCM network traffic. It is intended that the expected number of these devices will soon surpass. The goal of mMTC is to provide the ability to connect a large set of equipment that may occasionally send a small quantity of data. mMTC devices are categorized by having long battery life and can be deployed remotely [16].

3.3. The key feature of 5G

The 5G was defined by the 3rd generation partnership project (3GPP) as a collection of key functionalities with associated parameters. In addition to the technological and protocolling environment and architecture inherited from the 4G system. We can summarize those technologies as follows:

- 3D MIMO: to ensure high performance in terms of spectral effectiveness. MIMO was firstly implemented in the LTE system was improved than the third generation [18]. The old system was composed of traditional antennas and composed of 8 to 10 elements stacked in a column, but with the evolution, restraint adjustments have been made and provide a downward tilt angle for network convergence. In this situation, the identical radio frequency (RF) pulses are delivered to everyone on the antenna. In such a manner, with the implementation of 3D MIMO, the signal will be simultaneously monitored in both the horizontal and vertical levels.
- Software-Defined Air Interface (IA): the fifth generation intended to handle various use scenarios and a broad range of the spectrum with a signal AI. For this reason, a unified framework was defined to handle different usage scenarios.
- User-Centric Network: to ensure the most suitable operator environment on the entire network, based on user requirements and the network conditions. Therefore, to provide a better user-centric experience the 5G network should handle a user context awareness where the network can serve the user more effectively when there is more information or expertise which means that the more information is loaded onto the network, the better the customer is served. Also, it should have the MEC functionality that includes traffic, security, transport isolation computing, and storage resources. Last and not least it should have a core network perspective to plan the use of the base stations.
- SDN/NFV: This technology enables the management and coordination of various network tranches on a shared physical infrastructure [19].

4. 5G applications

4.1. 5G and Mobile Ad Hoc Network

Mobile Ad Hoc Network (MANET) refers to a signal type of cellular network that contains mobile wireless nodes for communication [20]. These nodes are dynamically organized in random and volatile topologies. The utility to use the 5G for wireless communication in the field of MANET is to give more efficiency to the network. With the implementation of the NG, the MANET network achieves high levels of flexibility and intelligence. Figure 4 illustrates an example of MANET using the 5G connection.

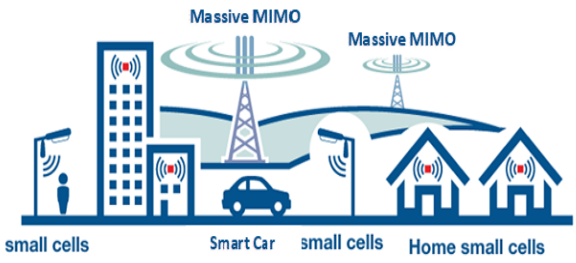


Figure 4: An example of MANET applied in 5G

The NG allows the connection of several devices and provides communication between multiple types of equipment. This communication can be classified into two different parts the device to device communication (D2D) and the vehicle for everything communication (V2X).

D2D communication: also known as proximity service communication (ProSe). The main objective of this functionality is to allow mobile devices to explore the presence of any other close devices within a radius of 500 meters and establish direct communication with it. The basic D2D procedures consist of three main steps:

- Device detection (ProSe discovery): detects if there are other devices nearby.
- Link parameters: establish a link between the associated devices.
- Data communication (ProSe communication): send or receive data via a set of links.

Figure 5 shows the D2D communication.

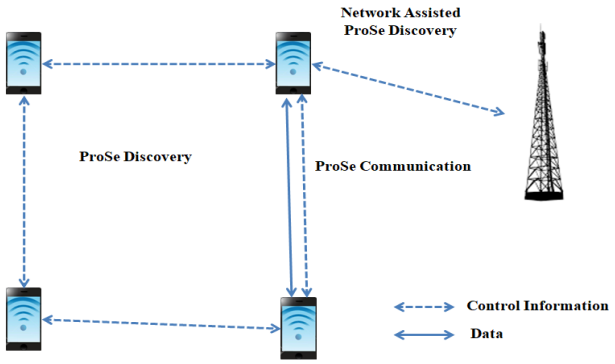


Figure 5: Basic ProSe procedure

V2X communication: are the exchanges between the vehicle and any device that may impact the vehicle, and the opposite. The X can for instance be on a different vehicle (V), an area of infrastructure (A/I), a pedestrian (P), a domicile (D), or any entity that communicates with the vehicle. We can summarize the cases of use of the vehicle with X as follows:

- Vehicular internet: user experience similar to that provided by their home and network.
- Pre-collision detection and mitigation: detect impending collisions and exchange relevant data from the vehicles involved.
- Cooperative vehicles: V2V and V2I communication to safely operate vehicles such as a train of autonomous cars on a highway.
- V2V information exchange: allows the vehicle to communicate on road to avoid any congestion.

5. 5G and V2X

The 5G mobile communication system will continue to emerge to meet unprecedented requirements beyond the capabilities of the past generations. The NG will not only be used for human interaction but also will be implemented in-vehicle communication. Therefore, the 5G has to support:

- Connection density: to handle million of demands and links across the square kilometer.
- Continuous latency from end to end.
- The density of road traffic volume: several dozens of terabytes created each second per square km.
- Mobility and maximum data rate.
- Network efficiency and high network capacity.

Continuous improvement and the increase in the number of applications have led to the development of new usages, in particular, that of connected objects. Vehicular communications or V2X communications allow vehicles to communicate in real-time with their environment to enhance road control, traffic performance, and energy savings. Figure 6 illustrates the different modes of communication, grouped under the term V2X.

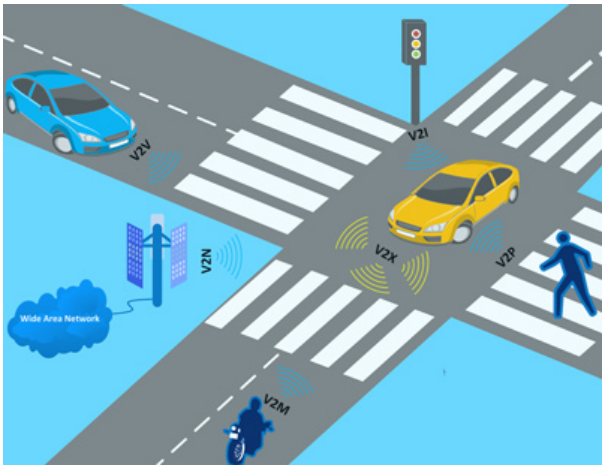


Figure 6: V2X communication modes

Embedded 5G will reform the experience of users in the transport sector (vehicles, subways, planes, trains, etc.), in particular by developing internet connectivity, which will make it possible to create an environment connected to the vehicle. V2X (Vehicle to Everything) communication technology will enable enormous changes the almost instantaneous exchange can be summarized as follows:

- Vehicle-to-vehicle (V2V): that permit inter-vehicle communication.
- Vehicle to infrastructure (V2I): it another type of communication that allows the vehicle to communicate with

its surrounding like the communication between the roadside unit and the vehicle.

- Vehicle to pedestrian (V2P); that permits the exchange between the vehicles and the pedestrian into the road.
- Vehicle to motorcycle (V2M): Communication between vehicles and motorcycles to decrease the high rate of collisions and deaths in motorcycle crashes.

The concept of a connected vehicle is based on sensors that operate in line-of-sight such as radars, cameras, or laser remote sensing. The 5G will play a key role in making road transport more efficient and safe. Indeed, the automotive industry will require a more robust, stable, and low-latency network.

5.1. 5G V2X architecture

The structure of the fifth generation and vehicle to everything architecture is based on four network levels:

- 5G Access Network;
- Network Edge;
- 5G Core Network;
- Data Network.

The 5G Access Network is composed of a Next Generation Radio Access Network (NG-RAN) or/and a non 3GPP access network, which interconnects the 5G Core Network and the User Equipment (UE). Figure 6 shows the 5G V2X architecture:

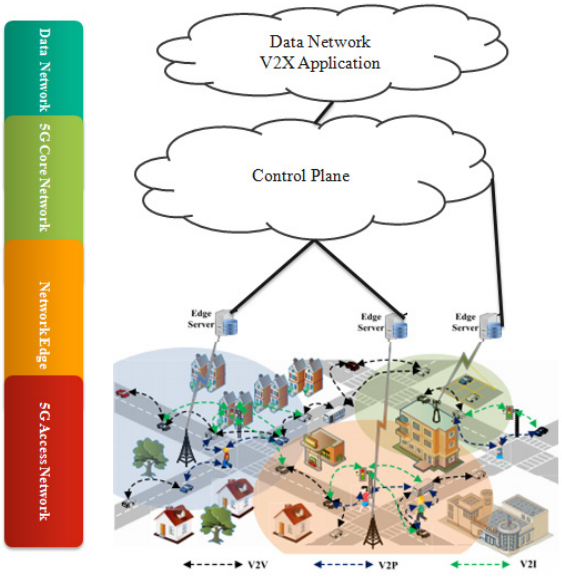


Figure 7: 5G V2X architecture

5.2. Security issues in 5G V2X services: problems and risks

Even though 5G guidelines are as yet being defined it will confront a few difficulties the moment of its implementation. The new radio (NR) will require a steady improvement in security. In this specified situation, portable security and oversaw security services are one of the most escalated areas regarding advancement, to react to the difficulties in the field of digital security inborn to computerized change for the communication service providers and their clients. Moreover, 5G vows to reshape services and advance development by providing an expanded transfer speed, multitenant edge processing, and a few creative services, high-data transmission networks, and high-bandwidth connectivity to machine learning and artificial intelligence (AI). The insurance of services in 5G V2X requires ground-breaking, versatile, and coordinated security. The implementation of 5G V2X brings with it several security threats. The attack on the vehicle can lead to serious problems and can affect all sides of the security requirement

- Attacks on availability: the attacks that focus on the availability of data are the riskiest in V2X communication since they have a real impact on the basic circumstances of well-being. We can find attacks such as the jamming attack which is a radio interference attack that can be carried out across any communication based on wireless technology.
- Attack on data integrity: where the attacker intercepts the data circulated into the network. the attacker can launch attacks such as GPS spoofing or even replay attacks.
- Attack on confidentiality: as an important security requirement in V2X communication and all sent data in the V2X should be secured. An attack such as eavesdropping and location tracking are launched in the network to break through the V2X confidentiality
- Attack on authenticity: Assuring authenticity consists of the process to verify that nodes are granted access to the network according to their identities using certificates and digital signatures. Some threats that infringed this requirement are certificate replication where the nodes use replicated certificates that were added to the blacklist. Moreover, the Sybil attacks where one node (the attacker node) pretends to be many identities in the road network.
- Attack on the non-repudiation: Non-repudiation is concerned with the ID of the effective node that is performing a particular behavior.

The table below shows some attacks that impact the security requirement:

Table 1: Threats of security for 5G V2X

Security Requirement	Attack	External	Internal
availability	Blackhole Grey hole	✓	×
	Flooding	✓	×
	Jamming	×	×
	Coalition	✓	×
data integrity	Messages Alteration	✓	×
	False messages injection	✓	×
	Replay	✓	✓
	GPS Spoofing	×	×
confidentiality	Eavesdropping	✓	×

	Location tracking	×	×
authenticity	Certificate replication	✓	✓
	Sybil	✓	✓
non-repudiation	Masquerading impersonation	✓	✓

5.3. Security solutions for 5G V2X communication

Vehicular communication experiences numerous difficulties, hence, guaranteeing safe communication is viewed as a complicated process due to the lack of safety features in both empowering innovations of the vehicle to everything interchanges, an additive safety scheme is required to secure vehicle networks. The fields of security for 5G V2X was not deeply discussed. In this section, we will try to gather the mean full strategy that has been implemented to secure the 5G V2X communication. In [21], they proposed a lightweight and protected navigation system for VANET. Each vehicle implements a cryptographic and digital signature to manage the secured correspondence process from the vehicle to the roadside unit (RSU). The system dealt with the replay attack and the man-in-the-middle attack. Another solution was proposed by [22], consisting of a trusted and smart forwarding protocol using double ciphering for data packets and implemented strong verification schemes to gauge the trust of the node. In [23], they proposed a new technique depends on a one-time Identity-Based confirmed asymmetric gathering key consent to make cryptographic blend areas that oppose malicious Eavesdroppers. There are several types of research have proposed potential solutions for 5G from a machine learning (ML) perspective. In [24], the authors suggested a monitoring system employing a Support Vector Machine (SVM) learning algorithm to prevent the replay attack. Also in [25] an anti-jamming methodology for auxiliary users to choose the correspondence channel and mobility. Another researcher [26], based on deciding the arrangements of conceivable interfacing neighboring vehicles, and design the boundaries reserving the joint V2V network. In [27], they have introduced a trusted and standard secure signature solution for collecting plain text for vehicle system applications. The authors in [28], have suggested an algorithm for VANET that delivers authenticate delivered messages. He gave oversight roles to some specified nodes referred to as verification nodes. The authors in [29], suggested a model framework as an IDS based on compartmental attacks. it addressed various attacks in VANET such as the black hole, wormhole, and Sybil attacks.

6. Discussion

This article presents the state of the art on 5G technology applied to transportation to make roads safer. The security of 5G V2X communication presents a great challenge, for this reason, a security layer should be added to strengthen security, and the communication must respect the CIA principle (confidentiality, integrity, and availability):

- Confidentiality concerns the condition under which the vehicle user has the right to know what information is sent by the On-Board Unit (OBU) and for how long this information exists.
- Integrity to verify that all data has been transferred.
- Availability, to ensure that the protocols must remain operational even in the event of a failure, and remains fault-tolerant and stabilize once the failure has been eliminated.

In addition to two other aspects: privacy, this technique implies the preservation of the privacy of the driver and the vehicle. Also, the authentication which consists of ensuring that the intended recipient receives messages from a real sender.

As we have already presented, many applications and solutions have been implemented to ensure the security of 5G V2X communication to preserve and maintain the security of users. We can summarize these in three categories:

- Identity-Based Cryptography (IBC): The concept of IBC is a cryptographic scheme very suitable for devices with limited computing resources. It has been presented in many articles, to encrypt and effectively sign the messages interchanged.
- Behavior-Based: The concept is quite simple this solution is based on the behavior of the attached vehicles communicating in the network and the messages transmitted by the RSU.
- Machine Learning-Based (ML): This provides threat and attack prediction based on suspicious network activities and anomaly detection. The implementation of ML enables automation, adaptability, and enables efficient orchestration and dynamic provisioning for scaled network resources cognition and prediction of mobility

Last and not least the security of such a network can be done depending on four parameters:

- Arrangement: shows if the model is mounted on a flat or cluster network.
- Structure: specifies if the model has used structured, distributed, or hybrid.
- Purpose: specify if the goal of the system is an IDS, the implementation of a safe route, or the protection of the message content.
- Targeted attacks: specifies the attacks handled in the model.

7. Open issues and future orientations

Open issues and future perspectives were highlighted in this section. Securing the 5G V2X is a quite difficult challenge, For this reason, both security solutions and possible cures for identified and unidentified menaces rely on 5G-V2X implementation approaches. We can divide two approaches for 5G V2X. The first one the V2X is implemented with non-autonomous-5G (NA-5G), possible assaults on LTE-V2X are real and can leverage 5G-V2X capabilities. However, with the second implementation Autonomous-5G (A-5G), the attack window diminishes and safeguarding from threats can be augmented. Network deployment plays a critical factor in the 5G-V2X security policy decision. for this purpose, we have encountered the security needs at two levels in the 5G V2X field:

- Secure network caching at the network edge: Storing commonly demanded data in the cache at the edges of storage is an important and effective way to decrease the delay of V2X operations. Yet, it is difficult to maintain the privacy of stored and buffered content, from which opponents can determine which are the relevant services for a given target. To ensure

network caching is secure, these key issues need to be resolved: knowing when and where to cache the desired data for a given vehicle, choosing trusted peripheral devices for data storage, and ensuring the confidentiality of buffered data when peripheral nodes are under attack. For this reason, cache placement could be well planned.

- Security-enhanced network slicing: The security and confidentiality of the grid segmentation have not yet gained a lot of attention. Therefore, it is essential to perform both authentication and authorization for a specified part of a network. For this reason, it is crucial to guarantee authentication with low latency to V2X services.

8. Conclusion and perspectives

Five generation and V2X services have expanded tremendously and have benefited our day-to-day lives. While providing a range of advantages to conductors and passengers on the road, 5G V2X's connectivity has as well raised several concerns about security, trust, and privacy. In this article, we have presented the fifth generation and its characteristics. Besides, we have identified the key challenges that face this new generation. We also briefly outlined the main issues in securing the 5G V2X by determining its issues of reliability, safety, and privacy. Finally, we brought some open issues and future orientations in this field. This latter should have more attention in the future our perspective is to build a system that will be able of securing the communication between the vehicle and everything basing on the 5G connection.

References

[1] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," IEEE Communications Magazine, **52**(5), 44–51, 2014, doi:10.1109/MCOM.2014.6815892.

[2] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, "5G radio access," Ericsson Review (English Edition), **91**(1), 42–47, 2014.

[3] B. Manale, M. Tomader, "A survey of different storage methods for NGN mobile networks," in Proceedings of 2019 International Conference of Computer Science and Renewable Energies, ICCSRE 2019, 2019, doi:10.1109/ICCSRE.2019.8807526.

[4] M. Iwamura, "NGMN view on 5G architecture," IEEE Vehicular Technology Conference, **2015**, 2015, doi:10.1109/VTCSpring.2015.7145953.

[5] T.L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," IEEE Transactions on Wireless Communications, **9**(11), 3590–3600, 2010, doi:10.1109/TWC.2010.092810.091092.

[6] A. Gupta, R.K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," IEEE Access, **3**, 1206–1232, 2015, doi:10.1109/ACCESS.2015.2461602.

[7] S. SEN, J. CLARK, J. TAPIADOR, Security Threats in Mobile Ad Hoc Networks, 2010, doi:10.1201/ebk1439819197-9.

[8] M.Á. Vázquez, L. Blanco, A. Pérez-Neira, "Sparse beamforming for mmWave spectrum sharing systems," European Signal Processing Conference, **2018-Septe**(1), 1517–1521, 2018, doi:10.23919/EUSIPCO.2018.8553148.

[9] E. Al Abbas, A.T. Mobashsher, A. Abbosh, "Polarization reconfigurable antenna for 5G cellular networks operating at millimeter waves," Asia-Pacific Microwave Conference Proceedings, APMC, 772–774, 2017, doi:10.1109/APMC.2017.8251562.

[10] M. Tayyab, X. Gelabert, R. Jantti, "A simulation study on handover in LTE ultra-small cell deployment: A 5G challenge," IEEE 5G World Forum, 5GWF 2019 - Conference Proceedings, 388–392, 2019, doi:10.1109/5GWF.2019.8911616.

[11] T. Varum, A. Ramos, J.N. Matos, "Planar microstrip series-fed array for 5G applications with beamforming capabilities," 2018 IEEE MTT-S International Microwave Workshop Series on 5G Hardware and System Technologies, IMWS-5G 2018, 2018, doi:10.1109/IMWS-5G.2018.8484697.

[12] E. Dahlman, S. Parkvall, "NR - The new 5G radio-access technology," IEEE Vehicular Technology Conference, **2018-June**, 1–6, 2018, doi:10.1109/VTCSpring.2018.8417851.

[13] X. Lin, A. Grovlen, K. Werner, J. Li, R. Baldemair, J.F.T. Cheng, S. Parkvall, D.C. Larsson, H. Koorapaty, M. Frenne, S. Falahati, "5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology," IEEE Communications Standards Magazine, **3**(3), 30–37, 2019, doi:10.1109/MCOMSTD.001.1800036.

[14] O. Teyeb, G. Wikström, M. Stattin, T. Cheng, S. Faxér, H. Do, "Evolving LTE to Fit the 5G Future," Ericsson Technology Review, 1–17, 2017.

[15] S. Parkvall, E. Dahlman, A. Furuskär, M. Frenne, "WIRELESS AND RADIO COMMUNICATION NR : The New 5G Radio Access Technology," IEEE Communications Standards Magazine, **1**(December), 24–30, 2017, doi:10.1109/MCOMSTD.2017.1700042.

[16] M.A. Abu-Rgheff, 5G Physical Layer Technologies, 2019, doi:10.1002/9781119525547.

[17] B. Badic, C. Drewes, I. Karls, M. Mueck, Rolling out 5G: Use cases, applications, and technology solutions, 2016, doi:10.1007/978-1-4842-1506-7.

[18] G. Liu, Y. Huang, F. Wang, J. Liu, Q. Wang, "5G features from operation perspective and fundamental performance validation by field trial," China Communications, **15**(11), 51–61, 2018, doi:10.1109/CC.2018.8543047.

[19] J. Ordonez-Lucena, P. Ameigeiras, Di. Lopez, J.J. Ramos-Munoz, J. Lorca, J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," IEEE Communications Magazine, **55**(5), 80–87, 2017, doi:10.1109/MCOM.2017.1600935.

[20] R. Bruzgiene, L. Narbutaite, T. Adomkus, "MANET Network in Internet of Things System," Ad Hoc Networks, (May 2017), 2017, doi:10.5772/66408.

[21] G. Li, M. Ma, C. Liu, Y. Shu, "A lightweight secure VANET-based navigation system," 2015 IEEE Global Communications Conference, GLOBECOM 2015, 2015, doi:10.1109/GLOCOM.2014.7417462.

[22] S.K. Bhoi, P.M. Khilar, "A Road Selection Based Routing Protocol for Vehicular Ad Hoc Network," Wireless Personal Communications, **83**(4), 2463–2483, 2015, doi:10.1007/s11277-015-2540-x.

[23] L. Zhang, "OTIBAAGKA: A New Security Tool for Cryptographic Mix-Zone Establishment in Vehicular Ad Hoc Networks," IEEE Transactions on Information Forensics and Security, **12**(12), 2998–3010, 2017, doi:10.1109/TIFS.2017.2730479.

[24] Q.G. Fan, L. Wang, Y.N. Cai, Y.Q. Li, J. Chen, "VANET Routing Replay Attack Detection Research Based on SVM," MATEC Web of Conferences, **63**, 4–7, 2016, doi:10.1051/matecconf/20166305020.

[25] G. Han, L. Xiao, H.V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2087–2091, 2017, doi:10.1109/ICASSP.2017.7952524.

[26] L.T. Tan, R.Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," IEEE Transactions on Vehicular Technology, **67**(11), 10190–10203, 2018, doi:10.1109/TVT.2018.2867191.

[27] M.S.I. Mamun, A. Miyaji, H. Takada, "A multi-purpose group signature for vehicular network security," Proceedings - 2014 International Conference on Network-Based Information Systems, NBiS 2014, 511–516, 2014, doi:10.1109/NBiS.2014.93.

[28] U. Khan, S. Agrawal, S. Silakari, "Detection of Malicious Nodes (DMN) in vehicular ad-hoc networks," Procedia Computer Science, **46**(Icict 2014), 965–972, 2015, doi:10.1016/j.procs.2015.01.006.

[29] H. Sedjelmaci, S.M. Senouci, "An accurate and efficient collaborative intrusion detection framework to secure vehicular networks," Computers and Electrical Engineering, **43**, 33–47, 2015, doi:10.1016/j.compeleceng.2015.02.018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1093-1098 (2020)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Posture Recognition Method for Caregivers during Postural Change of a Patient on a Bed using Wearable Sensors

Kodai Kitagawa[1,*], Koji Matsumoto[1], Kensuke Iwanaga[1], Siti Anom Ahmad[2], Takayuki Nagasaki[3], Sota Nakano[4], Mitsumasa Hida[1,5], Shogo Okamatsu[1,6], Chikamune Wada[1]

[1]*Graduate School of Life Science and System Engineering, Kyushu Institute of Technology, Kitakyushu, 8080196, Japan*

[2]*Malaysian Research Institute of Ageing (MyAgeing™), Universiti Putra Malaysia, Serdang, Selangor, 43400, Malaysia*

[3]*Department of Rehabilitation, Tohoku Bunka Gakuen University, Sendai, 9818550, Japan*

[4]*Department of Rehabilitation, Kyushu University of Nursing and Social Welfare, Tamana, 8650062, Japan*

[5]*Department of Physical Therapy, Osaka Kawasaki Rehabilitation University, Kaizuka, 5970104, Japan*

[6]*Department of Physical Therapy, Kitakyushu Rehabilitation College, Kanda, 8000343, Japan*

ARTICLE INFO

Article history:
Received: 09 September, 2020
Accepted: 10 October, 2020
Online: 20 October, 2020

Keywords:
Posture recognition
Caregiver
Patient handling
Lower back pain
Wearable sensor
Machine learning

ABSTRACT

Caregivers experience lower back pain due to their awkward postures while handling patients. Therefore, a monitoring system to supervise caregivers' postures using wearable sensors is being developed. This study proposed a postural recognition method for caregivers during postural change while handling a patient on a bed. The proposed method recognizes foot positions and arm movements by a machine learning algorithm using inertial data on the trunk and foot pressure data obtained from wearable sensors. An experiment was conducted to evaluate whether the proposed method could recognize three foot positions and three arm movements. Participants provided postural change for a simulated patient on a bed (patient: supine to lateral recumbent) under nine conditions, including different combinations of the three foot positions and three arm movements; the experiment was repeated ten times for each condition. Experimental results showed that the proposed method using an artificial neural network with all features obtained from an inertial measurement unit and insole pressure sensors could recognize arm movements and foot positions with an accuracy of approximately 0.75 and 0.97, respectively. These results suggest that the proposed method can be used in a monitoring system tracking a caregiver's posture.

1. Introduction

Recently, many caregivers have experienced lower back pain owing to frequent patient handling [1–2]. Previous studies have suggested that caregivers should maintain an appropriate posture based on body mechanics during patient handling [3–4]. Accordingly, previous studies have developed monitoring systems to supervise the postures of caregivers to prevent lower back pain [5–8].

Lin et al. developed a robot patient comprising an inertial measurement unit (IMU), angular position sensors, and servo motor encoders and found that the robot could classify correct and incorrect transfer methods [5]. Huang et al. developed an automatic evaluation method for patient transfer skills using two Kinect RGB-D sensors [6, 7]. Their study showed that this method could improve the patient transfer skill of students [7]. Itami et al. developed a monitoring system using electromyography, goniometer, and inclination sensor [8]. This system could monitor skills and lumbar loads of patient handling [8]. These systems are useful for monitoring caregivers' postures; however, these systems have limitations in terms of workspace and usability because they

*Corresponding Author: Kodai Kitagawa, Kyushu Institute of Technology, Japan, kitagawakitagawa156@gmail.com

require vision-based devices, multiple specialized sensors, or robots. Previous study reported that assistive devices for caregivers should be time-efficient, cost-effective, and suitable for various workspaces [9]. Hence, there is a need to develop wearables and simple monitoring systems for caregivers' postures.

PostureCoach is a wearable and simple system for preventing lower back pain among caregivers [10]. This system can provide real-time biofeedback using the trunk angle obtained from only two IMUs [10]. Previous investigations have shown that PostureCoach can reduce the lumbar spine flexion in a beginner who handles patients [10]. However, this system cannot suggest caregiver ways to adjust to an appropriate posture because the system measures only the trunk angle. Several factors, such as the foot position and arm movement, are useful to achieve suitable postures to prevent lower back pain during patient handling [11–13]. Therefore, a simple and wearable system that monitors the trunk angle, foot position, and arm movement is being develop [14].

Figure 1 shows the flow of the proposed monitoring system for caregivers' postures. The proposed system monitors a caregiver's posture using wearable sensors mounted on the caregiver's trunk and shoes (Step 1). The proposed method evaluates several factors such as trunk angle, foot positions, and arm movements using sensor data obtained from the wearable sensors (Step 2). If the evaluated factors are unsuitable, the proposed system informs the caregiver to adjust their posture (Step 3).

Our previous study proposed a postural recognition method during sit-to-stand assistive motions but could not consider other patient-handling aspects and various foot positions and arm movements [14]. Several patient-handling aspects, such as providing postural change to turn a patient on a bed, cause lumber load among caregivers [12, 15]. Such a type of patient handling causes lower back pain among caregivers as a patient is frequently repositioned on a bed [15]. As explained above, arm movements and foot positions are important factors for patient handling [11-13]. Thus, this study proposes the recognition method for various foot positions and arm movements of caregivers using a postural recognition method during postural change for a patient on a bed.

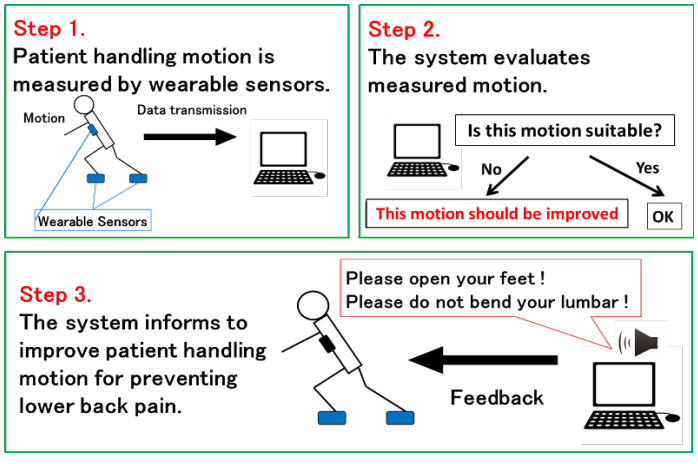


Figure 1: Monitoring system for patient-handling motion.

This section discusses the background, previous studies of this paper. The remainder of this paper is organized as follows. Section 2 mentions the objective and contribution of this paper. Section 3 explains the proposed method for postural recognition during postural change while handling a patient on a bed. Section 4 describes the experiments conducted to evaluate whether the proposed method can recognize a caregiver's posture during patient handling. Section 5 presents experimental results and discussions. Finally, section 6 presents conclusions.

2. Objective and Contribution

The objective of this study is proposal of a novel posture recognition method for wearable monitoring system to prevent lower back pain among caregivers. Furthermore, this study explores appropriate combination of wearable sensors and machine learning algorithms for the proposed method.

The contribution of this study is as follows. The proposed method will be applied to suggest an appropriate posture for caregivers because this method recognizes both arm movements and foot positions. This study presents the novel posture recognition method using sensor fusion consisted of IMU and insole pressure sensor. As a scientific contribution, this method using sensor fusion of wearable sensor has a potential to be applied to posture recognition related to occupational health. In addition, the appropriate combination of wearable sensors and machine learning algorithms shown in this study will be useful in the research area of computer science and wearable systems.

3. Proposed Method

3.1. Overview

Figure 2 shows a block diagram of the proposed postural recognition method. The proposed method measured postural information by wearable sensors. Machine learning algorithm using features obtained from wearable sensors generated function for postural recognition. Caregiver's posture such as arm movement and foot position were recognized by generated function. The details of each component are described below.

3.2. Input (wearable sensors)

The IMU attached on the trunk and insole pressure sensors measured postural information during patient handling. The IMU (Logical Product Co., Japan) measures the three-axis acceleration and angular velocity of the trunk for features of the machine learning algorithm. Insole pressure sensors are fabricated by arranging eight FlexiForce sensors (Tekscan, USA) on each foot. Because the FlexiForce sensors are thin and flexible, they are suitable for measuring forces between various surfaces, such as an insole [16]. Furthermore, the FlexiForce sensors can be used in real time applications owing to the lack of linearity, non-repeatability, and hysteresis [17]. Front and rear forces on each foot are measured using insole pressure sensors as features for the machine learning algorithm. These sensor data were measured at a 1-kHz sampling rate.

3.3. Machine learning-based recognition

Seven features (mean, maximum, minimum, standard deviation, root mean square, kurtosis and skewness) were extracted for machine learning algorithm. These features were extracted from each sensor data using a sliding window technique with a 1.0-s window size and 50% overlap. These features, window

size, and overlap were selected based on the previous researches about posture recognition using wearable sensors [18–19].

The machine learning algorithm recognized three arm movements and three foot positions during postural change while handling a patient on a bed. In this study, artificial neural network, decision tree, and support vector machine were selected as machine learning algorithm for the proposed method because these three algorithms were used for wearable sensor-based posture recognition [19-20]. Details of the recognition models were described below.

Artificial neural network is applied to complex relationship between input and output by nonlinear and flexible decision boundary [19-21]. The artificial neural network-based model consists of input layer, hidden layer and output layer [19-20]. In our proposed method, the features obtained from the wearable sensors were inputted to the input layer, and the recognized posture was outputted from the output layer. Hidden layer was set function for learning the relationship between features obtained wearable sensor and postures. The specifications and parameters in artificial neural network of proposed method is shown in Table 1.

Decision tree provides if-then rules which have free of ambiguity [22-24]. If-then rules generated by decision tree consist of root, internal and leaf nodes [22]. The root and internal nodes have threshold of features for recognition. Position and features of these nodes are determined based on comparison of entropy for all features [23]. Leaf nodes are defined as recognition output via root and internal nodes. When these nodes are too many and complex, recognition model has problem such as over fitting [24]. Thus, decision tree has several parameters to adjust number of nodes [24]. The specifications and parameters of decision tree in our proposed method is shown in Table 2.

Support vector machine is applied to nonlinear data without over fitting by large margin separation and kernel function [25]. Kernel function provides efficient calculation in nonlinear feature spaces [25]. Large margin separation is required for generalization performance of support vector machine [25]. The specifications and parameters of support vector machine in our proposed method is shown in Table 3.

Comparison of these machine learning algorithms was necessary in order to determine the appropriate algorithm depends on the target postures [19–20]. In this study, the proposed method was compared to investigate which algorithm could recognize patient handling motion using these wearable sensors.

Table 1: Specifications and parameters for artificial neural network of the proposed method.

Specification / Parameter	Status / Value	
Number of Node / Neuron	Input Layer	70 (based on features)
	Hidden Layer	70
	Output Layer	3 (based on postures)
Activation Function	Hidden Layer	Sigmoid
	Output Layer	Linear
Learning Rate	0.3	
Momentum	0.2	
Training	Back Propagation	

3.4. Output (recognized posture)

In the proposed method, arm movements and foot positions were recognized as postural factors. The monitoring system shown in previous section will provide caregiver the appropriate posture based on recognition results.

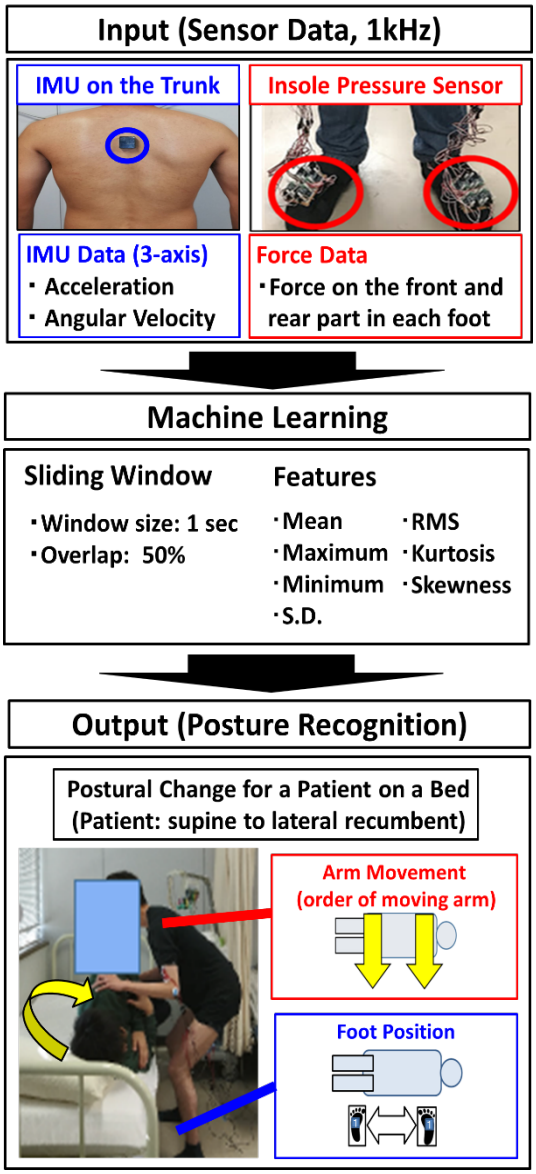


Figure 2: Proposed postural recognition method.

Table 2: Specifications and parameters for decision tree of the proposed method.

Specification / Parameter	Status / Value
Algorithm	J48 (entropy-based)
Confidence Factor for Pruning	0.25
Minimum Number of Samples per Leaf Node	2
Number of Folds for Data Segmentation	3

Table 3: Specifications and parameters for support vector machine of the proposed method.

Specification / Parameter	Status / Value
Training	Sequential Minimal Optimization
Kernel	Polynomial Kernel
c (weight for slack variable)	1.0

4. Experiment

Four young healthy males (24.75 ± 0.83 years, 1.72 ± 0.05 m, 67.00 ± 10.79 kg) were participants as caregivers. One young healthy male (25.00 years, 1.69 m, 70.00 kg) participated as a simulated patient. All participants provided their verbal informed consent to the experiment.

The four participants (caregivers) were asked to perform a postural change for the simulated patient on a bed under nine postural conditions. The nine conditions were a combination of the three arm movements and three foot positions shown in Figure 3. Each participant (caregiver) repeated this motion ten times for each condition. Data for the proposed method were measured by wearable sensors (IMU and insole pressure sensors) at 1-kHz sampling frequency for each trial.

The proposed method recognized the three arm movements and three foot positions by processing data obtained from the sensors using the machine learning algorithm. This experiment compared the recognition performance of three machine learning algorithms to determine the appropriate algorithm for the proposed method. Furthermore, the recognition performances of three feature patterns ("IMU," "Insole Sensor," and "All") were compared to verify the effectiveness of the sensors and features.

The metrics of accuracy and F-measure were used to evaluate the recognition performance of the algorithms. These evaluation metrics were calculated according to

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

where TP denotes true positive and represents the number of correct recognition for positive samples, TN denotes true negative and represents the number of correct recognition for negative samples, FP denotes false positive represents the number of samples incorrectly recognized as positive in actual negative samples, and FN denotes false negative and represents the number of samples incorrectly recognized as negative in actual positive samples. The accuracy of the proposed method was evaluated based on the recognition performance of each algorithm using the feature patterns. The F-measure of each arm movement and foot position was calculated using the harmonic mean of precision and recall as the overall performance. These metrics were calculated by ten-fold cross validation using data obtained from all trials and participants.

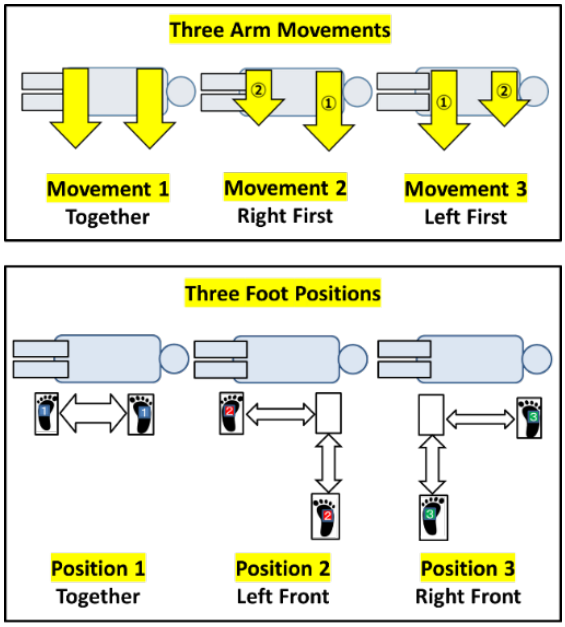


Figure 3: Postural conditions in the experiment.

5. Results and Discussion

Tables 4 and 5 show the accuracy and F-measure of the proposed method. The results show that the combination of artificial neural network and all features obtained from both IMU and insole pressure sensors achieved the best performance for postural recognition. In this combination, the proposed method could recognize arm movements and foot positions with an accuracy of approximately 0.75 and 0.97, respectively. These results suggest the possibility of using the proposed recognition method in monitoring systems to prevent lower back pain.

A comparison between the machine leaning algorithms showed that the artificial neural network exhibited the best performance for both accuracy and F-measure. Therefore, the artificial neural network was deemed the most suitable machine learning algorithm for the proposed method. The artificial neural network's best performance was attributed to its nonlinear and flexible decision boundary, which was effective for wearable sensor data [19-21].

A comparison between feature patterns showed that the proposed method yielded the best performance for accuracy and F-measure when all features obtained from both IMU and insole pressure sensors were used. These results indicated that both IMU and insole pressure sensors were effective and necessary for the proposed method. Moreover, in foot position recognition, the proposed method using only insole pressure sensors achieved a great performance. These findings indicate that the insole pressure sensors could detect the force distribution depending on the foot positions. However, the proposed method could not recognize arm movements using only the insole pressure sensors. Thus, IMU was also considered necessary for the proposed method. Table 6 shows the confusion matrices of the proposed method using the artificial neural network and all features for postural recognition. These results showed that the proposed method achieved good performance for foot position recognition. The reason for this

performance, as explained earlier, was that insole pressure sensors in the shoes could detect relevant features related to foot positions. However, the proposed method incorrectly recognized arm movement between movements 1 and 3 because there was no sensor on the upper limbs. Placing additional sensors on the upper limbs could obstruct care activities. Thus, the optimal position to place an IMU on the trunk to detect arm movement without additional sensors should be explored.

The limitation of this study is that the proposed method was not evaluated in an actual workspace and under various aspects of patient handling, such as assisting transfer. Moreover, the proposed method should be tested for female participants, patients of different ages, and caregivers. Furthermore, other feature selection patterns and algorithms should be examined to improve the performance of the proposed postural recognition method. Following further tests and improvements on the proposed method, the postural monitoring system to prevent lower back pain will be implemented based on the proposed postural recognition method.

Table 4: Accuracy of the proposed method.

Algorithm	Feature Pattern	Accuracy	
		Arm Movement	Foot Position
Artificial Neural Network	IMU	0.691	0.693
	Insole Sensor	0.508	0.971
	All	**0.755**	**0.979**
Decision Tree	IMU	0.616	0.635
	Insole Sensor	0.634	0.938
	All	0.645	0.930
Support Vector Machine	IMU	0.586	0.591
	Insole Sensor	0.364	0.958
	All	0.599	0.972

Table 5: F-measure of the proposed method.

Algorithm	Feature Pattern	F-measure					
		Arm Movement			Foot Position		
		1	2	3	1	2	3
Artificial Neural Network	IMU	0.637	0.760	0.674	0.675	0.629	0.776
	Insole Sensor	0.486	0.539	0.497	0.959	0.976	0.980
	All	**0.707**	**0.813**	**0.742**	**0.974**	**0.978**	**0.985**
Decision Tree	IMU	0.575	0.695	0.571	0.627	0.587	0.694
	Insole Sensor	0.595	0.681	0.622	0.915	0.954	0.945
	All	0.632	0.682	0.619	0.902	0.951	0.938
Support Vector Machine	IMU	0.593	0.646	0.510	0.617	0.438	0.700
	Insole Sensor	0.325	0.429	0.309	0.942	0.970	0.964
	All	0.591	0.653	0.552	0.962	0.975	0.980

In future work, the optimal placement of sensors will be explored to improve the accuracy of postural recognition in the proposed method. Moreover, other machine learning algorithms will be tested to determine a more suitable algorithm for the proposed method. Furthermore, a modified version of the proposed recognition method should be evaluated under various conditions. Finally, a monitoring system will be implemented based on the modified postural recognition method to prevent lower back pain due to patient handling.

Table 6: Confusion matrices of posture recognition (using an artificial neural network with all features).

(A) Arm Movement

		Recognized		
		Movement 1	Movement 2	Movement 3
Actual	Movement 1	259	42	72
	Movement 2	40	326	26
	Movement 3	63	35	277

(B) Foot Position

		Recognized		
		Position 1	Position 2	Position 3
Actual	Position 1	387	6	3
	Position 2	7	364	1
	Position 3	5	2	365

6. Conclusion

In this study, the posture recognition method for caregivers during postural change for a patient on a bed was proposed. The experimental results showed that the proposed method could recognize a caregiver's posture. The proposed method using novel sensor fusion with machine learning will also be useful for posture recognition in other various application.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The first author is grateful for a scholarship from the Nakatani Foundation for Advancement of Measuring Technologies in Biomedical Engineering.

References

[1] S. Kai, "Consideration of low back pain in health and welfare workers," Journal of Physical Therapy Science, **13**(2), 149–152, 2001, doi:10.1589/jpts.13.149.

[2] A. Holtermann, T. Clausen, M.B. Jørgensen, A. Burdorf, L.L. Andersen, "Patient handling and risk for developing persistent low-back pain among female healthcare workers," Scandinavian Journal of Work, Environment & Health, **39**(2), 164–169, 2013, doi:10.5271/sjweh.3329.

[3] R. Ibrahim, O.E.A.E. Elsaay, "The effect of body mechanics training program for intensive care nurses in reducing low back pain," IOSR Journal of Nursing and Health Science, **4**(5), 81–96, 2015, doi:10.9790/1959-04548196.

[4] A. Karahan, N. Bayraktar, "Determination of the usage of body mechanics in clinical settings and the occurrence of low back pain in nurses," International Journal of Nursing Studies, **41**(1), 67–75, 2004,

doi:10.1016/s0020-7489(03)00083-x.
[5] C. Lin, T. Ogata, Z. Zhong, M. Kanai-Pak, J. Maeda, Y. Kitajima, M. Nakamura, N. Kuwahara, J. Ota, "Development and Validation of Robot Patient Equipped with an Inertial Measurement Unit and Angular Position Sensors to Evaluate Transfer Skills of Nurses," International Journal of Social Robotics, 1–19, 2020, doi:10.1007/s12369-020-00673-6.
[6] Z. Huang, A. Nagata, M. Kanai-Pak, J. Maeda, Y. Kitajima, M. Nakamura, K. Aida, N. Kuwahara, T. Ogata, J. Otha, "Automatic evaluation of trainee nurses' patient transfer skills using multiple kinect sensors," IEICE TRANSACTIONS on Information and Systems, **97**(1), 107–118, 2014, doi:10.1587/transinf.E97.D.107.
[7] Z. Huang, A. Nagata, M. Kanai-Pak, J. Maeda, Y. Kitajima, M. Nakamura, K. Aida, N. Kuwahara, T. Ogata, J. Otha, "Self-help training system for nursing students to learn patient transfer skills," IEEE Transactions on Learning Technologies, **7**(4), 319–332, 2014, doi: 10.1109/TLT.2014.2331252.
[8] K. Itami, T. Yasuda, Y. Otsuki, M. Ishibashi, T. Maesako, "Development of a checking system for body mechanics focusing on the angle of forward leaning during bed making," Educational Technology Research, **33**(1–2), 63–71, 2010, doi: 10.15077/etr.KJ00006713267.
[9] S. Sivakanthan, E. Blaauw, M. Greenhalgh, A.M. Koontz, R. Vegter, R.A. Cooper, "Person transfer assist systems: a literature review. Disability and Rehabilitation," Assistive Technology, 1–10, 2019, doi: 10.1080/17483107.2019.1673833.
[10] M. Owlia, M. Kamachi, T. Dutta, "Reducing lumbar spine flexion using real-time biofeedback during patient handling tasks," Work, **66**(1), 41–51, 2020, doi: 10.3233/WOR-203149.
[11] K. Kitagawa, T. Nagasaki, S. Nakano, M. Hida, S. Okamatsu, C. Wada, "Optimal foot-position of caregiver based on muscle activity of lower back and lower limb while providing sit-to-stand support," Journal of Physical Therapy Science, **32**(8), 534–540, 2020, doi:10.1589/jpts.32.534.
[12] B. Schibye, A.F. Hansen, C.T. Hye-Knudsen, M. Essendrop, M. Böcher, J. Skotte, "Biomechanical analysis of the effect of changing patient-handling technique," Applied Ergonomics, **34**(2), 115–123, 2003, doi: 10.1016/S0003-6870(03)00003-6.
[13] Q. An, J. Nakagawa, J. Yasuda, W. Wen, H. Yamakawa, A. Yamashita, H. Asama, "Skill Extraction from Nursing Care Service Using Sliding Sheet," International Journal of Automation Technology, **12**(4), 533–541, 2018, doi:10.20965/ijat.2018.p0533.
[14] K. Kitagawa, T. Uezono, T. Nagasaki, S. Nakano, C. Wada, "Classification Method of Assistance Motions for Standing-up with Different Foot Anteroposterior Positions using Wearable Sensors," in 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT2018), 1–3, 2018, doi:10.1109/ICT-ROBOT.2018.8549912.
[15] H. Wardell, "Reduction of injuries associated with patient handling," Aaohn Journal, **55**(10), 407–412, 2007, doi:10.1177/216507990705501003.
[16] A.M. Almassri, W.Z. Wan Hasan, S.A. Ahmad, A.J. Ishak, A.M. Ghazali, D.N. Talib, C. Wada, "Pressure sensor: state of the art, design, and application for robotic hand," Journal of Sensors, **2015**, 846487, 2015, doi:10.1155/2015/846487.
[17] A.M. Almassri, W.Z. Wan Hasan, C. Wada, "Evaluation of a Commercial Force Sensor for Real Time Applications," ICIC Express Letters, Part B: Applications, **11**(5), 421–426, 2020, doi:10.24507/icicelb.11.05.421.
[18] S. Pirttikangas, K. Fujinami, T. Nakajima, "Feature selection and activity recognition from wearable sensors," in 2006 International Symposium on Ubiquitous Computing Systems (UCS2006), 516–527, doi:10.1007/11890348_39.
[19] M.F. Antwi-Afari, H. Li, Y. Yu, L. Kong, "Wearable insole pressure system for automated detection and classification of awkward working postures in construction workers," Automation in Construction, **96**, 433–441, 2018, doi:10.1016/j.autcon.2018.10.004.
[20] M.F. Antwi-Afari, H. Li, J. Seo, A.Y.L. Wong, "Automated detection and classification of construction workers' loss of balance events using wearable insole pressure sensors," Automation in Construction, **96**, 189–199, 2018, doi:10.1016/j.autcon.2018.09.010.
[21] S. Dreiseitl, L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," Journal of Biomedical Informatics, **35**(5–6), 352–359, 2002, doi:10.1016/S1532-0464(03)00034-0.
[22] Y.Y. Song, L.U. Ying, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, **27**(2):130-135, 2015, doi:10.11919/j.issn.1002-0829.215044.
[23] Y. Zhang, S. Lu, X. Zhou, M. Yang, L. Wu, B. Liu, P. Phillips, D. Wang, "Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine," Simulation, **92**(9), 861-871, 2016, doi:10.1177/0037549716666962.
[24] T. Wang, Z. Qin, Z. Jin, S. Zhang, "Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing, and pruning," Journal of Systems and Software. **83**(7), 1137-1147, 2010, doi: 10.1016/j.jss.2010.01.002.
[25] A. Ben-Hur, C.S. Ong, S. Sonnenburg, B. Schölkopf, G. Rätsch, "Support vector machines and kernels for computational biology," PLoS Comput Biol, **4**(10), p.e1000173, 2008, doi: 10.1371/journal.pcbi.1000173.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 1096-1102 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Development of Secondary Processing Data Methods under Single Point Thunderstorm Activity Monitoring

Anatoly Panyukov[*,1], Alexander Bogushov[2]

[1]*School of Electronic Engineering and Computer Science, South Ural State University, 454080, Russian Federation*

[2]*Institute of Natural Sciences and Mathematics, South Ural State University, 454080, Russian Federation*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 25 November, 2020
Online: 14 December, 2020

Keywords:
single-point lightning sensor
pre-storm radiation
equivalent electric dipole
thunderstorm simulation
passive monitoring
visualization

ABSTRACT

Forecasting the occurrence and development of thunderstorms near insured mobile objects, such as aircraft carriers, oil tankers, requires the usage of appropriate single-point monitoring systems (SPMS) of thunderstorm activity. These systems must be small and included in the equipment of the objects. A SPMS based on EH location finder meets the requirements of small dimensions. But the primary processing algorithms for them used to determine the location of the radiation source lead to errors in the calculation of the bearing per atmospheric discharge. This inevitably occurs due to the presence of an anomalous component of the magnetic field of an arbitrarily oriented dipole. The error depends on the orientation of the equivalent dipole of the radiation source. If the dipole moment is vertical, the error is minimal, otherwise the direction finding error is completely determined by the horizontal projection of the equivalent dipole and can be equal to π radians. Information of a thunderstorm activity collected by SPMS can be used by secondary processing algorithms to identify a real location of the thunderstorm foci. It is especially important when processing pre-storm radiation which consist of intra- and inter-cloud discharges. Secondary processing algorithms reduce errors by analyzing the parameters of all registered lightning discharges over a certain time interval. The proposed algorithms divide the entire space into a set of three-dimensional cells and determine the their grade of membership to a thundercloud or its projection. The presented methods for analyzing the location parameters of radiation sources defined across the entire set of received signals and ways to display them reduce the uncertainty of determining the coordinates of the thunderstorm location. Processing of pre-thunderstorm radiation is essential for predicting the development of thunderstorms.

1 Introduction

Predicting the occurrence and development of a thunderstorm near the insured object requires the development of appropriate single-point systems for monitoring thunderstorm activity. For mobile insured objects, such as aircraft carriers, oil tankers, these systems must be small and included in the equipment of the objects [1]. Primary processing algorithms which used to determine location of the radiation source can lead to errors in calculation the bearing of discharge. It inevitably occurs in a single-point monitoring systems (SPMS). The error depends on the orientation of the equivalent dipole of the radiation source. If the dipole moment is vertical then the error is minimal, otherwise the direction finding error is completely determined by the horizontal projection of the equivalent dipole and can be equal to π radians.

Information of a thunderstorm activity collected by a SPMS can be used by secondary processing algorithms to identify a real location of the thunderstorm foci. It is particularly critical when processing pre-storm radiation which consist of intra- and inter-cloud discharges. Secondary processing algorithms can reduce the uncertainty of determining the coordinates of the thunderstorm location by analyzing the location parameters of radiation sources that found for a time interval. This approach has been announced at "2019 Russian Open Conference on Radio Wave Propagation (RWP), Kazan, Russia" [2] in the field of study that deals with electromagnetic methods of lightning detection. This paper is an original extension of this work.

Each cloud and ground lightning flash consist of many individual physical processes. These processes can be associated with

[*]Corresponding Author: Anatoly Panyukov, Email: anatoly.panyukov@gmail.com

characteristic electric and magnetic fields. Lightning emits significant electromagnetic energy in the radio-frequency range from below 1 Hz to near 300 MHz. This electromagnetic radiation can be detected by lightning locating systems. Lightning data have utility in different areas of importance meteorological service organizations. Typically, meteorological organizations use this data to protect life and property. Commercial organizations use lightning data to provide weather warnings, forecasts to aviation, agriculture, energy, mass media.

There are plenty of lightning locating systems that uses different radio-frequency ranges, for example, the very low-frequency (VLF) range from 3 to 30 kHz, the low-frequency (LF) range from 30 to 300 kHz. Accurate lightning locating systems necessarily employ multiple sensors. But single-point passive monitoring system are also popular as they are inexpensive, mobile and do not require communication. Many companies are engaged in the production of such devices [3]–[6].

A challenging problem which arises in this domain is processing pre-thunderstorm radiation (that is before the first lightning flash). During the project ISTC 1822 [7] were recorded 2.5 million atmospherics and less than 10% were defined as a lightning discharge, the rest radiation were classified as pre-thunderstorm radiation. Algorithms used in multiple sensors systems require registration of the atmospherics by several sensors, it is not always possible. Otherwise, single-point systems generally filtered out this data or it processed incorrectly.

Algorithms to determine location of the radiation source uses the magnetic direction finding (MDF) technique to calculate the bearing of discharge. It works really well for multistation systems especially with combination with time of arrival (TOA) technique. But using MDF technique in SPMS leads to significant errors for the discharges whose equivalent dipole has an orientation different from vertical.

In general, the results of the project showed possibility of creating a new generation of passive thunderstorm detection systems to extend their functionality. One way to overcome these problems is not only revision of the technical characteristics of sensors, but also development of new mathematical models and algorithms for analyzing thunderstorm activity. The problem of identifying the position of thunderstorm foci is practically important to forecasting the development of thunderstorm. The registration of the cloud radiation by a single-point lightning protection system, for the purposes of forecasting thunderstorm development, can be reached by expansion of receiving hardware dynamic range and development of mathematical and software.

The standard model for the lightning discharge is based on a vertical dipole that cannot adequately represent all discharges. Modern methods of analyzing the electromagnetic radiation source (EMRS) field allow to determine the parameters of the EMRS location by many algorithms based on the dipole as a model of the EMRS [1], [8]–[10]. A more correct approach is to simulate a lightning discharge as a arbitrarily oriented dipole and it possible by using special a group of $E-H$ methods that allow arbitrary orientation of dipoles [11, 12]. These algorithms allow us to estimate the position parameters of this dipole: the distance to the radiation source and the pseudo bearing, and can be used for processing pre-thunderstorm radiation. In this case, signals from various sources of electromagnetic radiation are registered, the appropriate mathematical model of which is an arbitrarily oriented electric dipole. Processing signals from dipoles with a predominant horizontal projection leads to uncertainty in the estimation of the bearing on the [2, 12] discharge, which significantly reduces the quality of dipoles location detection. One way to reduce this uncertainty is to analyze the information received from all dipole sources.

The aim of the work is describing approach for the secondary processing of monitoring results of thunderstorm activity in VLF range, allowing to get plausible estimates of the location of a thunderstorm center based on its pre-storm radiation.

A review of the status of passive storm monitoring systems and the demonstration of the use of lightning-position detection systems for the passive radar of hazardous meteorological phenomena are presented in [8]. Modern methods for analyzing the field, allowing to determine the parameters of the EMRS, characterizing its location and orientation are presented in the works [12].

2 Primary processing of EMRS signal

Figure 1 shows the geometry of the problem: the earth's surface is plane Oxy, the observation point is origin O, the polar coordinates of the radiation source are r, ψ, θ , and the angular coordinates of the radiation source dipole moment are ψ_0, θ_0.

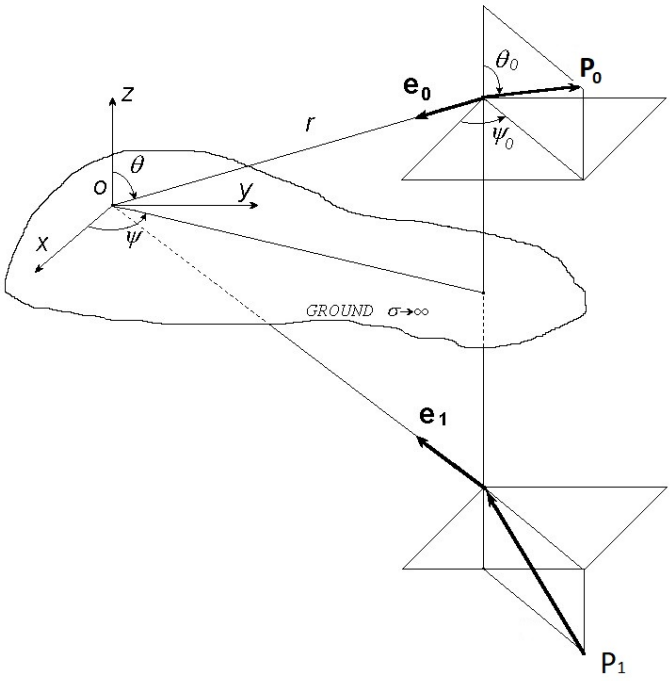


Figure 1: Geometry of the problem

There are a dipole $\mathbf{P_0}$ and an imaginary dipole $\mathbf{P_1}$ which is the mirror image of the dipole $\mathbf{P_0}$. $\mathbf{P_0}$ results a field $(\mathbf{E}, \mathbf{H})$ consisting of the field ($\mathbf{E}(\mathbf{P_0})$, $\mathbf{H}(\mathbf{P_0})$) of the source $\mathbf{P_0}$, and the reflected field represented by the field ($\mathbf{E}(\mathbf{P_1})$, $\mathbf{H}(\mathbf{P_1})$) of the $\mathbf{P_1}$.

The electric $\mathbf{E}$ and magnetic $\mathbf{H}$ components of the field in terms of the Cartesian system of coordinates with origin at the observation

point O and the O_z axis being the normal of the bounding plane have representation [9]

$$E_x = E_y = H_z = 0, \tag{1}$$

$$E_z = \frac{1}{2\pi\epsilon_0 c^2 r}\left\{p''[\cos\theta\sin\theta\sin\theta_0\cos(\psi-\psi_0) - \sin^2\theta\cos\theta_0] + \right.$$
$$\left. +(p'\alpha + p\alpha^2)[3\sin\theta\cos\theta\sin\theta_0\cos(\psi-\psi_0)]\right\} =$$
$$= \frac{w}{2\pi\epsilon_0 c^2 r}\left[p''u + (p'\alpha + p\alpha^2)(3u-2v)\right], \tag{2}$$

$$H_x = \frac{1}{2\pi c^2 r}(\sin\theta\cos\theta_0\sin\psi - \cos\theta\sin\theta_0\sin\psi_0)\,(p'' + p'\alpha) =$$
$$= \frac{w\sin\varphi}{2\pi c^2 r}\,(p'' + p'\alpha), \tag{3}$$

$$H_y = \frac{1}{2\pi c^2 r}(-\sin\theta\cos\theta_0\cos\psi + \cos\theta\sin\theta_0\cos\psi_0)(p'\alpha + p'') =$$
$$= -\frac{w\cos\varphi}{2\pi c^2 r}(p'\alpha + p''), \tag{4}$$

$$\alpha = \frac{c}{r}, \quad u = \sin\theta\cos(\phi-\psi), \quad v = \frac{\sin(\phi-\psi_0)}{\sin(\psi-\psi_0)}, \tag{5}$$

$$\tan\varphi = \frac{\sin\theta\cos\theta_0\sin\psi - \cos\theta\sin\theta_0\sin\psi_0}{\sin\theta\cos\theta_0\cos\psi - \cos\theta\sin\theta_0\cos\psi_0}, \tag{6}$$

c is speed of light in vacuum, ϵ_0 is permittivity of vacuum.

The values of the parameters r, φ, u, v are measured as a result of primary processing, for each pulse signal received from a radiation source [9]. The quality of estimating the location of an individual source is quite low. Really, a vertical dipole (that is when $\theta_0 = 0$) have

$$\varphi = \psi, \quad \sin\theta = u = \frac{1}{v}.$$

In other cases (that is for an inclined dipole) there is an uncertainty in determining the coordinates ψ and θ. As shown in the article in [9] they are bounded by

$$\arcsin u \le \theta \le \frac{\pi}{2}, \tag{7}$$

$$\varphi - \arccos u \le \psi \le \varphi + \arccos u. \tag{8}$$

The probability density for angles ψ and θ are

$$f_\psi(\eta) = \frac{s}{2\widehat{\psi}\,[s^2\cos^2(\varphi-\eta) + \sin^2(\varphi-\eta)]},$$
$$-\arccos u \le \eta - \varphi \le \arccos u, \tag{9}$$

$$f_\theta(\zeta) = \frac{su\cos\zeta\sin\zeta}{\widehat{\psi}\,(s^2u^2 - u^2 + \sin^2\zeta)\cdot\sqrt{\sin^2\zeta - u^2}},$$
$$\arcsin u \le \zeta \le \frac{\pi}{2}. \tag{10}$$

where

$$s = \left|\frac{1-vu}{uv}\right|, \quad \widehat{\psi} = \arctan\left(\frac{1}{s}\tan\arccos u\right). \tag{11}$$

It's easy to see that for $s < 1$, that is for θ_0 approaching the vertical,

$$\varphi = \arg\max_{\eta:\,-\arccos u\le\eta-\varphi\le\arccos u} f_\psi(\eta), \tag{12}$$

$$\arcsin u = \arg\max_{\zeta:\arcsin u\le\zeta\le\frac{\pi}{2}} f_\theta(\zeta), \tag{13}$$

in other words, the maximum likelihood estimates are the same as for the vertical dipole: $\psi = \varphi$ and $\theta = \arcsin u$; and for $s \ge 1$, that is for orientation with a significant horizontal component, poor angular resolution is gained.

An adequate mathematical model of the pre-thunderstorm EMRS is an inclined electric dipole with a predominance of the horizontal component. Processing signals from dipoles with a predominant horizontal projection leads to uncertainty in the estimation of the angle coordinates of the [2, 9] discharge, which significantly reduces the quality of dipoles location detection.

The multipoint system of passive monitoring of thunderstorm activity not only removes the uncertainty of location [9], but also allows to determine the orientation of the dipole [13, 14]. Nevertheless, the improvement of forecasting methods using single-point systems remains relevant [15, 16].

One way to reduce the uncertainty of thunderstorm location is to analyze the information received from all dipole sources for time T.

The purpose of secondary processing is to estimate the location of the thunderstorm center by information about all dipoles location parameters for time T.

3 Secondary processing of EMRS signals from thunderstorm foci

It is known that the Cartesian coordinates x, y of the possible position of the EMRS satisfy the condition [9]:

$$x\cos\varphi + y\sin\varphi = r\sin\theta\cos(\varphi-\psi) = ru, \tag{14}$$

which determines the OXY projection of the EMRS possible positions set. The coordinates of the possible location EMRS belong to the set of solutions of the system of equations

$$x\cos\varphi + y\sin\varphi = ru, \qquad x^2 + y^2 + z^2 = r^2. \tag{15}$$

Let S be set of all solutions of (15) system. In general, S is multiple.

Entire space divided into $4N^3$ cells:

$$C(i,j,k) = \left\{ (x,y,z) \;\middle|\; \begin{array}{l} i\Delta \le x \le (i+1)\Delta \\ j\Delta \le y \le (j+1)\Delta \\ k\Delta \le z \le (k+1)\Delta \end{array} \right\}, \quad k = 0,\,1,\,\ldots,\,N,$$
$$i,\,j = -N,\,-N+1,\ldots,-1,\,0,\,1,\,\ldots N. \tag{16}$$

The OXY plane is divided into flat cells

$$C(i,j) = \{(x,y):\; i\Delta \le x \le (i+1)\Delta,\; \Delta \le y \le (j+1)\Delta,$$
$$i,\,j = -N,\,-N+1,\ldots,-1,\,0,\,1,\,\ldots N\}, \tag{17}$$

and OZ axis of the coordinate system $OXYZ$ to be divided into segments

$$C(k) = \{z:\; k\Delta \le z \le (k+1)\Delta\}, \quad k = 0,\,1,\,\ldots,\,M.$$

In the future we will use the following notation:

$$C_{OXYZ}(r,u,\varphi) = \{(i,j,k) \mid C(i,j,k) \cap S \neq \emptyset\}$$

is set of cells in which the EMRS can be locate with a non zero probability;

$$C_{OXY}(r,u,\varphi) = \{(i,j) \mid \exists k:\ C[i,j,k] \cap S \neq \emptyset\}$$

is OXY projection of set $C_{OXYZ}(r,u,\varphi)$;

$$C_{OZ}(r,u,\varphi) = \{k \mid \exists C(i,j):\ C(i,j,k) \cap S \neq \emptyset\}$$

is OZ projection of set $C_{OXYZ}(r,u,\varphi)$.

For each cell containing solutions of the (15) system, we find the probability that the EMRS belongs to this cell when some hypothesis is fulfilled. In the future, we will use the found probabilities to build a three-dimensional map of the membership level of cells that are part of the thunderstorm center.

3.1 Hypothesis H_1: Uniform probability distribution on $C_{OXY}(r,u,\varphi)$

It is following from (15) that on the OXY plane, the valid points farthest from the origin of the coordinates are $(x_1,\ y_1)$ and $(x_2,\ y_2)$

$$x_{1,2} = r(u\cos\varphi \mp \sqrt{1-u^2}\sin\varphi), \tag{18}$$

$$y_{1,2} = r(u\sin\varphi \pm \sqrt{1-u^2}\cos\varphi, \quad z_{1,2} = 0. \tag{19}$$

The coordinates x, y for all possible ERMS positions belong to segment $[(x_1,y_1),\ (x_2,y_2)] \subset OXY$ between the points (x_1,y_1) and (x_2,y_2). The all possible radiation source locations can be represented in a parametric form as the set of points:

$$\mathcal{L}(r,u,\varphi) = \left\{ \begin{array}{c} x(\lambda) = ru\cos\varphi + \\ +r(2\lambda-1)\sqrt{1-u^2}\sin\varphi, \\ y(\lambda) = ru\sin\varphi + \\ +r(1-2\lambda)\sqrt{1-u^2}\cos\varphi, \\ z(\lambda) = 2r\times \\ \times\sqrt{\lambda(1-\lambda)(1-u^2)} \end{array} \middle| \ 0 \le \lambda \le 1 \right\}. \tag{20}$$

Let

$$\Lambda = \left\{0,\ \frac{1}{N},\ \frac{2}{N},\ \ldots,\ \frac{N-1}{N},\ 1\right\}.$$

It is following from (20) that points $(x(\lambda), y(\lambda)),\ \lambda \in \Lambda$ are evenly distributed on the segment $[(x_1,y_1),\ (x_2,y_2)]$, and

$$[(x_1,y_1),\ (x_2,y_2)] \quad \subset \bigcup_{(i,j)\in C_{OXY}(r,u,\varphi)} C(i,j).$$

If the hypothesis on uniform probability distribution on $C_{OXY}(r,u,\varphi)$ is accepted, there are equalities

$$C_{OXYZ}(r,u,\varphi) = \left\{(i,j,k)\ :\ i = \left\lfloor\frac{x(\lambda)}{\Delta}\right\rfloor,\ j = \left\lfloor\frac{y(\lambda)}{\Delta},\right\rfloor\ k = \left\lfloor\frac{z(\lambda)}{\Delta}\right\rfloor,\ \lambda\in\Lambda\right\},$$

$$C_{OXY}(r,u,\varphi) = \left\{(i,j)\ :\ i = \left\lfloor\frac{x(\lambda)}{\Delta}\right\rfloor,\ j = \left\lfloor\frac{y(\lambda)}{\Delta}\right\rfloor,\ \lambda\in\Lambda\right\}.$$

Let

$$N(i,j,k) = \mathrm{card}\,\{\lambda\in\Lambda\ :\ (x(\lambda),\ y(\lambda),\ z(\lambda)\in C(i,j,k))\},$$

$$N(i,j) = \sum_{k\in C_{OZ}(r,u,\varphi)} N(i,j,k), \quad .$$

Then, the probability of EMRS in cell $C[i,j,k]$ is equal to

$$\mathrm{P}_{OXY}(i,j,k) = \frac{N(i,j,k)}{N(i,j)\cdot N}.$$

3.2 Hypothesis H_2: Uniform probability distribution on $C_{OZ}(r,u,\varphi)$

It follows from (15) that the set of points of all possible locations of EMRS can be represented in a parametric form other than (20) :

$$\mathcal{L}(r,u,\varphi) = \left\{ \begin{array}{c} x(z) = ru\cos\varphi \pm \sqrt{r^2(1-u^2)-z^2}\sin\varphi, \\ y(z) = ru\sin\varphi \mp \sqrt{r^2(1-u^2)-z^2}\cos\varphi, \\ z:\ \ 0\le z\le \sqrt{r^2(1-u^2)} \end{array} \right\}. \tag{21}$$

Let

$$M = \left\lfloor\frac{\sqrt{r^2(1-u^2)}}{\Delta}\right\rfloor, \quad Z = \{0,\ \Delta,\ 2\Delta,\ \ldots,\ M\cdot\Delta\}. \tag{22}$$

It is following from (22) that points $z\in Z$ are evenly distributed on the segment

$$[0,\ \sqrt{r^2(1-u^2)}] \quad \subset \bigcup_{k:\in C_{OZ}(r,u,\varphi)} C[k].$$

If the hypothesis on uniform probability distribution on $C_{OZ}(r,u,\varphi)$ is accepted, there are equalities

$$C_{OXYZ}(r,u,\varphi) = \left\{(i,j,k)\ :\ i = \left\lfloor\frac{x(z)}{\Delta}\right\rfloor,\ j = \left\lfloor\frac{y(z)}{\Delta},\right\rfloor\ k = \left\lfloor\frac{z}{\Delta}\right\rfloor,\ z\in Z\right\},$$

$$C_{OXY}(r,u,\varphi) = \left\{(i,j)\ :\ i = \left\lfloor\frac{x(z)}{\Delta}\right\rfloor,\ j = \left\lfloor\frac{y(z)}{\Delta}\right\rfloor,\ z\in Z\right\}.$$

Let

$$M(i,j,k) = \mathrm{card}\,\{z\in Z\ :\ (x(z),y(z),z)\in C(i,j,k)\},$$

$$M(k) = \sum_{[i,j]\in C_{OXY}(r,u,\varphi)} M(i,j,k), \qquad k:\ k\cdot\Delta\in Z,$$

then

$$\mathrm{P}_{OZ}(i,j,k) = \frac{M(i,j,k)}{M(k)\cdot M}.$$

3.3 Hypothesis H_3: Uniform probability distribution on $C_{OXYZ}(r, u, \varphi)$

Since (20) and (21) give different representations of set $\mathcal{L}(r, u, \varphi)$, there are also two ways to represent a set of cells with a non-zero probability of having the EMRS:

$$C_{OXYZ}(r, u, \varphi) = \{(i, j, k) : \ \mathrm{P}_{OXY}(i, j, k) > 0\} = \\ = \{(i, j, k) : \ \mathrm{P}_{OZ}(i, j, k) > 0\}.$$

If the hypothesis on uniform probability distribution on $C_{OXYZ}(r, u, \varphi)$ is accepted, we have

$$\mathrm{P}_{OXYZ}(i, j, k) = \frac{1}{\operatorname{card} C_{OXYZ}(r, u, \varphi)}, \quad (i, j, k) \in C_{OXYZ}(r, u, \varphi).$$

3.4 Three-dimensional map of the level belonging of thunderstorm foci to space cells

So, for each cell containing solutions of the (15) system, we found the probability that the EMRS belongs to this cell when hypothesis $H \in \{H_1, H_2, H_3\}$ is fulfilled. Let us use the found probabilities to build a three-dimensional map of the membership level of cells that are part of the thunderstorm center when hypothesis H is fulfilled.

The set of discharges registered for the time period T is defined as $R(T)$. The cell grade membership to thunderstorm of the radiation source defined as the sum of all of radiation during this period.

$$\operatorname{grade}_H^{R(T)}(i, j, k) = \frac{\sum\limits_{d \in R(T)} \mathrm{P}_H(i, j, k)(d)}{\max\limits_{(l,m,n)} \sum\limits_{d \in R(T)} \mathrm{P}_H(l, m, n)(d)}. \qquad (23)$$

Further we demonstrate the possibility to estimate the thunderstorm center location by (23) formula applying simulation.

4 Simulation

Let's define two artificial thunderstorms $R_1(T)$ and $R_2(T)$ for modeling purposes. All discharges of the thunderstorms will have the same location $R = 60$ km, $\Delta = 10$ km, $\psi = \pi/4$, $\theta = \pi/3$.

Figures 3–8 demonstrate the observation point ($x = 0$, $y = 0$, $z = 0$) and the thunderstorm location cell is $C(i = 3,\ j = 3,\ k = 3)$ (marked with a blue square). Cell color shows the cell grade membership level to the thunderstorm (see figure 2).

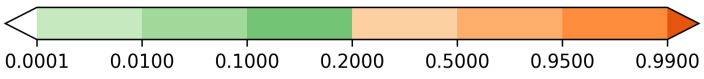


Figure 2: Scale of the level of belonging of a cell to a thunderstorm center

The $R_1(T)$ emits 256 pulse signals from dipole radiation sources with orientation $\psi_0 \in \{\pi k/16 : \ k = 1, 2, \ldots, 32\}$, $\theta_0 \in \{\pi k/8 : k = 0, 1, \ldots, 8\}$, that is EMRS are different incline dipoles, which is typical for a thunderstorm center. Location parameters values r, u, v, φ are calculated by (5)-(6) formulas.

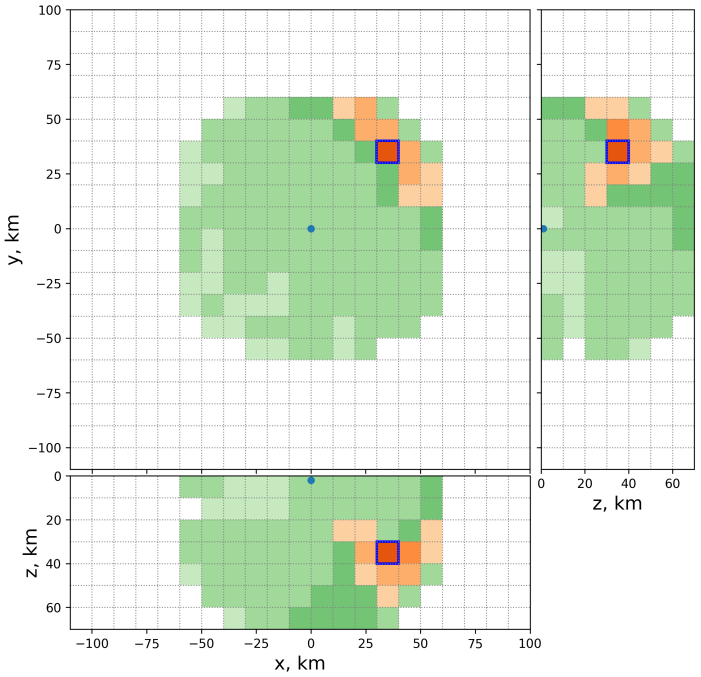


Figure 3: Map of $R_1(T)$ under hypothesis H_1

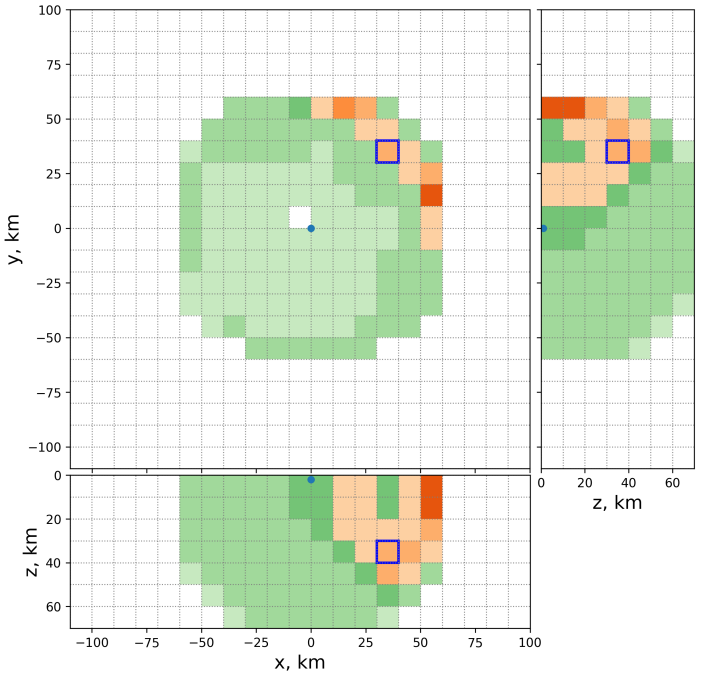


Figure 4: Map of $R_1(T)$ under hypothesis H_2

Figures 3–8 demonstrate three projections of the three-dimensional map of thunderstorm foci $R_l(T)$, $l \in \{1, 2\}$ level belonging to space cells under hypothesis $H \in \{H_1, H_2, H_3\}$

$$\operatorname{grade}_H^R(T)(i, j) = \sum_k \operatorname{grade}_H^R(T)(i, j, k), \qquad (24)$$

$$\operatorname{grade}_H^R(T)(i, k) = \sum_j \operatorname{grade}_H^R(T)(i, j, k), \qquad (25)$$

$$\text{grade}_H^R(T)(j,k) = \sum_i \text{grade}_H^R(T)(i,j,k). \quad (26)$$

Figure 3 demonstrates the belonging of desired cell to cell with maximum level of membership. Figure 4 demonstrates that the desired cell does not belong to cells with maximum level of membership. Figure 5 demonstrates the belonging of desired cell to cell with maximum level of membership.

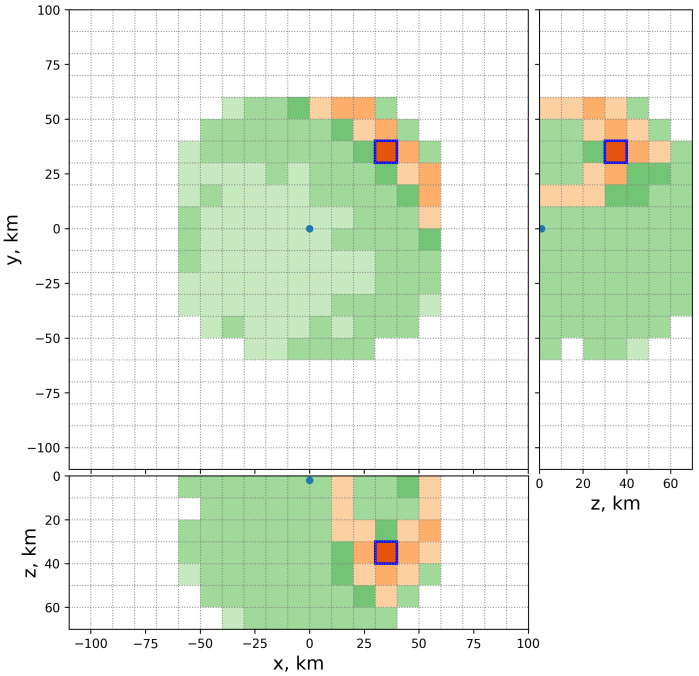


Figure 5: Map of $R_1(T)$ under hypothesis H_3

Figures 3, 5 make it possible to identify the desired cell $C(i = 3, j = 3, k = 3)$. Figure 4 does not allow to identify the desired cell at all.

Table 1 shows cells with grade membership level to the thunderstorm $R_1(T)$ no less 0.9 and similar to visual analysis of the projections hypothesis H_1 and H_3 found the desired cell.

Table 1: Cells with grade membership level to the thunderstorm $R_1(T)$ no less 0.9

Hypothesis	$C_{OXYZ}(r,u,\varphi) \cap \{(i,j,k) : \text{grade}_H^{R_1(T)}(i,j,k) > 0.9\}$
H_1	$\{(3,3,3),\ (3,2,3),\ (4,3,3)\}$
H_2	$\{(1,5,0),\ (1,5,1),\ (3,3,4),\ (5,1,0),\ (5,1,1)\}$
H_3	$\{(3,3,3)\}$

But the best hypothesis is hypothesis H_3, because it will allow to identify the cell with a thunderstorm center uniquely and accurately.

The $R_2(T)$ has the same location as $R_1(T)$ but emits only 100 pulse signals from dipole radiation sources with orientation $\psi_0 \in \{2\pi k/100 :\ k = 1,2,\ldots,100\}$, $\theta_0 = \pi/2$, that is EMRS are different horizontal dipoles, which is typical for a pre-storm center. Location parameters values r, u, v, φ are calculated by (5)-(6) formulas.

Figures 6–8 demonstrate three projections of the three-dimensional map of thunderstorm foci $R_2(T)$ level belonging to space cells under hypothesis $H \in \{H_1, H_2, H_3\}$.

Figure 6 demonstrates that the desired cell does not belong to cells with maximum level of membership.

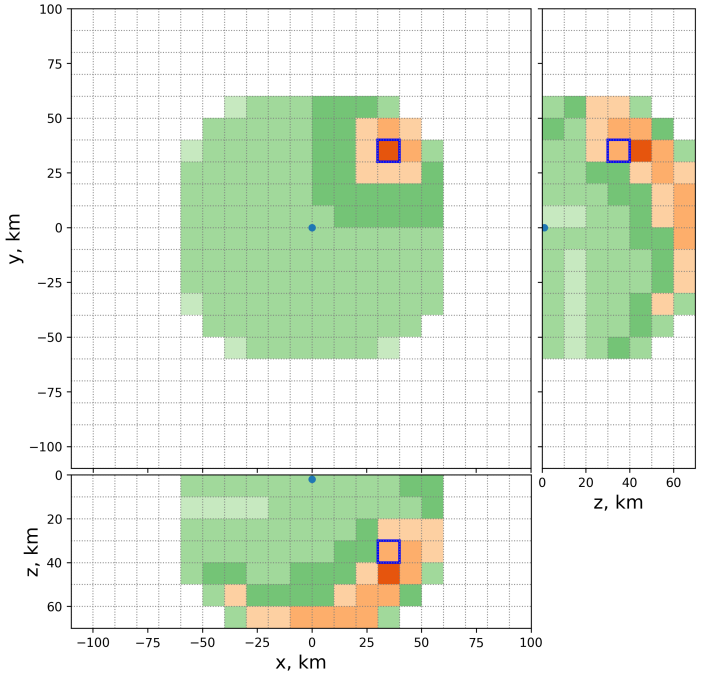


Figure 6: Map of $R_2(T)$ under hypothesis H_1

Figure 7 demonstrates that the desired cell does not belong to cells with maximum level of membership.

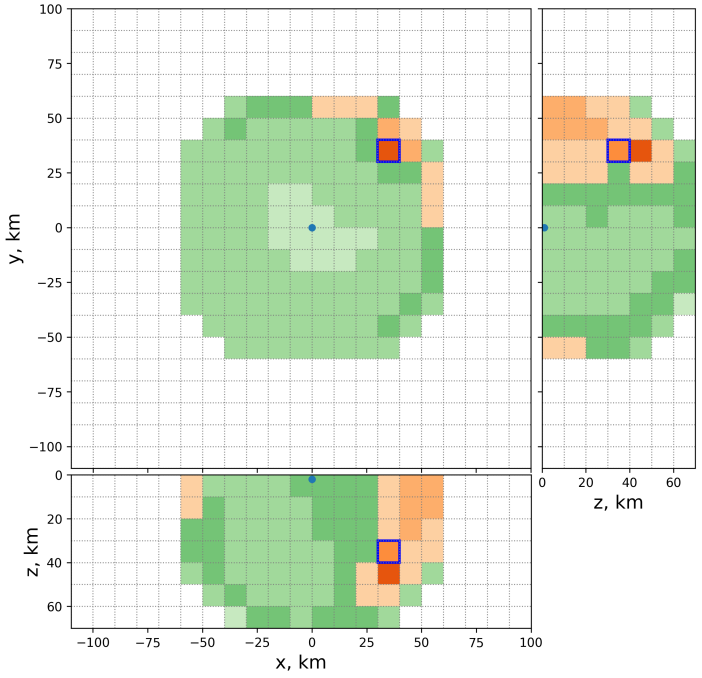


Figure 7: Map of $R_2(T)$ under hypothesis H_2

Figure 8 demonstrates that the desired cell belong to cells with maximum level of membership only.

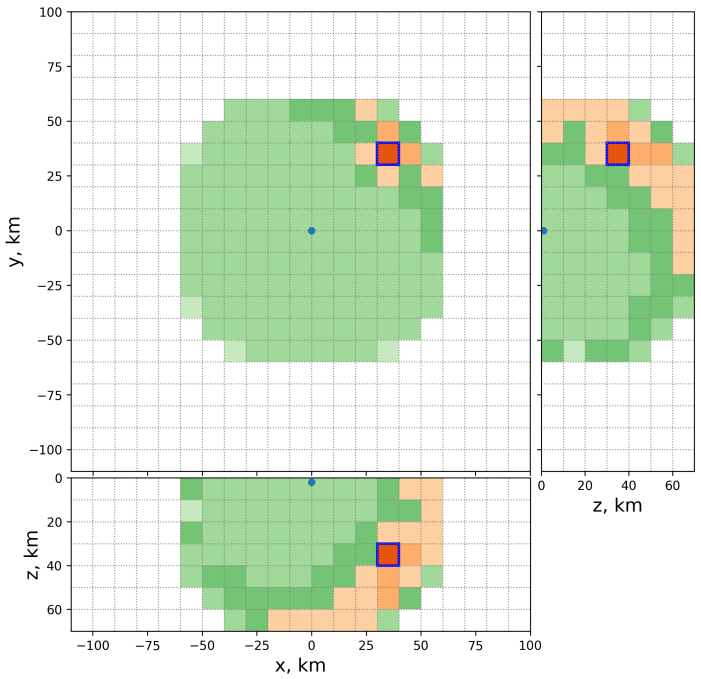


Figure 8: Map of $R_2(T)$ under hypothesis H_3

Table 2 also shows the best hypothesis is H_3. Hypotheses H_1 and H_2 maps do not contains the desired cell and does not allow to unambiguously establish the location of the foci.

Table 2: Cells with grade membership level no less 0.9 to the thunderstorm $R_2(T)$

Hypothesis	$C_{OXYZ}(r,u,\varphi) \cap \{(i,j,k):\ \text{grade}_H^{R_2(T)}(i,j,k) > 0.9$
H_1	$\{(3,3,4)\}$
H_2	$\{(3,3,4)\}$
H_3	$\{(3,3,3),\ (3,3,4)\}$

5 Conclusion

These findings provide a potential new technique to processing of the results of thunderstorm activity monitoring in the VLF range. The processing of all data from the SPMS during thunderstorm activity introduces a lot of uncertainty and makes it difficult to determine the real location of the thunderstorm center. The methods of analysis and visualization of the entire set of received signals presented in this paper reduce the uncertainty in determining the coordinates of the thunderstorm location. Processing of pre-thunderstorm radiation is significantly important to forecasting the development of thunderstorm.

The artificial single-point thunderstorm model described in the paper is a very simple, which is very convenient for research, however, a real thunderstorm can have a more complex structure, therefore it is important to carry out modelling with more complex models of thunderstorms, for example, check the proposed approaches for their ability to identification of multiple foci. This is an interesting topic for future work.

In future work, investigating another hypothesis of probability that the EMRS belongs to cell might be important.

References

[1] A. V. Panyukov, A. K. Bogushov, "Single-point system of positioning storms in close zone," RU Patent 2,599,775, 2016.

[2] A. V. Panyukov, A. K. Bogushov, "A Method For Processing Data of Thunderstorm Activity Monitoring," in 2019 Russian Open Conference on Radio Wave Propagation (RWP), 556–559, 2019, doi:10.1109/RWP.2019.8810345.

[3] "Thunderstorm Electric Field Mill EFM550," 2020.

[4] "Thunderstorm Advanced Total Lightning Sensor LS7002," 2020.

[5] "Boltek LD-350 Lightning Detector," 2020.

[6] "Biral BTD-300 Thunderstorm Detector," 2020.

[7] "The International Science and Technology Center. Project; 1822. Development and Investigation of the Single-Point System for Lightning Location in the Range of SuperLong Waves," 2000.

[8] I. Kononov, I. Yusupov, N. Kandaratskov, "Analysis of One-Point Methods for Lightning-Discharge Passive Location," Radiophysics and Quantum Electronics, **56**, 788–800, 2014, doi:10.1007/s11141-014-9481-1.

[9] A. V. Panyukov, "Estimation of the Location of an Arbitrary Oriented Dipole under Single-point Direction Finding," Journal of Geophysical Research, **101**, 14997–14982, 1996, doi:10.1029/96JD00067.

[10] B. Taibin, "An Approach to Define Parameters for Localization of Thunderstorm," IEEE Antennas and Propagation Magazine, **48**, 48–54, 2006, doi:10.1109/MAP.2006.1650816.

[11] A. V. Panyukov, A. K. Bogushov, "The Spectral Statistical Method for Determining the Location Parameters of a Dipole Source of Electromagnetic Radiation," Radiophysics and Quantum Electronics, **59**, 278–288, 2016, doi:10.1007/s11141-016-9696-4.

[12] A. V. Panyukov, "Inverse Problem for an Electrical Dipole and the Lightning Location Passive Monitoring System," in M. S. Kandelousi, editor, Electric Field, chapter 13, IntechOpen, Rijeka, 2018, doi:10.5772/intechopen.74264.

[13] S. He, M. Popov, V. Romanov, "Explicit full identification of a transient dipole source in the atmosphere from measurement of the electromagnetic fields at several points at ground level," Radio Science, **35**(1), 107–117, 2000, doi:https://doi.org/10.1029/1999RS002198.

[14] M. Popov, S. He, "Identification of a transient electric dipole over a conducting half space using a simulated annealing algorithm," Journal of Geophysical Research, **105**, 20821–20831, 2000, doi:10.1029/2000JD900261.

[15] A. Leal, V. Rakov, E. Alves, M. Lopes, "Estimation of –CG lightning distances using single-station E-field measurements and machine learning techniques," 1–8, 2019, doi:10.1109/SIPDA47030.2019.9004484.

[16] "Vaisala. ALARM Automated Lightning Alert and Risk Management System," 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1126-1136 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Tolerance of Characteristics and Attributes in Developing Student's Academic Achievements

Wongpanya Nuankaew[1], Pratya Nuankaew[2,*]

[1]*Faculty of Information Technology, Rajabhat Mahasarakham University, Maha Sarakham, 44000, Thailand*

[2]*School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand*

ARTICLE INFO

Article history:
Received: 03 July, 2020
Accepted: 15 October, 2020
Online: 22 October, 2020

Keywords:
Academic Achievement Attributes
Educational Data Mining
Lifelong Learning
Learning Strategies
Learning Analytics

ABSTRACT

The purpose of this research is to study the relevance of factors for the analysis of the effectiveness of suitable educational institutions that illustrate the significance of the characteristics and attributes of the student's academic achievements and to identify the acceptance and tolerance of each attribute, which supports lifelong learning. The data used in this research is 1109 students who used and tested the institution recommender system based on student context and educational institution application. The research methodology focuses on the study of user involvement and application analysis. There are six significant phases of the research: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The machine learning tools and data mining techniques are k-means, k-medoids, decision trees, cross-validation methods, and confusion matrix. From the data analysis, it can be concluded that the overall level of satisfaction with the application is accepted (average = 3.70, S.D. = 0.84). In addition, the prototype model has been developed for predicting and recommending appropriate institutions for the learner has moderate accuracy levels (92.25%), and the results of the self-test data model are very accurate at the highest level, which is equal to 93.78%. Finally, this research demonstrates the relevance and success of education engineering projects. It demonstrates a worthy accomplishment. For future research, the researchers aim to construct and develop applications that promote and support the findings of this research.

1. Introduction

Education and learner sustainability are the main goals of Thailand's government and the United Nations (UN) as defined in the Sustainable Development Goals (SDGs) [1,2], which are aimed and set as the world target success in 2030 [1]. The results are from an extensive research related to the development in quality of education. Numerous research studies showed a variety of educational outcomes [3–5], such as the development of learners' theories, learning styles, educational models, and educational technology.

Most of the research results is abstract and difficult to develop into a model or method that can be effectively implemented. In addition, there are many ways to take the next step towards student achievement, as shown in Figure 1. The question asked is "what is truly reasonable in theoretical practice for student characteristics?"

While there are many theories for the development of learning, there seems to be no technology and method to present reasonable and appropriate models, tools and strategies for learners. The combination of the two academic domains between engineering and education that had occurred is called Education Engineering (EE). In addition, it has completely changed the current radical educational system to be known as "Disruptive Technology" [6,7]. However, the researchers tried to study and carry out the research by applying the two knowledge together [8–14]. The discovery was that there were too many research areas that have yet to be explored. Figure 2 shows the studies and research related to the research results of the past researchers.

From Figure 2, the researchers gathered the hot issues, critical problems, data set, student information, research models, and research results in many perspective [2,8]. Preliminary studies have driven the researchers to conduct research studies to develop learners' potential and significantly increase the learners' academic achievement. Therefore, this research has important objectives and goals for studying the relevance and success of

*Corresponding Author: Pratya Nuankaew, School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand, +66 89 961 4832 & pratya.nu@up.ac.th

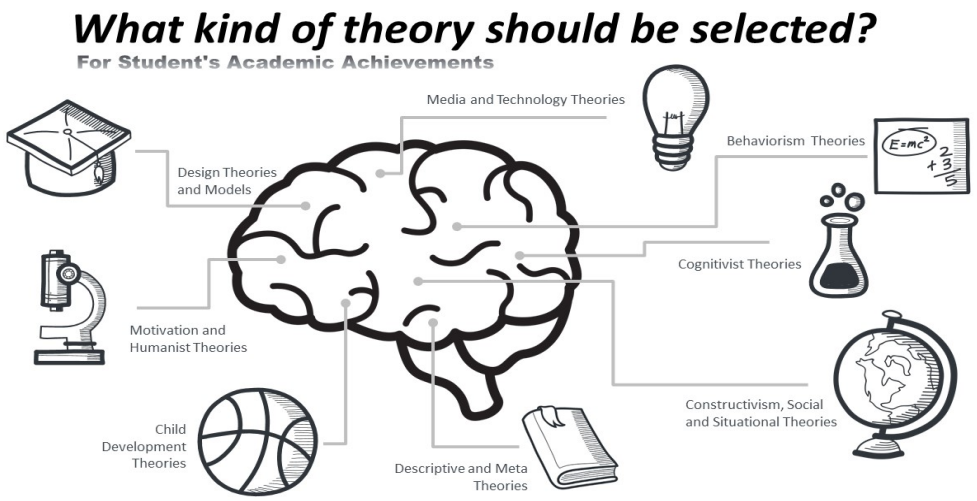


Figure 1: World of Learning Theory



Figure 2: Areas of research being offered

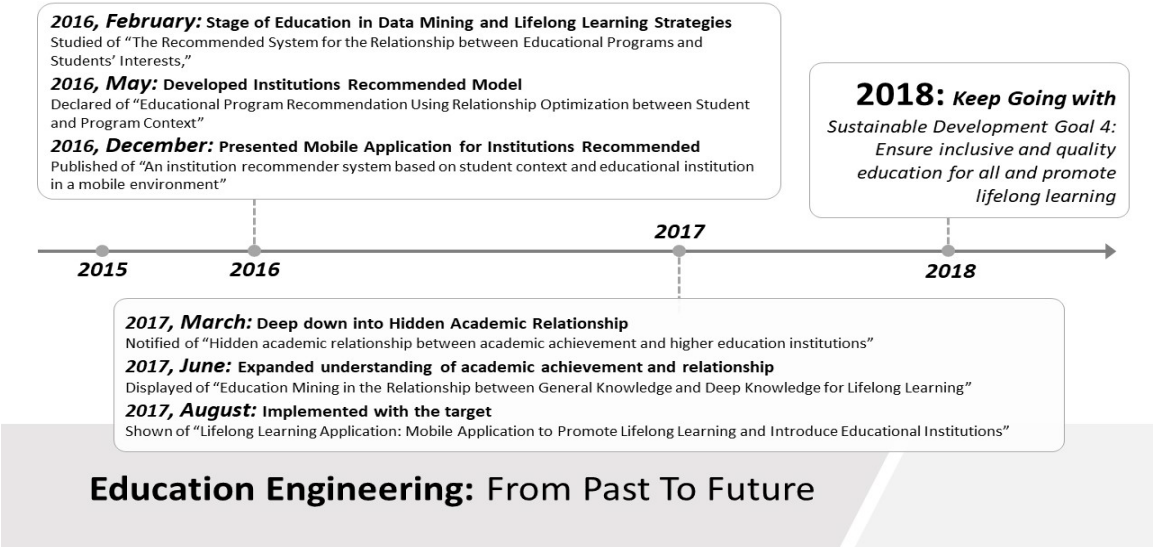


Figure 3: Research Progress

research that demonstrate the significant attributes of the student's academic achievements, and to identify the acceptance and tolerance of each attribute.

The important process of this research is therefore aimed at the model chosen from the previous research to improve the application that had been developed [10,15]. After that, the data collected from 1109 actual users were analyzed and searched for factors that supported and encouraged learners to meet the needs and suitability between the learners and the educational institutions. This is the main objective of the research.

The outline of this research is divided into four sections. Section 1 leads the reader to understand the problem. Section 2 is the presentation of literature, related work and research methods. Section 3 is a summary of the research results and discussions. Finally, the last section is the summarized conclusion.

2. Literature Reviews and Related Works

This section summarizes the overview of research related to the moment of research in engineering education. Areas of education engineering in the paper are scoped into five stages: traditional problems, students' academic achievement, engineering education, technology of education engineering, and how can education engineering work.

2.1. Thailand's educational policy and problems

Since 1999, education in Thailand has been improved and educational opportunities have presented for the public. The system provides the learning process for personal and social development with factors available conducive to continuous lifelong learning [5]. Everything seems beautiful for Thai students [6]. While twenty years have passed, Thailand's education system is still hopeful that there is light at the end of the tunnel. The failure of a test to measure students' achievement is a good example [16]. However, when considering the factors, learners are most important for developing because no matter how the system has changed, the end result should be the learner's achievement. Finding the identity of the learner to present what is worthy for her/him is the first priority [10].

2.2. Student's Academic Achievements

Certifications and degrees are not a measure of a person's achievements, but it is a tool that helps everyone find what they want and expect in life. Academic results are an opportunity for learners to find what they are. At the same time, grading the exam and testing can be used to determine the advancement and recognition of students' knowledge, however there is a need on considering what determines to be for success as the process may require more testing.

Student's academic achievements can be judged from many perspectives, such as the dimension of learning styles, student's emotions, academic motivation, and hidden academic relationship [2,9,17]. Considering the reasonableness of the individual is more important than anything.

2.3. Education Engineering

Education and Engineering has changed over time, where the desire to achieve "the best practices" requires a clear framework. For the best, finding the answer requires research, which also requires a method, technology, theoretical, and application that depends on the researcher's needs [17,18].

It is only a minority researcher that mentions engineering education [3,19]. To combine the education and engineering requires a step by step framework for developing the best practice models. In addition, in-depth research using modern technology also supports learning theory by finding and developing models that are sensible to learners. No matter how smart and important the technology is, the emphasis of educational development is to develop the learning processes, analytical thinking, and ability to apply their knowledge to solve problems and increase knowledge [10].

2.4. Education Engineering Technology

Educational engineering technology is the use of technology for effective education for learners. It resulted from a change known as disruptive technology [7, 8]. There are many programs and applications that support Thailand's education, which bring in novel and advanced technology for managing in order to achieve better student performance [4,20,21]. The results show different aspects of the development in the education process.

2.5. Implemented Education Technology

A good example is on how the application of engineering knowledge is integrated into the educational process [5]. It provides and explains the advanced technology application, which consists of five steps: requirement definition, system and software design, implementation and unit testing, integration and system testing, and operation and maintenance. All steps are the key principles in the development of knowledge, research, and discoveries to be practically usable [5, 11, 16].

2.6. Thailand's educational policy and problems

This research was conducted and studied logically, as illustrated in the study period in Figure 3. Their research has started to focus on education engineering since the year 2016 and has proposed concepts of applied engineering and education for the education models in Thailand. At this stage, it presented their research from the past to the future, which consists of six main steps. The first step is defining the research problems. The second step is finding the significant and appropriate attributes. The third step is development of the model. The fourth step is building the application. The fifth step is testing and deploying the application. Finally, the sixth step is applying the application to the targets

2.6.1. Defining the Research Problems

It started in early 2016, which presented the idea of matching students to universities [5]. The research problem is "how to introduce and match diverse students to different contexts of educational institutions" as shown in Figure 4.

The results of their research found that Thailand's education system and interest to study are important factors in choosing universities and influencing the decision of studying at a higher education level [16]. The data used in their research was 256 students from two universities: (1) the Rajabhat Mahasarakham University, Maha Sarakham, Thailand and (2) the University of Phayao, Phayao, Thailand.

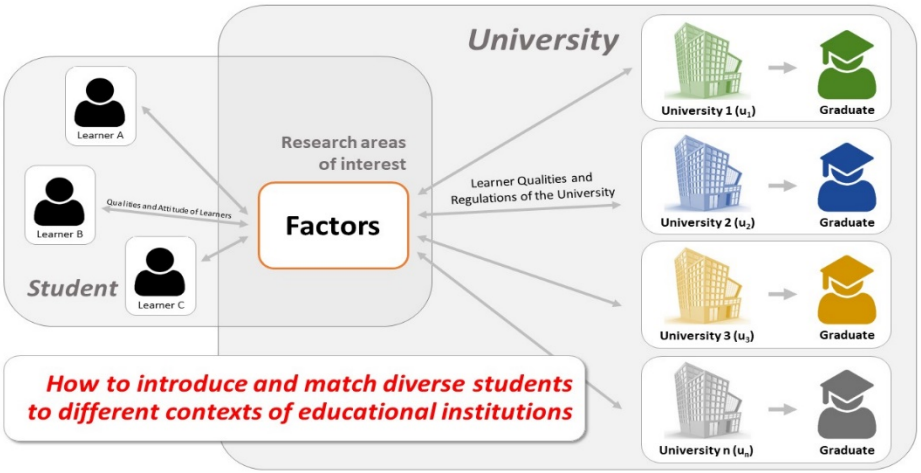


Figure 4: Matching Students to Universities

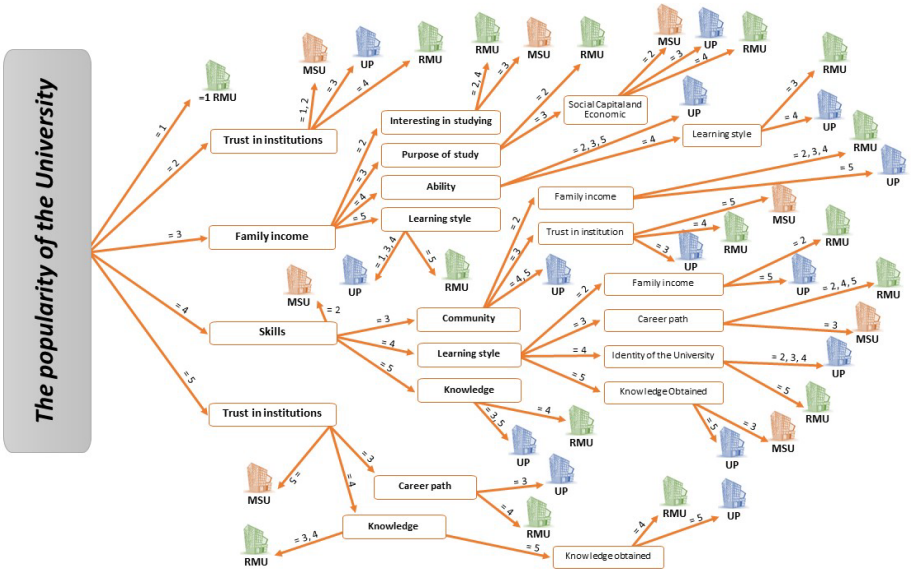


Figure 5: Appropriate Model with Decision Tree [23]

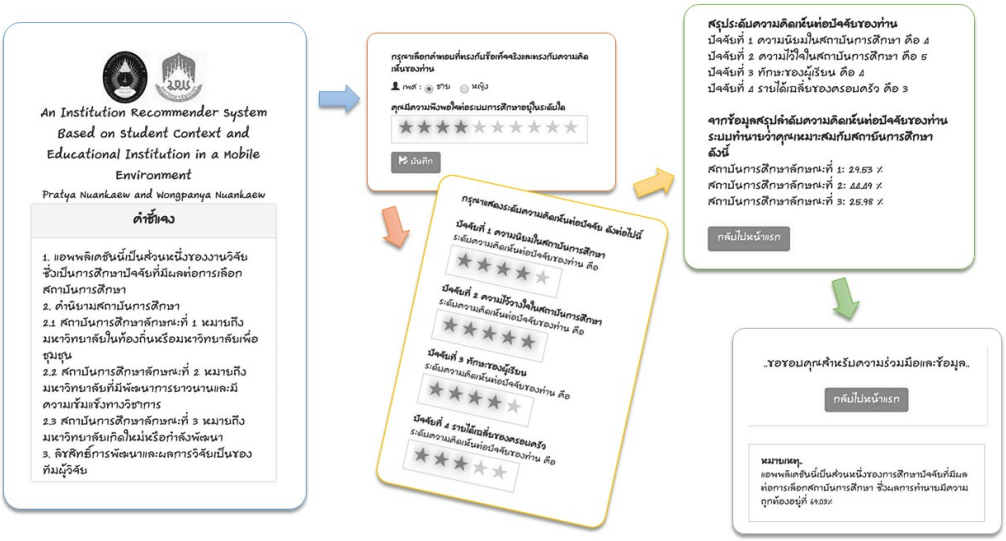


Figure 6: Institution Recommender Mobile Application

2.6.2. Finding Significant and Appropriate Attributes

After the success of the study in the problems of the decision to study at a higher level, the researchers conducted the study to find factors that are important matching components between the learners and the educational institutions [22]. The data used in the research was 885 students from three universities: (1) the Rajabhat Mahasarakham University, Maha Sarakham, Thailand, (2) the Maha Sarakham University, Maha Sarakham, Thailand, and (3) the University of Phayao, Phayao, Thailand.

The result is the acceptance of the factors studied. It was found that the factors associated with their research consisted of 19 significant factors including 15 factors of student's context and 4 factors of program's context, as shown in Table 1.

Table 1: Significant and Appropriate Attributes [23]

Perception and Perspective	Statement and Details
Student's Context	*Student's Interests*
	Stage 1.1 Interest in studying Stage 1.2 Career path Stage 1.3 Identity of the University Stage 1.4 Knowledge obtained Stage 1.5 Education system
	Student's Characteristics
	Stage 2.1 Student Abilities Stage 2.2 Student Skills Stage 2.3 Student Knowledge Stage 2.4 Learning Styles of Student
	Student's Environment
	Stage 3.1 Social capital and Economic Stage 3.2 Trust in institutions Stage 3.3 Personal decision Stage 3.4 Family income Stage 3.5 Community Stage 3.6 Purpose of study
Program's Context	*Program's Context*
	Stage 4.1 The popularity of the University Stage 4.2 Curriculum Stage 4.3 Education standard Stage 4.4 Quality assurance

2.6.3. Model Development

After studying and analyzing perceptions of factors that are important to the choice of continuing education between students and institutions [23], the data collected the satisfaction level with factors having been analyzed and developed by data mining techniques and machine learning tools. The result is a decision tree model that is shown in Figure 5. Figure 5 illustrates the decision tree obtained from the appropriate model.

All among 19 factors, 15 factors can determine whether the students are suitable to learn at that particular university.

2.6.4. The Application Development

After modeling success, the mobile applications have been developed [5]. Their application consists of three important steps: (1) introduction of the application, (2) prediction of interest at a higher education level, and (3) recommendation of an appropriate institution for the user. The structure and working process of the application is shown in Figure 6.

2.6.5. Testing and Deployment

The application in the previous section was searched for performance, which used a satisfaction questionnaire from 431 samples [5]. It consisted of 186 high school students, 191 university students, 12 staffs and 42 lectures in the universities. The testing process consisted of four main stages: functional requirement testing, functional testing, usability testing, and security testing. According to their results, it indicated that the total satisfaction toward the application are highest, which is equal to 4.32, and the groups with the highest levels of satisfaction are university students, which is equal to 4.43. It can be concluded that the application is appropriate and reasonable for students who wish to study at a higher education level [5].

2.6.6. Applied Application to the Targets

After the prototype application was created [5], it was applied to 1,109 students from the three universities; Rajabhat Mahasarakham University, Mahasarakham University, and University of Phayao to test the application [11]. Based on the data collected, they found that the relationships were more expected and exceeded that was supported by Thailand's educational system. It is defined as a "Nest Relationship" or "Hidden Relationship". The results of their research demonstrate the different levels of relationships that are influenced by the Grade Point Average (GPA), school size, community environment, and school friends. It is consistent with the hypothesis that was predicted as shown in Figure 7.

From the findings and additional information collected, the research studied on matching the right students with the most appropriate institutions. The research aimed on developing the knowledge and seeking the best tools for the learner, which have research related to the development tools in education engineering, machine learning, active learning theory, and data mining in education [2, 5, 9, 12].

3. Research Methodology

The purpose of this research is to study the relevance of factors for the analysis of the effectiveness of suitable educational institutions that illustrate the significant of the characteristics and attributes of the student's academic achievements, and to identify the acceptance and tolerance of each attribute, which supports lifelong learning. The data collection used in this research is 1109 students who used and tested the institution recommender system based on student context and educational institution application [11]. The research methodology focused on the study of user involvement and application analysis. It consisted of six main steps in data analysis for data mining management [23,24]: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

3.1. Business Understanding

The business understanding presents the viewpoint of research problems [24]. Understanding the overview of the problem will enable the research goals to be defined clearly. Thus, the research problem of this research is to screen the groups of testers who tested the applications [11], and to use the data collected to study the group's behavior.

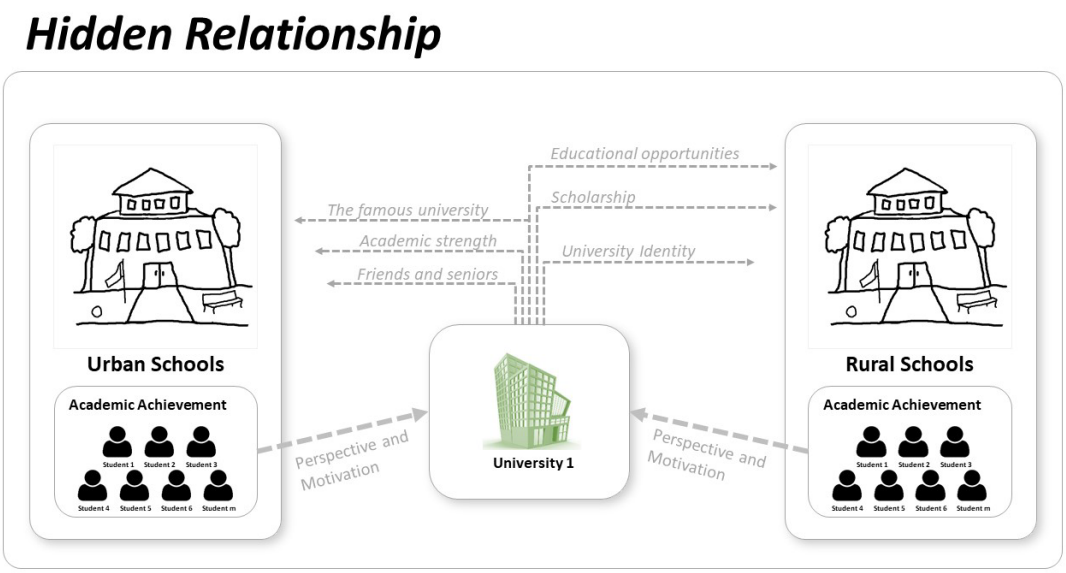


Figure 7: Thinking cycle of students toward the institution

3.2. Data Understanding

Data understanding is closely linked to the business understanding of activities which are aimed at identifying data quality problems, discovering the insight into the data, and examining the attractiveness of problems from hidden data [24]. Therefore, the research designed the data collection to be consistent with the research objectives. The characteristics of the data consisted of specifying the educational institution of the data providers and showing the level of attitude towards the factors that the researcher is studying.

3.3. Data Preparation

Data preparation is a process of data management for use in the development of prototype models. In addition, it prepares data in formats such as tables and records, including feature selection, feature management, data cleaning, construction of the new attributes, and so on [24]. In this research, it was compiled and prepared as summarized in Table 2 to Table 3.

Table 2: Data Collection

Institutions	Data Collection	
	Providers	*Percentage*
MSU: Maha Sarakham University	345	31.11%
RMU: Rajabhat Mahasarakham University	478	43.10%
UP: University of Phayao	286	25.79%
Total:	**1109**	**100%**

From Table 2, it shows that the largest number of tester who provided the data was Rajabhat Mahasarakham University, which had 478 tester or equal to 43.10 percent of all data providers. In addition, the summary of general statistics from data collection is shown in Table 3, referring to the accepted factors from Table 1.

From Table 3, it can be concluded that the overall perception and perspective on the overall factors averages at 3.70 (S.D. = 0.84), which is the most accepted point of view is program's context. It has an average of 4.00 (S.D. = 0.89). Therefore, it can be concluded that the providers accept all factors as a whole, with particular emphasis on the program's context.

Table 3: Data Analysis

Perception and Perspective	Data Collection (n = 1109)		
	Stages	*Means*	*S.D.*
Student's Context	*Student's Interests*		
	Stage 1.1	3.80	0.81
	Stage 1.2	3.74	0.85
	Stage 1.3	3.85	0.79
	Stage 1.4	3.82	0.74
	Stage 1.5	3.63	0.87
	Average	*3.77*	*0.82*
	Student's Characteristics		
	Stage 2.1	3.75	0.81
	Stage 2.2	3.63	0.74
	Stage 2.3	3.71	0.69
	Stage 2.4	3.40	0.78
	Average	*3.63*	*0.77*
	Student's Environment		
	Stage 3.1	3.61	0.81
	Stage 3.2	4.03	0.72
	Stage 3.3	3.27	0.90
	Stage 3.4	3.49	0.89
	Stage 3.5	3.20	0.84
	Stage 3.6	3.41	0.93
	Average	*3.50*	*0.89*
Program's Context	*Program's Context*		
	Stage 4.1	4.08	0.83
	Stage 4.2	4.12	0.74
	Stage 4.3	4.00	0.67
	Stage 4.4	3.79	0.70
	Average	*4.00*	*0.75*
	Total Average	*3.70*	*0.84*

3.4. Modeling

The modeling phase is the selection of tools suitable for the data collected to meet the research questions and goals [24]. Therefore, the researchers selected the machine learning tools and data mining techniques, which consisted of two types of tools: clustering tools consisting of k-means [25] and k-medoids [26], and classification tools consisting of decision tree [5, 11, 12].

The k-means is the most common algorithm used in an iterative refinement technique. It is also called Lloyd's algorithm [26], especially in the computer science community. The k-mean algorithm is performed by switching between two steps: (1) the assignment step, which assigns each observation to the cluster with the closest mean. (2) updates the procedure which is calculated a new means to be a centroid of observations in a cluster.

The k-medoids algorithm is a clustering algorithm that is associated with the k-means and the medoid shift algorithm. Both the k-means and k-medoids algorithms have some characteristics which divide the data set into groups. K-means is an attempt to reduce the total squared error, while the k-medoids reduce the sum of the dissimilarities between the points labeled in the cluster and the point that determines the center of that cluster [27].

The decision tree is one of the learning methods used in statistics, machine learning, and data mining. It works by determining data from observations and separating data for use in data consideration and finding predictive results. The benefit is getting important factors, which are caused by the nodes or the decision-making part of the model [5, 11, 12].

3.5. Evaluation

The goal of the evaluation is to assess the results and review the process [14,23]. The tools that are used in the research include cross-validation methods as shown in Figure 8, and confusion matrix as mentioned in Figure 9. Figure 8 displays the separation of data for evaluating the model. The cross-validation method divides the data into two parts. The first part is used for modeling and the remainder to test the model. In addition, model evaluation requires a tool called a confusion matrix [12, 14, 15, 24] to test the model's performance, with the principles shown in Figure 9.

Figure 9 presents the composition of the confusion matrix performance, which is composed of the actual class and the predicted class. An important benefit of the performance of the confusion matrix is the ability to determine the model's ability to predict results, such as the predictive ability or accuracy, model precision, model sensitivity, and model specificity (recall measurement). These values are used to determine the actual performance model. Moreover, Figure 9 also demonstrates the formulas and methods for calculating the various performance parameters in detail.

3.6. Deployment

The implementation is intended to be used in further applications [24]. The researcher aims to improve and use it in the next academic year (academic year 2021). It has been described in applying an activity recommendation as the following:

3.6.1. Testing Model Results

As mentioned above, the tools used to evaluate the model consist of two parts: cross-validation method, and confusion matrix. This section describes the implementation of assessment tools in the research.

The testing process divides the data into two parts according to the cross-validation method principles, but the division is divided into three types. The first type is 5-Fold cross-validation, which used 4-Fold (80%) for modeling and 1-Fold (20%) for testing. The second type is 10-Fold cross-validation, which used 9-Fold (90%) for modeling and 1-Fold (10%) for testing. The last type is leave-one-out cross-validation, which used 99% of data for modeling and 1% of data for testing. However, each time the cross-validation test is reported, the model results are also tested using the confusion matrix too.

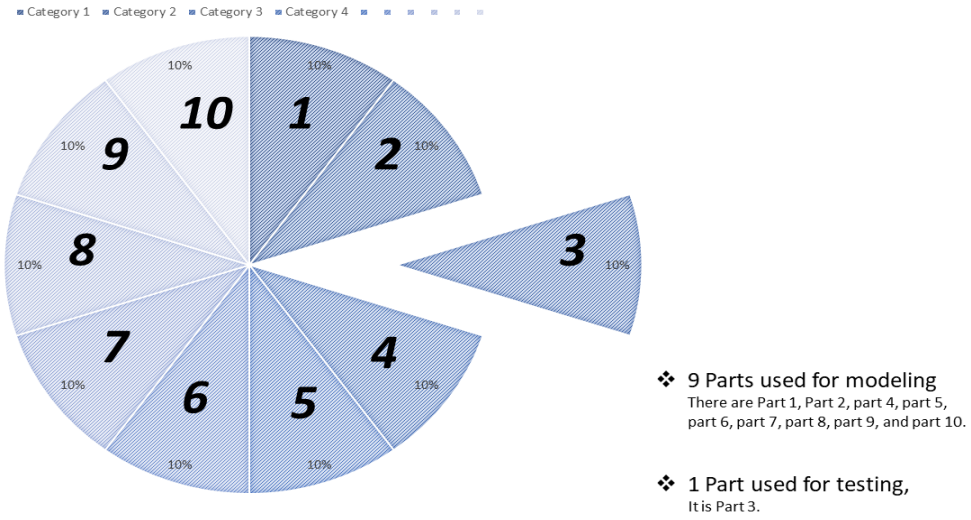


Figure 8: Cross-Validation Method

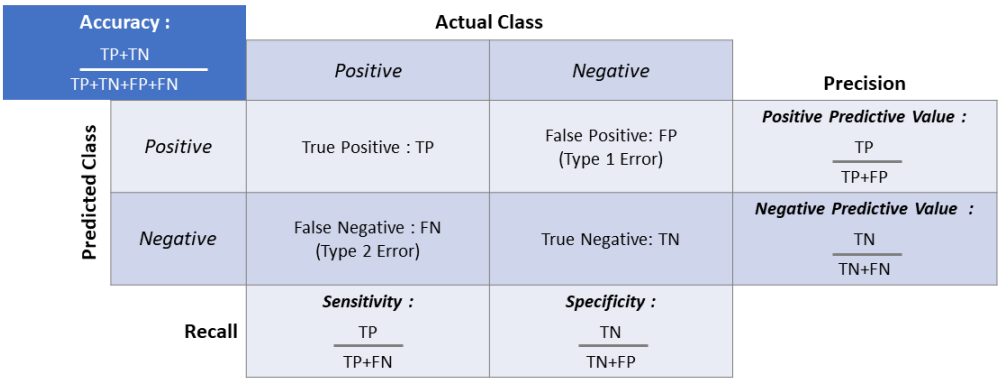


Figure 9: Confusion Matrix Performance

3.6.2. *Applying Model Results*

The applying model results are in three parts: generate a decision tree model for decision-making, test the decision tree model with data collection, and conclude the centroid of k-means. All details are presented on the section of research results and discussion.

4. Research Results and Discussion

In the research results, the researchers classified the research report into three parts which are modeling results, model testing results, and model applying results.

4.1. Modeling Results

Modeling results are the various models on different criteria, such as defining the unequal depth of the decision tree, and determining the different types of cross-validation method tests, which have the results shown in Table 4 and Table 5.

Table 4: Modelling Results from k-Means Clustering

Cluster Number	Decision Tree Depth	Cross-Validation Method Types		
		5-Fold	*10-Fold*	*Leave-one-out*
3 Clusters	Level 3	88.10%	88.46%	88.64%
	Level 4	88.73%	88.55%	89.00%
	Level 5	89.27%	89.18%	88.91%
	Level 6	89.63%	89.90%	89.90%
	Level 7*	91.07%	91.16%	92.25%*
	Level 8	91.52%	92.24%	92.25%
	Level 9	91.61%	92.24%	91.88%
4 Clusters	Level 3	47.61%	47.43%	47.07%
	Level 4	47.97%	47.88%	47.79%
	Level 5	58.97%	59.24%	59.24%
	Level 6	60.32%	61.59%	60.50%
	Level 7	65.10%	63.30%	63.39%
	Level 8	80.61%	79.44%	78.72%
	Level 9*	88.28%*	87.65%	87.65%
5 Clusters	Level 3	54.82%	54.73%	54.64%
	Level 4	55.46%	55.82%	55.91%
	Level 5	60.50%	59.87%	60.23%
	Level 6	63.12%	62.13%	60.50%
	Level 7	68.98%	66.72%	65.46%
	Level 8*	83.94%*	74.56%	66.28%
	Level 9	87.91%	85.12%	83.95%

From Table 4, it shows that the k-means model with the highest accuracy is the decision tree model that is classified into 3 clusters by dividing the model testing into the leave-one-out cross-validation method with a depth of 7 levels of the decision tree model and has an accuracy of 92.25%.

However, the test results classified by other clusters yielded a lower accuracy. For example, the 4 clusters with the highest results are 88.28% and the 5 clusters have the highest results of 83.94%.

Table 5: Modelling Results from k-Medoids Clustering

Cluster Number	Decision Tree Depth	Cross-Validation Method Types		
		5-Fold	*10-Fold*	*Leave-one-out*
3 Clusters	Level 3	78.00%	77.37%	76.01%
	Level 4	80.25%	81.24%	81.61%
	Level 5	84.67%	83.50%	77.37%
	Level 6	88.10%	86.83%	91.52%
	Level 7	90.62%	89.72%	90.35%
	Level 8	91.25%	91.25%	91.07%
	Level 9*	91.62%	91.97%*	91.61%
4 Clusters	Level 3	77.37%	76.38%	75.74%
	Level 4	78.08%	80.08%	81.33%
	Level 5	81.96%	83.32%	81.70%
	Level 6	84.49%	86.30%	86.93%
	Level 7	88.19%	89.18%	90.08%
	Level 8	90.26%	90.72%	91.52%
	Level 9	91.61%	91.62%*	90.62%
5 Clusters	Level 3	78.99%	78.54%	77.01%
	Level 4	82.23%	82.42%	81.15%
	Level 5	84.76%	85.39%	85.57%
	Level 6	86.65%	86.66%	86.38%
	Level 7	88.73%	88.73%	88.73%

Cluster Number	Decision Tree Depth	Cross-Validation Method Types		
		5-Fold	*10-Fold*	*Leave-one-out*
	Level 8	90.53%	89.36%	88.64%
	Level 9	91.96%*	90.81%	90.98%

From Table 5, it shows that the k-medoids model with the highest accuracy is the decision tree model that is classified into 3 clusters by dividing the model testing into the 10-fold cross-validation method with a depth of 9 levels of the decision tree model and has an accuracy of 91.97%.

However, the test results classified by other clusters yielded a lower accuracy. For example, the 4 clusters with the highest results are 91.62% and the 5 clusters have the highest results of 91.96%.

Table 6: Model Testing Results

Accuracy: 92.25%		Actual Class			*Precision Class*
		True Cluster_0	*True Cluster_1*	*True Cluster_2*	
Predicted Class	*Pred. Cluster_0*	185	*4*	*7*	94.39%
	Pred. Cluster_1	7	*423*	*31*	91.76%
	Pred. Cluster_2	21	*16*	*415*	91.81%
Recall Class		86.85%	95.49%	91.61%	

Table 7: Decision Tree Model

```
Stage 4.1 > 3.5
|  Stage 3.6 > 3.5
|  |  Stage 3.1 > 2.5
|  |  |  Stage 3.5 > 3.5: cluster_2 {cluster_2=268}
|  |  |  Stage 3.5 ≤ 3.5
|  |  |  |  Stage 3.3 > 3.5: cluster_2 {cluster_0=1, cluster_2=56, cluster_1=3}
|  |  |  |  Stage 3.3 ≤ 3.5
|  |  |  |  |  Stage 3.1 > 3.5: cluster_2 {cluster_2=11, cluster_1=3}
|  |  |  |  |  Stage 3.1 ≤ 3.5: cluster_1 {cluster_1=3}
|  |  Stage 3.1 ≤ 2.5: cluster_2 {cluster_0=1, cluster_2=3}
|  Stage 3.6 ≤ 3.5
|  |  Stage 1.4 > 2.5
|  |  |  Stage 1.3 > 2.5
|  |  |  |  Stage 3.2 > 2.5
|  |  |  |  |  Stage 3.5 > 3.5: cluster_2 {cluster_0=1, cluster_2=25, cluster_1=5}
|  |  |  |  |  Stage 3.5 ≤ 3.5: cluster_1 {cluster_0=6, cluster_2=30, cluster_1=423}
|  |  |  |  Stage 3.2 ≤ 2.5: cluster_0 {cluster_0=3}
|  |  |  Stage 1.3 ≤ 2.5
|  |  |  |  Stage 1.4 > 4: cluster_2 {cluster_2=2}
|  |  |  |  Stage 1.4 ≤ 4: cluster_0 {cluster_0=8}
|  |  Stage 1.4 ≤ 2.5: cluster_0 {cluster_0=4}
Stage 4.1 ≤ 3.5
|  Stage 4.4 > 4.5: cluster_2 {cluster_0=1, cluster_2=29, cluster_1=1}
|  Stage 4.4 ≤ 4.5
|  |  Stage 1.4 > 4.5
|  |  |  Stage 3.2 > 3.5: cluster_2 {cluster_2=9}
|  |  |  Stage 3.2 ≤ 3.5: cluster_0 {cluster_0=2,}
|  |  Stage 1.4 ≤ 4.5
|  |  |  Stage 2.3 > 4.5: cluster_2 {cluster_0=1, cluster_2=3}
|  |  |  Stage 2.3 ≤ 4.5
|  |  |  |  Stage 1.2 > 3.5
|  |  |  |  |  Stage 3.1 > 3.5: cluster_2 {cluster_0=10, cluster_2=16, cluster_1=1}
|  |  |  |  |  Stage 3.1 ≤ 3.5: cluster_0 {cluster_0=18, cluster_1=1}
|  |  |  |  Stage 1.2 ≤ 3.5: cluster_0 {cluster_0=157, cluster_2=1, cluster_1=3}
```

4.2. Model Testing Results

From the results of the prototype model development, it can be concluded that the model with the highest accuracy is the development of the model from k-means clustering, with the appropriate number of 3 clusters and the leave-one-out cross-validation result with an accuracy of 92.25%. Details of the testing developed model are shown in Table 6.

4.3. Model Applying Results

From the model that has been selected and demonstrated the performance of the model, this section shows the decision tree model in Table 7; the decision tree rules for self-testing is shown in Table 8, and the centroid of each cluster is shown in Table 9. In addition, a summary of the members from the collected datasets described by source classified by cluster. It is shown in Table 10.

Table 7 shows the decision tree models, which are the selected models from the tests in Table 8 and Table 9.

Table 8: Model Applying Results

Rule	Condition (If)	Prediction (Then)
1	If Stage 4.1 > 3.5 and Stage 3.6 > 3.5 and Stage 3.1 > 2.5 and Stage 3.5 > 3.5	Then, suitable for cluster_2 = 100%
2	If Stage 4.1 > 3.5 and Stage 3.6 > 3.5 and Stage 3.1 > 2.5 and Stage 3.5 ≤ 3.5 and Stage 3.3 > 3.5	Then, suitable for cluster_0 = 1.67%, suitable for cluster_1 = 5.00%, and suitable for cluster_2 = 93.33%.
3	If Stage 4.1 > 3.5 and Stage 3.6 > 3.5 and Stage 3.1 > 2.5 and Stage 3.5 ≤ 3.5 and Stage 3.3 ≤ 3.5 and Stage 3.1 > 3.5	Then, suitable for cluster_1 = 21.43%, and suitable for cluster_2 = 78.57%.
4	If Stage 4.1 > 3.5 and Stage 3.6 > 3.5 and Stage 3.1 > 2.5 and Stage 3.5 ≤ 3.5 and Stage 3.3 ≤ 3.5 and Stage 3.1 ≤ 3.5	Then, suitable for cluster_1 = 100%
5	If Stage 4.1 > 3.5 and Stage 3.6 > 3.5 and Stage 3.1 ≤ 2.5	Then, suitable for cluster_0 = 25.00%, and suitable for cluster_2 = 75.00%.
6	If Stage 4.1 > 3.5 and Stage 3.6 ≤ 3.5 and Stage 1.4 > 2.5 and Stage 1.3 > 2.5 and Stage 3.2 > 2.5 and Stage 3.5 > 3.5	Then, suitable for cluster_0 = 3.23%, suitable for cluster_1 = 16.13%, and suitable for cluster_2 = 80.65%.
7	If Stage 4.1 > 3.5 and Stage 3.6 ≤ 3.5 and Stage 1.4 > 2.5 and Stage 1.3 > 2.5 and Stage 3.2 > 2.5 and Stage 3.5 ≤ 3.5	Then, suitable for cluster_0 = 1.31%, suitable for cluster_1 = 92.16%, and suitable for cluster_2 = 6.54%.
8	If Stage 4.1 > 3.5 and Stage 3.6 ≤ 3.5 and Stage 1.4 > 2.5 and Stage 1.3 > 2.5 and Stage 3.2 ≤ 2.5	Then, suitable for cluster_0 = 100%
9	If Stage 4.1 > 3.5 and Stage 3.6 ≤ 3.5 and Stage 1.4 > 2.5 and Stage 1.3 ≤ 2.5 and Stage 1.4 > 4.0	Then, suitable for cluster_2 = 100%
10	If Stage 4.1 > 3.5 and Stage 3.6 ≤ 3.5 and Stage 1.4 > 2.5 and Stage 1.3 ≤ 2.5 and Stage 1.4 ≤ 4.0	Then, suitable for cluster_0 = 100%
11	If Stage 4.1 > 3.5 and Stage 3.6 ≤ 3.5 and Stage 1.4 ≤ 2.5	Then, suitable for cluster_0 = 100%
12	If Stage 4.1 ≤ 3.5 and Stage 4.4 > 4.5	Then, suitable for cluster_0 = 3.23%, suitable for cluster_1 = 3.23%, and suitable for cluster_2 = 93.55%.
13	If Stage 4.1 ≤ 3.5 and Stage 4.4 ≤ 4.5 and Stage 1.4 > 4.5 and Stage 3.2 > 3.5	Then, suitable for cluster_2 = 100%

Rule	Condition (If)	Prediction (Then)
14	If Stage 4.1 ≤ 3.5 and Stage 4.4 ≤ 4.5 and Stage 1.4 > 4.5 and Stage 3.2 ≤ 3.5	Then, suitable for cluster_0 = 100%
15	If Stage 4.1 ≤ 3.5 and Stage 4.4 ≤ 4.5 and Stage 1.4 ≤ 4.5 and Stage 2.3 > 4.5	Then, suitable for cluster_0 = 25.00%, and suitable for cluster_2 = 75.00%.
16	If Stage 4.1 ≤ 3.5 and Stage 4.4 ≤ 4.5 and Stage 1.4 ≤ 4.5 and Stage 2.3 ≤ 4.5 and Stage 1.2 > 3.5 and Stage 3.1 > 3.5	Then, suitable for cluster_0 = 37.04%, suitable for cluster_1 = 3.70%, and suitable for cluster_2 = 59.26%.
17	If Stage 4.1 ≤ 3.5 and Stage 4.4 ≤ 4.5 and Stage 1.4 ≤ 4.5 and Stage 2.3 ≤ 4.5 and Stage 1.2 > 3.5 and Stage 3.1 ≤ 3.5	Then, suitable for cluster_0 = 94.74%, and suitable for cluster_1 = 5.26%.
18	If Stage 4.1 ≤ 3.5 and Stage 4.4 ≤ 4.5 and Stage 1.4 ≤ 4.5 and Stage 2.3 ≤ 4.5 and Stage 1.2 ≤ 3.5	Then, suitable for cluster_0 = 97.52%, suitable for cluster_1 = 1.86%, and suitable for cluster_2 = 0.62%.

Correct: 1040 out of 1109 training examples (93.78%).

Table 8 shows the model test results using the collected data. It shows the validity and suitability of the model at 93.78%, which concludes nicely that the selected model is reasonable.

Table 9: Average within Centroid for each cluster

	Stage and Cluster			
	Stages	*Cluster_0*	*Cluster_1*	*Cluster_2*
Student's Context	*Student's Interests*			
	Stage 1.1	3.19	3.72	4.15
	Stage 1.2	2.94	3.62	4.23
	Stage 1.3	3.03	4.24	3.85
	Stage 1.4	3.11	3.81	4.16
	Stage 1.5	3.15	3.60	3.88
	Student's Characteristics			
	Stage 2.1	3.20	3.72	4.04
	Stage 2.2	3.15	3.52	3.98
	Stage 2.3	3.30	3.58	4.04
	Stage 2.4	2.81	3.06	4.03
	Student's Environment			
	Stage 3.1	2.87	3.48	4.09
	Stage 3.2	3.14	4.33	4.15
	Stage 3.3	2.94	2.88	3.81
	Stage 3.4	3.07	3.04	4.13
	Stage 3.5	3.00	2.67	3.83
	Stage 3.6	3.11	2.72	4.23
Program's Context	*Program's Context*			
	Stage 4.1	3.01	4.36	4.29
	Stage 4.2	3.54	4.18	4.33
	Stage 4.3	3.37	4.08	4.22
	Stage 4.4	3.23	3.67	4.17
Total Number of Items: 1109		*213 Items*	*443 Items*	*453 Items*

Table 9 shows the average within centroid for each cluster. The benefit of this study and Table 9 is that it can be used to recommend suitable educational institutions according to the learner interest clusters. In addition, a classification of interest clusters by educational institutions is presented in Table 10.

Table 10 shows the details of the data collected by classification for the groups analyzed in the study. It was discovered that clusters of members from the collected datasets were unclearly distributed. Therefore, in future research, the researcher should suggest research tools to improve the more balanced distribution of datasets.

Table 10: Details of Each Cluster from Data Collection

Institutions	Data Collection		
	Cluster_0	*Cluster_1*	*Cluster_2*
MSU: Maha Sarakham University	72 Items	171 Items	103 Items
RMU: Rajabhat Mahasarakham University	58 Items	223 Items	197 Items
UP: University of Phayao	83 Items	49 Items	153 Items
Total Number of Items: 1109	*213 Items*	*443 items*	*453 items*

However, the data collection were 1109 people who provided attitudes and satisfaction to the research. It can be concluded that the overall level of satisfaction with the study is accepted (average = 3.70, S.D. = 0.84), as shown in Table 3.

In addition, the prototype model has been developed for predicting and recommending the appropriate institutions for the learner, which has moderate accuracy levels (92.25%) shown in Table 4 and Table 6. Moreover, the results of the self-test data model are very accurate at the highest level, which is equal to 93.78% (Correct: 1040 out of 1109 training examples) as shown in Table 7. Therefore, it can be concluded that this study was successful. It can develop highly accurate models and also have high performance model testers.

5. Conclusions

From this research, the researcher can summarize according to the research objectives as follows: (1) to study the relevance of factors for the analysis of the effectiveness of suitable educational institutions, and (2) to identify the acceptance and tolerance of each attribute, which supports lifelong learning.

The conclusion was obtained from the study of the relevance of factors by using machine learning tools and data mining techniques in the development of prototype models. It was discovered that the developed model is highly effective, with the ability to summarize 11 predictive factors: stage 1.2, stage 1.3, stage 1.4, stage 2.3, stage 3.1, stage 3.2, stage 3.3, stage 3.5, stage 3.6, stage 4.1, and stage 4.4 as shown in Table 7. While studying factors acceptance, it was found that the overall level of acceptance was at a medium level (average = 3.70, S.D. = 0.84). When the researcher has considered all dimensions, it was found that all factors had been accepted, as shown in Table 3. With the research study by collecting a total data of 1109 people from three universities (Rajabhat Mahasarakham University, the Maha Sarakham University, and the University of Phayao), the researcher concluded that this research achieved its objectives with five machine learning tools and data mining techniques: k-means, k-medoids, decision tree, cross-validation methods, and confusion matrix.

For future research, the researchers aim to construct and develop applications that promote and support the findings of this research.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is supported by the two organizations: (1) The Faculty of Information Technology, Rajabhat Mahasarakham University, Maha Sarakham, Thailand. (2) The School of Information and Communication Technology, University of Phayao, Phayao, Thailand. The researchers would like to thank the advisor, lecturers, students, technicians, and all respondents for their entire support.

References

[1] U.N. Desa, "Transforming our world: The 2030 agenda for sustainable development," 2016.

[2] W. Nuankaew, P. Nuankaew, S. Bussaman, P. Tanasirathum, "Hidden academic relationship between academic achievement and higher education institutions," in 2017 International Conference on Digital Arts, Media and Technology (ICDAMT), 308–313, 2017, doi:10.1109/ICDAMT.2017.7904982.

[3] F. Bellotti, R. Berta, A. De Gloria, E. Lavagnino, A. Antonaci, F. Dagnino, M. Ott, M. Romero, M. Usart, I.S. Mayer, "Serious games and the development of an entrepreneurial mindset in higher education engineering students," Entertainment Computing, **5**(4), 357–366, 2014.

[4] S. Fungchomchoei, U. -mapor. Kardkarnklai, "Exploring the intercultural competence of Thai secondary education teachers and its implications in English language teaching," Procedia-Social and Behavioral Sciences, **236**, 240–247, 2016.

[5] A. Navy, "Higher Education Reform In Thailand: Toward Quality Improvement And University Autonomy," KATHA-The Official Journal of the Centre for Civilisational Dialogue, **9**(1), 1–16, 2013.

[6] M. Flavin, "Disruptive technologies in higher education," Research in Learning Technology, **20**, 2012, doi:10.3402/rlt.v20i0.19184.

[7] S. Gallagher, G. Garrett, "Disruptive education: Technology-enabled universities," The United States Studies Centre at the University of Sydney, 2013.

[8] P. Nuankaew, P. Temdee, "Online Mentoring Model by Using Compatible Different Attributes," Wireless Personal Communications, **85**(2), 565–584, 2015, doi:10.1007/s11277-015-2755-x.

[9] P. Nuankaew, P. Temdee, "Determining of compatible different attributes for online mentoring model," in 2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace Electronic Systems (VITAE), 1–5, 2014, doi:10.1109/VITAE.2014.6934434.

[10] K. Pupara, W. Nuankaew, P. Nuankaew, "An institution recommender system based on student context and educational institution in a mobile environment," in 2016 International Computer Science and Engineering Conference (ICSEC), 1–6, 2016, doi:10.1109/ICSEC.2016.7859877.

[11] P. Nuankaew, "Dropout Situation of Business Computer Students, University of Phayao," International Journal of Emerging Technologies in Learning (IJET), **14**(19), 115–131, 2019.

[12] P. Nuankaew, W. Nuankaew, K. Phanniphong, S. Imwut, S. Bussaman, "Students Model in Different Learning Styles of Academic Achievement at the University of Phayao, Thailand," International Journal of Emerging Technologies in Learning (IJET), **14**(12), 133–157, 2019.

[13] W.S. Nuankaew, P. Nuankaew, D. Teeraputon, K. Phanniphong, S. Bussaman, "Perception and Attitude Toward Self-Regulated Learning of Thailand's Students in Educational Data Mining Perspective," International Journal of Emerging Technologies in Learning (IJET), **14**(09), 34–49, 2019.

[14] P. Nuankaew, W. Nuankaew, K. Phanniphong, R. Fooprateepsiri, S. Bussaman, "Analysis Dropout Situation of Business Computer Students at University of Phayao," in: Auer, M. E., Hortsch, H., and Sethakul, P., eds., in The Impact of the 4th Industrial Revolution on Engineering Education, Springer International Publishing, Cham: 419–432, 2020, doi:10.1007/978-3-030-40274-7_42.

[15] P. Nuankaew, W. Nuankaew, T. Thamma, "The Recommended System for the Relationship between Educational Programs and Stu-dents' Interests," in International Conference on Digital Arts, Media and Technology, 2–3A–4: 34, 2016.

[16] J. Witte, "Education in Thailand after the crisis: A balancing act between globalization and national self-contemplation," International Journal of Educational Development, **20**(3), 223–245, 2000.

[17] N. Omar, M.M. Mohamad, A.N. Paimin, "Dimension of learning styles and students' academic achievement," Procedia-Social and Behavioral Sciences, **204**, 172–182, 2015.

[18] P.R. Medwell, P.N. Grimshaw, W.S. Robertson, R.M. Kelso, "Developing sports engineering education in Australia," Procedia Engineering, **34**, 260–265, 2012.

[19] A. Rezvanfar, M. Ghorbanian, F. Shafiee, "An investigation of the behaviour of agricultural extension and education engineering students in Tehran University towards employability," Procedia-Social and Behavioral Sciences, **152**, 65–69, 2014.

[20] S. Rongraung, S. Kanokorn, "Soft skills for private Ba," Procedia-Social and Behavioral Sciences, **112**, 956–961, 2014.

[21] W. Techataweewan, U. Prasertsin, "Development of digital literacy indicators for Thai undergraduate students using mixed method research," Kasetsart Journal of Social Sciences, **39**(2), 215–221, 2018.

[22] P. Nuankaew, W. Nuankaew, P. Temdee, "Institution recommendation using relationship optimisation between program and student context," International Journal of Higher Education and Sustainability, **2**(4), 279–302, 2019.

[23] R. Wirth, J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Citeseer: 29–39, 2000.

[24] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc, **9**, 13, 2000.

[25] G. Hamerly, J. Drake, Accelerating Lloyd's algorithm for k-means clustering, Springer: 41–78, 2015.

[26] H.-S. Park, C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," Expert Systems with Applications, **36**(2), 3336–3341, 2009.

[27] Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, **20**, 53–65, 1987, doi:10.1016/0377-0427(87)90125-7.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1137-1146 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Feature Extractors Evaluation Based V-SLAM for Autonomous Vehicles

Mounir Amraoui[*,1], Rachid Latif[1], Abdelhafid El Ouardi[2], Abdelouahed Tajer[3]

[1]*LISTI, ENSA Ibn Zohr University, Agadir, 80000, Morocco*

[2]*SATIE Paris-Saclay University, Orsay, 91405, France*

[3]*LISA, ENSA Cadi Ayyad University, Marrakech, 40140, Morocco*

A R T I C L E I N F O

Article history:
Received: 07 August, 2020
Accepted: 08 October, 2020
Online: 22 October, 2020

Keywords:
V-SLAM
Feature Extraction
Detectors
Descriptors
Bio-inspired algorithms
Front-end processing

A B S T R A C T

Visual Simultaneous Localization and Mapping known as V-SLAM, is an essential task for autonomous vehicles. It can be carried out using several sensors, in particular with on board cameras. To locate a vehicle, SLAM algorithms are based on two main tasks. The first task (front-end kernel) is intended to process images in order to provide features (called also landmarks or primitives) of the perceived environment. The second task (back-end kernel) is intended for localization and environment reconstruction.
This work focuses on the front-end task which uses extractors (detectors and descriptors) in a stereo-vision system. Several feature detectors and descriptors exist in the state of the art. The aim of this paper is to evaluate the possible combinations of detectors and descriptors to achieve a precise localization while considering the processing times. The study is extended to bio-inspired extractors. The evaluation is achieved with SLAM algorithms over well-known indoor and outdoor datasets. Experimental results highlight the performance of bio-inspired extractors and their potential integration in designing vision systems for real-time SLAM applications.

1. Introduction

Simultaneous localization and mapping (SLAM) [1] are a complex task aiming to reconstruct a map during a vehicle motion in parallel while localizing its position. This problem can be solved by using complicated algorithms, where many threads should be executed simultaneously, since the need is to recognize at the same time the vehicle pose and process previously detected landmarks for pose estimation and mapping. It is known that in ordinary circumstances a map cannot be achieved without knowing the precise position, while to know it, a map with landmarks is needed. The pose recognition is done by using feature extractors, that are a combination between feature detectors and descriptors. A SLAM algorithm can use input data issued from one or many sensors, to estimate the robot position. However, for a higher accuracy, a fusion of different sensors data is required.
Vision SLAM systems use a camera to detect a maximum of key Points (KP). Some software approaches are then developed to improve the quality of detected features, by enhancing description and detection algorithms. Both these tasks are the main ingredient for back-end kernels (localization, mapping and building environment) of any real-time vision SLAM application.

Despite some weakness like sensitivity to light, vision SLAM systems need a powerful architecture for image processing. This could be improved based on a hardware software co-design.

Back-end and front-end algorithms, both dependent as shown in figure 1, are continuously object of improvement and will keep being, as well as all other researches related to the embeddability on heterogeneous architectures and parallel implementation. The works in [2, 3] are among many similar studies, an example where the SMG-SLAM and EKF-SLAM, were accelerated on a field-programmable gate array (FPGA).

Most of feature extractors are mainly based on existing algorithm but are solving their drawbacks or improving their weaknesses, to be more robust against environment changes.

There are many ways to determine a key-point, also called a landmark in image processing depending on its type, since it could be a corner, a blob, an edge or a ridge. However, a recent survey [4] related to advances on feature extraction and description algorithms, show that the most used method is the corner detection, due to its algorithmic moderated complexity and the improved calculation adopted formulas, which has an impact on the number of memory access and CPU usage. Corners, called interest points, can be obtained by intersection of at least two edges.

[*]Corresponding Author: Mounir Amraoui, LISTI, ENSA Ibn Zohr University, E-mail: mounir.amraoui@edu.uiz.ac.ma

The corner detector is based on the edge detector. It is represented by a set of pixels that have rapid change in direction. Each of these pixels have at least two dominants and various directions of the corresponding point in a local area.

The quality of a corner detector can be evaluated by applying above criterions on the same corner, for multiple images and under different environment effects such as light changes, rotation, translation, and image resolution.

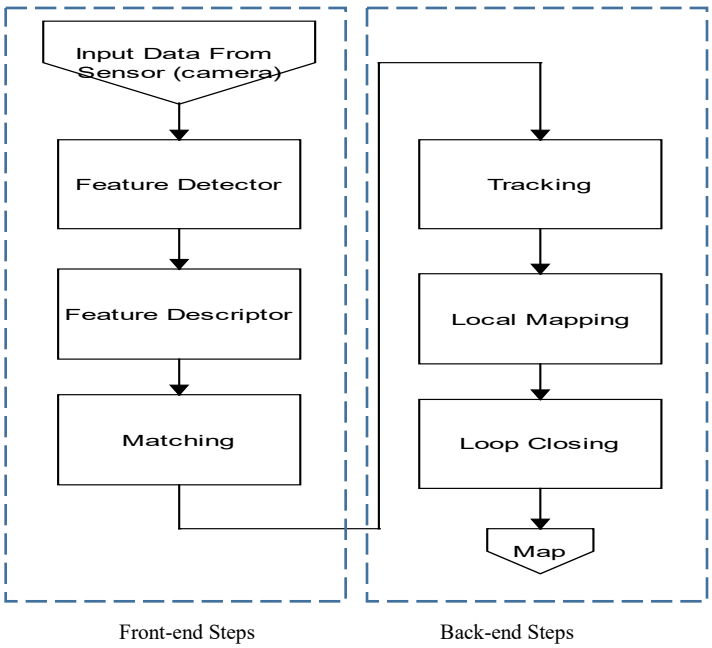


Figure 1: Front-End / Back-end main sequence for V-SLAM

As per [5] Harris represented by equation (1), was evaluated as the most stable, with the lower computational complexity compared to other corner detectors, such as Shi-Tomasi [6] and Forsner [7].

$$f(\Delta x, \Delta y) \approx (\Delta x, \Delta y)\, M \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \quad (1)$$

Where, in a given 2-dimensional grayscale image I, $(x, y)\epsilon\, w$ (window) are the image patch, $(\Delta x, \Delta y)$ is the shift and M is a Matrix structure tensor represented by:

$$M = \sum_{(x,y)\epsilon w} \begin{bmatrix} IxIx & IxIy \\ IxIy & IyIy \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{(x,y)\epsilon w} IxIx & \sum_{(x,y)\epsilon w} IxIy \\ \sum_{(x,y)\epsilon w} IxIy & \sum_{(x,y)\epsilon w} IyIy \end{bmatrix} \quad (2)$$

Hence based on [5], for lower threshold values Harris detects many features together with noisy data from the image, which is considered as a weakness. Since Harris cannot set multiple threshold values for each image feature detection, it needs to be enhanced.

To solve such issue an accelerated segment test model, mathematically simple, was developed and proved with better performance, since it can determine if the interest point is a corner or not, only by evaluating the point neighborhood. Based on this concept some very known detector was developed, such as Feature Accelerated Segment Test (FAST) [8], which is also considered as an enhancement of SUSAN [9] and Harris corner detectors.

There are other different approaches for corner detection, that can detect interest points using a genetic programming for automatic image operator synthesis, like the one developed by Trujillo and Olague [10] in 2011. This method can be competitive, but it is more used for object recognition and need to be trained based on a bag of features.

To perform the desired task in an optimal way, the representation of the complete initial image data must be reduced by eliminating redundant features. So, a descriptor is applied once the feature detection is done, to provide complementary attributes such as gradient, magnitude and orientation.

Several descriptors like SIFT, SURF, BRIEF, ORB or BRISK has been successfully applied for tracking and object detection tasks, but most of pervious evaluation works were related to the computational complexity of algorithms.

In this work, our contribution is a study of detectors and descriptors to select the suitable extractor combination dedicated to the V-SLAM. The study considers constraints of processing times and precision for autonomous vehicles. It is extended to bio-inspired extractors and their integration into a back-end V-SLAM system.

The following section is related to the state-of-the-art and the essential concepts, to highlight and explain the reasons behind each selection and present also advantages and drawbacks of the chosen feature descriptors and detectors. Then, the evaluation methodology will be presented in section 3. Section 4 shows the results related to the processing times and precision of the evaluated extractors. It presents also front-end experimental tests which lead us to select an extractor for a full SLAM evaluation (back-end evaluation). The last section gives a general holistic view point and a conclusion.

2. Related Work

2.1. Feature detectors

FAST: based on AST (Accelerated Segment Test) and stands for Feature Accelerated Segment Test, is a corner detector that exceeds the other detectors in both computational performance and repeatability as per [11]. This method consists on the examination of the intensity for a given central pixel p in a radius r circle following a mathematical representation, where the intensity value Ip of the corresponding pixel p and a threshold t are used in three intervals:

$$S_{p\to x} = \begin{cases} d, & I_{p\to x} \le I_{p-t} & (darker) \\ s, & I_{p-t} < I_{p\to x} < I_{p+t} & (similar) \\ b, & I_{p+t} \le I_{p\to x} & (brighter) \end{cases} \quad (3)$$

where, $S_{p\to x}$ is the state, $I_{p\to x}$ is the intensity of the pixel x and t is a threshold.

There is a way to make this process faster by evaluating *n* successive pixels from the circle. Based on the brightness and darkness of these pixels, the evaluated pixel can be identified as a feature or not.

Authors in [11] has optimized the algorithm to achieve higher computational efficiency with improved processing speed, by defining the order in which the test is done and applying a non-maximum suppression (NMS), to pixels that have successfully verified the segment test. The idea is mathematically represented by:

$$V = max \begin{cases} \Sigma(pixel\ values - p)\ if\ (value - p) > t \\ \Sigma(p - pixel\ values)\ if\ (p - value) > t \end{cases} \quad (4)$$

In this equation, *p* refers the central pixel, *t* is a threshold and "*pixel values*" correspond to the *N* contiguous pixels in the corresponding circle.

The AST processing speed depends mainly on the pixel from which the evaluation is starting first. Therefore, the decision tree can be computed based on the distribution of the learnt set of the corner configuration for a given environment.

AGAST: Adaptive and Generic Accelerated Segment Test, is a corner detector developed by Elmar Mair et al. [12] where a binary decision tree is computed. This method is common and does not require any adaptation to new environments. However, it is more memory consuming due to the number of memory access to weight the various pixels. The decision tree is optimal for a given probability of similar pixels in the AST mask.

Also, the corner detector is automatically adapted to the environment by combining two trees, to make, with only one-pixel delay, the optimal decision tree for the image region. Hence, it results in a corner detector that does not need a training task, while maintaining the same corner repeatability and response as FAST.

By varying the nucleus (*Ps)*, the decision tree can be balanced. The system is then represented as follow:

$$S_{n\to x} = \begin{cases} d, & I_{n\to x} < I_{n-t} & (darker) \\ \bar{d}, & I_{n\to x} \not< I_{n-t} \wedge S'_{n\to x} = u & (notdarker) \\ s, & I_{n\to x} \not< I_{n-t} \wedge S'_{n\to x} = \bar{b} & (simillar) \\ \bar{s}, & I_{n\to x} \not> I_{n+t} \wedge S'_{n\to x} = \bar{d} & (simillar) \\ \bar{b}, & I_{n\to x} \not> I_{n+t} \wedge S'_{n\to x} = \bar{b} & (notbrighter) \\ b, & I_{n\to x} > I_{n+t} & (brighter) \end{cases} \quad (5)$$

where, $S'_{n\to x}$ is the previous state, *I* is the pixel brightness and *u* refers to an unknown state.

Therefore, when the environment switches at consecutive pixels from homogenous to structured, AGAST can be less efficient than FAST. However, it is not possible practically due to the mirroring effect of dissimilar pixels, also because a random brightness distribution does not exist on a natural image.

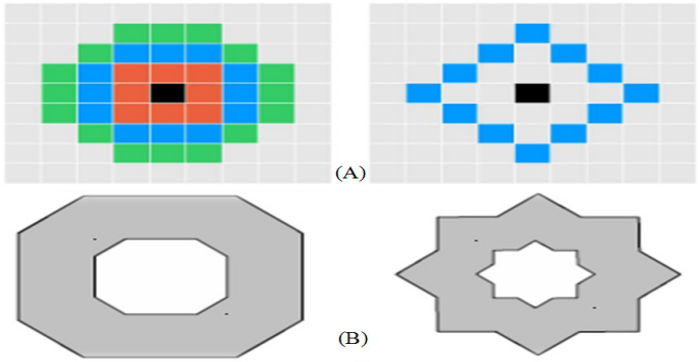
SURF: Stands for Speed Up robust Feature, it is inspired and developed based on SIFT [13]. The main motivation of this feature detector and descriptor is to overcome SIFT's low processing time and computational complexity. SURF [14] has been reported to be up to a few times faster than SIFT without compromising the performance.

For a given pixel, SURF calculate first the Haar-wavelet responses in x and y-direction, around the key-point in a circular neighborhood of radius 6. Integral images are also used on big scales wavelets for fast filtering. The second step in the process is to calculate the main orientation of the feature descriptor, by making the sum of vertical and horizontal wavelet responses in a scanning area, then changing the scanning orientation by adding π/3, and re-calculating until to find the largest sum value.

STAR: It is distinguished from SIFT and SURF which find extrema at sub-sampled pixels that consequently reduces accuracy of feature localization at larger scales. This feature detector is a derivative from CenSurE (Center Surround Extrema) feature detector [15]. It was integrated to the well-known Open source Computer Vision library.

Aiming to create a multiscale detector with full spatial resolution, this detector uses a bi-level approximation of the Laplacian of Gaussians (LoG) filter. So instead of the simple circular shape, the mask is represented by a circle with 2 overlapping squares: 1 upright and 1 45-degree rotated, resulting in a polygon with thick borders, to be invariant to rotations and to enable the use of integral images for efficient computation (see figure 2). This configuration is computationally better compared to the other scale-space detectors and presents real-time implementation possibilities.

The masks shape used in feature detection for each method is given by Figure 2.



ORB: Stands for Oriented FAST and Rotated BRIEF (Binary Robust Binary Robust Independent Elementary Features) [16]. The ORB detector is developed based on FAST to find key-point then applies Harris corner detector measuring method to give the top N points among them. But since FAST does not compute the orientation, the authors in [16] came up with enhancing modifications.

The ORB detector creates a multiscale image pyramid, with versions of image at different resolutions, where each level in the pyramid contains a subsampled version than the previous level. Once this process is finished, FAST detector is used to detect key-points at each level, and therefore ORB is partial scale invariant.

2.2. Features descriptors

BRIEF: Stands for Binary Robust Independent Elementary Features. This feature descriptor as per [17] uses a binary subsequent matching, with Hamming distance method to compute the descriptors similarities. This algorithm is much faster and computationally efficient.

Due to the high sensitivity to noise, BRIEF performs a smoothing to the image before applying the actual descriptor, using a simple averaging filter.

The value of each contributing bit to the descriptor, is given by the comparison of intensity values of two points inside an image segment centered on the currently described feature. The value of this bit is equal to 1 if the first point intensity of this pair is higher than the second point intensity value, otherwise it is equal to 0.

A few sampling strategies were tested by the authors to select the point pairs, show that sampling according to the Gaussian distribution centered on the described feature point, results in best performances.

For BRIEF, the initial smoothing is done with a 9x9 pixels rectangular averaging filter, the image patch is a 48x48, and the proposed descriptor is 512-bit long. The size of the window can be modified to be fitted to the application.

ORB: The Oriented FAST and Rotated BRIEF descriptor, uses a modified version of BRIEF (Binary Robust Binary Robust Independent Elementary Features).

Since Standard BRIEF descriptor performs weakly on rotation, ORB drives BRIEF depending on the orientation of key-points (KP's). The algorithm constructs a lookup table of pre-computed BRIEF patterns, to get a rotated version S_θ, where the angle is discretized by an incrementation of 12° degrees. then use the correct set of points S_θ to compute its descriptor.

BRISK: The Binary Robust Invariant Scalable Key-point (BRISK) [18], is a key point detector and descriptor inspired by AGAST [12] and BRIEF [17]. It uses AGAST for detection, which is an upgrading of FAST in term of processing time while preserving the similar detection performance.

To achieve scale invariance, the key-point is detected using a non-maxima suppression and interpolation across all scales space pyramid. To describe the features, authors have used symmetric patterns instead of using learned or random patterns.

For long-distance and to determine orientation, several sample point comparisons were used. The relative difference in intensity of the displacement vector is weighted and stored. These vectors are weighted. The computed average is used to find the major gradient direction of the patch.

FREAK: Fast Retina Key point [19], inspired from the biological human retina. The detection concept is based on summing the estimated local gradients over selected point pairs to provid the feature orientation.

To allow applying coarser discretization of rotation, a specific biologically inspired point sampling pattern is used which also results on saving memory space.

To reject false matches and accelerate the computation time, the feature description used is BREIF-based binary string. This allows a sampling pattern approach by comparing the most distinctive characteristics feature neighborhood to the point pairs carrying the information.

HOOFR: Hessian ORB-Overlapped FREAK [20] is a FREAK descriptor-based method combined with the ORB detector, with enhancements on the detection algorithm to speed up the process, improve memory allocation and reliability, the matching process timing is then reduced because the descriptor size is 256 instead of 512 compared to FREAK. HOOFR has three steps:

1- The first step consists on extracting key points from training data and building the description: a matrix M is created using all possible pairs where each key-point has its own descriptor. The number of key-points is equivalent to rows and the size of descriptor is equivalent to the number of columns.

2- The second step aims to define the variance value of the binary distribution, by computing the average for each column limited between 0 and 1. A mean value of 0.5 indicates the maximum variance desired to have a discriminant feature.

3- Finally, the third step consist on processing all columns to maintain only 256 columns that have the greatest variances.

ZEON: Defined by H. Angrish [21] as a new technique combining features of, SURF [14], BRISK [17] and FREAK [19], to overcome each method drawback. The main purpose of this improvement is to increase the extraction robustness and accuracy. This new proposed technique is intended to use features of SURF and BRISK as well as the detection capabilities of FREAK as a first step. This will provide a certain edge and will not add massive time strains to the new method timing.

3. Evaluation Methodology

First, we will develop and set many combinations of different feature detectors and descriptors. All extractors will be evaluated in terms of processing time and accuracy for different set of images of well-known outdoor dataset, using different lighting condition and at different camera rates. Results from the first step will be analyzed to select only the best candidate that could be used in a real time SLAM system for automotive application. For this reason, we will consider the first three high score combinations taking into consideration the compromise between processing time and accuracy. At the end, we will make a full implementation of the selected extractors with a V-SLAM algorithm, to be able to evaluate the behavior of each extractor and define the most robust solution versus back-end constraints related to rotation, filtering, type of landmarks, lighting, sensor noise or ground truth. The evaluation methodology will also take indoor scenes.

4. Experimental Results

4.1. Front-End Evaluation

The adopted method for front-end process evaluation, is using a calibrated monocular camera to get image input, focusing on the number of detected features, the processing time of each extractor combination and the matching time.

To make an objective comparison, the same image is used where the calibration settings described in Table 1 are similar for each set of images given by KITTI benchmark [22].

The hardware configuration shown in Table 1.

Table 1: Hardware and software configuration

CPU: Intel i5 core ™ @ 2.60 GHZ GPU : NVIDIA GeForce GTX 1060 Max-Q 6GiB RAM: 12 GB	
Scale factor	1.6
Initial threshold	20
Scale level	8
Min threshold	7
Max Feature	1000

The number of features is set to a fix value number to limit the image maximum interest points that can be found by using a descriptor, 1000 is the highest number found during experiment, so it is set to be a reference value to make the comparison and to calculate the detector accuracy. All values are given after calculating the means. Figure 3 gives examples of proceeded images with their respective resolutions.



(A) Street: Image resolution: 1392 x 512 pixels. Labels: 1 Cars, 0 Vans, 0 Trucks, 0 Pedestrians, 0 Sitters, 0 Cyclists, 0 Trams, 0 Misc.



(B) Cars: Image resolution: 1392 x 512 pixels. Labels: 5 Cars, 2 Vans, 1 Trucks, 0 Pedestrians, 0 Sitters, 0 Cyclists, 0 Trams, 0 Misc.

Figure 3: Outdoor images of KITTI dataset [22] used for evaluation.

Some extractors can be used, in the same time, as detectors and descriptors like ORB and SURF. Therefore, the description task is done in parallel with the detection task. From the other hand and since it is expected that the extractors will not have the same number of detected features, the processing time will be higher for the ones with high number of detections. So, the given value will consider the total processing time divided by the number of detected features as presented in Table 2 and Table 3.

Table 2: Processing time evaluation for different combinations of detectors-descriptors vs number of detected features using the image from figure 3 (A).

	Image (A): Street		
Descriptor	Detector	Number of detected features	Total detection and description processing time per feature (ms)
FREAK	FAST	956	0.046
	STAR	643	0.12
	ORB	213	0.291
	SURF	684	0.409
	AGAST	950	1.065
ORB	FAST	946	0.037
	STAR	660	0.106
	SURF	967	0.278
	ORB	500	0.036
	AGAST	939	1.399
BRISK	FAST	988	0.041
	STAR	660	0.126
	ORB	402	0.239
	SURF	896	0.355
	AGAST	980	1.357
BRIEF	FAST	951	0.035
	STAR	660	0.12
	ORB	500	0.144
	SURF	973	0.197
	AGAST	944	1.292
ZEON	FAST	3353	0.018
	STAR	654	0.124
	ORB	371	0.396
	SURF	971	0.346
	AGAST	3382	0.374
SURF	FAST	8115	0.049
	STAR	660	0.127
	ORB	500	0.57
	SURF	1334	0.31
	AGAST	8217	0.177

Table 3: Processing time evaluation for different combinations of detectors-descriptors vs number of detected features using the image from figure 3 (B)

	Image (B): Cars		
Descriptor	Detector	Number of detected features	Total detection and description processing time per feature (ms)
FREAK	FAST	960	0.05
	STAR	741	0.09
	ORB	206	0.30
	SURF	691	0.47
	AGAST	957	1.39
ORB	FAST	953	0.05
	STAR	760	0.10
	SURF	969	0.25
	ORB	500	0.19
	AGAST	947	1.45
BRISK	FAST	988	0.07
	STAR	760	0.11
	ORB	414	0.21
	SURF	897	0.28
	AGAST	984	1.49
BRIEF	FAST	956	0.04
	STAR	760	0.09
	ORB	500	0.15
	SURF	973	0.25
	AGAST	952	1.57

ZEON	FAST	982	0.06
	STAR	753	0.11
	ORB	373	0.37
	SURF	862	0.38
	AGAST	980	1.37
SURF	FAST	1000	0.06
	STAR	760	0.13
	ORB	500	0.55
	AGAST	1000	0.40

As shown in the above tables, it is clear that FAST as feature detector, is the most improved among all evaluated detectors because it provides the lower processing time with the maximum points of interest. Nonetheless, the total time is evaluated as the mean of the sum of all processing times needed for a feature detection and description for several iterations. Having a high number of detected features necessarily leads to a higher processing time needed for matching process. To evaluate this effect, the matching time and accuracy are then calculated according to the following formula:

$$\%Accuracy = \frac{\Sigma(True\ Positives + True\ negatives)}{\Sigma(\ Total\ cases)} * 100 \quad (6)$$

Table 4 gives the time spent per feature to have the corresponding match, based on image (B) from figure 3.

Table 4: Matching time using Brut force for SURF, and Hamming distance for the remaining descriptors

Descriptor	Detector	Number of matched Features	Matching time (ms)	% Acc
FREAK	FAST	869	0.139	90.52%
	STAR	683	0.035	92.17%
	ORB	193	0.049	93.69%
	SURF	638	0.055	92.33%
	AGAST	875	0.079	91.43%
ORB	FAST	891	0.061	93.49%
	STAR	720	0.064	94.74%
	SURF	885	0.040	91.33%
	ORB	471	0.038	94.20%
	AGAST	885	0.046	93.45%
BRISK	FAST	896	0.138	90.69%
	STAR	702	0.103	92.37%
	ORB	388	0.048	93.72%
	SURF	852	0.068	94.98%
	AGAST	897	0.113	91.16%
BRIEF	FAST	896	0.056	93.72%
	STAR	722	0.049	95.00%
	ORB	482	0.038	96.40%
	SURF	938	0.040	96.40%
	AGAST	894	0.065	93.91%
ZEON	FAST	870	0.155	88.59%
	STAR	681	0.086	90.44%
	ORB	348	0.054	93.30%
	SURF	632	0.101	73.32%
	AGAST	590	0.063	60.20%
SURF	FAST	845	0.187	84.50%
	STAR	658	0.120	86.58%
	ORB	463	0.062	92.60%
	AGAST	408	0.162	40.80%

Based on above result, a complete evaluation is conducted for front-end vision SLAM system where, we can conclude that if we seek for speed and low processing time, BRIEF-FAST is the best descriptor-detector combination to be selected. However, if the priority is given to higher accuracy, the BRIEF-ORB, and BRIEF-SURF is those that must be chosen. Also, as it is seen, the FAST algorithm is better in terms of processing time and number of detected features than all actual known detectors. Therefore, any evaluation will depend strongly on the associated descriptor and matching process. However, from another side, ORB descriptor, is the best descriptor in term of accuracy.

Additionally, the evaluation of the bio-inspired FREAK descriptor has proven that it has the lowest matching time performance, beside good results on both accuracy and processing time.

To narrow down and make the choice of the most adapted combination for automotive application, we must look for a combination that fulfil both the minimum processing time and the highest accuracy, which will leave us with 3 extractor combinations as shown below in Table 5. The remaining combination will be ignored, because they do not fulfill requirement and could lead to incorrect final mapping results.

This choice can be justified by the fact that any automotive vision SLAM application must be real time. Therefore, a low processing time has the advantage, but it does not mean that the precision can be high. The best combinations are the ones shown in the Table 5.

Table 5: High score combination leading to a compromise between the maximum accuracy percentage and the minimum processing time

Extractor	Total detection and description processing time per feature (ms)	Matching time (ms)	Total Processing time (ms)	% Acc
BRIEF-FAST	0.04	0.056	0.096	93.72%
ORB-FAST	0.05	0.061	0.111	93.49%
FREAK-STAR	0.09	0.035	0.125	92.17%

In a similar way to BRISK and BRIEF, FREAK uses also a binary string, but it has an optimized false matching algorithm which reduces the processing time during matching process, giving the advantage to be used for real time V-SLAM applications.

The second evaluation step is to apply the selected extractors to a full V-SLAM algorithm to evaluate the behavior considering back-end constraints (localization and mapping kernel).

4.2. Back-End Evaluation

For an objective evaluation, we implemented an algorithm based on ORB-SLAM2 and modify it to incorporate additional detectors and descriptors functional blocks. We used the already prepared class from OpenCV.

The matching method is based on the Brut force matcher which uses Hamming method to calculate distance. The intended stereo-vision system has three main parallel threads with an additional loop executed also in a separate thread.

The used datasets are an indoor sequence from Tum, fr2/desk [23] running at 10 FPS, and an outdoor sequence from the well-known KITTI dataset [24] running at 30 FPS. The reason behind this choice is because these datasets has a denser co-visibility graph and therefore the local map contains more key frames and points to evaluate object reconstruction and SLAM/odometry methods under different texture, illumination and structure conditions. The sequence is schematized in Figure 4. The obtained results are given by Figure 5.

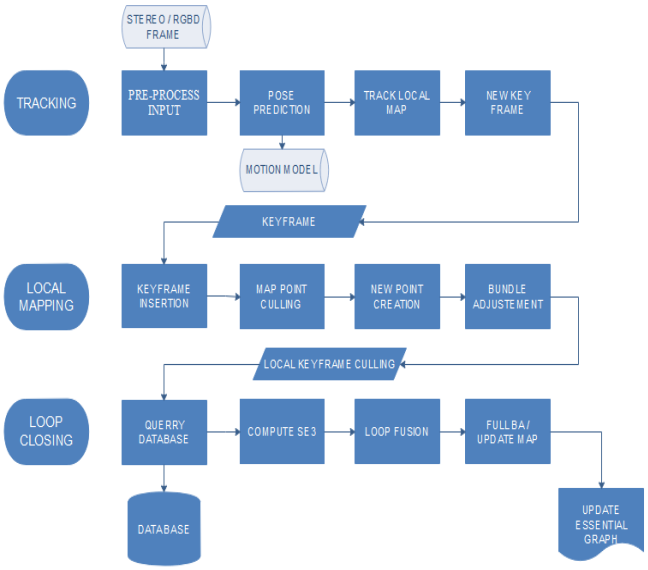


Figure 4: V-SLAM threads executed in parallel

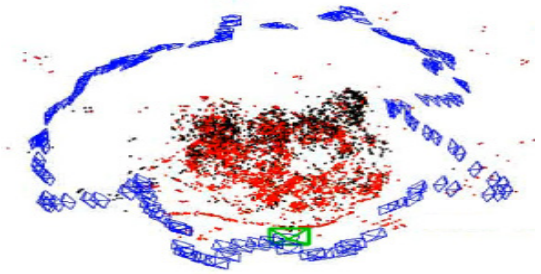
where:

- **Tracking:** a separate thread where every frame is localized by extracting feature and matching it to the local map. A local Mapping loop is then applied.
- **Local Mapping**: minimize the reprojection error, optimize and manage the local map, by performing a bundle adjustment (BA) loop.
- **Loop Closing**: it enhances the pose-graph by correcting the accumulated drift, within large loops.

To allow the system to continue loop detection and creating the map, a fourth thread of Full bundle adjustment, also called Full BA [25] is launched when the optimization process is aborted, and loop closing is finished.

Once the full BA finishes, the updated subset of key frames and points are merge and optimized, with the non-updated key frames and points that where inserted while the optimization was running. This is done by propagating the correction of updated key frames.

Indoor mapping results are shown in Figure 6 with the resulting path compared to the ground truth, that is provided initially with the dataset. Camera calibration data are also provided by [23].



(a)

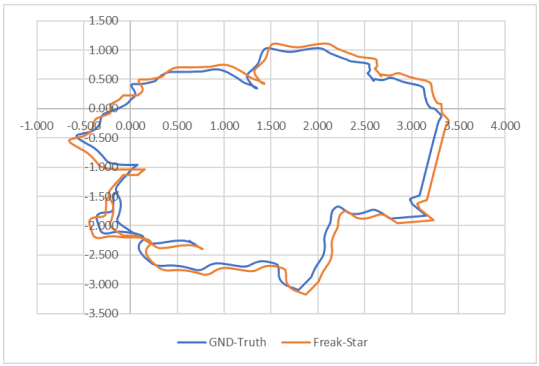


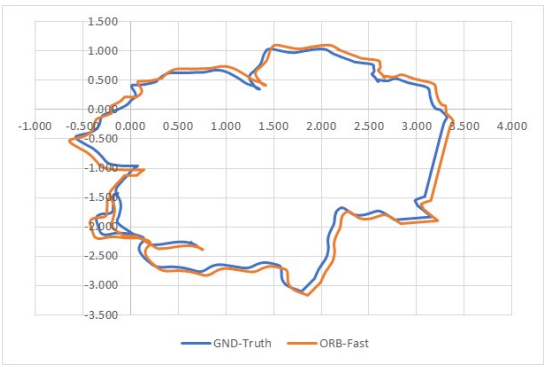
(b)



(c)

Figure 5: Dense point reconstruction from estimated keyframe poses: (a), sensor (camera) pose, (b) maps in TUM, fr2/desk. (c), KITTI_00 dataset tracked by the visual odometry system indicated in blue.





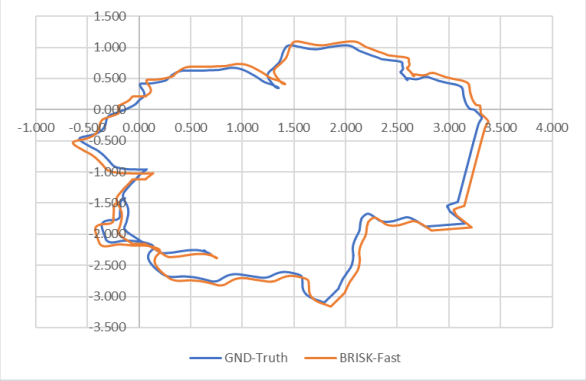


Figure 6: Pose-graph corresponding to indoor dataset, from up to down, for BRIEF-FAST, ORB-FAST, and FRAEK-STAR with comparison to ground-truth for TUM, fr2/desk

Bellow, the Figure 7 shows the evaluation using the outdoor dataset where the resulting path is compared to the ground truth given by [24].

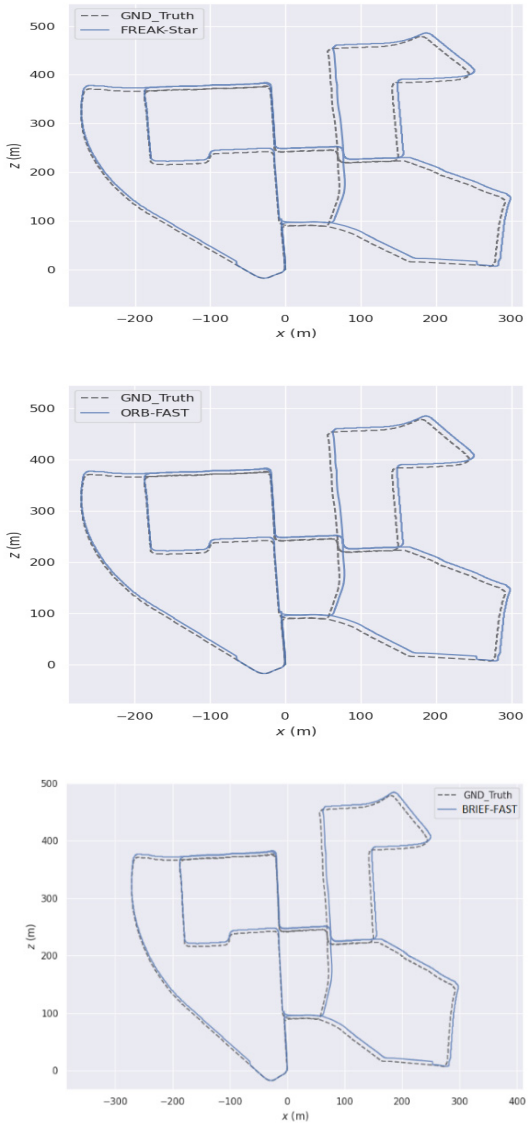


Figure 7: Pose-graph corresponding to outdoor dataset for BRIEF-FAST, ORB-FAST and FRAEK-STAR from up to down with comparison to ground-truth for KITTI dataset 00.

To compare both reconstructions, we used both metrics: the absolute trajectory error (ATE) and relative pose error (RPE) since they directly measure the difference between the ground-truth and the generated trajectory when using visual SLAM algorithms.

The estimated poses are associated to ground truth poses using the timestamps, then aligned using singular value decomposition. Finally, the difference between each pair of poses is computed. The output value is the mean, median, and the standard deviation of these differences, the RPE calculate the rotational errors. Results are given in Table 5 and Table 6 bellow.

Table 6: Translational and rotational errors calculated based on relative pose error (RPE) and absolute trajectory error ATE for 10000 pose pairs for TUM, fr2/desk indoor dataset.

FPS = 10		FREAK-STAR	ORB-FAST	BRISK-FAST
Translational Error (m)	RMSE	0.050	0.044	0.429
	Mean	0.045	0.028	0.027
	Median	0.047	0.017	0.017
	STD	0.021	0.034	0.033
	Min	0	0	0
	Max	0.103	0.148	0.143
Rotational Error (deg)	RMSE	9.85	3.73	3.618
	Mean	3.52	1.5	1.51
	Median	0.014	0.011	0.010
	STD	9.206	3.394	3.285
	Min	0	0	0
	Max	35.629	18.801	18.223

Table 7: Translational and rotational errors calculated based on relative pose error (RPE) and absolute trajectory error ATE for KITTI-00 outdoor dataset.

FPS = 30		FREAK-STAR	ORB-FAST	BRISK-FAST
Translational Error (m)	RMSE	8.623	7.914	6.707
	Mean	8.364	7.680	6.510
	Median	7.223	6.633	5.633
	STD	2.096	1.908	1.616
	Min	0.000	0.000	0.000
	Max	10.212	9.377	7.963
Rotational Error (deg)	RMSE	0.538	0.505	0.429
	Mean	0.294	0.278	0.250
	Median	0.264	0.250	0.227
	STD	0.451	0.422	0.349
	Min	0	0	0
	Max	14.079	12.562	10.161

Indoor dataset is considered as an environment with low lighting variations, where a robot is moving in a low speed. Additionally, the camera is held by a human, so the rotational effects are also much more important and unexpected. While the outdoor dataset is an environment that presents better other constraints like high speed, lighting changes and lower camera rotation, because it is fixed in the same position during the vehicle motion.

Considering above results, it is crystal clear that BRISK-FAST introduces the minimum rotational and translational error. Hence it has a better performance in term of trajectory precision in both shorter and longer distances.

From the other hand, FREAK-STAR extractor performance is acceptable for short distances and in environment with low variations, since it has the minimum standard deviation with

lowest maximum translational error, but the rotation can be considered as a weakness compared to the ORB-FAST and BRISK-FAST extractors.

To summarize, even if all above three methods results are very close to each other, related to timing performance, a bio-inspired extractor based on FREAK has a low complexity since it uses a binary string like BRISK and BRIEF. It has an optimized matching algorithm reducing processing time. It is suitable to be implemented on a GPU or FPGA due to the number of threads that could be executed in parallel. For these reasons a parallel implementation will give the advantage to the bio-inspired algorithm and make it the more suitable for a real time Visual SLAM application.

5. Discussion and Conclusion

FAST is still till today the most adjusted feature detector despite of its known drawbacks. The reason behind this statement is that, as per above results, the other descriptors were outperformed in terms of processing time and the number of detected features. Hence, due to this fact any selection of a descriptor merged with FAST will lead to the minimum detection time. However, this descriptor used in a SLAM application must run together with a descriptor and a matching process, which is affecting the global timing performance, as also demonstrated in this paper.

The matching process is a task that must be considered when selecting the combination of detector/descriptor, since the algorithm could lose more time during this step especially when the dataset has much more Key-points that need to be matched. This is clearly shown during experiments done in the current work, where pros and cons of every detector/descriptor combination are analyzed.

Additionally, a proof of consistency, accuracy and speed is made in this study with regards to the bio-inspired extractors, that still need to be improved to have lower rotational and translational error and the processing time. A new descriptor called HOOFR [20] try to improve theses weaknesses and an evaluation of this descriptor combined with different detectors in a full SLAM system could be achieved as a perspective with a parallel implementation on a heterogeneous architecture. This could be a solution leading to promising results.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is supported by Information Systems and Technologies Engineering Laboratory, LISTI. *ENSA Ibn Zohr University, Agadir, Morocco* and *SATIE Paris-Saclay University, Orsay, France*.

References

[1] H. Durrant-Whyte, T. Bailey, "Simultaneous localization and mapping: part I", IEEE Robotics & Automation Magazine. **13** (2), 99-110, 2006. doi:10.1109/mra.2006.1638022. ISSN 1070-9932

[2] D. T. Tertei, J. Piat and M. Devy, "FPGA design and implementation of a matrix multiplier-based accelerator for 3D EKF SLAM," 2014 International Conference on ReConFigurable Computing and FPGAs (ReConFig14), Cancun, 2014, pp. 1-6, doi: 10.1109/ReConFig.2014.7032523

[3] G. Mingas, E. Tsardoulias, L. Petrou "An FPGA implementation of the SMG-SLAM algorithm" Microprocessors and MicrosystemsVolume 36, Issue 3, May 2012, 190-204.

[4] E. Salahat, "Recent Advances in Features Extraction and Description Algorithms: A Comprehensive Survey" arXiv:1703.06376v1 [cs.CV] 19 Mar 2017.

[5] J. Sánchez, N. Monzón and A. Salgado "An Analysis and Implementation of the Harris Corner Detector". Image Processing on Line. 8: 305–328, 2018. doi:10.5201/ipol.2018.229

[6] H. Kadhim, Araheemah, W. "A Comparative Between Corner-Detectors (Harris, Shi-Tomasi&FAST) in Images Noisy Using Non-Local Means Filter". Journal of Al-Qadisiyah for Computer Science and Mathematics, 11(3), Comp Page 86-93, 2019.

[7] A.S. Eltanany et al., "Key Point Detection Techniques". In: Hassanien A., Shaalan K., Tolba M. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019. AISI 2019. Advances in Intelligent Systems and Computing, **1058**. Springer, Cham, 2020.

[8] E. Rosten "Machine learning for high-speed corner detection », in 9th European Conference on Computer Vision, Graz, Autriche, **7**(13), 430-443, 2006.

[9] X. Chen, et al. Corner detection and matching for infrared image based on double ring mask and adaptive SUSAN algorithm. Opt Quant Electron **50**, 194, 2018.

[10] L. Trujillo and G. Olague, "Automated design of image operators that detect interest points". Evolutionary Computation. **16** (4): Archived from the original on 2011-07-17, 2008.

[11] R. Edward; R. Porter; T. Drummond "FASTER and better: A machine learning approach to corner detection". IEEE Transactions on Pattern Analysis and Machine Intelligence. **32** (1): 105–119. arXiv:0810.2434, 2010. doi:10.1109/TPAMI.2008.275. PMID 19926902..

[12] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in Proceedings of the 11th European Conference on Computer Vision: Part II, Heraklio, 183–96, 2010.

[13] M. Safdari, P. Moallem and M. Satari, "SIFT Detector Boosted by Adaptive Contrast Threshold to Improve Matching Robustness of Remote Sensing Panchromatic Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **12**(2), 675-684, Feb. 2019, doi: 10.1109/JSTARS.2019.2892360.

[14] X. Qu, "Evaluation Of Sift And Surf For Vision Based Localization "International Archives of the Photogrammetry, Remote Sensing & Spatial InformationSciences . **41**(B3), 685-692, 2016.

[15] M. Agrawal, K. Konolige, M.R. Blas, "CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In: Forsyth D., Torr P., Zisserman A. (eds) Computer Vision – ECCV 2008. ECCV 2008. Lecture Notes in Computer Science, **5305**. Springer, Berlin, Heidelberg, 2008.

[16] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," 2011 International Conference on Computer Vision, Barcelona, 2011, 2564-2571, doi: 10.1109/ICCV.2011.6126544.

[17] M. Calonder, Lepetit, V., Strecha, C., Fua, P.: BRIEF: binary robust independent elementary features. ECCV'10 Proceedings of the 11th European Conference Computer Vision: Part IV, 2010.

[18] S. Leutenegger, M. Chli and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," 2011 International Conference on Computer Vision, Barcelona, 2011, 2548-2555, doi: 10.1109/ICCV.2011.6126542.

[19] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. doi: 10.1109/ICECCE47252.2019.8940756.

[20] D. Nguyen, A. El Ouardi, E. Aldea and S. Bouaziz, "HOOFR: An enhanced bio-inspired feature extractor," 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, 2977-2982, 2016.

[21] H. Angrish. Navjot Kaur. Raman Chadha." A Novel Technique to Detect Local Features in Digital Images" IJSTE - International Journal of Science Technology & Engineering **2**(3), September 2015.

[22] A. Geiger and Philip Lenz and Christoph Stiller and Raquel Urtasun," Vision meets Robotics: The KITTI Dataset", International Journal of Robotics Research (IJRR), 2013, Geiger2013IJRR.

[23] F. Steinbruecker, J. Sturm and D. Cremers, In Workshop on Live Dense Reconstruction with Moving Cameras at the Intl. Conf. on Computer Vision (ICCV), 2011.

[24] A. Geiger and Philip Lenz and Raquel Urtasun "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite", Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[25] D. P. Frost, O. Kähler and D. W. Murray, "Object-aware bundle adjustment for correcting monocular scale drift," 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, 2016, 4770-4776, doi: 10.1109/ICRA.2016.7487680.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1147-1153 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Effects of Resting Actions Using Smart Toys During Break Times on Concentration in E-learning

Takashi Ito[1], Kenichi Takahashi[*,2], Tomoko Kajiyama[2]

[1]*Aoyama Gakuin University, Department, Institute of Industrial and Systems Engineering, Kanagawa, 252-5258, Japan*

[2]*Hiroshima City University, Department of Intelligent Systems, Hiroshima, 731-3194, Japan*

ARTICLE INFO

Article history:
Received: 24 August, 2020
Accepted: 15 October, 2020
Online: 22 October, 2020

Keywords:
Refresh
Resting action
Smart toys
Effect
Break times

ABSTRACT

This study conducted experiments to investigate the effects of resting actions of e-learners during break times on keeping and improving the concentrations of learners in e-learning. Two smart toys (a dog-type robot "aibo" and a toy drone) were used for the resting actions. Two types of experiments were conducted to examine the effects: one is a work type experiment (simple mathematical calculation) and the other is a memorizing type experiment (learning of English words). In those experiments, the learners played with one of the smart toys during the resting times between the learning sessions and performed certain actions accordingly. The experimental results were combined with the previous study using a humanoid robot "RoBoHoN." A questionnaire was also employed to investigate the feeling of the learners. The experiments showed that the "playing with aibo" action refreshed the learners effectively in the mathematical experiment, while the "playing with RoBoHoN" action was effective in the learning experiment.

1. Introduction

In recent years, e-learning has been actively used with the development of ICT [1]. E-learning is the education and learning way that combines the Internet with PCs and mobile terminals. With e-learning, learners can deeply understand the content by learning iteratively and utilizing videos and voices. Also, learners can take advantage of good features of e-learning, such as learning at any time and place and furnishing functions for learners to select learning contents suitable to their knowledge level and to repeat lessons, as well as the variety of devices used for e-learning, such as tablets and smartphones. As a result, learning effect by e-learning is expected to be higher than the conventional learning using paper media. Additionally, with e-learning systems, the management and utilization of learning data become easy.

On the other hand, since e-learning is mainly used for individual learning, there is no instructor present and there is no opportunity that learners communicate with instructors and other students. The stimulation from instructors and friends is small, which can decrease their willingness to learn gradually. Thus, it is not easy for learners to keep motivation and concentration high, and learners can easily get bored to abandon learning, which is one of the important problems to be solved for e-learning [2].

In order to challenge the problem, various systems have been proposed, e.g., a system with a function to praise and scold students with images according to their learning performance, a system that measures student concentration and learning states during learning using Kinect, a system introducing game-based learning, intelligent collaborative e-learning systems, an e-learning platform, and portfolio assessment in elearning environment [3-6].

Another important factor is to provide appropriate break times in order to maintain motivation and concentration of learners. A study investigated the effect on student learning performance by introducing a break during a 90-minute lecture at a university [7]. It showed that taking a short break of about 8-10 minutes improved the learning performance: After the break, drowsiness of students decreased and their concentration increased. In addition, for VDT (visual display terminal) work, a study examined the lengths of an appropriate work time and the break time in order to reduce the burden on the mind and body and to improve work performance [8]. They suggested that one session of continuously working on a VDT is less than one hour and a 10–15 minute break between working sessions is desirable. Also, work on the way how office workers using VDTs spent their break times influenced their working efficiency was published [9], and Japan's Ministry of Health, Labor and Welfare recommends occupational health management guidelines regarding the use of

[*] Corresponding Author: Kenichi Takahashi, Hiroshima City University, +81-82-848-2872, 6961643@hiroshima-cu.ac.jp

VDTs [10]. Further, apps for supporting learners in the break time has been proposed [11]. The authors consider that it is important to introduce appropriate resting actions during break times for efficient and effective learning in e-learning that usually uses the VDT equipment.

In our research, a web-based e-learning system equipped with functions to improve or maintain learning motivation and concentration has been developed, such as a ranking function, a function that praises or scolds students appropriately with images and a sound, and a function for learners to present problems to one another [3]. In addition, a learning system equipped with games to investigate the effect of fun and relaxation of the games has been developed [12]. The games are one that refreshes learners by quick physical motion (shooting flies on a screen) and one that relaxes learners by taking care of a dog without heavy physical movement. These studies showed that refreshing by playing games during the break times between learning sessions is effective on improving learners' will to learn and maintaining the degree of concentration during e-learning. Furthermore, to examine what kinds of resting actions during break times keep or increase the concentration, we have conducted mathematical calculation experiments and English word memorizing experiments previously [13], [14]. The resting actions(behaviors) during the break times are: "playing the beat-fly game," "playing the dog-care game," "reading a book," "listening to music," and "doing nothing." We have showed that the type of these resting actions influences the degree of concentration.

Additionally, there was a study that introduced new resting actions and investigated the influence of the learner's motivation for learning and concentration. In this study, we performed the same experiments using two types of works as in our previous study: "simple work" (the learners solve continuously simple numeric calculation with four arithmetic operations for 15 minutes per work session) and "learning work" (they memorize English words for 30 minutes per learning session). In the study, we introduced the interaction with communication robots, "RoBoHoN" and "aibo", as new resting actions [15]. The study confirmed that the break time using "playing with aibo" had a good relaxing effect, and had a positive effect on simple learning such as numeric calculation with arithmetic operations. However, it is not appropriate for memorization learning such as English word learning, while a break time using "playing with RoBoHoN" gave better results.

In this study, we employ resting actions using a "toy drone" and aibo, and investigate how they affect the learning motivation, concentration, and learning results. The resting action using aibo in this study was the same as that in the previous study [15]. We also conducted two types of experiments for e-learning: simple work in which students try to calculate the numeric questions of four arithmetic operations continuously for 15 minutes and memorizing in which students try to remember new English words for 30 minutes. After the experiments, the learners were asked to answer a questionnaire and respond to statements, in order to examine how they felt the effect the resting actions had on their concentration in the experiments. We discuss the effect of the resting actions after combining these experimental results with the results in our previous study [15].

2. Smart Toys

Recently with the development of artificial intelligence technology, various robots become popular. For example, a robot can make dialogue and conversation with humans. We call those robots smart toys. As these smart toys become popular, they are used for children's learning, while the privacy and security problems occurs [16-18].

This study employed smart toys in order to refresh the feeling of the learners by the resting actions. In the experiments, the learners perform different types of resting actions during break times: "playing with a toy drone" (in which the learners manipulate a toy drone to finish a toy course), "playing with RoBoHoN" (in which the learners play five kinds of games with a humanoid robot), and "playing with aibo" (in which the learners play with a dog-shaped robot by talking and touching).

2.1. RoBoHoN

"RoBoHoN" is a fourth-generation mobile communication robot developed by Sharp Corporation [19]. RoBoHoN works as a smartphone and has functions to talk with users and to make actions through the original apps. In our experiments, using software called "Scratch Pack for RoBoHoN," four actions were programmed so that the learners can play with RoBoHoN during break times. We refer to these actions together as "playing with RoBoHoN."

The four actions used for RoBoHoN are as follows.

1) Raise a flag: RoBoHoN speaks a command such as "raise a red (or white) flag", and the learner tries to obey the command correctly.

2) Answer a riddle: RoBoHoN speaks randomly one of five programmed riddles, and the leaner tries to answer it.

3) Guess who-I-am: RoBoHoN pretends to be a thing, a person, or an animal, and gives hints, such as "I am a humanoid". The learner guesses what or who RoBoHoN is from those hints spoken by RoBoHoN. The game continues giving other hints until the learner hits the correct answer or the number of hints becomes a certain number.

4) Dance: RoBoHoN can dance randomly one of more than 20 dances with music.

2.2. aibo

A pet robot "aibo" is developed by Sony Corporation [20]. It can walk on four legs and behaves like a puppy. It can recognize persons and changes its behavior gradually as the person interacts it frequently. The latest version of "aibo" employs artificial intelligence techniques and network functions such as expressing emotions and online backup of data. The first "aibo" was developed in 1999, and now new "aibo" developed in 2017 was on sale. With the introduction of artificial intelligence, the technology has improved the robot to show more puppy-like gestures and reactions.

In our previous research, we conducted experiments based on the idea that it will be effective to interact with "aibo", a robot that imitates a puppy, as well as to interact with actual small animals, since a certain improvement effect on e-learning was obtained [21]. In this study, we also adopted "aibo" for one of the resting actions from the same viewpoint. We asked the learners to interact with "aibo" during the break times between sessions. "aibo" performs various actions (behaviors) in response to human words. Specific keywords cause "aibo" to behave as if a dog behaves such

as "hands" and "sitting", and unique behaviors such as "dance" and "high touch". In addition, there are a special ball and a bone toy that can be used for play. "aibo" can learn tricks gradually by being praised with the words "good" or "good dog" and by being stroked in the part of the head, chin, and the back of its body.

2.3. Toy Drone

A drone refers to a remote-controlled or autonomous unmanned aerial vehicle. In the present, it is actively utilized for industrial use in various scenes.

Toy drones are inexpensive to buy, often have a small and lightweight body, and are simple to operate. With these features, even a beginner can easily operate the drone. The toy drone "Tello" developed by Ryze Technology Co. was used in this research [22]. We can use a smartphone as a controller to operate it by connecting the dedicated app to the drone. It can also be operated by connecting the app to a dedicated controller via Bluetooth. A small camera mounted on the front of the drone constantly captures the scenes in front of the drone, which are displayed in real time on the screen of the smartphone. The flight mode can be selected by the app. Acrobatic flight and various aerial photography functions can be easily executed.

In this study, we adopted a resting action called "playing with a toy drone" for comparison with the resting actions of different types of smart toys: "aibo" and "RoBoHoN." They were used for resting to refresh the mood by communicating with the robot. On the other hand, the "playing with a toy drone" action presents "fun" by performing precise operations of the drone, which requires concentration.

We asked the learners to operate the toy drone during the break times between work or learning sessions. A dedicated controller linked to the app is used to operate the drone. We can operate the drone more easily by linking the app with the controller. In order to enable the learners to operate the drone in the 3-minute break time, we completed various connections between the app and the controller in advance. In addition, before the experiments, we distributed the document summarizing how to operate the drone to the learners. Since the experiments were conducted indoors, flight mode functions were limited by eliminating dangerous acrobatic flights.

We constructed a drone course to have the learners perform precise operations. A gimmick of arches, a tunnel, and a slalom were placed for the drone to finish the course in about 1 to 2 minutes per lap. As for the arches, we prepared two patterns: one with multiple arches placed in curve side by side, and the other with multiple arches placed in different altitudes on the floor and a chair. The tunnel was made with multiple arches placed at narrow intervals in the same altitude. The slalom was made of rods of different lengths so that the drone goes forward in zigzags avoiding rods. Figure 1 shows the course used for the experiment. The course shown in Figure 1 is a basic form. We changed the difficulty of the course by changing the layout if the learner desired.

3. Experiments Incorporating Resting Actions

In this Section, we describe the results of the mathematical experiment and the English-word learning experiment, and discuss how the different types of resting actions affect learning motivation and maintenance of concentration. We estimate the effects of the resting actions on simple work (mathematical calculation) and learning work (memorizing English words) through the experiments. In the experiments, ten male students majoring computer science and belonging to our laboratory in our university served as participants and were divided into 2 groups of 5 students. The each group performed one of the resting actions during break times, but it did the same resting action in the mathematical experiment and the learning experiment to examine the effect of the same resting action on different types of work. In addition, we compared and discussed the learning results of this research with the results of the previous study using "aibo" and "RoBoHoN." We employed the correct-answer rate and the average number of answers as metrics for the mathematical experiment, and test score improvement as the metric for the learning experiment. We divided the learners to one of the groups through preliminary experiments so that the ability of each group is roughly equal.

3.1. Mathematical Experiment

The mathematical experiment is composed of three work sessions and two breaks between the work sessions. Figure 2 shows the flow of the experiment. The work session is 15 minutes long, and the resting time is 3 minutes. Figure 3 shows an example of mathematical calculation. The learner continuously calculates similar problems for 15 minutes in a work session. The mathematical experiment provides simple work in which learners tackle continuously to calculate simple numeric equations for 15 minutes. The numeric questions to be solved were produced by the automatic formula generation program of the system.

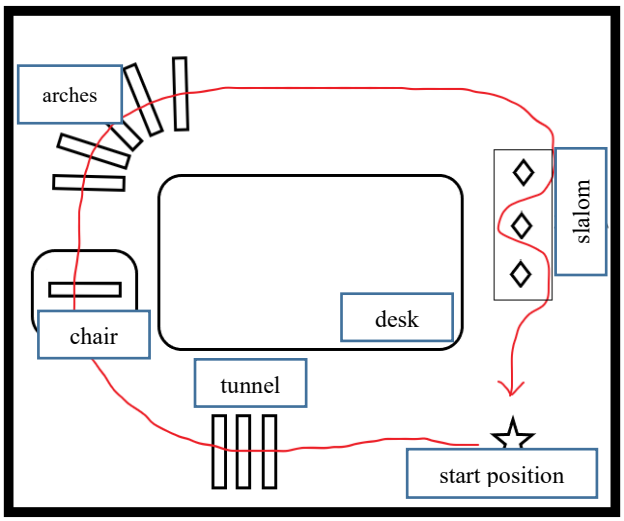


Figure 1: The basic course used for the toy drone

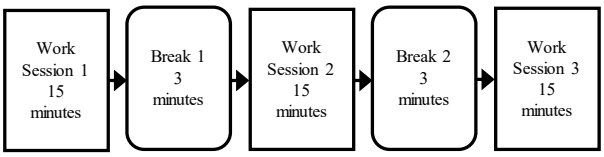


Figure 2: The flow of the mathematical experiment

Figure 4 illustrates the average correct-answer rates for the three actions. As for the aibo group, the result of the previous work [17] was incorporated with the result of the experiment in this study. From Figure 4, it can be seen that all groups tend to increase

the correct-answer rate from the first to second work sessions and to decrease the correct-answer rate from the second to third work sessions. Among them, the drone group shows that the degree of the tendency is large, whereas that of the RoBoHoN group is small. The correct-answer rate of the aibo group behaves between the RoBoHoN group and the drone group. Considering the decrease of concentration due to fatigue in the latter half of the work sessions, the average correct-answer rate of the drone group is significantly decreased in the second to third work sessions because the resting action using the drone operations heavily needs the concentration, while the average correct-answer rate of the RoBoHoN group is gently decreased. Also, the increase of the correct-answer rate of the RoBoHoN group from the first to second work sessions is small. Regarding the aibo group, the overall average correct-answer rate is low, while the tendency is between the RoBoHoN group and the drone group. Thus, we can say that the relaxing/refreshing effect of the resting action using the drone that influences learning motivation is low, because it needs the tense concentration.

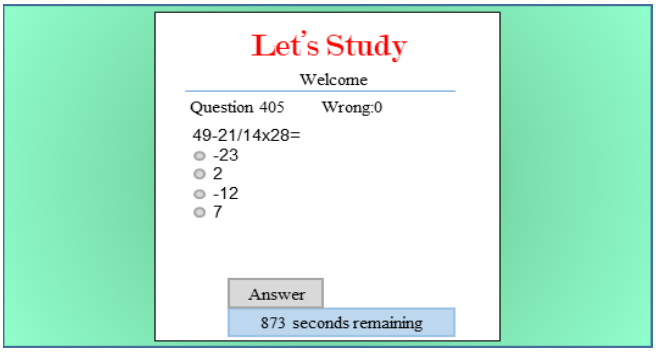


Figure 3: An example of mathematical calculation

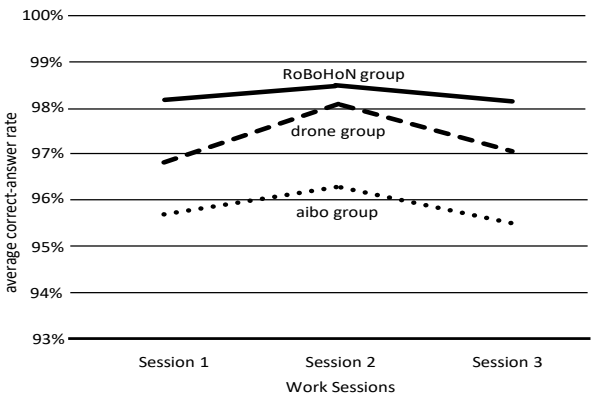


Figure 4: Correct-answer rates for RoBoHoN, aibo, and drone groups in the mathematical experiment

Figure 5 shows a graph summarizing the number of correct answers and the number of incorrect answers. Comparing the average number of answers shows that for all groups the number of answers in the second session is larger than that in the first one. Note that the increase of the number of answers by the drone group was large. However as the session proceeded, the learners became tired, and the number of answers decreased from the second to third sessions. Comparing the number of incorrect answers indicates that the numbers of the answers for all groups decreased except for the second to third sessions of the aibo group as the experiment proceeded, which suggests that the effect of the resting actions on learning was positive. Looking at Figure 5, it can be said that difference of the number of answers among the sessions is smallest for the aibo group.

In the mathematical experiment, the correct-answer rate for "playing with RoBoHoN" was a little higher than that for "playing with aibo," while the number of answers for "playing with aibo" was larger, indicating that "playing with aibo" was more effective to keep improve concentration. The authors believe that the resting behavior using aibo could provide a high refreshing/relaxing effect, and it can be considered as effective in improving learning motivation while maintaining moderate concentration.

3.2. Learning Experiment

In the English-word learning experiment, the learners were requested to memorize 200 English words for 30 minutes. In the experiment, two 30-minute learning sessions were held, and a 3-minute break was placed between the learning sessions, as shown in Figure 6. In the experiment, we asked the learners to learn 200 English words provided by the e-learning system. A post-learning test was administered to check to see whether they memorized English words or not, immediately after the second session. An example of a question for English word learning is shown in Figure 7. Note that the group of the resting action was the same group as the mathematical experiment in order to verify the effectiveness of the resting action.

Both the pre-learning test and the post-learning test consist of 50 words. The pre-learning test was performed about a week before the experiment to prevent the learners from memorizing the English words in the test. The score of the pre-learning test was used as a baseline for comparison.

We employed the score increase rate to evaluate the learning effect of the resting actions. The increase rate was calculated as follows, where IR is the score increase rate, P1 is the score on the pre-learning test, and P2 is the score on the post-learning test. The full score of each test was 50.

$$\mathrm{IR} = \frac{P2-P1}{50-P1} \times 100(\%) \qquad (1)$$

Figure 8 shows a graph summarizing the average increase rate of each group. We can see from Figure 8 that the RoBoHoN group had the highest average increase rate. In the memorization problem, it is necessary to remember the contents memorized in the learning sessions before the break time, and thus the resting action with high relaxation and refreshing effect is not appropriate, since the learners tend to forget the content. The authors considered that resting actions with a moderate refreshing effect is suitable. Also the positive effect was obtained for the aibo group in this experiment because the average increase rate was positive. On the other hand, the drone group has the lowest average increase rate. The resting action using the drone needs tense concentration for operating the drone, and had the high refresh effect, indicating that the resting action using the drone is not suitable.

Therefore, we can conclude that RoBoHoN-based resting action is suitable and effective for memorization learning such as English-word learning, because the score increase rate for the "playing with RoBoHoN" group was higher.

3.3. Questionnaire and Summary

At the time the learners finished each of the mathematical experiment and the learning experiment, we conducted a questionnaire. The questionnaire includes descriptive questions

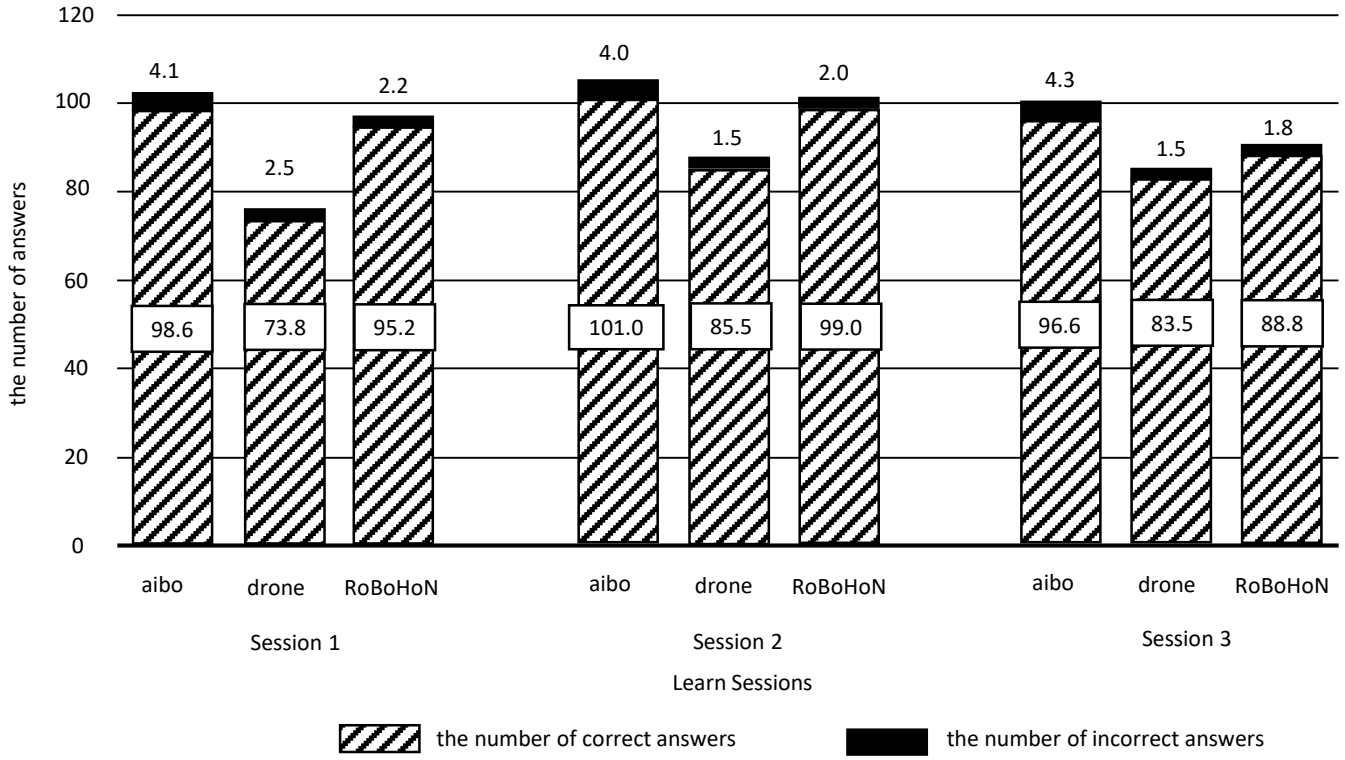


Table 1: Average Scores on Survey

Statement	RoBoHoN group		aibo group		drone group	
	Math	Eng	Math	Eng	Math	Eng
The game-playing with the smart toy during breaks were effective.	3.8	3.4	4.1	4.3	3.2	2.8
I felt refreshed after taking a break.	4.0	3.0	4.1	4.1	3.8	3.4
Playing with the smart toy improved my concentration.	3.6	3.0	3.7	3.8	3.4	2.8

and three selective questions. In the selective questions, the learners select a score on a five-point scale (5: strongly agree --- 1: strongly disagree). The questionnaire was conducted to investigate how the learners felt the effect of the resting actions.

Table 1 summarizes the evaluation results of the questionnaire obtained from aibo, drone, and RoBoHoN groups in the mathematical and the learning experiments. We can see from Table 1 that the aibo group gave the highest evaluation score for each question in both the mathematical experiment and the learning experiment, which shows higher refreshing and concentration effect than other groups did. The evaluation for the learning experiment was higher than the mathematical experiment, indicating that the resting action by aibo gave more refreshment in the learning experiment than in the mathematical experiment. We can see in Figure 5 that the number of answers was larger than those of the other groups, while in Figure 4 the correct-answer rate is the lowest, but the difference was less than 3%. Many learners said that they regarded aibo as a robot rather than as an animal. In the limited resting time, aibo didn't behave as ordered by the learners, and that may be stressful to them. We also see in Figure 8 that the score increase rate was positive, and the resting action worked well.

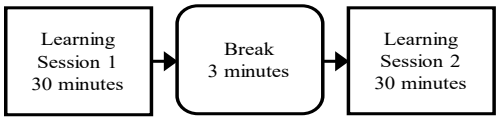


Figure 6: The flow of English-word learning experiment

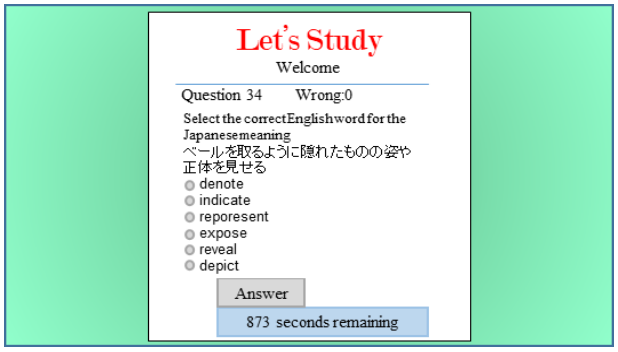


Figure 7: An example of an English-word question

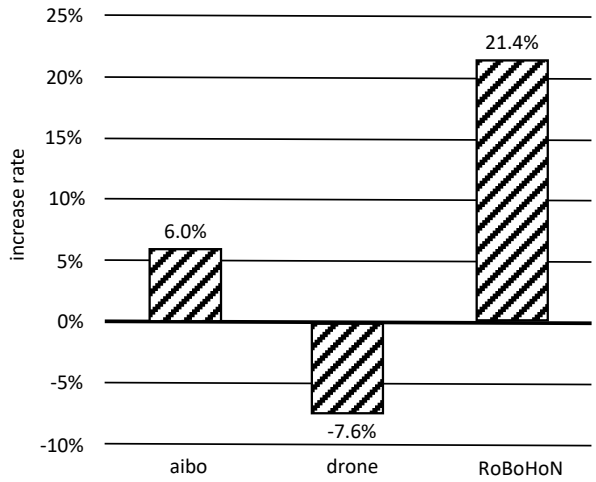


Figure 8: The average increase rates for RoBoHoN, aibo, and drone groups

The RoBoHoN group has the second highest rating scores. In this group, the evaluation of the mathematical experiment is also higher than that of the English experiments. We see in Figure 5 that the number of answers in the mathematical experiment was the second highest, while the increase rate in Figure 8 was highest in the English learning experiment, which shows that the resting action was effective in the learning experiment. The drone group has the lowest evaluation scores. They were refreshed in the mathematical experiment, while they gave evaluation score 2.8 and didn't concentrate in the learning experiment. The "playing with the drone" action needs concentration for precise operations, and that makes the learners feel stress. Thus, the authors considered that "playing with the drone" was not appropriate for the English word problem.

3.4. Summary

The "playing with aibo" action showed the best results as the resting action for the simple work regarding keeping and improving concentration, while the "playing with RoBoHoN" action shows the highly rated results in the English learning experiment.

The authors considered that the reason why the "playing with aibo" action was effective for the mathematical experiment was the high refreshment effect of the dog-shaped robot. The learners had to continue calculating simple mathematical problems for 15 minutes in the mathematical experiment and the "playing with aibo" highly refreshed the learners. On the other hand, in the English learning experiment, the "playing with RoBoHoN" action gained the highest score increase rate, although the evaluation scores of the "playing with aibo" group on the questionnaire were higher than those of the "playing with RoBoHoN" group. Thus, the feeling of refreshment by the learners themselves during the learning was different from the experimental results. Additionally, the "playing with RoBoHoN" action using language communication that has the moderate refreshing effect seems more suitable than the "playing with aibo" action that completely refreshed one's feeling. The "playing with a drone" behavior needs tense concentration on operation of the drone, and it was not appropriate for e-learning.

Future task includes further experiments with more participants of variety types and investigation from psychological point of view to evaluate the effect of the resting actions accurately, and the improvement of the learning system. It also includes experiments with other methods for relaxation than smart toys and discussion about the effect of those methods comprehensively in association with the type of learning.

4. Conclusion

In this study, we investigated the improvement of learner's concentration to learn and the learning will by using smart toys (either aibo or a toy drone) for e-learning, and discussed the effect by integrating the previous study using RoBoHoN . We compared the effect of the resting actions using aibo, RoBoHoN, and a toy drone among smart toys, since they have different features. We chose the functions that are considered to be effective for learning as the resting behavior and explained them so that the learners who touch the toys for the first time can operate the toys as they want. We found that that the "playing with aibo" action was effective for simple work like mathematical calculation, because the action had a well-balanced effect of improving concentration and learning motivation. Also we found that the "playing with RoBoHoN" action provides a moderate relaxing effect and was effective for English word memorizing. The "playing with a toy drone" action was not appropriate as a resting action, since some participants felt frustration to operate the drone.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank Mr. Masaki Ueta for collaboration on this work. This research was supported by a Grant-in-Aid for Scientific Research JP16K00485.

References

[1] Proseeds Inc., Merits and demerits of introducing e-learning, http://www.pro-seeds.com/trend/meritdemerit.html, Accessed in Aug. 2019 .

[2] A. Tominaga, C. Kogo, "Development and issues in Practical studies of e-learning", The Annual Report of Educational Psychology in Japan, **53**, 156–165, 2014 .

[3] M. Takeue, K. Shimada, K. Takahashi, M. Inaba, "Experiments of displaying images to keep the motivation in e-learning," Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 120–125, 2012, doi:10.1109/ICSMC.2012.6377687.

[4] T. Ito, S. Kotake, K. Takahashi, I. Inaba, M. Inaba, "Detecting Concentration of Students Using Kinect in E-learning," 450–456, 2017, doi:10.12792/icisip2017.082.

[5] T. Ito, K. Kamiya, K. Takahashi, M. Inaba, "A method for identification of students' states using kinect," Advances in Intelligent Systems and Computing, **935**, 341–350, 2019, doi:10.1007/978-3-030-19063-7_28.

[6] R. Takaoka, M. Shimokawa, T. Okamoto, "Special Section on Knowledge-Based Software Engineering A Development of Game-Based Learning Environment to Activate," **4**, 911–920, 2012.

[7] A.Yamazaki, "Effect of introducing small breaks during class hours on health promotion and learning performance," Journal of humanities and sciences, Nihon University, **24**(1-3), 33-58, 2019 .

[8] Y.Horie, "A study on optimum term of work hour with rest pause for VDT workers," Journal of Japan human factors and ergonomics society, Japan, **23**(6), 373-383, 1987 .

[9] T. Miki, T. Terada, et al., "Research of the effects of the break time on work performance," IPSJ SIG Technical Report, **2018-UBI-57**(3), 1–8, 2018 .

[10] Ministry of Health, Labour and Welfare, "Guidelines for occupational health management in VDT work," http://www.mhlw.go.jp/file/06-Seisakujouhou-11200000-Roudoukijunkyoku/0000184703.pdf, Accessed in Aug. 2019 .

[11] D. Hirama, A. Minaduki, "A development of an interval-supported application for VDT-work," Conference of Information processing sciety of Japan, 4-125 – 4.126,2014.

[12] M. Takeue, K. Takahashi, and M. Inaba, "Fundamental Experiments on Effects of Games in Rest Time between Learning," IPSJ SIG Technical Report, **2014-CE-124**(17) , 1–8, 2014 .
[13] K. Yamasaki, K. Takahashi, and M. Inaba, "Effect of the way of spending the rest time on learning," Conference of Japan Society for Fuzzy Theory and Intelligent Informatics, Kyushu Chapter, C-203, 83–86, 2015 .
[14] J. Kawaguchi, T. Ito, K. Takahashi, and M. Inaba, "Influence of How to Spend Recess in E-learning," Conference of IEICE Engineering Sciences Society, A-11-4, 151, Sapporo, Japan, 2016 .
[15] Y. Yasumura, T. Ito, K. Takahashi, T. Kajiyama, "Effects on Concentration of Different E-Learner Resting Behaviors in Reponse to Communication Robot Actions during Break Times," Proceedings of the 2020 14th International Conference on Ubiquitous Information Management and Communication, IMCOM 2020, 0–4, 2020, doi:10.1109/IMCOM48794.2020.9001742.
[16] L. Plowman, R. Luckin, "Interactivity, Interfaces, and Smart Toys," Computer, **37**(2), 98–100, 2004, doi:10.1109/MC.2004.1266302.
[17] F.D. Faraci, M. Papandrea, A. Puiatti, S. Agustoni, S. Giulivi, V. Drapuzzo, S. Giordano, F. Righi, O. Barberis, E. Thommen, E. Rossini, "AutoPlay: A smart toys-kit for an objective analysis of children ludic behavior and development," MeMeA 2018 - 2018 IEEE International Symposium on Medical Measurements and Applications, Proceedings, 2018, doi:10.1109/MeMeA.2018.8438636.
[18] S. Shasha, M. Mahmoud, M. Mannan, A. Youssef, "Playing with danger: A taxonomy and evaluation of threats to smart toys," IEEE Internet of Things Journal, **6**(2), 2986–3002, 2019, doi:10.1109/JIOT.2018.2877749.
[19] Sharp Corporation, "RoBoHoN," http://global.sharp/corporate/news/160414.html, Accessed in Aug. 2019.
[20] Sony corporation, "aibo," https://us.aibo.com/, Accessed in Aug. 2019.
[21] Mail Online, "The power of cute: How looking at pictures of baby animals can help improve your concentration levels," https://www.dailymail.co.uk/news/article-2210614/The-power-cute-How-looking-pictures-baby-animals-help-improve-concentration-levels.html, Accessed in Aug. 2019.
[22] Ryze Technology Co.,Ltd, "Tello," https://www.ryzerobotics.com/jp/tello/, Accessed in Aug. 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1154-1161 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN:2415-6698

A Proactive Mobile Edge Cache Policy Based on the Prediction by Partial Matching

Lincan Li[1], Chiew Foong Kwong[1,*], Qianyu Liu[2]

[1]*Department of Electrical and Electronic Engineering, University of Nottingham Ningbo China, Ningbo, 315100, China*

[2]*International Doctoral Innovation Centre, University of Nottingham Ningbo China, Ningbo, 315100, China*

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 15 October, 2020
Online: 22 October, 2020

Keywords:
Cache
Latency
Cache Hit Rate
Content Prediction
Location Prediction

ABSTRACT

The proactive caching has been an emerging approach to cost-effectively boost the network capacity and reduce access latency. While the performance of which extremely relies on the content prediction. Therefore, in this paper, a proactive cache policy is proposed in a distributed manner considering the prediction of the content popularity and user location to minimise the latency and maximise the cache hit rate. Here, a backpropagation neural network is applied to predict the content popularity, and prediction by partial matching is chosen to predict the user location. The simulation results reveal our proposed cache policy is around 27%-60% improved in the cache hit ratio and 14%-60% reduced in the average latency, compared with the two conventional reactive policies, i.e., LFU and LRU policies.

1. Introduction

This article is an extended version of a conference paper presented in 2018 at the Biomedical Engineering, Healthcare, Robotics, and Artificial Intelligence 2018 [1].

With the ubiquitous emergence of the smart mobile devices, e.g., the mobile phones, as well as the trend towards high data rate applications, exponential mobile data have been generated in the wireless network while the current network cannot support the fast growth of the mobile data traffic [2], [3]. As a result, the storage capacity of the wireless network needs to be expanded and the conventional way is the dense-deployment of the base stations (BSs). However, it costs a lot for the mobile operator to upgrade this infrastructure [4]. Therefore, a more cost-effective approach is needed and the caching technique is regarded as an ideal approach [5].

By deploying the cache devices at the edge of the wireless network in proximity to the users, e.g., BSs and user terminals (UTs), and storing the popular contents at the cache devices [6], [7], users can directly retrieve the contents from the edge nodes rather than the remote core network via the backhaul links [8]. Hence, the content access latency can be decreased due to the reduced content transmission distance [9]. In parallel, duplicated requests for the same contents from the core network to the edge of the network can be avoided, which reduces the potential data congestion of the network [10]. Furthermore, due to the increasingly decreased prices of the cache devices, the storage capacity of the wireless network can be more cost-effectively boosted compared to the conventional way [4].

Due to the limited storage capacity of the cache devices, only a part of the contents can be stored in the edge cache devices. Hence, multiple works are focusing on how to design an efficient cache content placement policy. The most common approaches are least frequently used (LFU) and least recently used (LRU), which are referred to as reactive cache policies that determine whether to cache a specific content after it has been requested [11], [12]. In detail, LRU always caches the most recently requested contents while LFU caches the most frequently requested contents [13]. While the reactive cache policy is not efficient during peak hours. Hence, the proactive caching strategy is introduced, by which the content can be cached before the request, and hence the users can access the preferred content immediately when they arrive in new areas.[14], [15].

There are many proactive schemes have been investigated. In [16], a threshold-based proactive cache scheme based on reinforcement is presented, aiming at minimising the average energy cost. In this case, the time variation of the content is considered, which means the content popularity is changed over time rather than static. In practice, only the content whose lifetime is not expired has the potential to be cached. In [17], a caching scheme is presented to improve the cache hit rate and reduce

*Corresponding Author: Chiew Foong Kwong, c.f.kwong@nottingham.edu.cn

energy consumption by predicting the content popularity distribution. In [18], a proactive cache based on the estimation of the content popularity is presented, targeting at increasing the cache hit rate and decreasing the content transmission expenditure. Motivated by deep learning which can improve the accuracy of the content prediction, many works utilise deep learning for proactive caching. In [19], deep learning is utilised to predict the future probability of the content and the predicted content with a high probability will be cached. In [20], a proactive cache policy is proposed based on a deep recurrent neural network model which can predict the future content requests.

Table 1: Summary of existing cache policies.

Policy	Contribution
[11] [12]	LFU and LRU are explained to manage the cache content updating.
[16]	A reinforcement learning-based proactive cache policy is proposed to minimise energy consumption. Here, content popularity is a time-varying variable and only the contents whose lifetime are not expired can be considered to be cached or not.
[17]	An accurate content popularity prediction is adopted to improve the cache hit rate and reduce energy consumption.
[18]	A proactive cache policy is proposed to increase the cache hit rate and decrease the content transmission cost. Here, transfer learning is applied to evaluate the content popularity, and a greedy algorithm is adopted to deal with the cache problem.
[19]	A deep learning algorithm is applied to predict the future probability of the content, and the content with a high predicted probability will be pre-cached.
[20]	A proactive cache policy is proposed to alleviate the data congestion and reduce the average latency, in which a deep recurrent neural network algorithm is adopted to predict future content requests.
[21]	A long-short term memory (LSTM) network is utilised to predict the direction of the moving vehicles, and the proactive cache problem is modeled as MDP and solved by a heuristic ε_n-greedy algorithm.
[22]	A two-layer cache network consisting of several MSBs and SBSs is proposed to improve the cache hit ratio and reduce the average latency. Here, the adjacent SBSs can communicate with each other. Besides, the users with different moving speeds are clustered into different layers, i.e. the MBS or the SBS.
[23]	A cooperative cache framework is proposed to increase the cache hit ratio and reduce access latency. Here, a PPM algorithm is adopted to predict vehicles' probability of arriving in the hot areas.

Besides, in [21], a proactive cache policy for the vehicular network is proposed, where the roadside units (RSUs) are equipped with the cache capability under high mobility of the moving vehicles. There, a long-short term memory (LSTM) network is utilised to predict the direction of the moving vehicles. Then the proactive cache problem is modeled as a Markov decision process (MDP) problem and solved by a heuristic ε_n-greedy algorithm. In [22], Gao *et al*. design a proactive cache scheme for the hierarchical network where each small base station (SBS) can perceive the user mobility of its adjacent small base stations (SBSs), aiming at maximising the cache hit rate and minimising the transmission latency. In specific, the users with different moving speeds are clustered into different layers and the cached content deployment problem is solved by a genetic algorithm. In [23], a cooperative cache framework is introduced to increase the cache hit rate and minimise the access latency, in which the prediction by partial matching (PPM) is utilised to predict the vehicles' probability of arriving at the hot areas. The vehicles with long sojourn time in a hot spot are equipped with cache capability and are regarded as cache nodes. A summary of the aforementioned works is shown in Table 1.

Different from the aforementioned works singly considering the prediction of the content popularity or the users' location, this extended paper designs a proactive cache policy jointly considering the prediction of the user preference and the user location to minimise the average latency and maximise the cache hit rate, which to the best of our knowledge has not been considered in the prior research works. In detail, a practical scenario is considered, in which the BSs are distributed and the users are mobile. A backpropagation (BP) neural network, one of the deep learning methods, is applied to predict the user preference based on the historical content requests. Furthermore, the user's future location is predicted via PPM which has been introduced in our previous work [1], and the user's preferred content is pre-cached at the location in which the user will highly arrive. The main contributions of this paper are as follows:

- This paper focuses on minimising the average latency and maximising the cache hit rate by jointly considering the content popularity prediction and user location prediction.
- The BP neural network is applied to predict the content popularity, and PPM is chosen to predict the user location.
- The effect of the several parameters on the cache performance is investigated, i.e., the Zipf parameter, the content size, the transmission rate, the distance of the backhaul link, and the distance between the user and the BS.

The remainder of this paper is organised as follows. The system model and the problem formulation are shown in section 2. Section 3 introduces the proactive cache policy. We show the simulation results in section 4 and conclude in section 5.

2. System model and problem formulation

In this section, we describe the system model, state the assumption, and formulate the problem.

2.1. System model

For each time slot t whose period is one hour, the proposed proactive cache policy adopts the PPM algorithm to obtain the probability of the user arriving at different locations. The location with the highest value is regarded as the future location. In parallel, the prediction of the user preference is trained via the BP neural network. Once the predicted user preference and the future location are obtained, the popular contents in the user preference are pre-cached at the future location. Consequently, once the user arrives in this location in the next time slot (t+1), the user can immediately obtain the requested content. However, if the prediction is not accurate, the BS needs to retrieve the requested content from the core network and then send it to the users, which imposes a more latency consumption issue.

As shown in Figure 1, the distributed cache architecture consists of the following network equipment (NE): a core network, $\mathcal{M}$ cache-enabled BSs, and $\mathcal{N}$ mobile users. The $\mathfrak{m}^{th}$ BS is denoted by BS_m for $1 < \mathfrak{m} < \mathcal{M}$, the nth user is denoted by U_n for $1 < n < \mathcal{N}$, the circular coverage area of BS_m is denoted as A_m, and the set of the users served by BS_m is denoted as U_m. Let $\mathcal{F}_n=\{\mathfrak{f}_{n1}, \mathfrak{f}_{n2},\ldots, \mathfrak{f}_{n\mathfrak{k}}\}$ denotes the set of $\mathfrak{k}$ contents requested by U_n. The content popularity distribution of a user, *i.e.* user preference, follows Zipf distribution law [24]. The popularity of the $\mathfrak{R}^{th}$ content requested by U_n is characterized as:

$$\mathfrak{p}_n(\mathfrak{R}, \mathfrak{M}, \mathfrak{U}) = \frac{\frac{1}{\mathfrak{R}^{\mathfrak{M}}}}{\sum_1^{\mathfrak{U}} \frac{1}{\mathfrak{U}^{\mathfrak{M}}}}, \quad (1)$$

where $\mathfrak{R}$ is the rank of the content in $\mathcal{F}_n$, $\mathfrak{M}$ is the Zipf parameter for $0< \mathfrak{M}<1$, and $\mathfrak{U}$ is the total number of contents in $\mathcal{F}_n$.

Let $\mathcal{F}_{U_m}$ represents the set of the contents requested at BS_m, $\mathfrak{p}_{U_\mathfrak{m}}$ represents the content popularity at BS_m and $\mathcal{C}=\{1,2,\ldots, \mathcal{H}\}$ represents the library of all the contents requested by $\mathcal{N}$ mobile users served by $\mathcal{M}$ BSs. Assume each BS can store $\mathfrak{k}$ contents at most for $\mathfrak{k}< \mathcal{H}$, and each content has the same size $\mathcal{B}$. Besides, one user only requests one content at most for each time slot *t*.

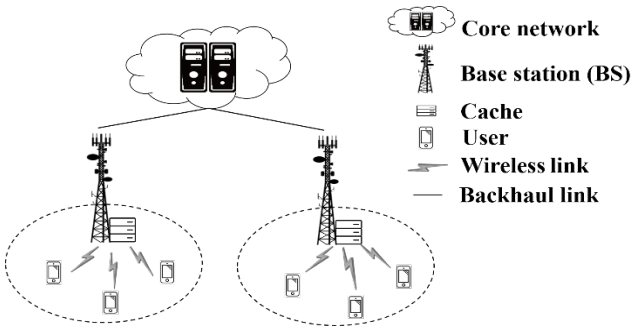


Figure 1: The cache-enabled network.

2.2. Problem formulation

Based on the mention before, our target is to minimise the access latency $\mathfrak{T}$, which is comprised of the transmission latency and propagation latency [25]. The transmission latency is caused by transmitting the content from EN_i to EN_j [26], in which EN_i and EN_j are any two network equipment. According to [27], the transmission rate $\mathbb{R}_{(i,j)}$ is calculated as:

$$\mathbb{R}_{(i,j)}= \mathrm{B}\log_2(1+\frac{\rho\mu}{\sigma^2}), \quad (2)$$

where B (Hz) is the available spectrum bandwidth, ρ is the transmitted power, σ^2 is the noise power and μ is the channel gain between EN_i and EN_j.

Therefore, the transmission latency $\mathfrak{T}_{t_{(i,j)}}^{c}$ based on the size of the requested content c and the transmission rate is derived as:

$$\mathfrak{T}_{t_{(i,j)}}^{c}=\frac{\mathcal{S}}{\mathbb{R}_{(i,j)}}, \quad (3)$$

where $\mathcal{S}$ is the size of requested content c.

The propagation latency $\mathfrak{T}_{p_{(i,j)}}^{c}$ is defined as the time of propagating the requested content c from EN_i to EN_j. Affected by the propagation speed of the electromagnetic wave and the distance between the EN_i and EN_j, the propagation latency $\mathfrak{T}_{p_{(i,j)}}^{c}$ is expressed as:

$$\mathfrak{T}_{p_{(i,j)}}^{c}=\frac{\mathcal{J}_{(i,j)}}{v}, \quad (4)$$

where v is the propagation speed of the electromagnetic wave in the corresponding channel, $\mathcal{J}_{(i,j)}$ is the distance between EN_i and EN_j.

Therefore, the access latency is expressed as:

$$\mathfrak{T} = \mathfrak{T}_{t_{(i,j)}}^{c} + \mathfrak{T}_{p_{(i,j)}}^{c} \quad (5)$$

$$= \frac{\mathcal{S}}{\mathbb{R}_{(i,j))}} + \frac{\mathcal{J}_{(i,j)}}{v}. \quad (6)$$

In detail, the content can be directly retrieved from BS if it is hit at the BS, i.e., the content is cached at the BS. Hence the latency $\mathfrak{T}_{hit}$ of cached content is shown as:

$$\mathfrak{T}_{hit}= \frac{\mathcal{S}}{\mathbb{R}_{(u,b)}} + \frac{\mathcal{J}_{(u,b)}}{v_{hit}}, \quad (7)$$

where $\mathbb{R}_{(u,b)}$ is the transmission rate between a user and a BS, $\mathcal{J}_{(u,b)}$ is the distance between a user and a BS, and v_{hit} is the propagation speed of the electromagnetic wave in the air.

Otherwise, the content needs to be retrieved from the core network via the backhaul links if the content is missed at the BS, i.e., the content is not cached at the BS. According to [28], the transmission rate $\mathbb{R}_{(u,core)}$ from the core network to the BS is shown as

$$\mathbb{R}_{(b,core)}= \mathrm{R}^*, \quad (8)$$

where R^* is the maximal transmission rate of the network.

Therefore, the latency of a missed content $\mathfrak{T}_{missed}$ consisting of the transmission latency $\mathfrak{T}_{t_{(u,core)}}^{c}$ and the propagation latency of a missed content is expressed as:

$$\mathfrak{T}_{t_{(u,core)}}^{c} = \frac{\mathcal{S}}{\mathrm{R}*}+\frac{\mathcal{S}}{\mathbb{R}_{(u,b)}}, \quad (9)$$

$$\mathfrak{T}_{p_{(u,core)}}^{c} = \frac{\mathcal{J}_{(u,b)}}{v_{hit}} + \frac{\mathcal{J}_{(b,core)}}{v_b}, \quad (10)$$

$$\mathfrak{T}_{missed} =\frac{\mathcal{S}}{\mathbb{R}_{(u,b)}} + \frac{\mathcal{S}}{\mathrm{R}*} +\frac{\mathcal{J}_{(u,b)}}{v_{hit}} + \frac{\mathcal{J}_{(b,core)}}{v_b}, \quad (11)$$

where $\mathcal{J}_{(u,b)}$ is the distance between the user and the BS, $\mathcal{J}_{(b,core)}$ is the distance between the BS and the core network, and v_b is the propagation speed of the electromagnetic wave in the backhaul link.

Therefore, the average system latency $\overline{\mathfrak{T}}$ is calculated as:

$$\overline{\mathfrak{T}}=\frac{\sum_1^N[\left(\frac{\mathcal{S}}{\mathbb{R}_{(u,b)}}+\frac{\mathcal{J}_{(u,b)}}{v_{hit}}\right)+(1-\mathcal{Z})(\frac{\mathcal{S}}{\mathrm{R}*}+\frac{\mathcal{J}_{(b,core)}}{v_{bk}})]}{N}, \quad (12)$$

where $\mathcal{Z}$ is the cache hit rate and is calculated as follows:

$$\mathcal{Z} = \frac{\sum_{n=1}^{\mathcal{N}} \sum_{1}^{\mathcal{G}} F(R_{i_{U_n}})}{\sum_{n=1}^{\mathcal{N}} \mathcal{G}}, \quad (13)$$

where $R_{i_{U_n}}$ is the content requests of U_n, $\mathcal{G}$ is the number of request times of $R_{i_{U_n}}$. The $F(R_{i_{U_n}})$ is calculated as

$$F(R_{i_{U_n}}) = \begin{cases} 1, & R_{i_{U_n}} \text{ is cached} \\ 0, & R_{i_{U_n}} \text{ is not ccahed} \end{cases} \quad (14)$$

The problem of minimising the average system latency is modeled as follows

$$\mathbf{P\_1}: \min \overline{\mathfrak{T}} \quad (15)$$

$$\text{s.t. } 0 < v_{bk} < v_{hit} \leq 3 \times 10^8 m/s \quad (16)$$

$$0 \leq \mathcal{Z} \leq 1 \quad (17)$$

3. The proactive cache based on the content popularity prediction and future location prediction

In this section, a proactive cache policy is proposed to address **P_1**. Firstly, the user preference is predicted according to the backpropagation (BP) neural network. Besides, we introduce the future location prediction based on the prediction by partial matching (PPM) algorithm. The proposed cache policy minimises the average system latency by pre-caching the predicted popular content at the correspondingly predicted location.

3.1. The content popularity prediction based on backpropagation neural network

User preference is the content probability distribution of individual user and content popularity is the content probability distribution of a cluster of users. Due to the characteristic of the user preference that a small number of contents account for most of the data traffic, the cache policy considers caching the popular content to reduce the complexity of the computation. Hence, the set of the popular contents of U_n is denoted as $\mathbb{P}_{U_n} = \{y_{U_n}^1, y_{U_n}^2, \ldots, y_{U_n}^k\}$, which contains k samples by choosing the top k contents with the highest probability from the user preference. Therefore, the set of the popular contents at BS$_m$ is denoted as $\mathbb{P}_{BSm} = \{\mathbb{P}_1, \mathbb{P}_2, \ldots \mathbb{P}_{U_n} \ldots, \mathbb{P}_{U_m}\}$.

After obtaining the popular content database of BS$_m$, the BP neural network, as shown in Figure 2, is applied to predict the content popularity. The proposed neural network is comprised of three layers, namely the input layer, hidden layer, and output layer. The number of the neuron cells in the input layer and the output layer is equal to the cache storage k. The content requests of U_n are collected each hour and denoted as a training data set. Besides, two continuous training data sets are chosen to optimise the parameter of the neural network. The value y_{in} for the input layer is the request times of the top k popular contents in the former training data set. The value y_{real} is the request times of the top k popular contents in the latter training data set. Furthermore, mean squared error (MSE) is utilised as the loss function in the content prediction. The *MSE* is formulated as

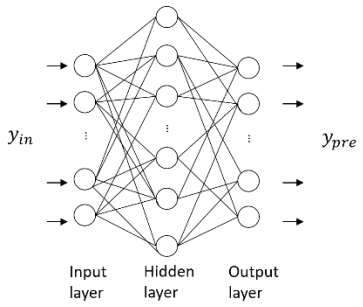
$$\text{MSE} = \frac{\sum_{1}^{k} (y_{real} - y_{pre})^2}{k}, \quad (18)$$

where y_{pre} is the value of the output layer.

Besides, the Relu function is chosen as the activation function, which is expressed as

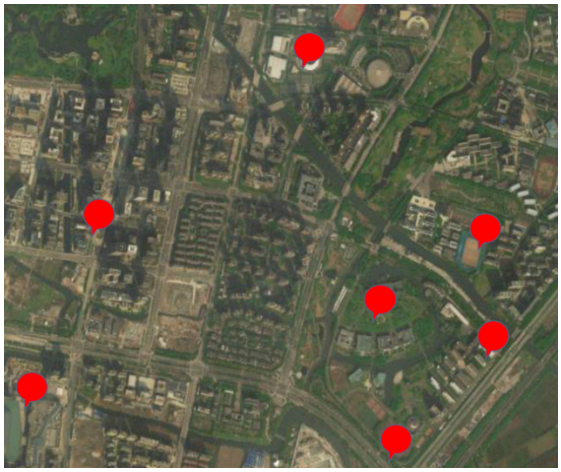
$$Relu(y_{in}) = \begin{cases} 0, & y_{in} < 0 \\ y_{in}, & y_{in} \geq 0 \end{cases}. \quad (19)$$

With the help of stochastic gradient descent (SGD), the proposed neural network can optimally predict the content popularity after enough training.



3.2. The future location prediction based on a prediction by partial matching

Before the location prediction, the historical location information is collected from a real environment model as shown in Figure 3. The areas labeled by red symbols are regarded as the hot spots with long sojourn time. The historical location information sequence is denoted as $\mathcal{L}$ which is related to the hot spots.



After obtaining the historical location information $\mathcal{L}$, PPM is applied to predict the user's future location. PPM is a data compression method based on the finite context and it has been proven effective for the location prediction [23]. The probability of the future location y appearing after the given context *Con* is model as *P(ycon),* where *Con* is the sequence of the location and the length of the sequence is called order [29]. Furthermore, PPM proposes an escape mechanism to deal with the zero-frequency problem [30]. When escape occurs, *i.e. y* is missed after *Con*. Then the PPM outputs an escape probability defined as

Pesc(esc|Con). The computation of PPM is shown in Algorithm 1. Firstly, PPM checks whether *y* appears after *Con*. If *y* appears, PPM records the number of appearing times and outputs the probability $P(y|Con)$, otherwise, PPM outputs the escape probability *Pesc(esc|Con)*. Under the escape situation, PPM restarts to check whether *y* appears after the new *Con* (the order of which is the original order minus 1). The process is finished until *y* appears after *Con* or the order is -1. The predictive probability of the future location is the multiple of the sub-probabilities and the calculation is shown as:

$$P = \prod_1^{step} P_i, \tag{20}$$

$$P_i = \begin{cases} P(y|Con) = \frac{N_y}{N_{esc}+N_{Con}}, \ y\ appears\ after\ Con \\ Pesc(esc|Con) = \frac{N_{esc}}{N_{esc}+N_{Con}}, \ y\ escapes\ after\ Con \end{cases}, \tag{21}$$

where P_i is the probability of step *i*, N_y represents the number of the times of *y* appearing after *Con*, N_{esc} represents the number of the characters appearing after *Con*, and N_{Con} represents the number of the times of all the characters appearing after *Con*.

Once the probabilities of the possible locations are obtained via PPM, these obtained probabilities are ranked in descending order. The location with the highest probability is regarded as the future location.

Algorithm 1 PPM algorithm

Input: historical data $\mathcal{L}$

Output: the probability *P* of the future location *y*

```
Initialize P=0
order = h, j =1
if y appears after Con
    P_j=P(y|Con)
    Process finished
else
    P_j=P(esc|Con)
    h - =1, j + =1
    Restart the step 3-7
The process is finished until y appears after Con or h =-1.
Output j, and P=∏_1^j P_j
```

Here is an example to help understand PPM computation by giving a user path $\{L_1, L_3\}$ and the future location L_4 in the historical data sequence $\mathcal{L}$ ={$L_1, L_2, L_3, L_4, L_5, L_1, L_3, L_1, L_4, L_1, L_2, L_4, L_3, L_4, L_1$}. First, since the sequence{ L_1, L_3, L_4 } cannot be found from the historical data sequence, the escape probability P(esc| L_1, L_3) is outputted based on *Pesc(esc|Con)* in Eq. (21), as shown in Eq. (22). Then the new order is 1 and consequently, the new context is { L_3 }. The new sequence { L_3, L_4 } can be found from the historical date sequence, and therefore the probability P(L_4| L_3) is obtained based the $P(y|Con)$ in Eq. (21), as shown in Eq. (23). Finally, the probability P(L_4| L_1, L_3) is obtained based on Eq. (20), as shown in Eq. (24).

$$\mathrm{P}(\mathrm{esc}|\ L_1, L_3) = \frac{N_{(\mathrm{esc}|\ L1,L3)}}{N_{(\mathrm{esc}|\ L1,L3)}+N_{(L1|\ L1,L3)}} = \frac{1}{1+1} = \frac{1}{2}, \tag{22}$$

where $N_{(\mathrm{esc}|\ L1,L3)}$ is the number of the characters appearing after $\{L_1, L_3\}$, and $N_{(L1|\ L1,L3)}$ is the number of the times of all the characters appearing after $\{L_1, L_3\}$ since only L_1 appears after $\{L_1, L_3\}$.

$$\mathrm{P}(L_4|\ L_3) = \frac{N_{(L4|\ L3)}}{N_{(\mathrm{esc}|\ L3)}+N_{(L4|\ L3)}+N_{(L1|\ L3)}} = \frac{2}{2+2+1} = \frac{2}{5}, \tag{23}$$

where $N_{(L4|\ L3)}$ is the number of L_4 appearing after L_3, $N_{(\mathrm{esc}|\ L3)}$ is the number of the characters appearing after L_3 and $N_{(L1|\ L3)}$ is the number of L_1 appearing after L_3. The sum of $N_{(L4|\ L3)}$ and $N_{(L1|\ L3)}$ is called the number of the times of all the characters appearing after L_3.

$$P(L4|L1, L3) = P(esc|L1,L3) \times P(L4|L3) = \frac{1}{2} \times \frac{2}{5} = \frac{1}{5}, \tag{24}$$

3.3. The pre-deployment of the popular content at the future location

In each time slot *t*, the users' future locations in which users will highly arrive at the next time slot *t+1* are predicted via PPM. In parallel, the user preference at *t+1* is predicted via BP neural network. The top *w* contents with the highest number of request times are regarded as the popular contents in the future. After that, these popular contents are pre-deployed at the corresponding future location. Hence, in the next time slot *t+1*, if the prediction is correct, users can immediately obtain their preferred contents, which extremely reduces the average system latency.

4. Simulation results and analyzation

In this section, we consider a distributed BS caching network which consists of 10 BS, 30 users, and 6 locations. The number of content requests of each user is 3000. The comprehensive simulation shows the performance of our proposed policy, LFU, and LRU in terms of the average latency and cache hit rate. The specific parameter settings are shown in Table 2. The program is modeled via PyTorch language in Pycharm software.To further show the improvement of our proposed policy in terms of the cache hit rate and the reduction of our proposed policy in terms of the cache hit rate compared with LFU and LRU policies, we propose the growth ratio $\mathcal{P}_G$ and the reduction ratio $\mathcal{P}_R$, which are expressed as:

$$\mathcal{P}_G = \frac{\mathcal{C}_{our} - \mathcal{C}_{existing}}{\mathcal{C}_{existing}}, \tag{25}$$

$$\mathcal{P}_R = \frac{\overline{\mathfrak{T}}_{existing} - \overline{\mathfrak{T}}_{our}}{\overline{\mathfrak{T}}_{existing}}, \tag{26}$$

where $\mathcal{C}_{our}$ and $\mathcal{C}_{existing}$ is the cache hit rate of our proposed policy and any one of the LFU and LRU policies, respectively. $\overline{\mathfrak{T}}_{our}$ and $\overline{\mathfrak{T}}_{existing}$ is the average latency of our proposed policy and any one of the LFU and LRU policies, respectively.

Table 2: The simulation parameter settings

symbol	value

v_{bk}	1×10^7 m/s
v_{hit}	3×10^8 m/s
$\mathbb{R}_{(i,j)}$	10~50 Mbps
σ^2	1W
ρ	3W
$\mathcal{J}_{(u,b)}$	10~50km
$\mathcal{J}_{(bs,core)}$	100km
$\mathcal{B}$	30Kb~450Kb
$\mathfrak{M}$	1.1~1.8
$\mathcal{Q}$	1%~10%
δ	2%~10%

Figure 4 reveals the cache hit rate (represented in percentage) of our proactive policy and the conventional reactive policies, *i.e.,* LFU and LRU. The number of the total content requests is 6000, the Zipf parameter of each user varies between 1.7 and 1.8. Besides, to demonstrate the effect of the cache capacity on the cache performance, we introduce the cache capacity ratio $\delta = \frac{\mathcal{h}}{\mathcal{H}}$. And in this simulation, we assume δ= 2%, 4%, 6%, 8% and 10%. Horizontally, the cache hit rates of LFU, LRU, and our proposed policy increase with the larger cache capacity ratio. The tendency demonstrates that increasing the cache capacity can improve the cache hit rate since more popular contents can be cached. We also notice that our proactive policy has the highest cache hit rate, which is around 10-25% higher than that of LFU and LRU policies, no matter how the Zipf parameter varies. Therefore, our proposed policy outperforms the other two policies.

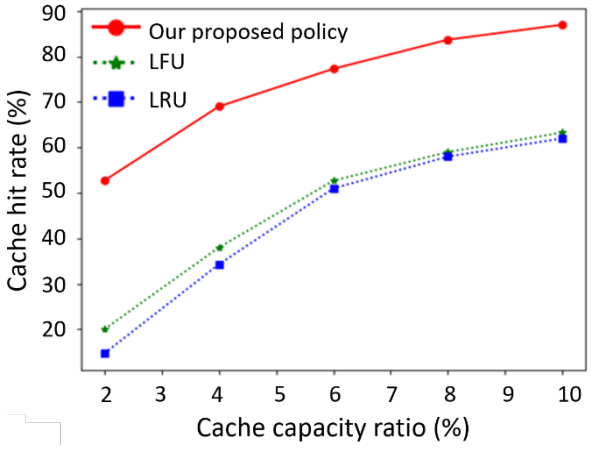


Figure 4: The cache hit rate vs. cache capacity ratio $\delta = \frac{\mathcal{Q}}{\mathcal{H}}$.

Figure 5 investigates the effect of the Zipf parameter $\mathfrak{M}$ on the cache hit rate of our proposed policy with the other two policies as mentioned before. We assume δ is 10%, and the Zipf parameter of each user varies in the range [1.1, 1.2], [1.2, 1.3], [1.3, 1.4], [1.4, 1.5], [1.5, 1.6], [1.6, 1.7] and [1.7, 1.8]. As the Zipf parameter grows, the cache hit rates of all the cache policies increase. The reason is that fewer contents are taking up more content requests as the Zipf parameter grows, and hence the popular content becomes more popular. Considering the fixed number of the total content request, the number of content reduces. With the same capacity, the cache has a higher chance to store more contents and the cached contents are more popular, which contributes to a higher cache hit rate. Furthermore, the slopes of the three curves are gradually reduced. The reason is with the larger Zipf parameter, the newly cached popular contents have fewer content requests compared with the initially cached contents. We also notice that the two reactive policies have a relatively close cache hit rate, and the cache hit rate of our proposed policy is around 24%-38% higher than that of the two reactive policies.

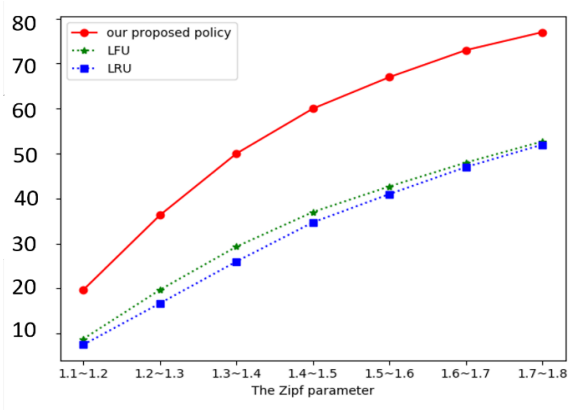


Figure 5: The cache hit rate vs. the Zipf parameter.

The relation between the average latency and the size of the content is displayed in Figure 6. Here, the size of the content is 30Kb, 200Kb, 200Kb, 250Kb, 300Kb, 350Kb, and 450Kb, respectively. Besides, we set the cache capacity ratio δ is 10% and the fluctuation of the Zipf parameter is between 1.7 and 1.8, the distance between the user and the BS is 10km and the distance of the backhaul link is 100km. As the size of the content grows, the average latencies of all the policies increase. The reason is the transmitter consumes more time to send the content into the channel as the size of the content grows. Vertically, the average latency obtained by our proposed policy is around 60% reduced compared with LFU and LRU regardless of the size of the content, which implies our proposed policy outperforms the two reactive policies.

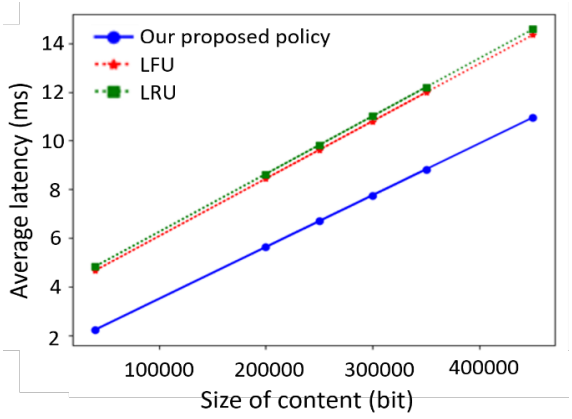


Figure 6: The average latency vs. the size of the content.

Figure 7 shows the relationship between the average latency and the transmission rate between the user and BS. Here, the content size is 400Kb, the storage capacity ratio is 10%, the distance between the user and the BS is 10km and the distance of the backhaul link is 100km. The transmission rate between user and BS is 10Mbps, 20Mbps, 30Mbps, 40Mbps, and 50Mbps, respectively. As the transmission rate between user and BS grows,

the average latencies of all the policies reduce. The reason is that, with the larger transmission rate, the latency between the user and the BS is reduced. Also, the average latency of our proposed policy is 31%-64% reduced compared with the other two policies.

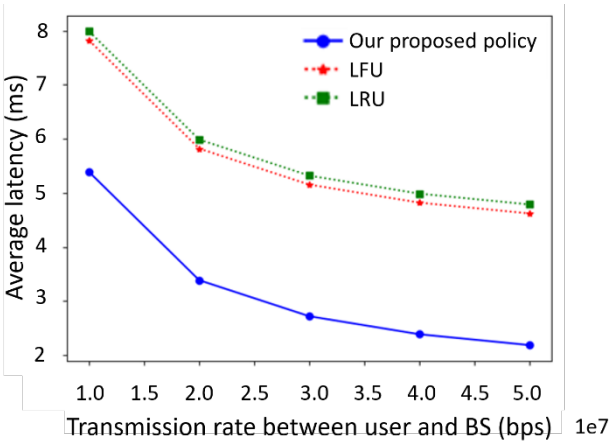


Figure 7: The average latency vs. transmission rate between the user and the BS.

As shown in Figure 8, the average latency is plotted as a function of the Zipf parameter. Here, the Zipf parameter of each user varies in the range [1.1, 1.2], [1.2, 1.3], [1.3, 1.4], [1.4, 1.5], [1.5, 1.6], [1.6, 1.7] and [1.7, 1.8], respectively. Besides, the transmission rate between the user and the BS is 50Mbps, the content size is 400Kb, the storage capacity ratio δ is 10%, the distance between the user and the BS is 10km and the distance of the backhaul link is 100km. As the Zipf parameter increase, the average latencies of three policies are reduced. The reason is that, with the increase of the Zipf parameter, more contents are cached locally, and hence fewer contents need to be retrieved from the remote core network. And the latency from the BS is lower than from the core network. Also, as the Zipf parameter grows, the slopes of the three curves gradually decrease. The tendency is caused since the newly cached contents are less popular than the initially cached contents. Furthermore, our proposed policy is around 14%-53% reduced in terms of the average latency compared with the two reactive policies.

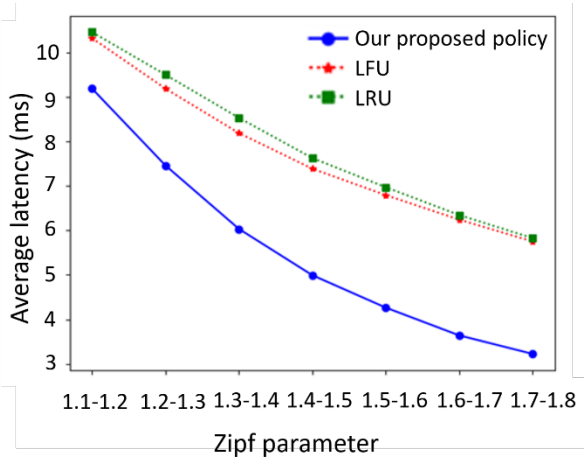


Figure 8: The average latency vs. Zipf parameter.

The effect of the cache capacity ratio δ on the average latency is shown in Figure 9. In this simulation, we assume δ= 2%, 4%, 6%, 8% and 10%. Besides, the content size is 400Kb, the transmission rate is 50Mbps, the distance between the user and the BS is 10km and the distance of the backhaul link is 100km. The cache capacity ratio δ is varied from 2% to 10%. It can be noticed that the average latencies of three policies decrease with the increment of the cache capacity ratio. The fact is that a larger cache capacity means more contents can be cached. As a result, more long-distance propagation time consumption from the core network to the BS can be avoided. Also, the average latency of our proposed policy is around 35%-55% reduced compared with the LFU and LRU.

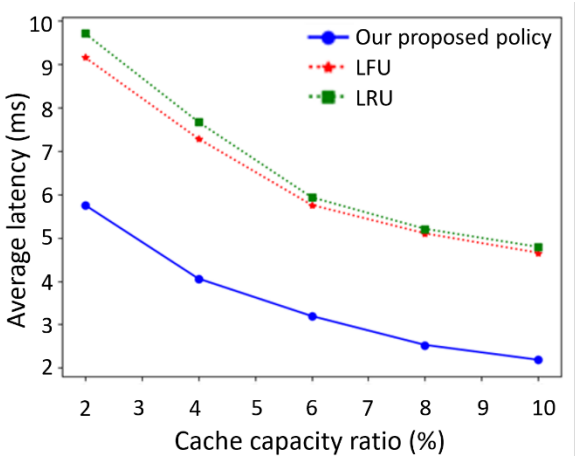


Figure 9: The average latency vs. cache capacity ratio.

5. Conclusion

In this paper, a proactive cache policy is proposed in a distributed manner to minimise the average latency, as well as maximising the cache hit rate. An accurate prediction is achieved to make sure the proactive cache policy can have a high cache performance. In specific, a BP neural network is applied to predict the content popularity, and a PPM algorithm is applied to predict the user location. The simulation results (Fig.4 and Fig.5 simulations) reveal our proposed cache policy is around 10%-38% improved in terms of the cache hit rate no matter how the cache capacity and Zipf parameter vary, compared with LFU and LRU policies. As for the average latency, our proposed policy has at least 14% decrease no matter how parameters change, i.e., the variation of the content size (Fig.6 simulation), the transmission rate between the user and BS (Fig.7 simulation), the Zipf parameter (Fig.8 simulation) and the cache capacity (Fig.9 simulation). Consequently, our proposed policy outperforms LFU and LRU policies.

References

[1] L. Li, C.F. Kwong, F. Chen, Q. Liu, J. Wang, 'Predicting future location in mobile cache based on variable order of prediction-by-partial-matching algorithm', in IET Conference Publications, 4–4, 2018, doi:10.1049/cp.2018.1729.

[2] E. Bastug, M. Bennis, E. Zeydan, M.A. Kader, I.A. Karatepe, A.S. Er, M. Debbah, 'Big data meets telcos: A proactive caching perspective', Journal of Communications and Networks, **17**(6), 549–557, 2015, doi:10.1109/JCN.2015.000102.

[3] Cisco, 'Cisco Annual Internet Report (2018–2023)', Cisco, 1–41, 2020.

[4] R. Wang, X. Peng, J. Zhang, K.B. Letaief, 'Mobility-aware caching for content-centric wireless networks: Modeling and methodology', IEEE Communications Magazine, **54**(8), 77–83, 2016, doi:10.1109/MCOM.2016.7537180.

[5] Y. Mao, C. You, J. Zhang, K. Huang, K.B. Letaief, 'A Survey on Mobile Edge Computing: The Communication Perspective', IEEE Communications Surveys and Tutorials, **19**(4), 2322–2358, 2017, doi:10.1109/COMST.2017.2745201.

[6] F.M. Modesto, A. Boukerche, 'An analysis of caching in information-centric vehicular networks', IEEE International Conference on Communications, 2017, doi:10.1109/ICC.2017.7997019.

[7] Y. Li, C. Zhong, M.C. Gursoy, S. Velipasalar, 'Learning-based delay-aware caching in wireless D2D caching networks', IEEE Access, **6**, 77250–77264, 2018, doi:10.1109/ACCESS.2018.2881038.

[8] S. Zhang, N. Zhang, P. Yang, X. Shen, 'Cost-Effective Cache Deployment in Mobile Heterogeneous Networks', IEEE Transactions on Vehicular Technology, **66**(12), 12264–12276, 2017, doi:10.1109/TVT2017.2724547.

[9] Z. Luo, M. LiWang, Z. Lin, L. Huang, X. Du, M. Guizani, 'Energy-Efficient Caching for Mobile Edge Computing in 5G Networks', Applied Sciences, **7**(6), 557, 2017, doi:10.3390/app7060557.

[10] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, W. Wang, 'A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications', IEEE Access, **5**, 6757–6779, 2017, doi:10.1109/ACCESS.2017.2685434.

[11] H. Ahlehagh, S. Dey, 'Video-aware scheduling and caching in the radio access network', IEEE/ACM Transactions on Networking, **22**(5), 1444–1462, 2014, doi:10.1109/TNET.2013.2294111.

[12] S. Shukla, O. Bhardwaj, A.A. Abouzeid, T. Salonidis, T. He, 'Proactive Retention-Aware Caching with Multi-Path Routing for Wireless Edge Networks', IEEE Journal on Selected Areas in Communications, **36**(6), 1286–1299, 2018, doi:10.1109/JSAC.2018.2844999.

[13] Y. Jiang, M. Ma, M. Bennis, F.C. Zheng, X. You, 'User preference learning-based edge caching for fog radio access network', IEEE Transactions on Communications, **67**(2), 1268–1283, 2019, doi:10.1109/TCOMM.2018.2880482.

[14] N. Golrezaei, A. Molisch, A.G. Dimakis, G. Caire, 'Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution', IEEE Communications Magazine, **51**(4), 142–149, 2013, doi:10.1109/MCOM.2013.6495773.

[15] E. Bastug, M. Bennis, M. Debbah, 'Living on the edge: The role of proactive caching in 5G wireless networks', IEEE Communications Magazine, **52**(8), 82–89, 2014, doi:10.1109/MCOM.2014.6871674.

[16] S.O. Somuyiwa, A. Gyorgy, D. Gunduz, 'A Reinforcement-Learning Approach to Proactive Caching in Wireless Networks', IEEE Journal on Selected Areas in Communications, **36**(6), 1331–1344, 2018, doi:10.1109/JSAC.2018.2844985.

[17] M. Yan, C.A. Chan, W. Li, L. Lei, Q. Shuai, A.F. Gygax, I. Chih-Lin, 'Assessing the energy consumption of 5G wireless edge caching', 2019 IEEE International Conference on Communications Workshops, ICC Workshops 2019 - Proceedings, **7**, 2019, doi:10.1109/ICCW.2019.8756642.

[18] T. Hou, G. Feng, S. Qin, W. Jiang, 'Proactive Content Caching by Exploiting Transfer Learning for Mobile Edge Computing', GLOBECOM 2017 - 2017 IEEE Global Communications Conference, 1–6, 2017, doi:10.1109/GLOCOM.2017.8254636.

[19] K. Thar, N.H. Tran, T.Z. Oo, C.S. Hong, 'DeepMEC: Mobile edge caching using deep learning', IEEE Access, **6**, 78260–78275, 2018, doi:10.1109/ACCESS.2018.2884913.

[20] L. Ale, N. Zhang, H. Wu, D. Chen, T. Han, 'Online proactive caching in mobile edge computing using bidirectional deep recurrent neural network', IEEE Internet of Things Journal, **6**(3), 5520–5530, 2019, doi:10.1109/JIOT.2019.2903245.

[21] L. Hou, L. Lei, K. Zheng, X. Wang, 'A Q -Learning-Based Proactive Caching Strategy for Non-Safety Related Services in Vehicular Networks', IEEE Internet of Things Journal, **6**(3), 4512–4520, 2019, doi:10.1109/JIOT.2018.2883762.

[22] N. Gao, X. Xu, Y. Hou, L. Gao, 'A Mobility-aware Proactive Caching Strategy in Heterogeneous Ultra-Dense Networks', IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, **2019**–**Septe**, 2019, doi:10.1109/PIMRC.2019.8904364.

[23] L. Yao, A. Chen, J. Deng, J. Wang, G. Wu, 'A Cooperative Caching Scheme Based on Mobility Prediction in Vehicular Content Centric Networks', IEEE Transactions on Vehicular Technology, **9545**(c), 1–10, 2017, doi:10.1109/TVT.2017.2784562.

[24] D. Liu, C. Yang, 'Caching at Base Stations with Heterogeneous User Demands and Spatial Locality', IEEE Transactions on Communications, **67**(2), 1554–1569, 2019, doi:10.1109/TCOMM.2018.2876424.

[25] F. Jiang, Z. Yuan, C. Sun, J. Wang, 'Deep Q-Learning-Based Content Caching With Update Strategy for Fog Radio Access Networks', IEEE Access, **7,** 97505–97514, 2019, doi:10.1109/access.2019.2927836.

[26] Y. Li, M.C. Gursoy, S. Velipasalar, 'A delay-aware caching algorithm for wireless D2D caching networks', 2017 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2017, 456–461, 2017, doi:10.1109/INFCOMW.2017.8116419.

[27] H. Wu, J. Zhang, Z. Cai, F. Liu, Y. Li, A. Liu, 'Towards Energy-Aware Caching for Intelligent Connected Vehicles', IEEE Internet of Things Journal, **4662**(c), 1–1, 2020, doi:10.1109/jiot.2020.2980954.

[28] J. Zhang, X. Hu, Z. Ning, E.C.H. Ngai, L. Zhou, J. Wei, J. Cheng, B. Hu, V.C.M. Leung, Joint resource allocation for latency−sensitive services over mobile edge computing networks with caching, **6**(3), 4283−4294, doi:10.1109/JIOT.2018.2875917.

[29] I. Burbey, T.L. Martin, 'Predicting future locations using prediction-by-partial-match', Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-Less Environments - MELT '08, 1, 2008, doi:10.1145/1410012.1410014.

[30] S.K. Pulliyakode, S. Kalyani, 'A Modified PPM Algorithm for Online Sequence Prediction Using Short Data Records', IEEE Communications Letters, **19**(3), 423–426, 2015, doi:10.1109/LCOMM.2014.2385088.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 1162-1171 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Improvised E-Rickshaws for Indian Roads by Effective Battery-Ultracapacitor Hybridization

Shimin Vayal Veetil[*,1], Varsha Shah[1], Makarand Lokhande[2]

[1]*Department of Electrical Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, 395007, India*

[2]*Department of Electrical Engineering, Visvesvaraya National Institute of Technology, Nagpur, 440010, India*

ARTICLE INFO

Article history:
Received: 23 July, 2020
Accepted: 28 September, 2020
Online: 22 October, 2020

Keywords:
Battery-Ultracapacitor
Electric Vehicles
Energy Management
Fuzzy Logic
Regenerative Braking

ABSTRACT

The revolution in the electrification of conventional vehicles due to increased petroleum prices and environmental concerns has had its impact on three-wheeler vehicles as well. Motorized three-wheeled vehicles, known as Auto-rickshaws, are a standard mode of transportation in India. Existing battery operated electric rickshaws (known as E-rickshaws) available in major cities of India faces challenges like insufficient charging facility, low driving range, high battery cost, battery replacement/disposal, etc. A battery-ultracapacitor (UC) hybrid energy source is proposed in this paper to overcome these issues. Two energy sources of complementary characteristics, when operated in tandem, can enhance the overall performance of a vehicle in terms of weight, volume, and efficiency. An Erickshaw with battery-UC hybrid energy sources is modeled in Matlab-Simulink in this paper. The vehicle dynamics are calculated real-time based using a GPS based PerformanceBox tool of VBOX motorsport for Surat city. A fuzzy logic-based approach was employed for efficient energy management between these two sources with efficient utilization of regenerative power generated while braking. The system was later tested on Real-Time Hardware-in-Loop (HIL) environment to validate the simulation results. It was observed that the addition of an additional source with complementary characteristics had enhanced the performance of vehicle operation.

1 Introduction

India, holding the second largest population in the world [1], is also home to some of the most polluted cities across the globe [2]. Fourteen of the twenty most polluted cities in the world are in India according to a recent study based on $PM_{2.5}$ particulate matter [3]. A major contributor towards pollution in these cities are emissions from motorized vehicles [4]. The rise in air pollution has escalated the research towards improving air quality across the country. Also incompetent petroleum resources which kept the growth of economy in check for the past few decades have raised the necessity to identify alternate fuel propelled vehicles for transportation. Transportation electrification was hence identified as a potential breakthrough to answer all these concerns faced by the country in the present decade and for future.

Three wheeled motorized vehicles called as Auto-rickshaws, capable of carrying three to four passengers plays a vital role in public transportation in India along with trains and buses. Unlike the other two, auto-rickshaws are preferred for shorter traveling distances (preferably inside cities) due to its compact size. A Tier-I city (with more than 4 million population) accommodates more than 50,000 auto-rickshaws [5] in India. That can account to 4-5% or sometimes even high as 10% of total vehicle population of a city. These vehicles can be a considerable source of air pollutants and is also a major customer of petroleum resources [6]. Electrification of auto-rickshaws can thereby lead to a better and cleaner mode of urban public transportation and can also urge the public to opt for them due to the improvisations in comfort, safety and cost to travel [7].

Battery operated electric vehicles are gaining public and industry interest in the past couple of decades [8]. A battery operated auto-rickshaw (termed as E-rickshaw) was first introduced in the streets of New Delhi in 2010. Many researches on implementation of E-rickshaws/vehicles in different states and cities across the country and their economical and environmental impacts are available in literature. The technical and commercial issues hindering the

[*]Corresponding Author: Shimin Vayal Veetil, Department of Electrical Engineering, S. V. National Institute of Technology, Surat, 395007, India, +91 83477 68150 & ds14el004@eed.svnit.ac.in

popularity of electric vehicles in Pune was discussed in [9]. The merits, demerits and the challenges faced by adopting E-rickshaw as a possible mode for public transportation in the roads of West Bengal was discussed in [10]. A 24% fuel economy with hybridization and optimization was observed by [11] in their study on parallel electric hybrid E-rickshaws. A compact, robust and affordable hybrid system was developed and tested on a Bajaj RE auto-rickshaw by [12] and identified that the fuel consumption and CO_2 emissions was significantly reduced by 21%.

There are many techno-commercial areas for improvisation in the design and development of E-rickshaws in India. Driving range is a major issue faced by e-rickshaws [13] along with waste management of used battery [14]. Another concern is method of power generation for charging batteries employed in E-rickshaws. One major aspect is the use of renewable energy for charging of the vehicle batteries which will lead to complete zero emissions [15]. Using conventional energy sources for electricity generation may result in zero tail-pipe emissions but will have less to no impact in global control of air pollution. Many literature is available on bonding renewable sources with electric vehicles. Incorporating solar energy as one of the energy source for electric vehicles and their economic impacts were studied in [16] and [17]. A correlative analysis study of a solar assisted three-wheeled rickshaw with various drive trains of hybrid configurations was carried out in [18]. A fuel cell based E-rickshaw which employs a fuzzy logic controller was proposed in [19]. In [20], a control scheme for solar rickshaw with maximum power point tracking algorithm (MPPT) was designed.

A rigorous review on various electrical energy storage technologies compatible for vehicular applications was carried out in [21] and [22]. The battery technologies available for electric vehicles including Li-ion batteries, lead acid batteries, Ni-MH batteries and ultracapacitors are discussed in [14]of which lithium-ion battery yields the lowest operating costs and GHG emissions (compared to lead-acid batteries). The research on lead-acid batteries, Li-ion batteries and ultracapacitor are gaining considerable significance in India as well [23]. Applications of different energy sources in Indian context was studied in [24]. Hybridization of energy sources with complementary characteristics to improve performance of electric vehicles [25] is attracting considerable interest these days. Hybridization of energy sources to achieve improved performance is been practiced in various applications [26, 27]. Hybridization enhances the performance of each sources involved, but with introducing complexity in its control. A study on possible methods for coupling a battery with an ultracapacitor for enhanced performance in electric vehicles was carried out in [28]. A study on different energy sources available for electric vehicles, their hybridization and energy management strategies was carried out in [29]. Performance improvement of a mild hybrid vehicle by hybridization of battery with ultracapacitor was suggested in [30].

Energy management between the hybrid sources employed in an electric vehicle calls for special care for ensuring optimum performance of both. The different control strategies viz. Rule based control strategy, Fuzzy rule based control strategy, Linear programming, Dynamic programming, Stochastic control strategy, Genetic algorithm, Particle swarm optimisation etc., for power split in a hybrid electric vehicle was discussed in [31]. A model predictive control strategy for sharing of power between a battery and ultracapacitor for a plug-in hybrid electric vehicle (PHEV) was proposed in [32] and has achieved significant improvement in performance. A data fusion approach for energy management in a fuel cell hybrid electric vehicle was proposed in [33]. The energy management strategy for a Fuel cell/Battery/Ultracapacitor hybrid electric vehicle is achieved by implementing an intelligent control technique constructed based on Fuzzy Logic Control (FLC) in [27]. A Fuzzy logic based energy management for a parallel hybrid electric vehicle was proposed by [34] considering torque demand and battery SOC as its inputs.

Presently, there is little to no work carried out in hybridization of energy sources in E-rickshaws in India which is the motivation for this work. In this paper the vehicle dynamics of an E-rickshaw is modeled in Matlab-Simulink for a standard Indian City Driving Cycle (IDC) which was calculated using a GPS based PerformanceBox tool of VBOX motorsport for Surat city. The system was later tested on a Real-time simulation environment to validate the results. The paper is divided into following sections. In section 2, the vehicle dynamics of an E-rickshaw is modeled to calculate the total tractive force. System architecture of an E-rickshaw is carried out in section 3 for a standard Indian City Driving Cycle (IDC). In section 4, the energy management between battery and ultracapacitor using a fuzzy logic based approach was discussed and the vehicle was tested for different driving conditions. The Matlab-Simulink simulation results are discussed in section 5 followed by a brief discussion on validation of the system implementation in Real-time environment in section 6. The paper concludes with discussions and conclusion in section 7.

2 Vehicle Dynamics of an E-rickshaw

The structure of a conventional auto-rickshaw is shown in Figure 2. The specifications of the auto-rickshaw is shown in Table 1 [12] whose parameters are taken as reference to design the E-rickshaw power-train.

Table 1: Specifications of Conventional Auto-rickshaw (Model: Bajaj RE 2S Petrol

Specifications	Values
Gross vehicle weight, M	610 Kg
Kerb weight	272 Kg
Tyres, r_w	4.00-8, 4PR
Roll resistance, C_r	0.015
Air drag coefficient, C_d	0.44
Frontal surface area, A_f	2.0 m^2

Load power requirement is calculated based on the drive cycle and vehicle dynamics. Energy source sizing is carried out based on this power requirement.

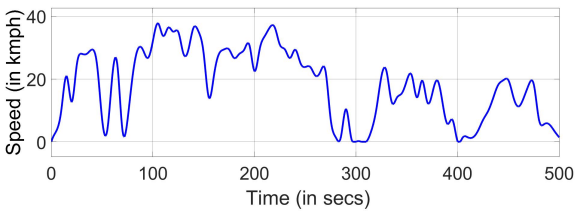


Figure 1: Indian City Drive Cycle (IDC) for Surat City

The drive cycle was calculated using a GPS based Performance-Box tool of VBOX motorsport for Surat city and is chosen because the E-rickshaw is desired to be operated in cities. The load power requirement was calculated for the calculated standard Indian city driving cycle shown in Figure 1. Battery and ultracapacitor was sized based on this power requirements. Modeling of vehicle dynamics is explained in detail in the following sections.

2.1 Total Tractive force

The forces acting on a vehicle needs to be understood for an effective modeling of vehicle dynamics. The forces acting on an auto-rickshaw while hill climbing is shown in Figure 2. The tractive force generated by the electric motor of the vehicle has to experience and overcome these forces namely (a) Rolling resistance force (b) Aerodynamic drag (c) Grade resistance force and (d) Acceleration force [35].

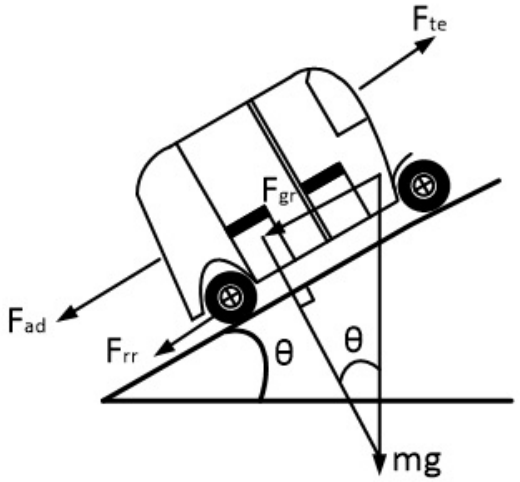


Figure 2: Forces acting on an auto-rickshaw moving along a slope

The major forces acting on an E-rickshaw and their expressions are shown in Table 2. Here, M is the mass of the vehicle in kg, g is the gravitational constant, i.e., 9.81 m/s^2, C_r is the coefficient of rolling resistance, ρ is the air density in kg/m^3, A is the frontal area of vehicle in m^2, C_d is the drag coefficient, v is the velocity of the vehicle in m/s, θ is the clockwise angle made by the slope of the road with the horizontal plain and a is the acceleration of the vehicle in m/s^2.

The total tractive force is the force required to propel the vehicle forward by overcoming all the above mentioned forces acting on the vehicle. Neglecting all other forces acting directly or indirectly on the vehicle, the total tractive force can be obtained from the Eq 1

$$F_{te} = F_{rr} + F_{ad} + F_{gr} + F_{af} \quad (1)$$

3 E-rickshaw System Architecture

The sizing of every components like energy source,electric motor, power electronic converters etc., of an E-rickshaw depends on the vehicle parameters and driving patterns. The total tractive force required by the vehicle for propulsion was calculated using Eq 1. System parameters like mass of the vehicle, frontal area etc., is available from Table 1. Vehicle velocity and acceleration is obtained from drive cycle data. Required load power is calculated as the product of total tractive force and velocity. The energy flow in a conventional electric-rickshaw is as shown in Figure 3. The vehicle dynamics of an auto-rickshaw is modeled in Matlab-Simulink environment.

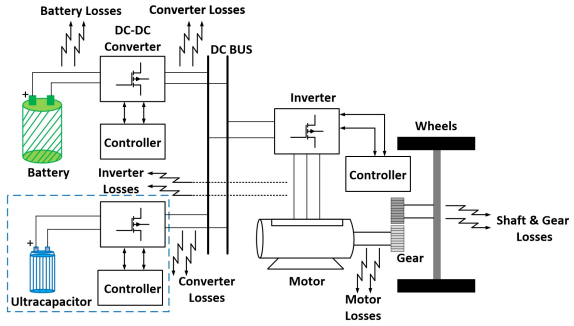


Figure 3: Energy flow diagram of an E-rickshaw with losses

The E-rickshaw is designed for the Indian City Drive Cycle (IDC) as shown in Figure 1 measured for a total time of 500 secs. The vehicle, being operated in the cities are desired to be run at a speed not more than 40 kmph due to the frequent starts and stops in traffic. The peak velocity is observed to be nearer at 38 kmph.

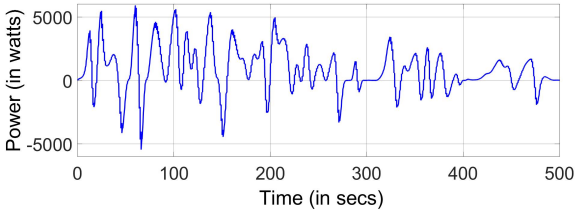


Figure 4: Load Power Requirement for Indian Drive Cycle (IDC)

The load power requirement for Indian City Drive Cycle (IDC) is shown in Figure 4. The peak power requirement was found to be approximately 5.0 kW. Considering the losses at each stages of transmission of power (source, converter and inverter), the required power to be provided by the E-rickshaw energy source was calculated. Assuming a constant source voltage of 48V (most of the E-rickshaws in India are operated with 48V Lead acid battery), the source current is calculated and is shown in Figure 5.

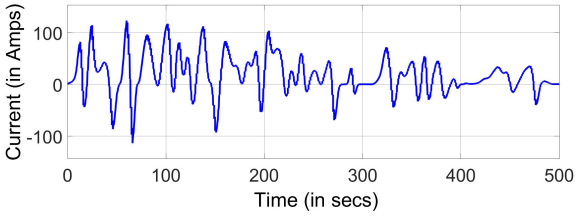


Figure 5: Source current for IDC

The proposed E-rickshaw is expected to have a driving range of 75 kms and the total distance travelled a day is expected to be 150 kms, there by calling for twice charging a day. The normal charging period is 2.5-3 hours. Figure 6 (a) shows the average speed at which E-rickshaw commutes inside a city. The vehicle covers a total distance of 2.548 kms at the end of 500 secs (Figure 6 (b)) at an average speed of 18.3 kmph.

Table 2: Forces Acting on an E-rickshaw

Force	Definition	Expression
Rolling Resistance force, F_{rr}	The force experienced by the vehicle due to friction between the tyres and the running surface.	C_r.M.g
Aerodynamic drag, F_{ad}	The force experienced by the vehicle due to friction with the surrounding air.	$\frac{1}{2}.\rho.\text{A}.C_d.V^2$
Grade Resistance force, F_{gr}	The force required to drive the vehicle upward on a slope.	M.g.Sinθ
Acceleration force, F_{af}	The force which is needed to accelerate the vehicle for different running velocities.	M.a

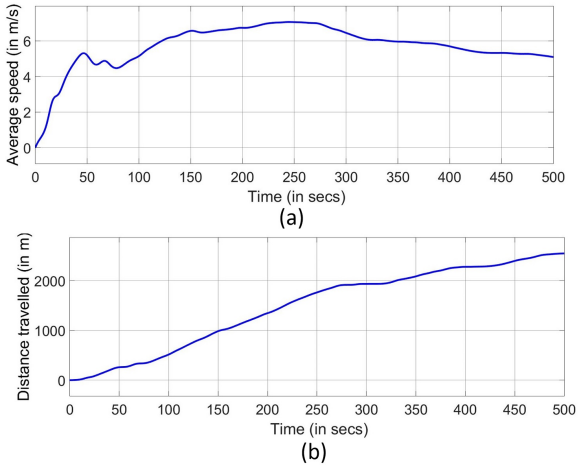


Figure 6: (a) Average Speed of an E-rickshaw (b) Total Distance travelled at the end of 500 secs

4 Fuzzy Logic based Energy Management

A fuzzy logic based approach for energy management was adopted in this paper due to its adaptive nature to the non-linear environment of electric vehicles. The source current shown in Figure 5, the battery and ultracapacitor state of charge (SOC) are taken as reference to split the energy requirement between battery and ultracapacitor for propelling the E-rickshaw as shown in Figure 7.

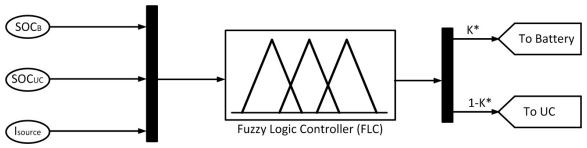


Figure 7: Fuzzy Logic controller

The three fundamental rules adopted for ensuring an optimum energy management was;

- Battery will be providing the average power requirements (cruising).
- Ultracapacitor will be providing the peak power requirements (accelerating).
- Ultracapacitor will be absorbing the generated braking power which shall be followed by battery, if necessary (regenerative braking).

Battery, being a high energy density device is designed to provide the average power requirement and ultracapacitor is designed to provide the peak power requirement as it is a high power density device and can charge/discharge at a higher rate. The battery is designed not to discharge at more than 2C rate where C-rate is the rate at which a battery is discharged with respect to its maximum capacity. Based on the power requirements, the battery and ultracapacitor ratings are chosen as shown in Table 3

Table 3: Ratings of Battery and Ultracapacitor

Function	Battery	UC
Type	Li-ion	Maxwell
Rating	48V, 20 AH	48V, 165F
Cell Specification	3.7 V	2.7 V
Cell Combination	13-series	18-series
Stored Energy	960 Wh	53 Wh
Absolute Maximum Current	60 A	1900 A
Weight	8 KG	14.2 KG
Charge temperature	-40 to 65°C	0 to 45°C
Discharge temperature	-40 to 65°C	-20 to 60°C

4.1 Cruising mode of operation

When an E-rickshaw is running at a constant speed thereby consuming a non-varying power, the vehicle can be said to be in cruising or coasting mode of operation. The source current requirement will be average in this case whereby enabling the battery to supply the total power required (case-I; see Table 4). Depending on the state of charge (SOC) of battery and ultracapacitor, either of them is chosen for acting as the propelling power source of the vehicle (case-II). A threshold SOC of 20% is chosen for both battery and ultracapacitor for the state of health (SOH) concerns of both. Falling of SOC below the threshold value will alert the driver for an immediate stop of the vehicle and signals external plug-in charging of the battery (case-III). An average current of 17.6A was found to be drawn from the source during cruising mode of operation. Hence an upper threshold of 40A (considering 2C-rate of discharge) was set as the maximum possible current to be drawn from the battery.

4.2 Accelerating mode of operation

During acceleration the power requirement of the e-rickshaw will reach its peak thereby making the battery unable to supply the same.

Table 4: Fuzzy rules for Optimum Energy Management between Battery and Ultracapacitor

Mode	Case	(if) Source current	(and) Battery SOC	(and) UC SOC	Battery	UC
Cruising	I	$0 < i \leq 40$	$20\% \leq SOC \leq 100\%$	XX	1	0
	II	$0 < i \leq 40$	$0\% \leq SOC < 20\%$	$20\% \leq SOC \leq 100\%$	0	1
	III	$0 < i \leq 40$	$0\% \leq SOC < 20\%$	$0\% \leq SOC < 20\%$	0	0
Accelerating	IV	$40 < i \leq I_{MAX}$	XX	$20\% \leq SOC \leq 100\%$	0	1
	V	$40 < i \leq I_{MAX}$	XX	$0\% \leq SOC < 20\%$	0	0
Braking	VI	i < 0	XX	$0\% \leq SOC < 100\%$	0	1
	VII	i < 0	$0\% \leq SOC < 100\%$	= 100%	1	0
	VIII	< 0	= 100%	= 100%	0	0

Ultracapacitor (UC) hence will take care of the peak power requirements of the vehicle being a high power density device (case IV). The UC is designed based on the maximum current the device has to support (here approximately 120A) for a short duration of time. The capacitance of UC is calculated using Eq 2.

$$dV = \frac{i}{C}.dt + iR \tag{2}$$

where, dV is the change in voltage in volts
i is the current in Amperes
dt is the total discharge time in seconds and,
R is the equivalent series resistance in ohms
The change in voltage is the desired maximum voltage drop of an ultracapacitor. Since 75% of energy in a UC gets consumed within 50% of terminal voltage drop, the change in voltage is kept as 24V (i.e., 50% of 48V). The total discharge time is kept equal to the vehicle test time i.e., 500 secs. Assuming the time constant RC=0.7, substituting for R in Eq 2 yields,

$$C = \frac{i}{dV}.dt + i.\frac{0.7}{dV} \tag{3}$$

For an optimum calculation of capacitance of the ultracapacitor, the regenerative power gained during deceleration/braking is also considered. Hence the capacitance is calculated considering the average time and peak power requirement during the stages (acceleration/regenerative braking) it is needed by the vehicle. If the UC can no longer meet the power requirements of the vehicle (case-V), an external/internal charging of the same has to be done for efficient performance of the vehicle.

4.3 Regenerative Braking mode of operation

E-rickshaw, while decelerating/braking will be having a considerable amount of energy stored in it in the form of kinetic energy. In conventional vehicles this energy is wasted as heat energy in brake drums through mechanical braking as there is no provision to store it. In an E-rickshaw this energy can be stored in both battery and ultracapacitor with ultracapacitor being a more suitable choice due to its higher charge/discharge rate (case-VI). Battery cannot be exposed to higher currents for long time as it effects its state of health which makes it a lesser preferable choice for absorbing the regenerative braking power. Depending on the state of charge (SOC) of both battery and ultracapacitor, this recovered energy can be stored in either of them. If the UC is completely charged then the remaining regenerative power will be absorbed by battery depending on its SOC (case-VII). If both energy sources are fully charged and still there is a flow of regenerative power then it will be dissipated across the available mechanical brakes (case-VIII).

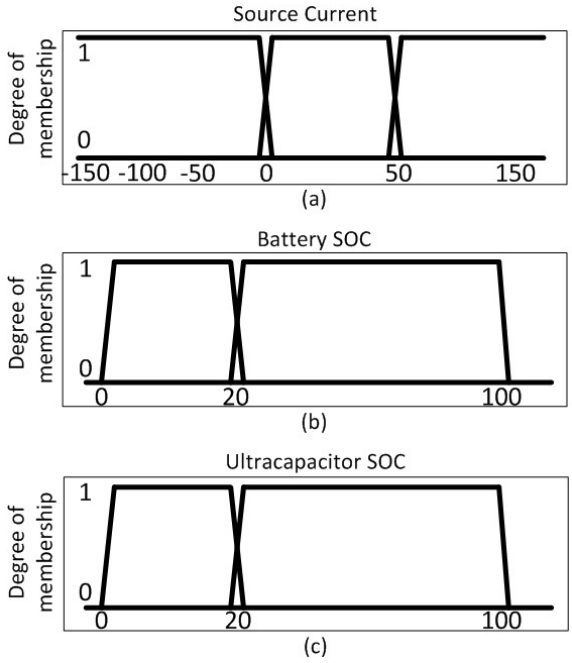


Figure 8: Membership functions for (a) Source current (b) Battery SOC and (c) Ultracapacitor SOC

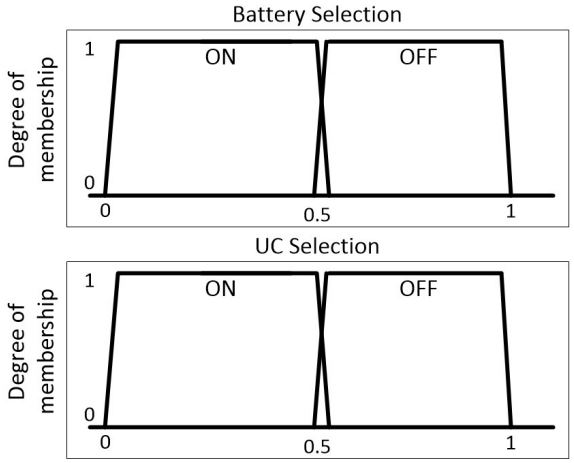


Figure 9: Membership functions for (a) Battery Selection and (b) Ultracapacitor Selection

The fuzzy rules set for an optimum energy management considering the efficient absorption of regenerative power is discussed in Table 4. The input membership functions of the fuzzy logic controller namely source current, battery SOC and ultracapacitor SOC are shown in Figure 8. The output membership functions for selection of energy sources are shown in Figure 9.

5 Simulation Results

Th vehicle has been tested in Matlab-Simulink environment for different conditions of battery and ultracapacitor. The results are discussed in detail below.

5.1 Case-I: Both Battery and UC fully charged (Batt-SOC=100% & UC-SOC=100%)

Both battery and ultracapacitor is assumed to be fully charged at start (i.e., Initial SOC = 100%). Depending on the power requirement, battery SOC and ultracapacitor SOC, the load power is shared between battery and ultracapacitor as shown in Figure 10.

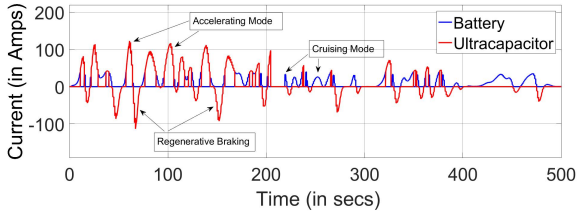


Figure 10: Case-I: Power split between Battery and UC

It is evident that during the cruising mode of operation battery is supplying the required power and during peak power requirement (accelerating mode) and while regenerative braking ultracapacitor supplies/absorbs power respectively.
The state of charge (SOC) of the battery and ultracapacitor is shown in Figure 11.

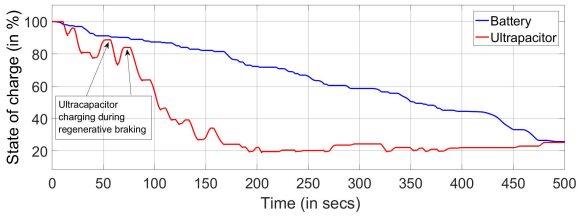


Figure 11: Case-I: Battery and UC SOC comparison

The terminal voltage characteristics (V_T) of the battery and ultracapacitor when operated at Indian drive cycle is shown in Figure 12.

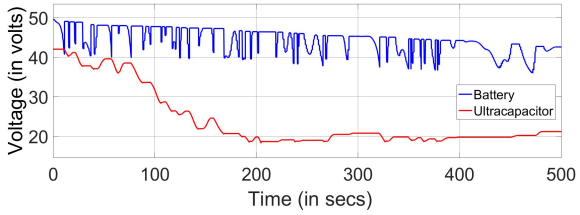


Figure 12: Case-I: Battery and UC Voltage (V_T) comparison

The vehicle is able to operate during all three modes of driving i.e., cruising, accelerating and regenerative braking. During regenerative braking the ultracapacitor is being charged as shown in Figure 11. Battery is able to supply the required power for a longer duration when hybridized with an ultracapacitor when compared to operated alone thereby improving its state of health. Ultracapacitor has successfully absorbed the generated power during braking thereby improving vehicle performance and energy efficiency.

5.2 Case-II: Battery fully charged but UC less charged (Batt-SOC=100% & UC-SOC=25%)

Battery is assumed to be fully charged at start (i.e., Initial SOC = 100%) and ultracapacitor is assumed to be 75% discharged.

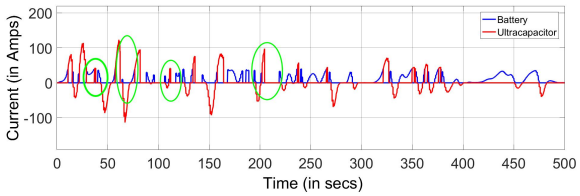


Figure 13: Case-II: Power split between Battery and UC

Depending on the power requirement, battery SOC and ultracapacitor SOC, the load power is shared between battery and ultracapacitor as shown in Figure 13 for a standard Indian drive cycle (IDC).
During the cruising mode of operation battery is supplying the required power but during peak power requirement (accelerating mode) the ultracapacitor is unavailable to meet the system requirements once its SOC level reaches below 20% (see the highlighted areas in Figure 13). While regenerative braking, ultracapacitor absorbs power which is later used to meet the vehicle requirements while accelerating mode of operation. The state of charge (SOC) of the battery and ultracapacitor is shown in Figure 14.

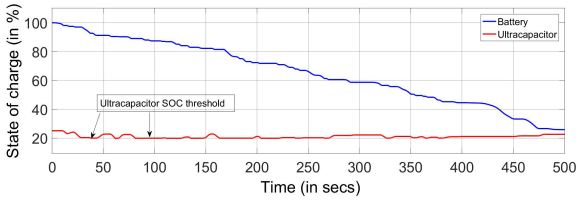


Figure 14: Case-II: Battery and UC SOC comparison

The terminal voltage characteristics (V_T) of the battery and ultracapacitor is shown in Figure 15.Since the ultracapacitor is discharged already by 75%, its terminal voltage will be 24V, as the depth of discharge will be 75% in an ultracapacitor by the time its terminal voltage falls by half.

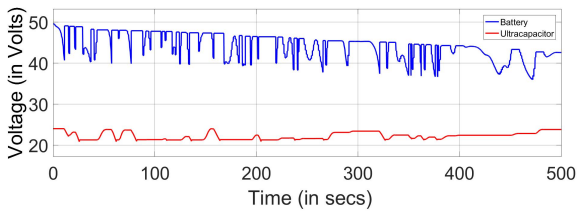


Figure 15: Case-II: Battery and UC Voltage (V_T) comparison

The vehicle is able to operate only during cruising mode of driving. While accelerating the ultracapacitor is only able to meet the demands of the vehicle till its SOC reaches the bottom threshold of 20%. Once the SOC of the ultracapacitor falls below its lower threshold, the vehicle may no longer be able to accelerate thereby hindering its performance hugely. The driver may have to charge the ultracapacitor externally (Plug-In) or internally (through battery) to ensure smooth operation of the vehicle. Interestingly, during regenerative braking the ultracapacitor can still be charged as shown in Figure 14. Once the SOC level reaches above the lower threshold of 20% it can resume it duties and can contribute to the vehicle operation as before.

5.3 *Case-III: UC fully charged but Battery less charged (Batt-SOC=25% & UC-SOC=100%)*

Battery is assumed to be initially charged at 25% (i.e., Initial SOC = 25%) and ultracapacitor is assumed to be fully charged.

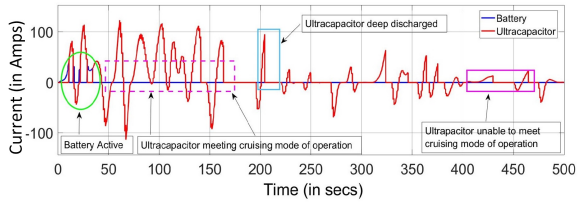


Figure 16: Case-III: Power split between Battery and UC

Depending on the power requirement, battery SOC and ultracapacitor SOC, the load power is shared between battery and ultracapacitor as shown in Figure 16.

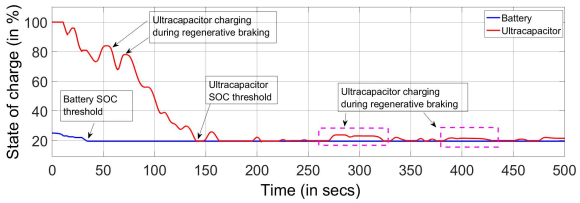


Figure 17: Case-III: Battery and UC SOC comparison

Since battery is very low charged, it cannot meet the vehicles' demands after its SOC falls below the threshold value of 20%. Once the battery SOC is below 20% the ultracapacitor may continue to meet the vehicle average and peak demands till its SOC reaches 20% (see the highlighted areas in Figure 16). While regenerative braking, ultracapacitor absorbs power which is later used to meet the vehicle requirements during accelerating mode of operation. The state of charge (SOC) of the battery and ultracapacitor is shown in Figure 17.

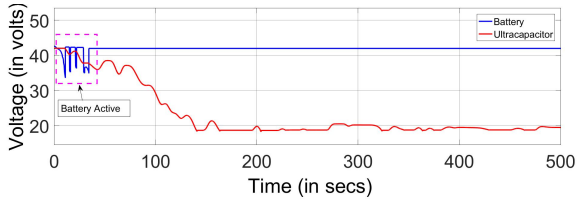


Figure 18: Case-III: Battery and UC Voltage (V_T) comparison

The terminal voltage characteristics (V_T) of the battery and ultracapacitor when operated at Indian drive cycle is shown in Figure 18. Since the battery is discharged already by 75%, its terminal voltage will be non-linearly reduced initially. The battery is active till its SOC reaches the lower threshold value. Once both the energy sources are below its lower threshold value of SOC, the driver has to perform an external charging of battery accompanied by an external (Plug-In) or internal (through battery) charging of ultracapacitor to ensure smooth operation of the vehicle. While regenerative braking, ultracapacitor absorbs power which is later used to meet the vehicle requirements during accelerating mode of operation.

6 Implementation Results

The proposed fuzzy logic based power-split between battery and ultracapacitor was experimentally tested in laboratory environment using the Xilinx made Spartan3 FPGA board (XC3S5000). A two-channel Agilent make DSO-X2002A is used to record the performance in real-time. The Indian city drive cycle generated from actual measured data is recreated in lab and is shown in Figure 19.

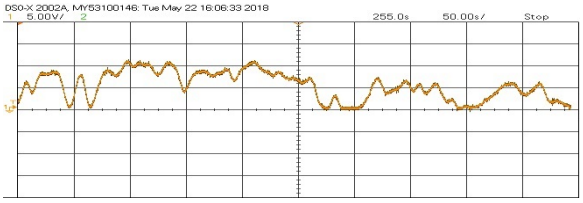


Figure 19: Real-time Indian City Drive Cycle at laboratory environment

The current to be drawn from the energy source(s) to meet the vehicle demands are shown in Figure 20.

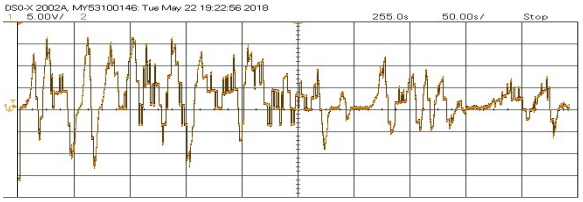


Figure 20: Source current requirement of E-rickshaw

The vehicle simulated in Matlab-Simulink was successfully recreated in a Real-time environment. The current drawn from the energy source for various modes of vehicle operation was observed to be closely matched with simulation results.
Al the three cases of vehicle operation based on battery and ultracapacitor state of charge (SOC) tested earlier with Matlab-Simulink was tested in laboratory environment. The results were found to be matching and is discussed briefly below.

6.1 *Case-I: Both Battery and UC fully charged (Batt-SOC=100% & UC-SOC=100%)*

Both battery and ultracapacitor is fully charged. As seen earlier, the vehicle is able to operate during all three modes of driving i.e., cruising, accelerating and regenerative braking. Ultracapacitor is being charged during regenerative braking as shown in Figure 21.

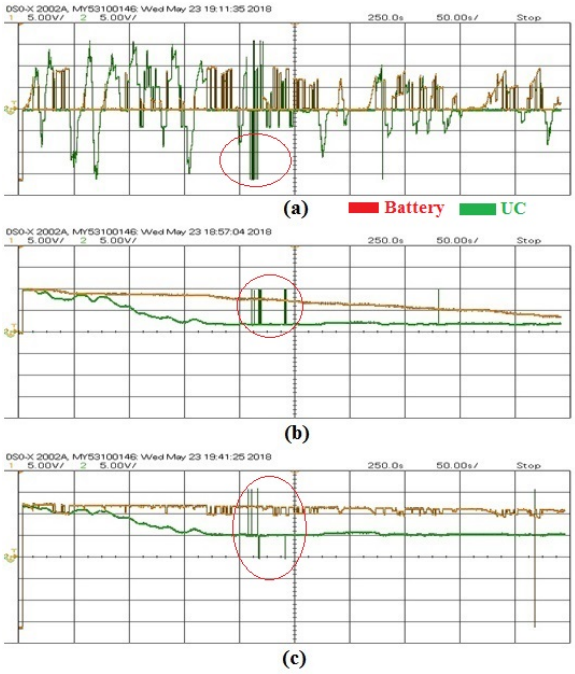


Figure 21: Case-I: Battery and UC (a) Current (b) SOC and (c) Voltage (V_T) comparison in Real-time laboratory environment

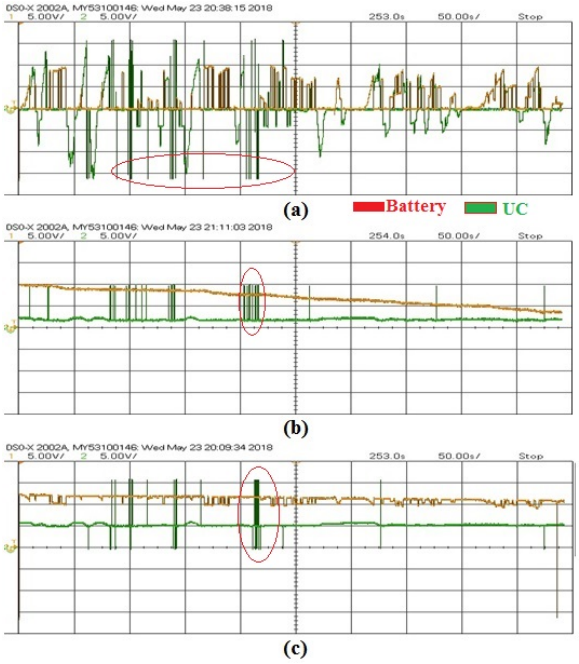


Figure 22: Case-II: Battery and UC (a) Current (b) SOC and (c) Voltage (V_T) comparison in Real-time laboratory environment

As the system is run for 500 seconds, few noises are present in the results which are highlighted in the results plotted. For average power requirements battery is acting as the energy source and for peak power requirements ultracapacitor is the source. Vehicle performance and energy efficiency is improved as ultracapacitor has successfully absorbed the generated power during braking.

6.2 *Case-II: Battery fully charged but UC less charged (Batt-SOC=100% & UC-SOC=25%)*

Battery is full charged but ultracapacitor is charged at 25% SOC. It can be seen that during the cruising mode of operation battery is supplying the average power required but during accelerating mode the ultracapacitor is unavailable to meet the system peak power requirements once its SOC level reaches below 20% as shown in Figure 22. Ultracapacitor absorbs the regenerative braking power. Ultracapacitor is only able to meet the demands of the vehicle till its SOC reaches the bottom threshold of 20%. Once the SOC falls below its lower threshold, the vehicle may no longer be able to accelerate thereby hindering its performance hugely calling for an external/internal charging based on system architecture.

6.3 *Case-III: UC fully charged but Battery less charged (Batt-SOC=25% & UC-SOC=100%)*

Battery is charged at 25% whereas ultracapacitor is fully charged. Since battery is very low charged, it cannot meet the vehicles' demands after its SOC falls below the threshold value of 20%. Ultracapacitor may continue to meet the vehicle average and peak demands till its SOC reaches 20% once the battery SOC is below 20% as shown in Figure 23. Ultracapacitor absorbs power while regenerative braking.

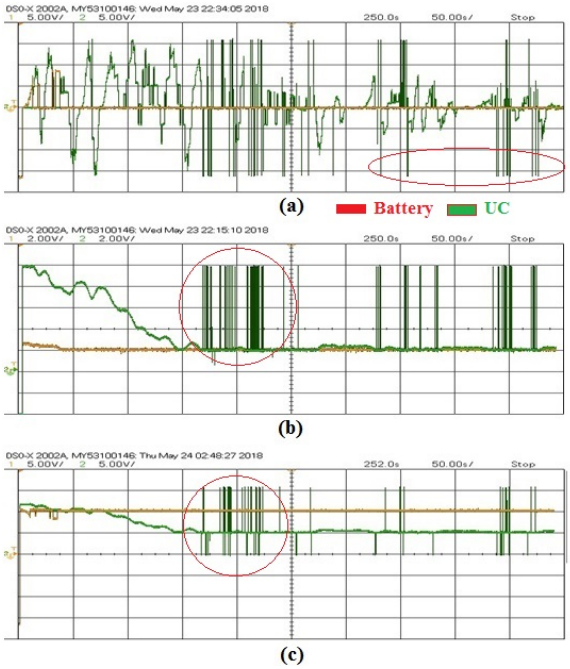


Figure 23: Case-III: Battery and UC (a) Current (b) SOC and (c) Voltage (V_T) comparison in Real-time laboratory environment

The battery is active energy source till its SOC reaches the lower threshold value of 20%. Once both the energy sources are below its lower threshold value of SOC, the driver has to perform an external charging of battery accompanied by an external (Plug-In) or internal (through battery) charging of ultracapacitor ensure smooth operation of the vehicle.

7 Discussion and conclusion

Rickshaws, even-though being a common mode of transportation is subjected to very less research to improve its performance. Conventional E-rickshaws which employ a single energy source, has a lot of limitation in terms of driving range. In Indian scenario, E-rickshaws‘ success will be always measured by its driving range and cost. It is identified that combining a high power density device (UC) for sharing peak power demand with the traditional high energy density device (battery) will improvise the vehicle performance which was not earlier attempted in E-rickshaws. An improved SOC and terminal voltage characteristics can increase the driving range of the E-rickshaw. Also since the UC is able to meet the peak power demands, battery can be sized for a much lesser average current demand. This will considerably reduce the size and cost of the vehicle. Since the battery in a traditional e-rickshaw occupies almost 15-20% of its kerb weight, any minute reduction in weight and size of the battery will always result in better performance. Also UC possess faster charge/discharge characteristics than battery which can be utilized in absorbing regenerative power during deceleration and braking in a better way.
This paper have attempted for the hybridization of energy sources for an E-rickshaw system architecture in Indian scenario. The performance improvement was justified using simulation and implementation in real-time results in this paper. Successful implementation of a fuzzy logic based energy management algorithm has enhanced the performance of E-rickshaw. The vehicle dynamics was calculated real-time based on an Indian city drive cycle calculated using a GPS based PerformanceBox tool of VBOX motorsport. The vehicle was tested for three different driving modes of operation and three different battery and ultracapacitor state of charge conditions. It was found that the algorithm was working perfectly with less computation time and complexity. Implementation of a hybrid energy source for an E-rickshaw can extend its driving range. Also successful absorption of the generated regenerative power can further enhance the vehicle performance.

References

[1] J. Bremner, A. Frost, C. Haub, M. Mather, K. Ringheim, E. Zuehlke, et al., “World population highlights: Key findings from PRB’s 2010 world population data sheet,” Population Bulletin, **65**(2), 1–12, 2010.

[2] S. K. Guttikunda, R. Goel, P. Pant, “Nature of air pollution, emission sources, and management in the Indian cities,” Atmospheric environment, **95**, 501–510, 2014, doi:https://doi.org/10.1016/j.atmosenv.2014.07.006.

[3] BBC, “Indian cities dominate world air pollution list,” 2018.

[4] H. Mayer, “Air pollution in cities,” Atmospheric environment, **33**(24-25), 4029–4037, 1999, doi:https://doi.org/10.1016/S1352-2310(99)00144-2.

[5] A. Mani, M. Pai, R. Aggarwal, “Sustainable urban transport in India: role of the auto-rickshaw sector,” 2012.

[6] J. Shah, N. Iyer, “Module 4c: two-and three-wheelers,” Sustainable transport: a sourcebook for policy-makers in developing cities. GTZ, Eschborn, 2004.

[7] J. Pucher, N. Korattyswaroopam, N. Ittyerah, “The crisis of public transport in India: overwhelming needs but limited resources,” Journal of public transportation, **7**(3), 5, 2004, doi:http://doi.org/10.5038/2375-0901.7.4.1.

[8] M. S. Kumar, S. T. Revankar, “Development scheme and key technology of an electric vehicle: An overview,” Renewable and Sustainable Energy Reviews, **70**, 1266–1285, 2017.

[9] V. Gulhane, M. Tarambale, Y. Nerkar, “A scope for the research and development activities on Electric Vehicle technology in Pune city,” in Electric and Hybrid Vehicles, 2006. ICEHV’06. IEEE Conference on, 1–8, IEEE, 2006.

[10] D. Majumdar, T. Jash, “Merits and Challenges of E-Rickshaw as An Alternative form of Public Road Transport System: A Case Study in the State of West Bengal in India,” Energy Procedia, **79**, 307–314, 2015, doi: https://doi.org/10.1016/j.egypro.2015.11.492.

[11] S. Tatipamula, “Study of a Parallel Electric Hybrid Three-Wheeled Motor Taxi,” IJCA Proceedings on International coriference on Green Computing and Techn% gy ICGCT, **1**, 38–41, 2013.

[12] T. Hofman, S. Van Der Tas, W. Ooms, E. Van Meijl, B. Laugeman, “Development of a micro-hybrid system for a three-wheeled motor taxi,” World Electric Vehicle Journal, **3**(3), 1, 2009.

[13] M. Mruzek, I. Gajdáč, L. Kučera, T. Gajdošík, “The possibilities of increasing the electric vehicle range,” Procedia engineering, **192**, 621–625, 2017.

[14] S. Manzetti, F. Mariasiu, “Electric vehicle battery technologies: From present state to future systems,” Renewable and Sustainable Energy Reviews, **51**, 1004–1012, 2015.

[15] A. Poullikkas, “Sustainable options for electric vehicle technologies,” Renewable and Sustainable Energy Reviews, **41**, 1277–1287, 2015.

[16] R. Kumar, “Evolution of solar rickshaw technology vis-à-vis economic feasibility,” in Advances in Electrical, Electronic and Systems Engineering (ICAEES), International Conference on, 646–651, IEEE, 2016.

[17] M. Sameeullah, S. Chandel, “Design and analysis of solar electric rickshaw: A green transport model,” in Energy Efficient Technologies for Sustainability (ICEETS), 2016 International Conference on, 206–211, IEEE, 2016.

[18] P. Mulhall, S. M. Lukic, S. G. Wirasingha, Y.-J. Lee, A. Emadi, “Solar-assisted electric auto rickshaw three-wheeler,” IEEE transactions on vehicular technology, **59**(5), 2298–2307, 2010.

[19] M. A. Mallouh, L. McInnes, B. Surgenor, B. Peppley, “Intelligent Control for optimal performance of a fuel cell Hybrid Auto Rickshaw,” Energy Procedia, **29**, 367–376, 2012.

[20] B. Masood, R. A. Naqvi, R. M. Asif, “Designing of a control scheme for the solar rickshaw in comparative study with conventional auto rickshaw,” in Engineering Technology and Technopreneuship (ICE2T), 2014 4th International Conference on, 324–329, IEEE, 2014.

[21] M. Hannan, M. Hoque, A. Mohamed, A. Ayob, “Review of energy storage systems for electric vehicle applications: Issues and challenges,” Renewable and Sustainable Energy Reviews, **69**, 771–789, 2017.

[22] G. Ren, G. Ma, N. Cong, “Review of electrical energy storage system for vehicular applications,” Renewable and Sustainable Energy Reviews, **41**, 225–236, 2015.

[23] A. K. Rohit, K. P. Devi, S. Rangnekar, “An overview of energy storage and its importance in Indian renewable energy sector: Part I–Technologies and Comparison,” Journal of Energy Storage, **13**, 10–23, 2017.

[24] A. Singh, P. Karandikar, N. Kulkarni, “Feasibility of energy storage systems in different applications for sustainable energy,” in Emerging Technological Trends (ICETT), International Conference on, 1–6, IEEE, 2016.

[25] A. F. Burke, “Batteries and ultracapacitors for electric, hybrid, and fuel cell vehicles,” Proceedings of the IEEE, **95**(4), 806–820, 2007.

[26] H. Fathabadi, “Fuel cell hybrid electric vehicle (FCHEV): Novel fuel cell/SC hybrid power generation system,” Energy Conversion and Management, **156**, 192–201, 2018.

[27] S. Ahmadi, S. Bathaee, A. H. Hosseinpour, “Improving fuel economy and performance of a fuel-cell hybrid electric vehicle (fuel-cell, battery, and ultracapacitor) using optimized energy management strategy,” Energy Conversion and Management, **160**, 74–84, 2018.

[28] L. Kouchachvili, W. Yaïci, E. Entchev, “Hybrid battery/supercapacitor energy storage system for the electric vehicles,” Journal of Power Sources, **374**, 237–248, 2018.

[29] N. Sulaiman, M. Hannan, A. Mohamed, E. Majlan, W. W. Daud, “A review on energy management system for fuel cell hybrid electric vehicle: Issues and challenges,” Renewable and Sustainable Energy Reviews, **52**, 802–814, 2015.

[30] S. S. Kulkarni, N. Gandhi, N. Chaithanya, S. Govindarajan, “Ultra-Capacitor based Hybrid Energy Storage and Energy Management for Mild Hybrid Vehicles,” Technical report, SAE Technical Paper, 2014.

[31] W. Enang, C. Bannister, “Modelling and control of hybrid electric vehicles (A comprehensive review),” Renewable and Sustainable Energy Reviews, **74**, 1210–1239, 2017.

[32] S. Zhang, R. Xiong, F. Sun, "Model predictive control for power management in a plug-in hybrid electric vehicle with a hybrid energy storage system," Applied Energy, **185**, 1654–1662, 2017.

[33] D. Zhou, A. Al-Durra, F. Gao, A. Ravey, I. Matraji, M. G. Simões, "Online energy management strategy of fuel cell hybrid electric vehicles based on data fusion approach," Journal of Power Sources, **366**, 278–291, 2017.

[34] M. Dawei, Z. Yu, Z. Meilan, N. Risha, "Intelligent fuzzy energy management research for a uniaxial parallel hybrid electric vehicle," Computers & Electrical Engineering, **58**, 447–464, 2017.

[35] J. Larminie, J. Lowry, Electric vehicle technology explained, John Wiley & Sons, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1171-1177 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

PV Integrated Recursive Least Mean Square Estimation Based Shunt Active Power Filter

Ragam Rajagopal[*,1], K. Palanisamy[1], S. Paramasivam[2]

[1]*School of Electrical & Electronics Engineering, VIT Vellore, 632014, India*

[2]*Functional Head (R&D), Danfoss Drives, Chennai, 602105, India*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 19 November, 2020
Online: 14 December, 2020

Keywords:
Shunt Active Filter
Adaptive Linear Control
Harmonics
Power Quality

ABSTRACT

This paper proposes a photovoltaic (PV) integrated shunt active power filter (SAPF) for reactive power compensation, uneven loading and compensation of harmonic current. The control of SAPF is implemented by using adaptive based current estimation algorithm to maintain power factor near to unity and source current sinusoidal at source side. The weights are calculated online by recursive least mean square algorithm once the three-phase load current is detected. Secondly, the unit vector templates (UVT) are multiplied with these weights and thus the load current with fundamental active component is derived. PV system enhanced by Maximum Power Point Tracking (MPPT) plays the role of maintaining the dc link voltage constant. Finally, a current control scheme based on hysteresis is used for converter switching and thus the source current to follow reference current precisely. The superiority of the proposed method of compensation of harmonic current and reactive power using SAPF with Recursive Least Mean Square (RLS) algorithm is proved by the simulation results based on MATLAB/SIMULINK.

1. Introduction

The pervasive use of converters based on power electronics has given rise to increasing use of non-linear and inductive loads. The adjacent loads are negatively affected by the voltage variation at point of common coupling (PCC) caused by the harmonic current component of non-linear loads linked to the same network [1-4]. The increased demand of reactive power diminishes the flow capability of active power, and increases feeder losses, whereas the ordinary operation of transformer and generator are adversely affected by the unbalanced system. IEC-61000 and IEEE-519-2014 standards are put into practice to minimize such problems. The introduction of custom power device technology such as unified power quality controller (UPQC), dynamic voltage regulator (DVR), distributed static compensator (D-STATCOM) are effective for the electrical distribution network that faces these issues (unbalanced load, reactive power and harmonics). It is found that SAPF is suitable to efficiently compensate the afore mentioned issues [5].

The control method used to determine the efficacy of the SAPF presented. It mainly relies on the fundamental current extraction algorithm and converter switching generation using extracted reference currents.

The concept here to is to separate load current into three constituent factors. They are fundamental active (i_{Lfp}), reactive(i_{Lfq}) and harmonic(i_{Lh}) components of load current. References [6-9] have reported the low pass and band pass filter (LPF-BPF) based fundamental active current extraction scheme, per phase computation, scheme based on neural network, instantaneous reactive power (IRP) theory, rotating reference frame and instantaneous symmetrical components. Artificial neural network-based extraction scheme requires training of load current datasets for different load conditions which makes system complex and suitable for very specific loads. IRP theory works very well under sinusoidal supply voltage but fails to separate out working and wattless components of the load current under distorted supply condition.

Adaptive filtering techniques are proposed in latest literature because of its merits such as robustness, instinctive monitoring and tracing, less real-time calculation and a fast-dynamic response even under complex load conditions [10-11]. References [12-16] propose various adaptive filtering techniques such as adaptive notch filter, improved sinusoidal notch filter, fuzzy based adaptive control and synchronous adaptive filters. Reference [17, 18] presents least mean square (LMS) based fundamental active current extractor scheme. Least mean squares (LMS) algorithms has the least difficulty and most effectively applied versatile

[*]Corresponding Author: Ragam Rajagopal, ragamrajagopal@gmail.com

algorithms. The disadvantage of using LMS algorithm is its slow convergence rate. Excellent performance and prominent fidelity make the recursive least square algorithm (RLS) superior, however they accompany expanded intricacy and computational expense. In adaptive filtering applications, RLS approaches Kalman filter with reduced computational complexity in the signal processor. At the cost of requiring more computations, the RLS approach has quicker convergence characteristics compared to LMS algorithm and less error with respect to the unknown system. Reference [19] have reported PV integrated system-based voltage source converter (VSC) to keep a constant voltage at dc link. The development of PV technology has given motivation to the concept of combining clean and free solar energy for SAPF.

This paper proposes a PV integrated 3-phase shunt active filter with recursive least mean square based extractor scheme for fundamental active current extraction by recursively updating the weight of the input signal. The recursive approach finds the filter coefficients that minimize a weighted linear least squares cost function relating to the input signals. Minimization of the total weighted squared error between the desired signal and the output is the main objective of this approach.

It also uses variable forgetting factor to get fast merging rate and less steady state error. The usage of forgetting factor makes the de-emphasis of previous old data as compared to the new data. Moreover, PV system is integrated to keep the voltage of dc link constant. The performance analysis of RLS based SAPF reveals IEEE standard 519 compliant functionality under various operating conditions.

2. System Configuration

The Shunt active filter which act as voltage source converter is coupled to PCC through interfacing inductor L_f. The key functions of the SAPF include a) the elimination of harmonic component of load current thereby maintaining sinusoidal nature of source current with THD<5% b) maintaining source current in phase with pcc voltages achieving near unity pf and c) unbalanced load current compensation by maintaining source currents balanced irrespective of load conditions. The implementation of SAPF requires accurate connection of interlinking inductors, capacitor in the dc link, current and voltage sensors. The control algorithm must be fast enough for the successful functioning of the SAPF. The adaptive RLS algorithm used in this work has higher convergence rate as compared to other variants of least mean square algorithms. Figure 1 indicates the typical illustrative figure of 3-phase SAPF. Notations used in Figure 1 are described in section 2.1.

Three phase (i.e. a, b and c) load currents are determined by means of current sensors. Similarly, source currents, i_{sa}-i_{sb}-i_{sc} are measured. Three voltage sensors are used to measured three phase voltages v_a-v_b-v_c at PCC. PLL will determine θ, based on PCC voltages which in turn is used to generate UVTs for the three phase system. Fundamental active current extractor which makes use of RLS based adaptive algorithm as illustrated in figure 2 utilize this θ, to determine i_{Lfp}.

The product of UVT and i_{Lfp} results in the production of the reference source currents, i_{saR}-i_{sbR}-i_{scR}, [19]. These reference source currents along with the instantaneous values of the three phase source currents are fed into the hysteresis controller for the production of gate signals. These gate signals are in turn fed to voltage source inverter (VSI) for the generation of compensating currents which are injected into the system. DC link voltage is regulated by means of the PV array. Economic reliability is enhanced since the VA rating of photovoltaic panel used need not be very high as required for the injection of active power since the compensating currents injected need to provide only for harmonic currents and reactive power.

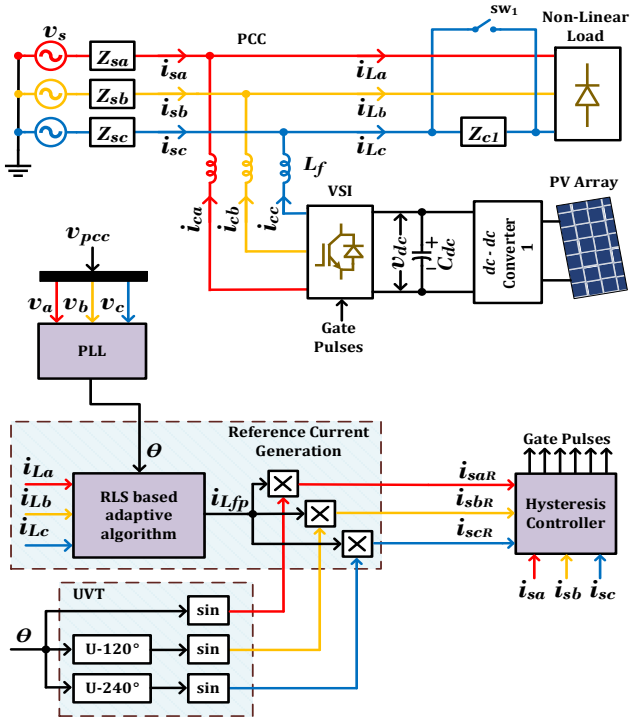


Figure 1: Block diagram of 3-phase SAPF

2.1. Nomenclature

a,b,c : Three Phases

v_{sa}-v_{sb}-v_{sc} : Supply voltage

v_a-v_b-v_c : PCC voltages

θ : instantaneous phase angle of fundamental component of supply voltage

Z_{sa}-Z_{sb}-Z_{sc} : Line impedances

i_{Lh},i_{Lfp},I_{Lfq}, : Harmonic load current component, Fundamental active and reactive components of load current

i_{La}-i_{Lb}-i_{Lc} : Load currents

u_{pa},u_{pb},u_{pc} : Unit vector template

i_{ca}-i_{cb}-i_{cc} : Compensating currents

i_{saR}-i_{sbR}-i_{scR} : Reference source currents

v_{dc} : DC link voltage

C_{dc} : Capacitor at dc link

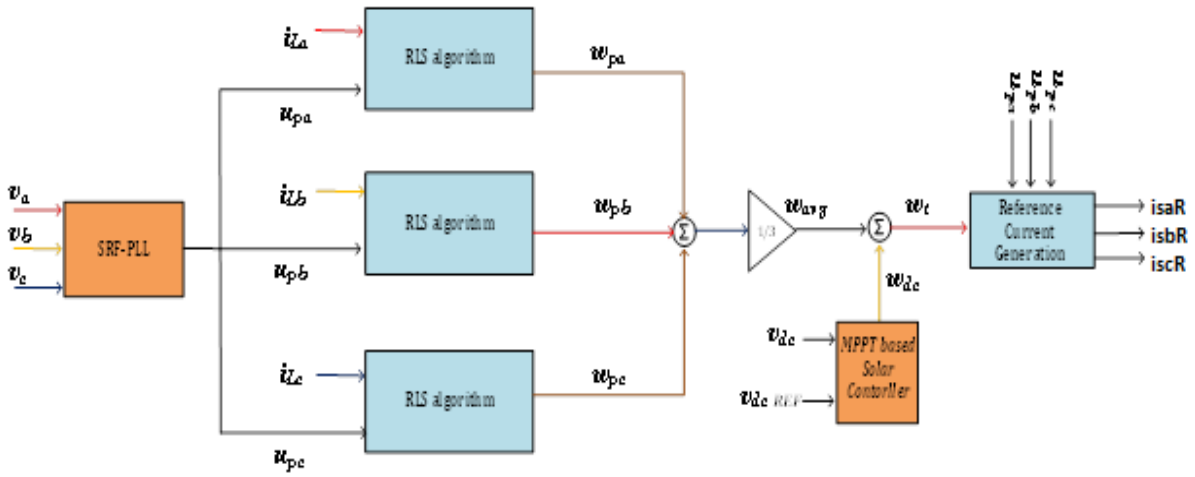


Figure 2: Control diagram for 3-phase SAPF

i_{Lp} : Fundamental active load current component's peak value

i_{Lq} : Fundamental reactive load current component's peak value

i_{Lm5}, i_{Lm7} : Peak value of 5th and 7th harmonic components of load current

i_{c1} : Fundamental compensating current component's peak value

i_{cm5}, i_{cm7} : Peak value of 5th and 7th harmonic components of compensating current

i_{sm1} : Fundamental source current component's peak value

i_{sm5}, i_{sm7} : Peak value of 5th and 7th harmonic components of source current

3. Control Scheme

The LS estimate of the filter coefficient w(n-1) at iteration n-1 is the basic technique behind RLS algorithm. This type of algorithm is known as recursive least mean square algorithm. This algorithm may be viewed as a special case of Kalman filter. Recursive filter is implemented by using the available starting conditions and then update the weight based on new arrived data. This leads to the minimization of cost function, where observed data's variable length is notated as n.

Cost function $J(n) = \sum_{k=1}^{n} w_n e^2(n)$ where w_n is weighting factor and e(n) is error.

This error e(n) can be expressed as,

$$e_j(n) = i_{Lj}(n) - i_{sj}(n) \quad (1)$$

where, j=a, b and c phase

$$e_j(n) = i_{Lj}(n) - w^T(n-1)x(n-1) \quad (2)$$

Weight updating equation for RLS control is given as,

$$g(n) = P(n-1)x(n)[\lambda + x^T(n)P(n-1)x(n)]^{-1} \quad (3)$$

$$P(n) = \lambda^{-1}P(n-1) - g(n)x^T(n)\lambda^{-1}P(n-1) \quad (4)$$

$$w(n) = w(n-1) + e(n)g(n) \quad (5)$$

where P(n) and g(n) represent the inverse input correlation matrix and Kalman gain vector. For algorithm learning, the term λ which is the forgetting factor plays a very major role, whose value is between 0 and 1.

The value of forgetting factor is closely related with convergence rate and tracking capability. In fixed forgetting factor based RLS algorithm, as λ increases and its value moves closer to 1, convergence rate increases and at the same time tracking speed will reduce. To improve the tracking capability of the algorithm λ value must be small which will diminish the convergence rate. These conflicting requirements are met by the usage of a variable forgetting factor RLS algorithm.

3.1. Generation of switching pulses for the operation of SAPF

Based on the three phase supply voltages, the control scheme shown in figure 2 will determine θ. This θ generates the three unit vector templates for the system. This UVT along with three phase load currents are the inputs to the control algorithm. The algorithm will generate weights corresponding to the three phase components of active power.

The mean weight magnitude equivalent to the components of active power (w_{pa}, w_{pb} and w_{pc}) from figure 2 is obtained by,

$$w_{avg} = \frac{w_{pa} + w_{pb} + w_{pc}}{3} \quad (6)$$

The dc bus voltage is sensed with the aid of a voltage sensor. This in turn is compared with the set point voltage of the dc bus (v_{dc}), and leads to the generation of the error signal (v_e). The MPPT based PI controller's tuned output (w_{dc}) ensures the steady maintenance of the dc bus voltage at its desired value.

The output of the PI controller at nth sampling instant is indicated below,

$$w_{dc}(n) = w_{dc}(n-1) + k_p(v_e(n) - v_e(n-1)) + k_i\, v_e(n) \quad (7)$$

where the gains of PI controller are notated as k_p and k_i

The sum of mean weight (w_{avg}) and PI controller's tuned output (w_{dc}) is utilized to find the active component of load current's weight.

$$w_t = w_{avg} + w_{dc} \quad (8)$$

UVT is used to calculate the reference source currents and denoted in (8) can be represented as,

$$i_{saR} = w_t u_{pa}, i_{sbR} = w_t u_{pb}, i_{scR} = w_t u_{pc} \quad (9)$$

The computation of VSC compensating current is carried out by subtracting the reference source currents (9) from load currents.

4. Performance Validation

MATLAB/SIMULINK platform is used to manifest the performance of the RLS based control algorithm. For various load conditions, validity of the proposed scheme has been checked as mentioned in the next section. The simulation parameters are detailed in Table 1. The SAPF comes into play at t=0.1 sec and the results of the three phase SAPF are as displayed in Figure 3.

Case 1: Steady state performance for Load-I

As shown in the Figure 3, from t=0.1s to 0.2s, undistorted v_{PCC}, supplies Load–I which has a Total Harmonic Distortion (THD) of 1.53% and RMS value of 230V. THD and max. value of load current is 27.16% and 1.84 A respectively.

Current harmonics is controlled by forcing peak value of fifth and seventh harmonic of compensating current to take a value as shown in Table 2. This will lead to a reduction in THD to 3.05%. Hence peak value of fifth and seventh harmonic of source current are found to be negligible. By ensuring source side pf at unity, enforcement of reactive power compensation is done by controlling the max. value of the compensating current's fundamental component to be at 0.17A. Allowing for a p-p ripple of 0.2 V, voltage at dc link, v_{dc} is maintained at a mean value of 220.2 V.

Table 1: System Parameters

System Parameters	Value
v_{pcc}(undistorted)/phase	230V, 50Hz
Load – I:	3-phase rectifier having load : (140+j31.42) Ω ;S_w closed; balanced load
Load – II:	3-phase inductive load having (90+j31.42) Ω ;S_w closed; balanced load
Load – III	3-phase rectifier having (90+j31.42)Ω ;S_w open; unbalanced load
Unbalanced load, Z_{c1}	(10+j3.14) Ω
Source impedance, Z_s	(0.02+j0.0314) Ω
DC link voltage, V_{dc}	220 V
Interfacing Inductor, L_f	4mH

Case 2: Dynamic performance under step change of load

Figure 3 manifests step load change from I to II at 0.2 sec, which demonstrates the SAPF performance under the condition of dynamic load change with the RLS control algorithm. From figure 3, peak value of source current under the transient condition is 2.02A. Values of all parameters in the new steady state condition with Load II is as indicated in Table 2. Compensation ensures that THD becomes well within the limit and also pf being improved to unity.

Under the sudden step load changing condition, voltage at dc link, v_{dc} droops to 215.8 V. The action of PV integration along with the MPPT control algorithm make sure that within 0.07 sec, voltage at dc link is restored to its set point value. The presented algorithm displays a satisfying performance by efficient tracking of the fundamental load current component, i_{Lfp} change from 1.28A to 1.90A.

Case 3: Steady state performance for unbalanced loading condition

Figure 3 shows the steady state unbalanced loading condition performance of 3-phase SAPF operating with RLS control scheme. Switch S_w is open at t=0.3sec to introduce the unbalanced loading scenario. The values related to unbalanced load currents and compensating currents for each phase are as indicated in Table 3. Unbalanced loading results in unbalanced compensating currents for the three phases. Each phase caters to the harmonic current and reactive power requirement set forth by the load. Once the load currents under unbalance are compensated, the source currents will automatically become balanced. The THD and max. magnitude for the source's currents i_{sa}, i_{sb} and i_{sc} are 3.21% and 1.61 A. The voltage at dc link, v_{dc} with p-p ripple of 0.21 V is sustained at a mean value of 221V. Also, the source power factor is kept at unity. This set up ensures unbalanced load compensation by the 3 phase SAPF with the presented control algorithm.

The results for the three cases are summarized in Table 2. From the table, it can be clearly understood that $i_{cm5} \cong i_{Lm5}$ and $i_{cm7} \cong i_{Lm7}$. This reveals harmonic current compensation with source current THD maintained well below the 5% limit. Unity power factor is well kept at supply end via reactive power compensation which is evident from the fact that $i_{cm1} \cong i_{Lq}$. Also unbalanced load current compensation is achieved by the algorithm which ensures that source current is balanced even under unbalanced load conditions. Thus the presented control scheme of 3 phase SAPF based on adaptive RLS algorithm is providing the necessary compensating currents, reactive power and unbalanced load compensation which leads to THD kept within 5% limit and unity power factor (UPF) at supply end.

From table 3, it is clear that convergence time, % steady state error and % THD is less for RLS based algorithm compared to LMS algorithms. Computational complexity is high for RLS algorithm since it considers the estimate of previous samples of output signal, error signal and filter weight. To include the effect of losses, distortion power should be taken into account because of the harmonic inducing nature non- linear loads connected to grid side. Thus the product of displacement power factor and distortion power factor gives the true power factor.

Table 2: Performance of the SAPF with the proposed control algorithm

Case	Category of load	v_{pcc}	i_L (A)				i_c (A)			i_s (A)			%THD	
			I_{Lp}	I_{Lq}	I_{Lm5}	I_{Lm7}	I_{cm1}	I_{cm5}	I_{cm7}	I_{sm1}	I_{sm5}	I_{sm7}	i_L	i_s
1	Load – I (balanced)	Undistorted (THD=1.53%)	1.22	0.14	0.25	0.13	0.16	0.24	0.13	1.28	0.01	0.01	27.16	3.05
2	Load – II (balanced)	Undistorted (THD=1.53%)	1.93	0.31	0.41	0.27	0.33	0.41	0.26	1.90	0.0	0.0	23.20	2.30
3	Load – III (unbalanced)	Undistorted (THD=26.5%)	1.78	0.21	0.28	0.34	0.25	0.27	0.34	1.61	0.01	0.01	31.6	3.21
			1.33	0.31	0.37	0.36	0.35	0.37	0.36	1.61	0.01	0.01	34.9	3.21
			0.45	0.2	0.13	0.31	0.40	0.12	0.31	1.61	0.01	0.01	34.8	3.21

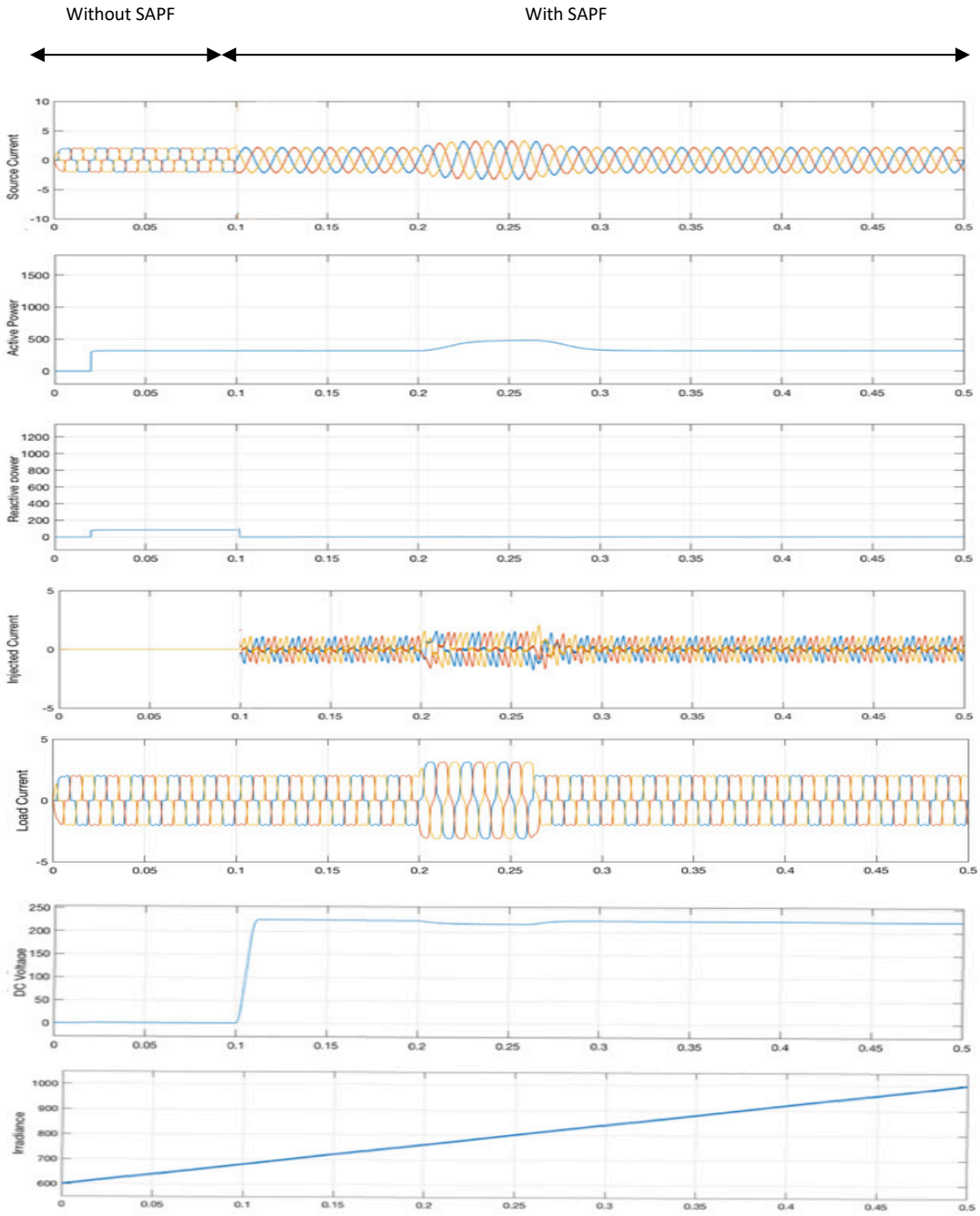


Figure 3: Simulation results for operation of the three phase SAPF with proposed algorithm

Table 3: Performance comparison of RLS algorithm with LMS algorithm [18]

Function	NLMS	VSSLMS	RLS
Convergence time	0.3 sec	0.12 sec	0.06 sec
Steady state error	2.11	0.85	0.56
THD	4.32	2.13	1.48
Complexity	Less	Less	High

Table 4: Power factors under varying load conditions

Case	Load	Distortion Factor	Displacement factor	True Power Factor
1	Load – I (balanced)	0.9995	0.9998	0.999
2	Load – II (balanced)	0.9997	0.9997	0.999
3	Load – III (unbalanced)	0.9995	0.9998	0.999

The equations (10-12) implies that the overall power factor gets diminished by the term cos γ. This leads to the conclusion that the average power transferred to the load gets diminished by cos γ, which depends on harmonic distortion induced in load current by the non-linear loads.

$$DF = \cos\gamma = \frac{I_{1(rms)}}{I_{(rms)}} \tag{10}$$

$$DPF = \cos\phi_1 = \frac{P_1}{S_1} \tag{11}$$

$$\begin{aligned} Power\ \ factor &= \frac{P}{S} = \frac{VI_1\cos\phi_1}{VI} \\ &= \left(\frac{I_1}{I}\right)\cos\phi_1 = \cos\gamma \times \cos\phi_1 \end{aligned} \tag{12}$$

The proposed SAPF configuration, maintains displacement power factor at 0.999 and the distortion power factor at near unity, for all the analysed load conditions. This leads to a magnitude of 0.999 for the true power factor. Table 3 manifests the true power factor under varying load conditions considered here.

5. Conclusion

This paper illustrates the performance aspects of PV powered RLS based shunt active power filter control utilizing the MATLAB/SIMULINK platform. Fundamental active components reference from the load currents are separated by the control law, based on the load currents, phase angle and reference currents of the previous sample. Load current's fundamental active component is only drawn from the source side and this is performed by the control system by ensuring that source currents track the reference currents with minimum steady state error. This makes the SAPF satisfy the reactive and harmonic power requisite of the considered non- linear load. Even under unbalanced loading condition, the source currents are in perfect balance. The put forward theory is validated by the results which shows that the source currents are drawn at UPF and they comply with the IEEE standard 519. Also, it is shown that dc voltage is maintained constant using PV integration into the network. With RLS algorithm, the 3-phase SAPF ensures sinusoidal and balanced supply currents which are maintained at UPF at various load conditions along with THD maintained under the 5% limit.

References

[1] B. Singh, A. Chandra and K. Al-Haddad, "Power Quality: Problems and Mitigation Techniques", John Wiley and Sons, 2015.

[2] R. Rajagopal, K. Palanisamy, S. Paramasivam, "A Technical review on control strategies of active power filters", IEEE 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), TIST, Cochin. 1-6, 2018, doi: 10.1109/ICETIETR.2018.8529008.

[3] K.S. Rajesh, S.S. Dash, Ragam Rajagopal, R.Sreedhar "A review on control of ac microgrid", Renewable & Sustainable Energy Reviews, **71**, 814-819, 2017, doi: 10.1016/j.rser.2016.12.106.

[4] K.S. Rajesh, S.S. Dash, Ragam Rajagopal, "Hybrid improved firefly-pattern search optimized fuzzy aided PID controller for automatic generation control of power systems with multi-type generations", Swarm and Evolutionary Computation, **44**, 200-211, 2019, doi:10.1016/j.swevo.2018.03.005.

[5] S. R. Arya and B. Singh, "Performance of DSTATCOM using leaky lms control algorithm", IEEE J. Emerging Selected Topics in Power Electronics, **1**(2), 104-113, 2013, doi: 10.1109/JESTPE.2013.2266372.

[6] B. Singh, K. Kant, and S. R. Arya, "Notch filter-based fundamental frequency component extraction to control distribution static compensator for mitigating current-related power quality problems," IET Power Electronics, **8**(9), 1758-1766, 2015., doi: 10.1049/iet-pel.2014.0486.

[7] B. N. Singh, B. Singh, A. Chandra, and K. Al-Haddad, "Design and digital implementation of active filter with power balance theory," IEE Proceedings-Electric Power Applications, **52**(5), 1149–1160, 2005, doi: 10.1049/ip-epa:20050097.

[8] D. Chen, "A Novel Cost-Effective Two-Level Inverter with Combined Use of Thyristors and IGBTs", Journal of Electrical Engineering and Technology, **13**(1), 152-159, 2018, doi:10.5370/JEET.2018.13.1.152

[9] K. Karaarslan, B Arifoglu, E Beser, S Camur, "Level Number Effect on Performance of a Novel Series Active Power Filter Based on Multilevel Inverter", Journal of Electrical Engineering and Technology, **13**(2), 711-721,2018, doi: 10.5370/JEET.2018.13.2.711.

[10] B. Singh , J. Solanki, "An Implementation of an Adaptive Control Algorithm for a Three-Phase Shunt Active Filter", IEEE Transactions on Industrial Electronics, **56**(8), 2811-2820, 2009, doi: 10.1109/TIE.2009.2014367.

[11] M. Qasim , Parag Kanjiya , Vinod Khadkikar, "Optimal Current Harmonic Extractor Based on Unified ADALINEs for Shunt Active Power Filters", IEEE Transactions on Power Electronics, **29**(12), 6383-6393, 2014, doi: 10.1109/TPEL.2014.2302539.

[12] M. Cirrincione, M. Pucci, G. Vitale, "A single-phase DG generation unit with shunt active power filter capability by adaptive neural Filtering," IEEE Trans. Ind. Electron., **55**(5), 2093- 2110, 2008, doi: 10.1109/TIE.2008.918642.

[13] R. R. Pereira , C. H. da Silva , L. E. Borges da Silva , G. Lambert-Torres, "Application of adaptive filters in active power filters", 2009 Brazilian Power Electronics Conference, Bonito-Mato Grosso do Sul, 770-774, 2009, doi: 10.1109/COBEP.2009.5347654.

[14] K. Sebasthirani , M. Manimegalai, "Recent Advances in Shunt Active Power Filter with Recursive Least Square Algorithm and its Applications", 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, 1288-1293, 2018, doi: 10.1109/ICOEI.2018.8553806.

[15] R.K. Agarwal, I. Hussain, B. Singh, "Dual-function PV-ECS integrated to 3P4W distribution grid using 3M-PLL control for active power transfer and power quality improvement", IET Renewable Power Generation, **12**(8), 920-927 ,2018, doi: 10.1049/iet-rpg.2016.0723.

[16] M. Badoni , A. Singh , B. Singh, "Comparative Performance of Wiener Filter and Adaptive Least Mean Square-Based Control for Power Quality Improvement", IEEE Transactions on Industrial Electronics, **63**(5), 3028-3027,2016, doi: 10.1109/TIE.2016.2515558.

[17] R. Rajagopal, K Palanisamy, S.Paramasivam, "Comparative Performance of Normalized LMS and Variable Step Size LMS based Control for Shunt Active Filtering," 2019 9th International Conference on Advances in Computing and Communication (ICACC), Kochi, India , 136-141, 2019, doi: 10.1109/ICACC48162.2019.8986184.

[18] R. Belaidi, A. Haddouche, M. Fathi, M. M. Larafi and G. M. Kaci, "Performance of grid-connected PV system based on SAPF for power quality improvement," 2016 International Renewable and Sustainable Energy Conference (IRSEC), Marrakech, 542-545, 2016, doi: 10.1109/IRSEC.2016.7984050.

[19] R. Rajagopal, K. Palanisamy and S. Paramasivam, "Shunt Active Filter Based on 7-Level Cascaded Multilevel Inverter for Harmonic and Reactive Power Compensation" Journal of Electrical Engineering & Technology, Springer, **14**(4), 1543-1551, 2019, doi: 10.1007/s42835-019-00201-1.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1172-1183 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Advanced Control Strategies on Nonlinear Testbench Dynamometer System for Simulating the Fuel Consumption

Marika Fanesi, David Scaradozzi*

Dipartimento di Ingegneria dell'Informazione (DII), Università Politecnica delle Marche, Ancona, 60121, Italy

ARTICLE INFO

Article history:
Received: 26 June, 2020
Accepted: 28 September, 2020
Online: 22 October, 2020

Keywords:
Engine-in-the-loop
Nonlinear dynamometer system
Testbench
Electric motor
Automotive

ABSTRACT

The adoption of Engine-in-the-loop technology shows real behaviour. This study presents a test runs simulation platform with real engine data. In addition, a test bench model is a demand approach that offers a significant potential to provide an excellent reproducibility of test runs. The platform includes the data integration to upgrade tests run and a comparison with previous results using the advancing control techniques designed. The dynamometer system presents significantly non-linearity. The adaptive control approach, integrated into the Model Predictive Control on the vehicle, allows increasing the tests run performance. The results show how the real data can improve performance and the validation of the system integrating the updated driving cycle and maintaining EiL approach. The conclusion showed the significant benefits regarding the control methods used.

Abbreviation

AMPC	Adaptive Model Predictive Controller
CO2	Carbon dioxide
DSM	dynamometer system motor
DUT	Device Under Test
EiL	Engine-in-the-Loop
EM	Electric Motors
EUDC	Extra-Urban Driving Cycle
FTP	Federal Test Procedure
ICE	Internal Combustion Engines
IM	Induction Motor
KF	Kalman Filter
LS	Least Square Algorithm
MHEV	Mild Hybrid Electric Vehicles
MIMO	Multiple Input Multiple Output
NEDC	New European Driving Cycle
PID	Proportional Integral Controller
UDC	Urban Driving Cycles
XiL	X-in-the-Loop
WLTP	World Light-duty vehicles Test Procedure

1. Introduction

Nowadays, in automotive industries, analysis and testing of Electric Motors (EMs) and Internal Combustion Engines (ICEs) are the key elements. Accuracy and efficiency are further improved and allows to evaluate the vehicles' production. However, to achieve these, the test benches are fundamentals. A novel data approach for advanced control strategies applied on a modelled test benches was introduced in a paper originally presented at the 23rd International Conference on System Theory, Control and Computing (ICSTCC) in 2019 [1], for which this work is an extension.

The tests are not limited to the EMs or ICEs in that the integration and combination of components are relevant. In the vehicles, brakes [2-4], energy management [5, 6], power devices [7, 8], sensors [9, 10], batteries [11-13], front steering [14, 15], are strongly analysed to increase vehicles' performances. Every component needs to be tested determining robustness, reliability, fault-tolerant [16-21] and, in case of traditional and Mild Hybrid Electric Vehicles (MHEVs), the engine has to respect the limits and the restriction regarding fuel consumption and, consequently, emissions [22-25]. To reduce costs, the virtualisation and the introduction of simulation's platform are included in the industry process [26, 27]. Considering the relevance of real data, in the last decade is introduced a new approach that includes the use of real components in the virtual tests. These techniques are called X-in-the-Loop (XiL) and include Engine-in-the-Loop, Hardware-in-the-Loop, Software-in-the-Loop, Model-in-the-Loop and Human-in-the-loop [28-32]. The XiL is used in different fields, as marine systems [33, 34], power units [36], home automation and robots [37], educational science [38].

*Corresponding Author: David Scaradozzi, Università Politecnica delle Marche, Email: d.scaradozzi@univpm.it

Engine-in-the-Loop (EiL) techniques allow testing of the electric motors and internal combustion engines. EiL is widely used to ensure precision on testbench. This type of simulations allows to reduce simulation timing and to simulate testing scenarios leading to higher efficiency. Another significant benefit is the reduction of costs that this technology enables. The strengths of this approach include the adaptability of the system with different hardware and the reliability of the measured data deriving from the real engine. The EMs and the ICEs tested, most of the time, are prototypes [39, 40] and this excludes the possibility to have a reliable model of the EM or the ICE to be included in the simulation. The Eil technique allows the integration of EMs or ICEs, and it is significant in that the modelling of these demonstrates a different behaviour from the real. The solution is identified in the EiL approach [41, 42]. Based on that, the simulation and control of the Induction Motor (IM) on the testbench become essential, and to efficiency, the parametric model of IM must be identified. The overall model is affected by accuracy and nonlinear behaviour [43-45]. The partial linearization of the motor helps to define the real-non-linear behaviour. A first step of calibration is needed to assure proper performance. The introduction of control techniques in these systems increase precision and real-behaviour and point to manage tests easily. To guarantee reliability and security, final design and implementation of the control strategies are required. Generally, this phenomenon is described and controlled utilising methods to achieve more accurate results. An adaptive control leads to higher efficiency and performance compared to conventional control [46-49].

With the paper, the authors describe a possible solution to design an EiL architecture and an improvement of the control quality in a real environment. This paper aims to compare the results obtained, adding real-driving conditions and upgrading the driving cycles presented. The work is organised as follow: in the next paragraph (Background and review) is shown the background related to two critical elements that influence the automotive industry, the market and the legislation concerning driving cycles and the Engine-in-the-loop state of the art; in the third paragraph (Nonlinear system) the models of the testbench and others components are presented including a section with the calibration test; then the adaptive control strategies are introduced and integrate to the system to increase the robustness and dynamics (Control strategy). The results are presented in the Simulation and discussion paragraph that shows an adaptive control used in the simulation of the overall system in a real environment applied to the new driving cycles.

2. Background and review

In this section, the authors present a review about the background, including the market reasons that guided the choice of development and implemented a specific configuration model. In this section is introduced also the driving cycles used in the simulation for the validation of the system and includes an Engine-in-the-loop methods review.

2.1. Automotive Market

In the automotive market, electric motors have been introduced recently. In the industries, the most critical challenge is the reduction of carbon dioxide (CO_2). The combination of engine and the electric motor in the vehicles guarantees excellent performance and the car's production has adapted to the growing market. These represent a radical innovation in the industries and allowed an expansion of the market. The key challenges are:

- Complexity and reduction of costs
- Different market
- Virtualisation process
- Industries landscape

Based on these four points, it is clear that in the future, the changes are fast and in different directions [50, 51].

Nowadays, vehicles include a lot of components, and the complexity is growing. Nevertheless, the market requests the reduction of costs and lower price. For this reason, a solution developed with the introduction of new technologies and features is to enter in the electrification process. Thus, a conventional vehicle with ICE become a hybrid vehicle [52]. At the same time, are developing new systems and safety, and performances are increased. This electrification process is essential to the point of becoming an evaluation factor. However, the main challenge remains the costs reduction. The governments help to achieve this target with new stricter limits and regulations regarding the emissions and pollutants, and these restrictions reflect in differentiation of the markets [53]. The differentiation of the markets has developed the adaptability of the automotive industries that need to supply vehicles based on sales. To be competitive, the approach uses the virtualisation process [54, 55]. This process combines the connectivity with accessibility and increases the use of digital resources. The virtualisation is a potential technique and represents an opportunity applies to the driving experience. Another added value in the car is the communication and interaction. Besides, the development of new powertrain technologies and innovative solutions for active safety and infotainment change and it is based on the industry landscape. Europe needs to be reconstructed and adapt the capacity of industries; East Asia is a competitive emerging market. Each challenge is dependent and interconnected from the others and to capture future growth strategic choice, appropriate investments and resources are essential [56-58].

2.2. Driving Cycles development

The driving cycles are evolved and upgraded based on government and legislation of every countries and are founded on the market's choices. The studied methodologies are not limited to the choice of driving cycles capable of making consumption real but include also to monitor energy consumption. An important aspect is the central role of driving cycles in emission measurement [59]. The pollutants, carbon dioxide, volatile organic compounds, nitrogen oxides and particulate are the results of the combustion of fuel and are regulated by countries directives. Emissions depend on different parameters and for this reason is classified based on vehicles types: cars, vans, buses, trucks and motorcycles.

In this paper, the authors illustrate driving cycles which had considered explicitly for the presented testbench. The driving cycles are described throughout some features shown in Table 1.

Table 1: Driving Cycles features

Feature	Detail
Distance	Total distance
Time	Total time Driving Time Acceleration Time Braking Time
Speed	Average Driving Speed Minimum Speed Maximum Speed
Acceleration	Average Driving Acceleration Minimum Acceleration Maximum Acceleration
Stops	Number of Stops Average Stop duration

The tests most used in the last ten years are:

- UDC or ECE-15: The Urban Driving Cycles represent a typical driving condition in Europe in a busy city with traffic and with a maximum speed of 50 km/h;
- EUDC: The Extra-Urban Driving Cycle is the high-speed road in the European cities with a maximum speed of 120 km/h;
- NEDC: The New European Driving Cycle is a combination of the UDC and EUDC, with four repetitions of UDC and once of EUDC;
- FTP-72 is the Federal Test Procedure used in US to simulate an urban road with multiple stops;
- FTP-75: Federal Test Procedure derived from FTP-72 is a city driving cycle with are a series of tests defined by the US Environmental Protection Agency (EPA),
- ARTEMIS Urban: is chassis dynamometer system procedure used in Europe with driving patterns derived from the analysis of a database containing real data and it is considered the urban track.
- ARTEMIS Rural: is chassis dynamometer system procedure as ARTEMIS Urban with the different that it is considered the rural road;
- ARTEMIS Motorway: as ARTEMIS Urban is chassis dynamometer system procedure and it is considered the motorway track.

These driving cycles are significant for new cars in that allow to test and know emission and pollutants. Lastly, as presented in the introduction, the validation tests assure that the results reflect real-world emission. This driving cycle is integrated into the system proposed in the following section to enhance the results on [1].

2.3. The engine in the loop

Engine-in-the-loop is a recent technique that makes possible to run a modelled testbench with an ICE reproducing the same conditions as when the engine is mounted in the real vehicle. Eil is a vehicle simulation linked and, in this technique, physical hardware called control unit, and an engine are integrated in a model of a testbench or a vehicle and driver model. The virtual testbench environment includes:

- High power
- Low inertia in the DSM

These perform a reasonable powertrain control and permit a detail emission evaluation.

The simulation of other components brings many benefits as repeatability and flexibility and can brush up the whole system. For this reason, the Eil approach results in an ideal choice for testbenches in that the accuracy, time-to-run and performance are highly improved [60, 61]. At the other hand, the request of this technique is the projecting of the entire system and for the high-speed real-time controller to have an adequate manage of the models implemented [62-66]. The fast response of the dynamometer loads the connected engine and the virtual vehicle follow the specific driving cycle target track as in the real world. Another influencer of the acceleration and speed is the vehicle pedal effect by the driver. In detail, the system includes an input-output behaviour deriving from the physical engine system and requests reasonable accuracy and response [67, 68]. The dynamics of the engine subsystem, along with the main subsystem testbench model and variables influence, is analysed to prevent fault. In automotive, the control strategy needs to be integrated into the control unit parameterising the whole virtual vehicle modelled and driveline configurations. The effects on the engine and performance are easily analysed as fuel consumption. In this use case, the relevant step is the calibration of the system in the virtualisation process and to reach a well-configurated testbench [69-71]. In automotive, the Eil is used during the development of new engine in that the testing of prototypes reduces the bugs and fastly improve the overall number. The development uses advanced simulation and modelling tools to realise a system with precise modelled functions that work together with physical hardware and processes. The virtualisation allows testing and validation of parameters and function guarantying highly quality [72-74]. However, several factors, as real-time computational complexity and calibration efforts, might lead to select faster and less-expensive techniques. The various combination and choice of vehicle calibration can limit in terms of management and testing functionalities, and the calibration steps must fit the engine behaviour. For these reasons and due to the architectural complexity of the Mild Hybrid Electric Vehicles, designing models for specific EM and ICE an adapt the entire system including interactions to the model is extremely difficult if even possible [75-77]. The major challenge is the satisfaction of the requirements during the Eil simulation.

3. Nonlinear system

The testbench detailed schema used in the present paper is shown in Figure 1. This collects some components: an inverter, the dynamometer motor (DSM) and the Device Under Test (DUT). On the EiL approach, the DUT could be an ICE (Internal Combustion Engine) or an EM (Electric Motor).

3.1. The Dynamometer System model

A general detailed model of the Dynamometer System is often too costly due to the induction machine typical non-linearity. Moreover, the parametrisation of the IM non-linearity is highly demanding for a standard automotive testbench commissioning process. Furthermore, the inverter includes a static controller with an unknown structure. The DSM is designed as a low-pass filter with fast dynamics with a simplified nonlinear mathematical model of DUT. The plant considered includes the inertia and the two parts of the testbench is treated as two mass oscillators. The model can be expressed as:

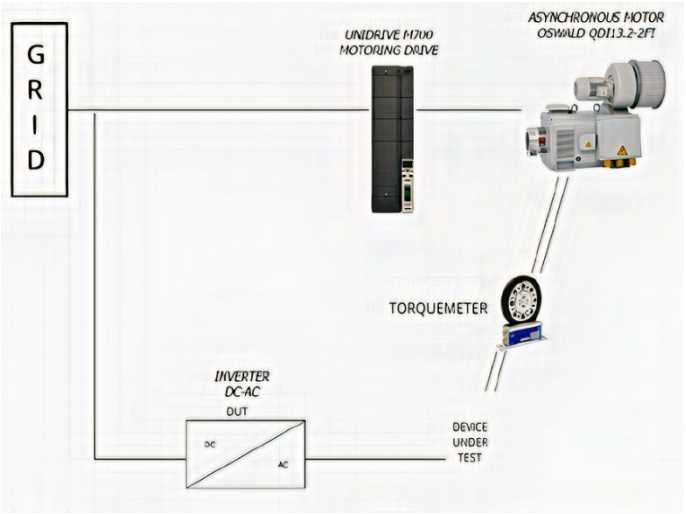


Figure 1: Dynamometer System

$$\Delta\dot{\varphi} = \omega_{EM} - \omega_D \tag{1}$$

$$\theta_{EM}\dot{\omega}_{EM} = T_{EM} - c\Delta\varphi - d(\omega_{EM} - \omega_D) \tag{2}$$

$$\theta_D\dot{\omega}_D = c\Delta\varphi + d(\omega_{EM} - \omega_D) - T_D \tag{3}$$

where:

$_{EM}$ is referred to DUT and can be consider as Engine or E-Motor,

$_D$ is applied to the dynamometer,

ω_x is the speed,

θ_x the inertia,

Δ_φ the torsion of the shaft,

T_x the estimated or measured torque,

c is the stiff constant

d is the damping.

For run the testbench model, the first tread is the calibration of the system as explained in the next paragraph.

3.2. Calibration Process

The calibration plays a crucial role in a testbench. The parameters setting is often a necessary step which influences the tests. The calibration process is a signal that allows assessing the exact initial configuration. The computation is real-time for the DSM. Parameters in the first step are set with an offline estimation. Then, the torque signal starts with a fixed amplitude and increasing frequency, and it is evaluated the response. The previous calibration phase is repeated increasing magnitude and testing the new value of amplitude for every frequency. The online calculation provides time constant for the system. This dynamic calibration has been achieved considering the transfer function between reference torque and IM air gap torque and shaft speed. In a MIMO (Multiple Input Multiple Output) systems, a controller with low overshoot is significant for improving dynamics compensation and consider the different steady-state. Considering that the reference for the inverter is a torque signal, the test is run in the torque mode. This calibration process is essential for estimate the parameters, evaluate the results obtained, select the model parameters and helps to interpret the well-fit model flexibility. To assure a good fit the evaluation of the model is essential and there are many concepts to guarantee the single-best fit [78]. However, the authors choose to focus this paper on the Engine-in the loop approach and on the advanced control strategies applied.

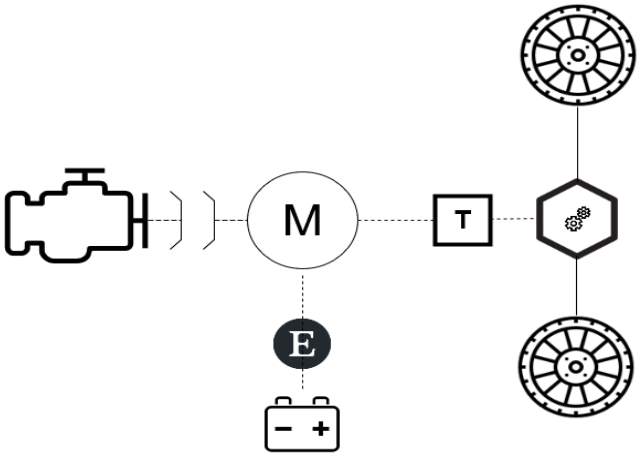


Figure 2: P2 Configuration: in order ICE, clutch, EM with Electronics and Battery pack, Transmission, Differential and Wheel.

3.3. P2 configuration

A testbench can simulate different types of ICE and EM; furthermore, the configuration of the vehicles can be changed, and this requires a different set of parameters and connected device. The structure considered is the parallel hybrid P2 of a Mild Hybrid Electric Vehicle. The P2 configuration is shown in figure 2. The EM (M) is placed in front of the transmission (T) and provide torque to compensate torque for dragging the ICE. The output torque derives from both ICE and EM. The vehicle can operate in pure electric mode, charge-battery mode, engine mode and parallel mode. Based on the way, the behaviours and the performances are

different. In the designed testbench it is possible to test every way during a driving cycle. The virtual model of the vehicle implemented with the integration of the Eil techniques is shown in the next section.

3.4. Vehicle Model

Based on the virtual vehicle developed presented in [1], lateral and vertical dynamics are included. The analysis of the static and dynamic behaviour is necessary to have a stable model of the vehicle. A multitude of parameters influences the stability and a representation of the main components is required to have a reliable system. For define the static behaviour, the steady-state cornering equations are introduced. The yaw velocity defines the steady steady-state $\dot{\Psi}$ and the sideslip angle β that are constant.

For the steady-state case, the torque balance influenced by the front and the rear wheel is:

$$F_{sf} \cdot I = -m \cdot a_y \cdot I_r \tag{4}$$

$$F_{sr} \cdot I = -m \cdot a_y \cdot I_f \tag{5}$$

with the tyre side forces:

$$F_{sf} = c_{sf} \cdot \alpha_f \tag{6}$$

$$F_{sr} = c_{sr} \cdot \alpha_r \tag{7}$$

where for small slip angles α results:

$$\alpha_f = \delta + \beta - \frac{I_f \cdot \dot{\Psi}}{v} \tag{8}$$

$$\alpha_r = \delta + \beta - \frac{I_r \cdot \dot{\Psi}}{v} \tag{9}$$

The equations become:

$$c_{sf} \cdot \left(\delta + \beta - \frac{I_f \cdot \dot{\Psi}}{v}\right) = \frac{I_r}{I} \cdot m \cdot a_y \tag{10}$$

$$c_{sr} \cdot \left(\beta + \frac{I_r \cdot \dot{\Psi}}{v}\right) = \frac{I_f}{I} \cdot m \cdot a_y \tag{11}$$

In the case of steady-state cornering, (10) and (11) are equal and the sideslip angle β are constant, so $\dot{\beta}$=0. The prediction of driving behaviour results:

$$\delta = \frac{I}{r} + \frac{m}{I} \cdot \left(\frac{I_r}{c_{sf}} - \frac{I_f}{c_{s_r}}\right) \cdot a_y \tag{12}$$

And based on (12), considering that $\frac{1}{r} = \frac{\dot{\psi}}{v}$ and $a_y = v \cdot \dot{\Psi}$, the stationary yaw amplification factor can be written as:

$$\frac{\dot{\psi}}{\delta} = \frac{v}{I + \frac{m}{I} \cdot \left(\frac{I_r}{c_{sf}} - \frac{I_f}{c_{sr}}\right) \cdot v^2} \tag{13}$$

$$\frac{\dot{\psi}}{\delta} = \frac{v}{I + \frac{d\delta}{da_y} \cdot v^2} \tag{14}$$

where:

F is the force
F_{sf} is side front
F_{sr} is side rear
I is the inertia
m is the mass of the vehicle
a_y is the acceleration
$\dot{\Psi}$ is the yaw velocity
β is the sideslip angle
v is the driving speed
c is the stiffness of tyre slip
α is the slip angle front or rear
δ is the wheel steering angle
r is the instantaneous curve radius.

The static behaviour shows the output variable $\dot{\Psi}$ divide for input variable δ that represent the stationary yaw amplification factor (13) that is related to the driving speed v. These equations are significant to define the vehicle's model and its behaviour. An essential prediction about the driving behaviour of a car while cornering results from these equations. The necessary steering angle input while navigating around a corner is composed of two parts: one part depends on geometric data, and the second part describes the influence of lateral acceleration. Driving at higher speeds in addition to low rates necessary input steering angle includes the steering angles at higher speeds which can increase or reduce the steering angle. This is very important for the interactions between the driver inputs and vehicle handling in the driver and vehicle control loop. While cornering, the driver must input a steering angle, which depends on the curve radius, and the present lateral acceleration. The reason that the steering angles depend on the lateral acceleration is that at the vehicle wheels while experiencing lateral force, a slip angle results, which usually differ for the front and rear wheel.

3.5. Electric Motor and Battery

In a Mild Hybrid Electric Vehicle, the electric motor can work as motor or generator based on the situation. The model of an EM is composed by an electronics unit setting on torque control mode, the EM and a battery pack. However, there are some constraints in terms of nominal value and maximum value of speed and torque. The operating range of motors in the electric vehicles is directly linked to the battery capacity. The battery has an enhanced capacity, compared with conventional cars, and this increases the operating range and decreases the emissions.

4. Control strategy

To face the non-linearities of the system, the feedback is supervised by an Adaptive Model Predictive Controller (AMPC). Its aim is performing robustness and reliability. The thought control pattern permits to improve the performance of the vehicle under study and the testbench behaviour concerning dead time compensation typical of such complex system. The controller presented strategy is necessary to guarantee disturbances rejection modified by the nonlinear behaviour of the vehicle. The schema in fig. 3 represents the action of the AMPC controller. The torque reference is the input of the dynamometer system placed as input of the inertia model. The control loop process handles the action

of driver and the vehicle reaction. Therefore, the steering behaviour in the single-track road is included in the response of the control vehicle as to the static behaviour.

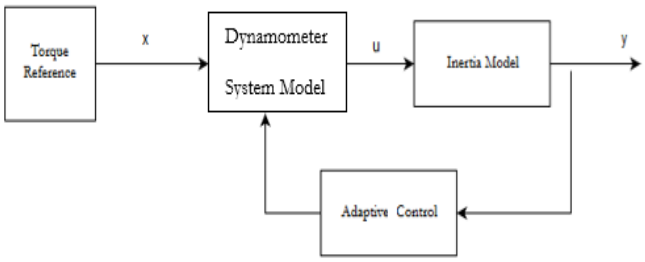


Figure 3: Control schema

The implemented control follows the same strategy explained in [1]. The Kalman filter inside is implemented as:

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (15)$$

where x_{k+1} and x_k are the discrete-time instant system state variables at k+1 and k respectively. u_k is the control vector and A,B are the matrices that link the state variables at time k to k+1 and w is the weight at time k.

The equations for the online state estimation on MATLAB/Simulink are as follow. The state prediction uses the matrices that represent an optional extension for the system, which are linearised in a non-stationary operating point.

State prediction:

$$X(.|k) = \Psi \cdot x(k) + Y \cdot u(k-1) + \theta \cdot \delta u(.|k) + Xi$$

Output prediction:

$$Y(.|k) = \gamma \cdot X(.|k) \quad (16)$$

Free response:

$$f(.|k) = \gamma \cdot \Psi \cdot x(k) + Y \cdot u(k-1) \quad (17)$$

Free control error:

$$e(.|k) = r(.|k) - f(.|k) \quad (18)$$

This implemented function generates the matrices used for the prediction. The adaptive controller executes in parallel an online Kalman Filter (KF) state estimator combined with a Least Square Algorithm (LS) for parameters estimation.

$$Min\ J(x) = \big(Z - h(x)\big)^T W \big(Z - h(x)\big) \quad (19)$$

Based on the nonlinear state estimation model. The objective function is the minimisation of the weighted square of error.

$$J(x) = \sum \frac{(z_i - h_i(x))^2}{\sigma_i^2} \quad (20)$$

The process includes:

- State update
- Update variance of the state error
- Gain matrix of the Kalman Filter
- Update estimation state, the variance disturb on state and loop.

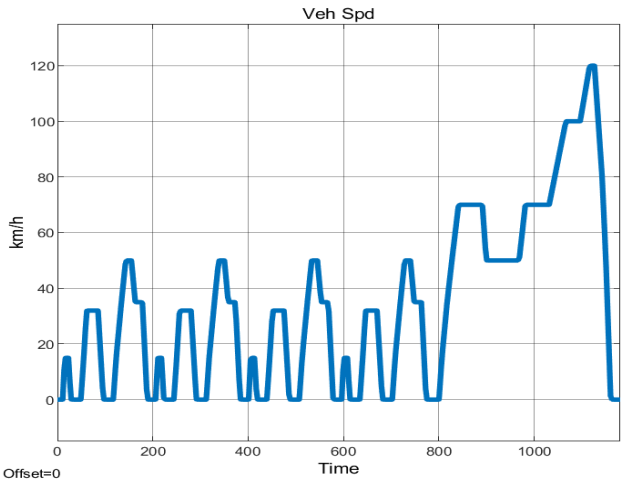
In detail, the evaluated estimation of the target is adjusted for each time, and the parameters are evaluated until the fulfilled qualities are gained. The KF is utilised for the dead-time compensation of DSM. In the KF and the A(k) and B(k) lattice are updated each progression in the calculation by the LS. The controller is additionally adjusted by the online estimation of the parameter by the LS. The evaluation of disturbance with the dead-time compensation is critical to address the distortion and the non-linearity in the general model. The undesirable deformation and significant suppression of oscillations is accomplished utilising this methodology.

5. Simulations and discussions

To verify the proposed method, the DSM model and the controller is developed using the MathWorks Matlab/Simulink software. The preliminary outcomes obtained simulating the vehicle on the testbench is shown in [1]. Based on these previous results, the simulations are enhanced, adding the updated driving cycles presented in section 2.

Table 2: Use Case table

Driving Cycle	Description			
	Time (s)	Distance (m)	Average Speed (km/h)	Fuel Consumption (gr CO_2/km)
NEDC	1180	11023	33.6	47,49
WLTP	1477	23262	44.5	50,95
FTP-72	1369	12070	31.5	52,81
FTP-75	2474	17770	34.12	49,14
ARTEMIS Urban	993	4870	17.6	102,45
ARTEMIS Rural	1082	17272	57.5	32,48
ARTEMIS Motorway	1068	28735	96.9	16,74



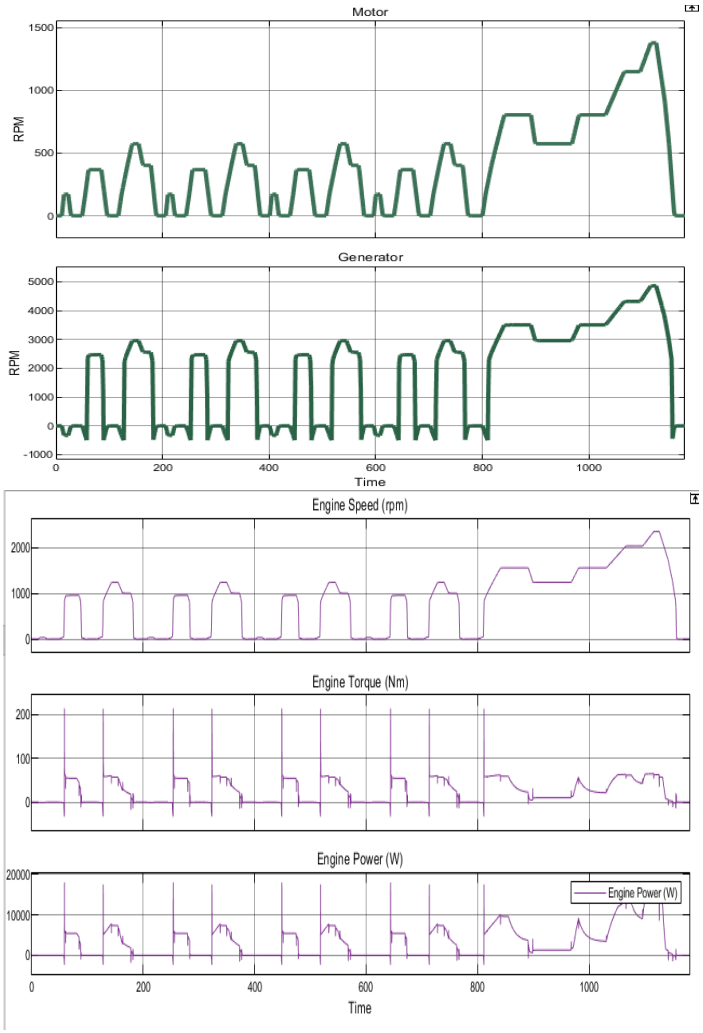
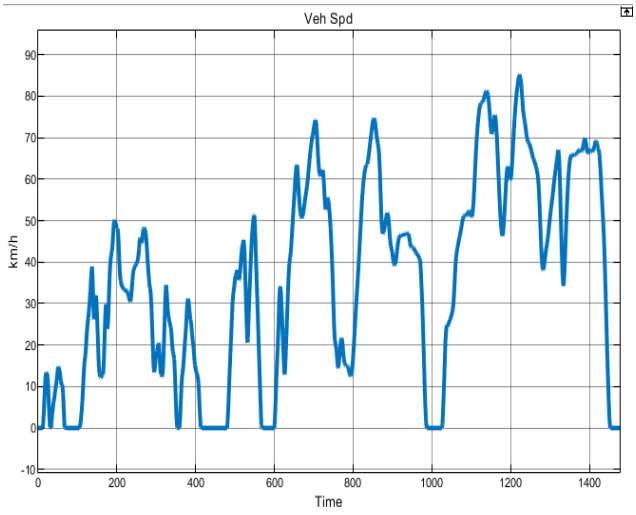


Figure 4: NEDC Cycle

Indeed, the AMPC performances exceed those of the conventional PID controller. Including the real-world disturbances on the vehicle, the PID controller no guarantees a fast, accurate response. For this reason, the authors test the considered driving cycles with AMPC controller by adding noise derived from the road.



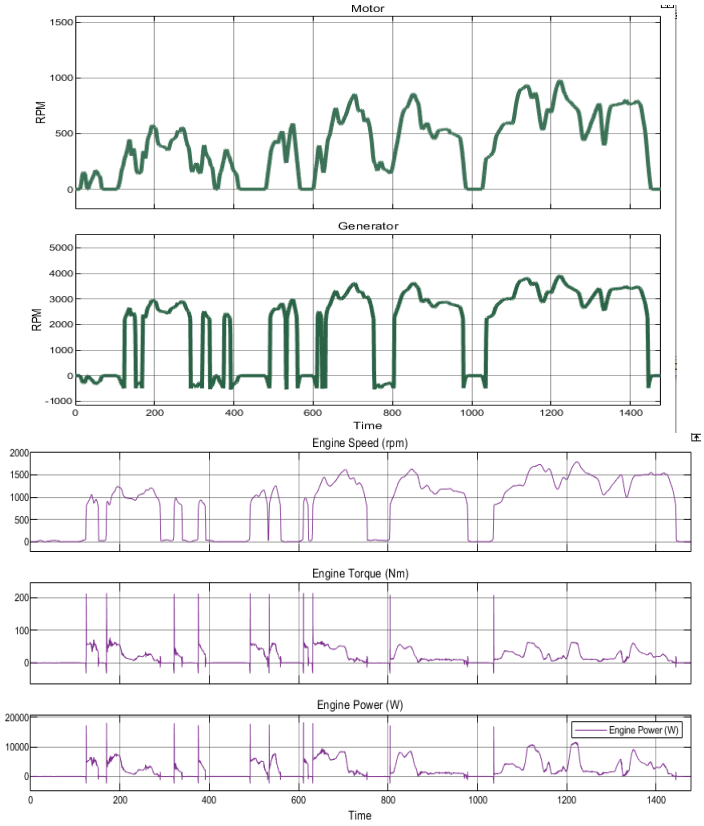


Figure 5: WLTP cycle

The torque request is based on the gas pedal value. The driving profiles are employed to examine the EiL test and to characterise the control system. Non-quantifiable amounts are assessed by filtering or with validated models. The AMPC controller results progressively strong, and the error is contained within 1.7 rpm to 2.8 rpm that is in the requirement of 3 rpm. This requirement derives from the possibility to measure not less than 3 rpm that results in the sensibility of the inverter. The error range is not changed, and this confirms and validates the control strategy. Every test procedure provides information about the behaviour of the control vehicle. Using Eil approach is important to maintain a defined speed minimum quantifiable with the physical process. The AMPC controls the reference quickly to the single setpoints while fulfilling the limitations. In the MHEVs this configuration is strongly used and the AMPC supply the ideal torque requested by the motor speed controller with more precision and control the values and references. The accuracy demonstrates through using AMPC improve the vehicle system and the reliable. The driving cycles are presented as a plot of the vehicle's speed: NEDC (Fig. 4), WLTP (Fig. 5), FTP-72 (Fig. 6), ARTEMIS Urban (Fig.7), Rural (Fig.8), Motorway (Fig.9). Furthermore, are shown the behaviour of electric motor modelled as motor or generator and the response of the engine, including the speed, the torque and the power. The description of every driving test cycle is in Table 2. The results are not only limited to the validation of the AMPC analysing the behaviour and the satisfaction of the requirements in that the emission for each test cycle is calculated. Table 2 shows the emission value for the driving cycles. The tests were chosen to cover a broad working range to investigate performance, and the controller results precise and accurate.

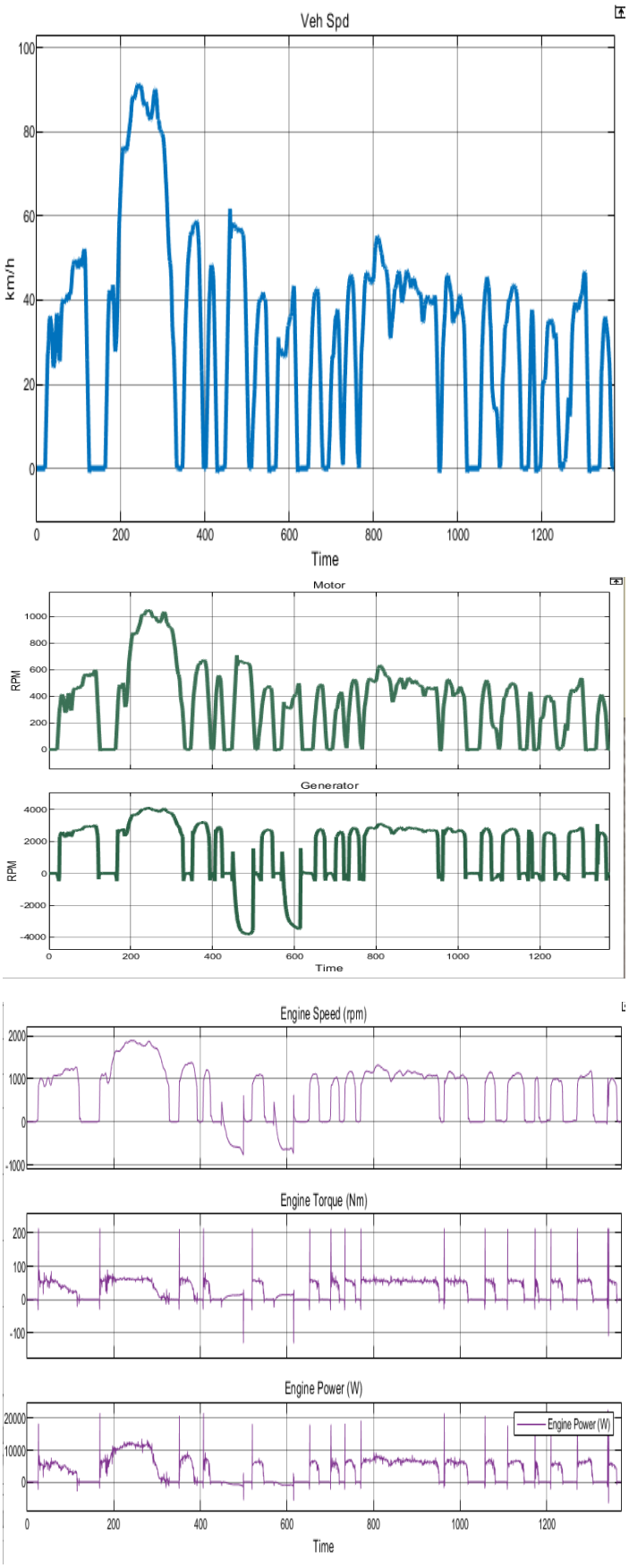


Figure 6: FTP-72 cycle

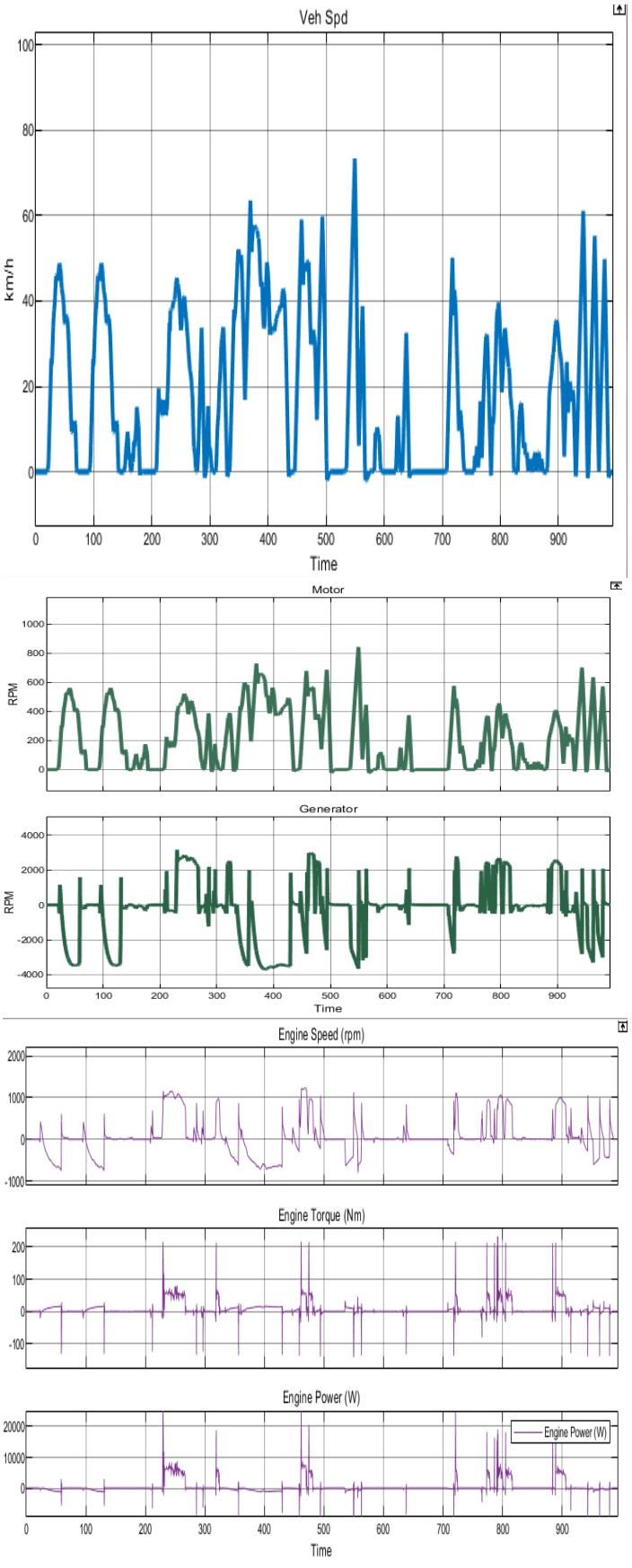


Figure 7: ARTEMIS Urban

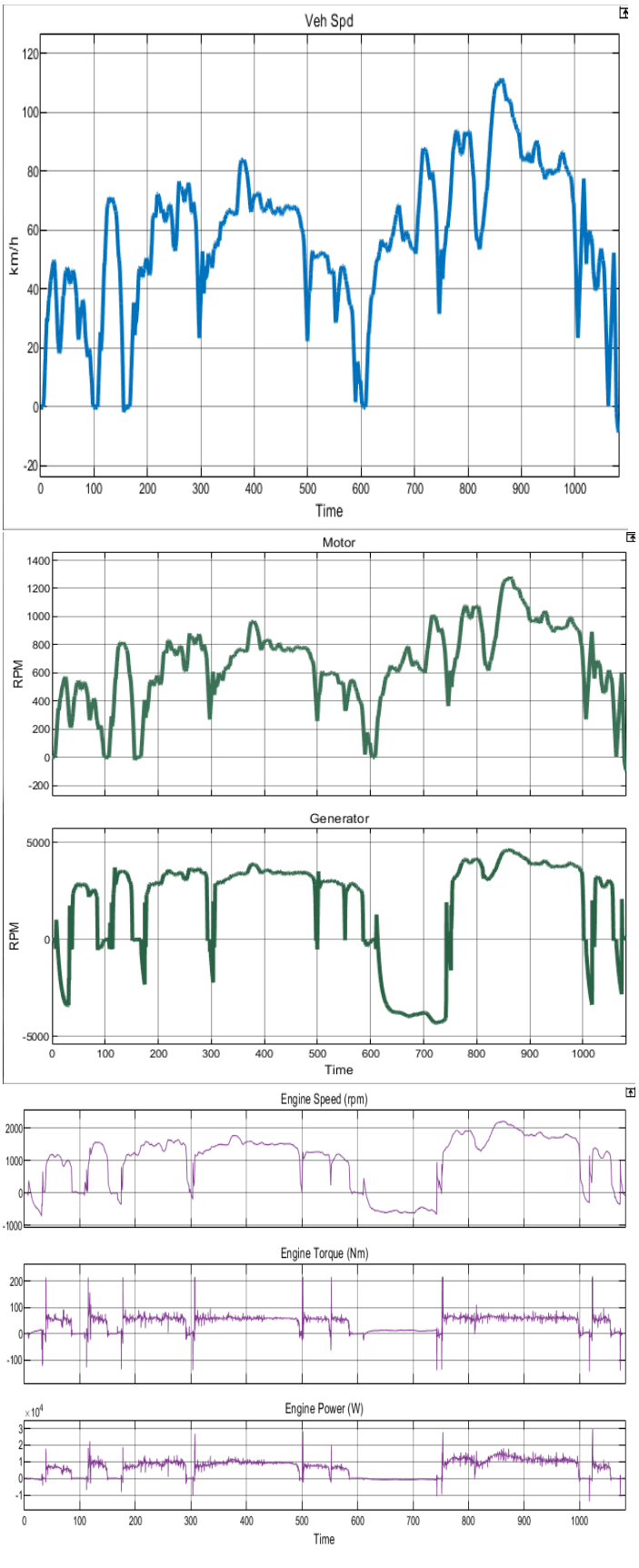


Figure 8: ARTEMIS Rural

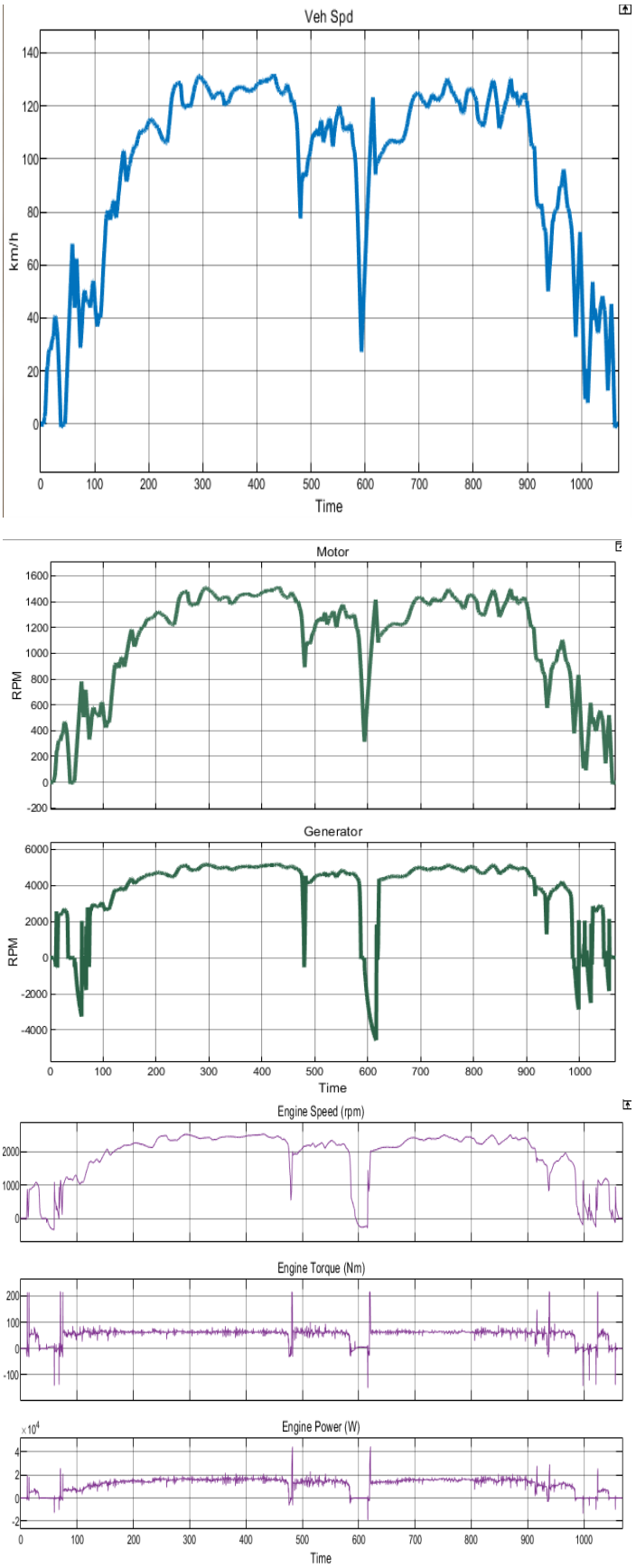


Figure 9: ARTEMIS Motorway

6. Conclusions and outlook

The results confirm the notable performances of the control strategy that allow characterising, with real feedback, the vehicles feature adding the disturbance to the road, a conventional control does not guarantee an accurate and fast response. Nevertheless, the AMPC disturbance rejection is strong, and the outcomes on the real system show excellent efficiency. This output is enhanced by measuring fuel consumption, and the results strongly validate the potential of this control strategy. Furthermore, the extension of driving cycles of the different country shows the R&R of the system. Overall, the research findings show that the nonlinear control strategies demonstrate outstanding disturbance rejection qualities. The flexibility of the testbench model can be investigated to find the single-best fit as the methods for assessing it. Moreover, these results can be extended to different vehicle's configuration as P3 and P4 and full electric vehicles in order to improve systems efficiency.

References

[1] M. Fanesi, D. Scaradozzi, "Adaptive control for non-linear test bench dynamometer systems," in 2019 23rd International Conference on System Theory, Control and Computing, ICSTCC 2019 - Proceedings, Institute of Electrical and Electronics Engineers Inc.: 768–773, 2019, doi:10.1109/ICSTCC.2019.8885558.

[2] J. Ko, S. Ko, H. Son, B. Yoo, J. Cheon, H. Kim, "Development of Brake System and Regenerative Braking Cooperative Control Algorithm for Automatic-Transmission-Based Hybrid Electric Vehicles," IEEE Transactions on Vehicular Technology, **64**(2), 431–440, 2015.

[3] F. Naseri, E. Farjah, T. Ghanbari, "An Efficient Regenerative Braking System Based on Battery/Supercapacitor for Electric, Hybrid, and Plug-In Hybrid Electric Vehicles With BLDC Motor," IEEE Transactions on Vehicular Technology, **66**(5), 3724–3738, 2017.

[4] S. Khastgir, "The simulation of a novel regenerative braking strategy on front axle for an unaltered mechanical braking system of a conventional vehicle converted into a hybrid vehicle," in 2013 Eighth International Conference and Exhibition on Ecological Vehicles and Renewable Energies (EVER), 1–6, 2013.

[5] H. Wang, Y. Huang, A. Khajepour, "Cyber-Physical Control for Energy Management of Off-Road Vehicles With Hybrid Energy Storage Systems," IEEE/ASME Transactions on Mechatronics, **23**(6), 2609–2618, 2018.

[6] H. Fathabadi, "Plug-In Hybrid Electric Vehicles: Replacing Internal Combustion Engine With Clean and Renewable Energy Based Auxiliary Power Sources," IEEE Transactions on Power Electronics, **33**(11), 9611–9618, 2018, doi:10.1109/TPEL.2018.2797250.

[7] C. Sun, S.J. Moura, X. Hu, J.K. Hedrick, F. Sun, "Dynamic Traffic Feedback Data Enabled Energy Management in Plug-in Hybrid Electric Vehicles," IEEE Transactions on Control Systems Technology, **23**(3), 1075–1086, 2015, doi:10.1109/TCST.2014.2361294.

[8] Z. Amjadi, S.S. Williamson, "Power-Electronics-Based Solutions for Plug-in Hybrid Electric Vehicle Energy Storage and Management Systems," IEEE Transactions on Industrial Electronics, **57**(2), 608–616, 2010, doi:10.1109/TIE.2009.2032195.

[9] X. Li, Q. Xu, "A Reliable Fusion Positioning Strategy for Land Vehicles in GPS-Denied Environments Based on Low-Cost Sensors," IEEE Transactions on Industrial Electronics, **64**(4), 3205–3215, 2017, doi:10.1109/TIE.2016.2637306.

[10] S. Saponara, P. Tisserand, P. Chassard, D. Ton, "Design and Measurement of Integrated Converters for Belt-Driven Starter-Generator in 48 V Micro/Mild Hybrid Vehicles," IEEE Transactions on Industry Applications, **53**(4), 3936–3949, 2017, doi:10.1109/TIA.2017.2687406.

[11] R. Romagnoli, L.D. Couto, M.M. Nicotra, M. Kinnaert, E. Garone, "Computationally-efficient constrained control of the state-of-charge of a Li-ion battery cell," in 2017 IEEE 56th Annual Conference on Decision and Control (CDC), 1433–1439, 2017, doi:10.1109/CDC.2017.8263855.

[12] J. Shen, A. Khaligh, "A Supervisory Energy Management Control Strategy in a Battery/Ultracapacitor Hybrid Energy Storage System," IEEE Transactions on Transportation Electrification, **1**(3), 223–231, 2015, doi:10.1109/TTE.2015.2464690.

[13] E. Kamal, L. Adouane, "Hierarchical Energy Optimization Strategy and Its Integrated Reliable Battery Fault Management for Hybrid Hydraulic-Electric Vehicle," IEEE Transactions on Vehicular Technology, **67**(5), 3740–3754, 2018, doi:10.1109/TVT.2018.2805353.

[14] S. Yim, J. Choi, K. Yi, "Coordinated Control of Hybrid 4WD Vehicles for Enhanced Maneuverability and Lateral Stability," IEEE Transactions on Vehicular Technology, **61**(4), 1946–1950, 2012, doi:10.1109/TVT.2012.2188921.

[15] S. Di Cairano, H.E. Tseng, D. Bernardini, A. Bemporad, "Vehicle Yaw Stability Control by Coordinated Active Front Steering and Differential Braking in the Tire Sideslip Angles Domain," IEEE Transactions on Control Systems Technology, **21**(4), 1236–1248, 2013, doi:10.1109/TCST.2012.2198886.

[16] A.H. Hajimiragha, C.A. Canizares, M.W. Fowler, S. Moazeni, A. Elkamel, "A Robust Optimization Approach for Planning the Transition to Plug-in Hybrid Electric Vehicles," IEEE Transactions on Power Systems, **26**(4), 2264–2274, 2011, doi:10.1109/TPWRS.2011.2108322.

[17] C. Quinn, D. Zimmerle, T.H. Bradley, "An Evaluation of State-of-Charge Limitations and Actuation Signal Energy Content on Plug-in Hybrid Electric Vehicle, Vehicle-to-Grid Reliability, and Economics," IEEE Transactions on Smart Grid, 3(1), 483–491, 2012, doi:10.1109/TSG.2011.2168429.

[18] H. Yu, Z. Liu, "Fault Analysis and Fault-Tolerant Control of Electric Motor Drive System in HEV," in 2012 Fifth International Conference on Intelligent Computation Technology and Automation, 177–180, 2012, doi:10.1109/ICICTA.2012.51.

[19] L. Zhang, Y. Fan, C. Li, C. Liu, "Design and Analysis of a New Six-Phase Fault-Tolerant Hybrid-Excitation Motor for Electric Vehicles," IEEE Transactions on Magnetics, **51**(11), 1–4, 2015, doi:10.1109/TMAG.2015.2447276.

[20] A. Berthon, F. Gustin, M. Bendjedia, J.M. Morelle, G. Coquery, "Inverter components reliability tests for hybrid electrical vehicles," in 2009 IEEE 6th International Power Electronics and Motion Control Conference, 763–768, 2009, doi:10.1109/IPEMC.2009.5157487.

[21] A. Drosu, G. Suciu, A. Scheianu, I. Petre, "An Analysis of Hybrid/Electric Vehicle Monitoring Systems and Parameters," in 2019 Electric Vehicles International Conference (EV), 1–5, 2019, doi:10.1109/EV.2019.8892923.

[22] Y. Cheng, C. Lai, J. Teh, "Application of Particle Swarm Optimization to Design Control Strategy Parameters of Parallel Hybrid Electric Vehicle with Fuel Economy and Low Emission," in 2018 International Symposium on Computer, Consumer and Control (IS3C), 342–345, 2018, doi:10.1109/IS3C.2018.00093.

[23] S.A. Zulkifli, S. Mohd, N. Saad, A.R.A. Aziz, "Influence of motor size and efficiency on acceleration, fuel economy and emissions of split-parallel hybrid electric vehicle," in 2013 IEEE Symposium on Industrial Electronics Applications, 126–131, 2013, doi:10.1109/ISIEA.2013.6738981.

[24] S.A. Rahman, Nong Zhang, Jianguo Zhu, "A comparison on fuel economy and emissions for conventional hybrid electric vehicles and the UTS plug-in hybrid electric vehicle," in 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), 20–25, 2010, doi:10.1109/ICCAE.2010.5451533.

[25] A.Y. Saber, G.K. Venayagamoorthy, "Plug-in Vehicles and Renewable Energy Sources for Cost and Emission Reductions," IEEE Transactions on Industrial Electronics, **58**(4), 1229–1238, 2011, doi:10.1109/TIE.2010.2047828.

[26] C.M. Martinez, X. Hu, D. Cao, E. Velenis, B. Gao, M. Wellers, "Energy Management in Plug-in Hybrid Electric Vehicles: Recent Progress and a Connected Vehicles Perspective," IEEE Transactions on Vehicular Technology, **66**(6), 4534–4549, 2017, doi:10.1109/TVT.2016.2582721.

[27] V. Delafosse, S. Stanton, T. Sekisue, Junsik Yun, "A methodology to use simulation at every stage of a hybrid vehicle design," in 2012 IEEE Vehicle Power and Propulsion Conference, 1134–1138, 2012, doi:10.1109/VPPC.2012.6422618.

[28] H. Zhang, Y. Zhang, C. Yin, "Hardware-in-the-Loop Simulation of Robust Mode Transition Control for a Series–Parallel Hybrid Electric Vehicle," IEEE Transactions on Vehicular Technology, **65**(3), 1059–1069, 2016, doi:10.1109/TVT.2015.2486558.

[29] Zheng Li, Zetao Ma, Shumei Cui, "Design and research on engine emulation system of HEV power train hardware-in-the-loop simulation platform," in 2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC Asia-Pacific), 1–6, 2014, doi:10.1109/ITEC-AP.2014.6941031.

[30] N. Shidore, A. Ickes, T. Wallner, A. Rousseau, M. Ehsani, "Evaluation of ethanol blends for PHEVs using engine-in-the-loop," in 2011 IEEE Vehicle Power and Propulsion Conference, 1–8, 2011, doi:10.1109/VPPC.2011.6043041.

[31] Wang Lei, Yang Yalian, Peng Zhiyuan, Yang Guo, Hu Xiaosong, "Research on hybrid electrical vehicle based on human-in-the-loop simulation," in 2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC Asia-Pacific), 1–5, 2014, doi:10.1109/ITEC-AP.2014.6940738.

[32] Wu Kai, Zhang Tong, Zhu Maotao, Zhou Jianhao, "Hardware-in-the-loop simulation for multi-energy management controller of Hybrid Electric Vehicle in a virtual car environment," in 2011 International Conference on Electric Information and Control Engineering, 5061–5064, 2011, doi:10.1109/ICEICE.2011.5776933.

[33] J. Zhou, G. Ouyang, M. Wang, "Hardware-in-the-Loop Testing of Electronically-Controlled Common-Rail Systems for Marine Diesel Engine," in 2010 International Conference on Intelligent Computation Technology and Automation, 421–424, 2010, doi:10.1109/ICICTA.2010.40.

[34] Y. Liu, M. Steurer, P. Ribeiro, "A novel approach to power quality assessment: real time hardware-in-the-loop test bed," IEEE Transactions on Power Delivery, **20**(2), 1200–1201, 2005, doi:10.1109/TPWRD.2005.844251.

[35] J. Wang, H. Lee, J. Wang, C. Lin, "Robust Environmental Sound Recognition for Home Automation," IEEE Transactions on Automation Science and Engineering, **5**(1), 25–31, 2008, doi:10.1109/TASE.2007.911680.

[36] A.M. Pavan, S. Castellan, G. Sulligoi, "An innovative photovoltaic field simulator for hardware-in-the-loop test of power conditioning units," in 2009 International Conference on Clean Electrical Power, 41–45, 2009, doi:10.1109/ICCEP.2009.5212086.

[37] J. Domaszewicz, S. Lalis, A. Pruszkowski, M. Koutsoubelias, T. Tajmajer, N. Grigoropoulos, M. Nati, A. Gluhak, "Soft Actuation: Smart Home and Office with Human-in-the-Loop," IEEE Pervasive Computing, **15**(1), 48–56, 2016, doi:10.1109/MPRV.2016.5.

[38] M. Karakose, "Hardware, software, and human in the loop education supported with social network analysis and mining," in 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET), 1–5, 2016, doi:10.1109/ITHET.2016.7760721.

[39] M. Hafner, T. Finken, M. Felden, K. Hameyer, "Automated Virtual Prototyping of Permanent Magnet Synchronous Machines for HEVs," IEEE Transactions on Magnetics, **47**(5), 1018–1021, 2011, doi:10.1109/TMAG.2010.2091675.

[40] X. Zeng, S. Zheng, D. Song, "Modeling and dynamic simulation of a virtual prototype for applying automobile differential into hybrid electric vehicle as power-slit device," in 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), V3-82-V3-87, 2010, doi:10.1109/ICCASM.2010.5620147.

[41] R.S. Vadamalu, C. Beidl, "Explicit MPC PHEV energy management using Markov chain based predictor: Development and validation at Engine-In-The-Loop testbed," in 2016 European Control Conference (ECC), 453–458, 2016, doi:10.1109/ECC.2016.7810326.

[42] Y. Kim, A. Salvi, A.G. Stefanopoulou, T. Ersal, "Reducing Soot Emissions in a Diesel Series Hybrid Electric Vehicle Using a Power Rate Constraint Map," IEEE Transactions on Vehicular Technology, **64**(1), 2–12, 2015, doi:10.1109/TVT.2014.2321346.

[43] Liyong Yang, Xiaolin Peng, Zhengxi Li, "Induction motor electrical parameters identification using RLS estimation," in 2010 International Conference on Mechanic Automation and Control Engineering, 3294–3297, 2010, doi:10.1109/MACE.2010.5535653.

[44] J. Xiaochun, Y. Geng, W. yunfei, "A Parameter Identification Method for General Inverter-fed Induction Motor Drive," in 2006 CES/IEEE 5th International Power Electronics and Motion Control Conference, 1–5, 2006, doi:10.1109/IPEMC.2006.4778331.

[45] B. Zhu, K. Rajashekara, H. Kubo, "Comparison between current-based and flux/torque-based model predictive control methods for open-end winding induction motor drives," IET Electric Power Applications, **11**(8), 1397–1406, 2017, doi:10.1049/iet-epa.2016.0517.

[46] M. Pakmehr, T. Yucelen, "Adaptive control of uncertain systems with gain scheduled reference models and constrained control inputs," in 2014 American Control Conference, 691–696, 2014, doi:10.1109/ACC.2014.6859326.

[47] D. Bourlis, J.A.M. Bleijs, "Gain scheduled controller with wind speed estimation via Kalman filtering for a stall regulated variable speed wind turbine," in 2009 44th International Universities Power Engineering Conference (UPEC), 1–5, 2009.

[48] T. Ertuğrul, M.A. Adli, M.U. Salamci, "Model reference adaptive control design for helicopters using gain scheduled reference models," in 2016 17th International Carpathian Control Conference (ICCC), 182–187, 2016, doi:10.1109/CarpathianCC.2016.7501090.

[49] E. Gruenbacher, L. del Re, "Robust inverse control for combustion engine test benches," in 2008 American Control Conference, 2852–2857, 2008, doi:10.1109/ACC.2008.4586926.

[50] C. Yang, M. Zha, W. Wang, K. Liu, C. Xiang, "Efficient energy management strategy for hybrid electric vehicles/plug-in hybrid electric vehicles: review and recent advances under intelligent transportation system," IET Intelligent Transport Systems, **14**(7), 702–711, 2020, doi:10.1049/iet-its.2019.0606.

[51] J.E. Siegel, D.C. Erb, S.E. Sarma, "A Survey of the Connected Vehicle Landscape—Architectures, Enabling Technologies, Applications, and Development Areas," IEEE Transactions on Intelligent Transportation Systems, **19**(8), 2391–2406, 2018, doi:10.1109/TITS.2017.2749459.

[52] M. Traub, A. Maier, K.L. Barbehön, "Future Automotive Architecture and the Impact of IT Trends," IEEE Software, **34**(3), 27–32, 2017, doi:10.1109/MS.2017.69.

[53] T. Sivakumaran, F. Köhne, M. Toth, "Identification of critical success factors for emerging market entry planning processes in the automotive industry," in 2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 1694–1698, 2015, doi:10.1109/IEEM.2015.7385936.

[54] L. Hanwu et al., "Regularity of Current Dispersal in Different Kinds of Grounding Electrode," in 2018 IEEE International Conference on High Voltage Engineering and Application (ICHVE), 1–4, 2018. doi: 10.1109/ICHVE.2018.8642240

[55] A. Balluchi, L. Benvenuti, M.D. di Benedetto, C. Pinello, A.L. Sangiovanni-Vincentelli, "Automotive engine control and hybrid systems: challenges and opportunities," Proceedings of the IEEE, **88**(7), 888–912, 2000, doi:10.1109/5.871300.

[56] H. A. Raza et al., "Analysis the effect of 500kv High-Voltage Power Transmission Line on the Output Efficiency of Solar-Panels," in 2019 International Conference on Electrical, Communication, and Computer Engineering, 1–6, 2019. doi: 10.1109/ICECCE47252.2019.8940803

[57] A.R. Mahayadin, I. Ibrahim, I. Zunaidi, A.B. Shahriman, M.K. Faizi, M. Sahari, M.S.M. Hashim, M.A.M. Saad, M.S. Sarip, Z.M. Razlan, M.F.H. Rani, Z.M. Isa, N.S. Kamarrudin, A. Harun, Y. Nagaya, "Development of Driving Cycle Construction Methodology in Malaysia's Urban Road System," in 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), 1–5, 2018, doi:10.1109/ICASSDA.2018.8477619.

[58] A. Charadsuksawat, Y. Laoonual, N. Chollacoop, "Comparative Study of Hybrid Electric Vehicle and Conventional Vehicle Under New European Driving Cycle and Bangkok Driving Cycle," in 2018 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific), 1–6, 2018, doi:10.1109/ITEC-AP.2018.8432599.

[59] A.R. Salisa, N. Zhang, J.G. Zhu, "A Comparative Analysis of Fuel Economy and Emissions Between a Conventional HEV and the UTS PHEV," IEEE Transactions on Vehicular Technology, **60**(1), 44–54, 2011, doi:10.1109/TVT.2010.2091156.

[60] J. Liu, L. Zhang, Q. Chen, S. Quan, R. Long, "Hardware-in-the-loop test bench for vehicle ACC system," in 2017 Chinese Automation Congress (CAC), 1006–1011, 2017, doi:10.1109/CAC.2017.8242913.

[61] M.H. Salah, T.H. Mitchell, J.R. Wagner, D.M. Dawson, "A Smart Multiple-Loop Automotive Cooling System—Model, Control, and Experimental Study," IEEE/ASME Transactions on Mechatronics, **15**(1), 117–124, 2010, doi:10.1109/TMECH.2009.2019723.

[62] J. Zhao, J. Wang, "Adaptive Observer for Joint Estimation of Oxygen Fractions and Blend Level in Biodiesel Fueled Engines," IEEE Transactions on Control Systems Technology, **23**(1), 80–90, 2015, doi:10.1109/TCST.2014.2313003.

[63] A. Pratt, M. Ruth, D. Krishnamurthy, B. Sparn, M. Lunacek, W. Jones, S. Mittal, H. Wu, J. Marks, "Hardware-in-the-loop simulation of a distribution system with air conditioners under model predictive control," in 2017 IEEE Power Energy Society General Meeting, 1–5, 2017, doi:10.1109/PESGM.2017.8273757.

[64] D. Michalek, C. Gehsat, R. Trapp, T. Bertram, "Hardware-in-the-loop-simulation of a vehicle climate controller with a combined HVAC and passenger compartment model," in Proceedings, 2005 IEEE/ASME International Conference on Advanced Intelligent Mechatronics., 1065–1070, 2005, doi:10.1109/AIM.2005.1511151.

[65] S. Jiang, M. Smith, J. Kitchen, A. Ogawa, "Development of an Engine-in-the-loop Vehicle Simulation System in Engine Dynamometer Test Cell," in SAE Technical Papers, 2009, doi:10.4271/2009-01-1039.

[66] T. Jung, M. Kötter, J. Schaub, C. Quérel, S. Thewes, H. Hadj-amor, M. Picard, S.-Y. Lee, Engine-in-the-Loop: A Method for Efficient Calibration and Virtual Testing of Advanced Diesel Powertrains: Antriebsentwicklung im digitalen Zeitalter 20. MTZ-Fachtagung, 209–224, 2019,

doi:10.1007/978-3-658-25294-6_12.
[67] F.C. Nemtanu, I.M. Costea, D. Buretea, L.G. Obreja, "Hardware in the loop simulation platform for intelligent transport systems," in 2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME), 247–250, 2017, doi:10.1109/SIITME.2017.8259901.
[68] R. Isermann, J. Schaffnit, S. Sinsel, "Hardware-in-the-Loop Simulation for the Design and Testing of Engine-Control Systems," IFAC Proceedings Volumes, **31**(4), 1–10, 1998, doi:https://doi.org/10.1016/S1474-6670(17)42125-2.
[69] M. Nasri, M. Kargahi, M. Mohaqeqi, "Scheduling of Accuracy-Constrained Real-Time Systems in Dynamic Environments," IEEE Embedded Systems Letters, **4**(3), 61–64, 2012, doi:10.1109/LES.2012.2195294.
[70] B.J. Bunker, M.A. Franchek, B.E. Thomason, "Robust multivariable control of an engine-dynamometer system," IEEE Transactions on Control Systems Technology, **5**(2), 189–199, 1997, doi:10.1109/87.556024.
[71] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. DOI: 10.1109/ICECCE47252.2019.8940756
[72] R. Jayaraman, A. Joshi, V. To, G. Kaid, "Fidelity Enhancement of Power-Split Hybrid Vehicle HIL (Hardware-in-the-Loop) Simulation by Integration with High Voltage Traction Battery Subsystem," 2018, doi:10.4271/2018-01-0008.
[73] H. Holzmann, K. Hahn, J. Webb, O. Mies, "Simulation-Based ESC Homologation for Passenger Cars," ATZ Worldwide, **114**, 40–43, 2012, doi:10.1007/s38311-012-0218-5.
[74] U. Baake, K. Wüst, M. Maurer, A. Lutz, "Testing and simulation-based validation of ESP systems for vans," ATZ Worldwide, **116**, 30–35, 2014, doi:10.1007/s38311-014-0021-6.
[75] G. Siva Sankar, R. Shekhar, C. Manzie, T. Sano, H. Nakada, "Fast Calibration of a Robust Model Predictive Controller for Diesel Engine Airpath," 2018.
[76] A. White, G. Zhu, J. Choi, "Hardware-in-the-Loop Simulation of Robust Gain-Scheduling Control of Port-Fuel-Injection Processes," IEEE Transactions on Control Systems Technology, **19**(6), 1433–1443, 2011, doi:10.1109/TCST.2010.2095420.
[77] H. Schuette, M. Ploeger, "Hardware-in-the-Loop Testing of Engine Control Units - A Technical Survey," SAE Transactions, **116**, 86–107, 2007.
[78] V.D. Veksler, C.W. Myers, K. Gluck, "Model flexibility analysis.," Psychological Review, 122 4, 755–769, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1178-1189 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Defeating Anti-Debugging Techniques for Malware Analysis Using a Debugger

Jong-Wouk Kim[1], Jiwon Bang[2], Mi-Jung Choi[1,2,*]

[1]*Department of Computer Science, Kanwon National University, 24341, Republic of Korea*

[2]*Interdisciplinary Graduate Program in Medical Bigdata Convergence, Kangwon National University, 24341, Republic of Korea*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 25 November, 2020
Online: 14 December, 2020

Keywords:
Anti-Debugging
Malware Analysis
Malware

ABSTRACT

Cyberattacks such as spear phishing and malspam pretending to be companies, institutes, and government officials are increasing and evolving. Malware has a variety of purposes, such as collecting personal information and illegal access to the system. New types of malware are increasing every day, and many malware programs spread all over the Internet, causing severe problems. To analyze such malware effectively, analysts first need to understand the inner structure of the malware. They can try to analyze malware manually and automatically. However, attackers who create malware use many different kinds of techniques, such as anti-reverse engineering, to hinder and delay analysis. They also extend malware life through a combination of different techniques, such as social engineering and anti-debugging. These techniques make the malware more sophisticated; thus, it is hard for an analyst to detect the malware. Anti-debugging, one way to protect malware, is a deadly poison to malware analysts because it makes the analysis more difficult by detecting a debugger or debugging environments. Therefore, this paper describes malware's anti-debugging techniques and how to defeat them through anti-anti-debugging mechanisms. It applies its findings to analyze a sample program, packed files, and actual malware with anti-debugging modules and performs various experiments to verify the proposed techniques. After the experiments, it confirms whether its countermeasure is useful for malware analysis.

1. Introduction

In the past, the development of malware was mainly used for malware programmers to show off. However, malicious software(*malware*) such as virus and keylogger in recent years illegally accesses server systems in various ways to steal personal and financial information and computing power. Fatal damage, such as the theft of essential data of institutes, corporations, public institutions, and personal user information, occurs. The malware damage can take many forms, from stealing personal information such as a user's financial information to destroying industrial facilities such as Stuxnet [1]. According to Symantect's "Internet Security Threat Report 2019", financial Trojans accounted for 16% of the many malware programs spread in 2018. In addition, Trend Micro's "Unseen Threats, Imminent Losses" said that a vulnerability was found in the Supervisory Control And Data Acquisition (SCADA) system, one of the industrial systems' components. Also, attackers can penetrate the system through malware that attacks this vulnerability and illegally controls the SCADA system.

New types of malware appear every day and spread over the Internet, causing serious problems frequently. The malware also uses anti-reverse engineering techniques such as Self-Modifying code, code compression, packing, and anti-debugging methods to hide internal code structures and interfere with the analysis. This paper is an extension of work initially presented in the *International Conference on ICT Convergence* [2] to solve these kinds of problems. It is challenging to analyze malware with these anti-reverse engineering techniques accurately and requires considerable time. Security experts reinforce security by using anti-virus software and various equipment types to minimize malware damage, while analysts scrutinize malware using various techniques, it is more likely that anti-virus software or an automated analysis tool fails to detect it. The workflow of malware analysis is as in Figure 1. Usually, when an analyst analyzes malware, an analyst gets information from automated and static tools [3]. For example, if malware has anti-reverse engineering analysis. After static analysis, the analyst analyzes malware behavior through dynamic analysis and investigates its behavior in more detail through advanced analysis. However, anti-debugging techniques are hinder all analysis steps especially advanced

*Corresponding Author: Mi-Jung Choi, 305 Hanbit-Gwan, 1 Gangwondaehakgil, Chuncheon-Si, Gangwon-Do, 24341 Repblic of Korea, mjchoi@kangwon.ac.kr

analysis. Most of the tools in the advanced analysis are insufficient to deal with anti-debugging techniques. The relevant APIs and structures such as PEB and HEAP could be modified to defeat anti-debugging techniques as soon as advanced analysis begins. This step makes it more convenient for analysts to analyze malware, and it could be analyzed more quickly because the malware interference is gone.

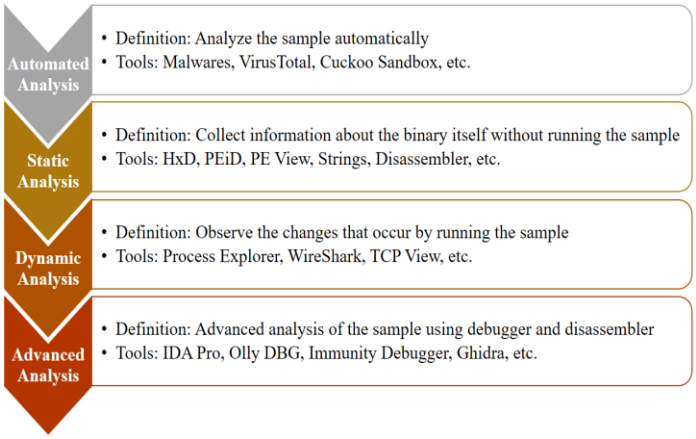


Figure 1: Workflow in safe and accurate analysis of malware

When analyzing malware in virtual environments such as QEMU, VMware, and VirtualBox, the virtual environments can be constructed similarly to the physical environments. The virtual environment analysis has advantages that are vastly similar to the actual execution, and does not have to worry about infecting a physical computer while performing malware analysis. However, there is a disadvantage in that analysis may be disrupted through the anti-VM (Virtual Machine) techniques. Anti-VM techniques disable malware analysis using CPU emulators or virtualization tools. The Dynamic Binary Instrumentation (DBI) tool can bypass anti-debugging by inserting or replacing code chunks into the operation command [4]. However, DBI has a disadvantage because it is not convenient, and it cannot execute huge or complex programs correctly.

The static analysis uses tools such as a decompiler and disassembler to analyze the binary or code of malware to infer the type of malware. However, this approach becomes more difficult and complicated by evasion techniques such as code obfuscation and packer [5]. A dynamic analysis examines the behavior of the target file. However, dynamic analysis tools cannot provide both accuracy and convenience when analyzing complex malware. The debugger, widely used to analyze malware dynamically, can examine changes in the machine's state while executing each command. However, the disadvantage of using a debugger is that there are anti-debugging techniques that can detect and interfere with the debugging process. Thus, if analysts can defeat anti-debugging techniques that interfere with the debugger, it becomes one of the most powerful methods for analyzing malware. Therefore, this paper introduces anti-debugging techniques that attack a debugger's vulnerabilities to interfere with analysis, and explains how to bypass analysis. In addition, it proposes anti-anti-debugging techniques that can defeat anti-debugging techniques, and verify the proposed method applying them to a sample program, packed files, and actual malware with anti-debugging modules. This paper aims to give direction for researchers who analyze malware or those who dream of being analysts to develop their methods or scripts that can defeat anti-debugging techniques when using a debugger.

This paper is organized as follows. Section 2 introduces research related to anti-debugging. Section 3 proposes anti-debugging techniques based on artifacts and time and describes correspondent anti-anti-debugging mechanisms. Section 4 presents the experiments performed and verifies anti-anti-debugging techniques with a sample program, packed files, and actual malware. Section 5 concludes the study.

2. Related Work

Analysts have continuously reviewed malware detection and binary code analysis. On the other hand, anti-reverse engineering research does not draw much attention from researchers except for some category topics [6], [7]. There are various topics in anti-reverse engineering such as anti-debugging, anti-VM, and code encryption. To understand anti-analysis, first, malware analysis must be briefly described. There are two different types of malware analyses, namely static analysis and dynamic analysis. Static analysis is a method of analyzing codes or binaries without executing them. Analysts can observe the file's operation code with a decompiler or disassembler and analyze the program flow and internal structure. Dynamic analysis probes its behavior by executing them using a sandbox, debugger, or emulator. Debugging refers to the process of detecting and eliminating errors or bugs when a process operates unexpectedly or crashes. A debugger is a tool or software that helps with this, therefore, it is a useful tool for understanding malware's inner workings. It likewise helps to observe changes in registry, memory, and stack closely while executing instructions in the target file (*debuggee*) line by line. Also, analysts use debuggers because they can manipulate elements such as code, environment, and memory. The sandbox case runs the file to record and analyze the process and network changes over a while [8].

However, malware developers use anti-analysis techniques against each analysis. Therefore, anti-debugging techniques delay malware detection and increase its life span by bypassing the anti-virus software. Anti-analysis is a series of techniques in which malware remains concealed, prevents detection, interferes with analysis, and makes analyzing it challenging. It is analyzed 6,222 samples of malware and studied how malicious behaviors differed in an environment with virtualization and a debugger [9]. As results, they found about 40 percent malware in an environment with a debugger performing less malicious behavior. Also, only two percent of the samples in a virtual environment showed malicious behavior when executed. This study shows malware with anti-VM and anti-debugging can bypass dynamic malware analysis tools such as a debugger, sandbox, and emulator.

Examples of anti-analysis techniques that interfere with static analysis include packer [10]-[12], code obfuscation [5], [13], [14], and opaque constants [15]. Packer is a technique that compresses the executable file and creates a new program entirely different from the original program. However, the new packed file has the same or similar behavior with various techniques such as Portable Executable (PE) relocation and anti-analysis. In the case of code obfuscation and opaque constants, these are the process of encrypting text, binary data, or code to make it difficult to understand. For these reasons, the malware detection method

based on signature became ineffective. Thus, dynamic analysis emerged to solve these kinds of problems. Debugging, which analyzes malware, carefully examines details such as behaviors, registers, and stacks while executing malware instructions by line. However, dynamic analysis using the debugger suffers from anti-analysis, such as anti-debugging, anti-VM, code compression, Self-Modifying, and blocking input devices. Among the anti-analysis, anti-debugging is an attempt by the debuggee to find or detect evidence of an analysis environment or debugger's presence. It is provided an overview of the analysis disruption techniques used by malware such as anti-debugging, anti-VM, and obfuscation [16]. Shang Gao and Qian Lin gave an overview of Windows's debugging mechanism, categorized according to the debugging methods used by the debugger, and explained the debugging functions and anti-debugging through exceptions [17]. It is classified general malware, targeted malware to attack specific systems and investigate how many anti-debugging techniques applied to it [18]. Most malware had anti-debugging techniques, however, it is found that the general malware had more anti-debugging techniques than targeted malware.

To defeat anti-analysis techniques, analysts can defeat most anti-debugging techniques using many CPU emulator and virtualization tools. However, these tools are not suitable for analyzing malware because they can analyze only a specific process or program. Dynamic Binary Instrumentation (DBI) framework can insert, modify, and delete code at the arbitrary address during the program's running [4], [19], [20], [21]. Therefore, the DBI can bypass the anti-debugging techniques by code insertion. However, analysis is inconvenient, and some DBIs cannot execute complex programs correctly [22]. Recently It is proposed the Apate framework to analyze malware [23]. Apate is a framework that hides the debugger from anti-debugging. The spectrum of powerful anti-debugging techniques is analyzed and classified into 79 attack vectors and six categories.

Apate is designed to defeat anti-debugging techniques by: (1) performing just-in-timed disassembling based on single-stepping and; (2) monitoring the execution flow of the debuggee and modifying the state of the debuggee to hide WinDbg. It is defined anti-debugging rule sets and proposed replacing the matching part of the binary with other instructions. The method proposed and modified the part that calls the anti-debugging functions in the debugger to other instructions [24], [25]. It is proposed a new framework for classifying, detecting and bypassing reverse engineering prevention techniques used by malware and protecting systems [9]. It is proposed UBER, a new system to neutralize anti-sandbox [26]. Sandbox has many artifacts of analysis environment. The anti-sandbox looks for various system artifacts expected to exist in the system for identifying the sandbox environment. UBER does not replicate artifacts or folders in use or directly simulate user behavior. Instead, it generalizes the user's computer usage pattern with an abstract behavioral profile, adopts the profile to simulate user and artifact creation operations, and then replicates the system with the created artifacts into a sandbox. Through this, UBER can defeat the anti-sandbox (i.e., usage artifacts analysis), while malware in the Windows or macOS malware is increasing. Since the tools for analyzing macOS based on malware are very limited. It is proposed Mac-A-Mal, a framework for analyzing Mac based malware [27]. It is developed a kernel extension to monitor malware behavior and bypass several evasion prevention techniques used in the wild, which uncovered 74 unknown malware programs.

3. Defeating anti-debugging techniques

This section first finds out how anti-debugging techniques work in Windows. Afterward, it explains how to defeat them [28].

3.1. Process Environment Block (PEB) Structure

The Process Environment Block (PEB) structure is a data structure in the Windows NT operating system. Each process has this structure and it contains each process information. Malware can detect the presence of debugger without using a specific function such as `IsDebuggerPresent`(). If implemented manually, there is no way to detect this in a static analysis or initial analysis. The anti-debugging based on PEB structure protects programs not only by malware but also many by many packers such as `PECompact`, `ASPack`, and `ASProtect` [29].

It confirms the `BeingDebugged`, `NtGlobalFlag`, and HEAP structure to perform anti-debugging among many PEB structure members. `BeingDebugged` is set to 0x0 if not debugged, and 0x1 if debugged. `NtGlobalFlag` has a value of 0x0 if not debugged, and is set to a value of 0x70 if debugged. Table 1 shows the flags of `NtGlobalFlag` while debugging. Not only the PEB structure, but HEAP structure can also detect the presence of a debugger.

Table 1: Flag configurations of `NtGlobalFlag` member in the PEB structure

Flag	Value
FLG_HEAP_ENABLE_TAIL_CHECK	0x10
FLG_HEAP_ENABLE_FREE_CHECK	0x20
FLG_HEAP_VALIDATE_PARAMETER	0x40

Table 2: Flag configurations of `Flags` member in the HEAP structure

Flag	Value
HEAP_GROWABLE	0x2
HEAP_TAIL_CHECKING_ENABLED	0x20
HEAP_FREE_CHECKING_ENABLED	0x40
HEAP_SKIP_VALIDATION_CHECKS (only in Windows XP or under)	0x10000000
HEAP_VALIDATE_PARAMETERS_ENABLED	0x40000000

Table 3: Flag configurations of `ForceFlags` member in the HEAP structure

Flag	Value
HEAP_TAIL_CHECKING_ENABLED	0x20
HEAP_FREE_CHECKING_ENABLED	0x40
HEAP_VALIDATE_PARAMETERS	0x40000000

Anti-debugging can detect debuggers using the `Flags` and `ForceFlags` members of the HEAP structure. `Flags` has a value of 0x2 if not debugged. It has a value of 0x50000062 in Windows XP or 0x40000062 in Windows 7 or later versions, if debugged. `ForceFlags` has a value of 0x0 if not debugged, and 0x40000060 if debugged. Table 2 shows the configuration flags of `Flags` and Table 3 shows the configuration flags of `ForceFlags`. To bypass anti-debugging based on PEB structure, analysts can defeat anti-debugging by modifying each member to an appropriate value, such as 0x0 in `BeingDebugged`.

3.2. IsDebuggerPresent()

IsDebuggerPresent() is the function in which software can easily detect the presence of a debugger. Malware and general software use this function to protect itself from reverse engineering. This function is also one of the anti-debugging techniques that most packers such as UPX and PECompact can apply to the original program. This function returns the BeingDebugged member of the PEB structure. To defeat this function, one must modify the code to return 0x0 or change the value of BeingDebugged member of the PEB structure. This paper uses PUSH and POP instructions to store 0x0 in the EAX register, shown in Figure 2. The original code of IsDebuggerPresent() is the left side of Figure 2, and the modification code of IsDebuggerPresent() is the right side of Figure 2.

```
MOV EAX,DWORD PTR FS:[18]          MOV EAX,DWORD PTR FS:[18]
MOV EAX,DWORD PTR DS:[EAX+30]      PUSH 0
MOVZX EAX,BYTE PTR DS:[EAX+2]      POP EAX
RETN                               RETN
```

Figure 2: The original code (left) and modification code (right) of IsDebuggerPresent()

```
MOV EDI,EDI                                  MOV EDI,EDI
PUSH EBP                                     PUSH EBP
MOV EBP,ESP                                  MOV EBP,ESP
CMP DWORD PTR SS:[EBP+8],0                   MOV EAX,DWORD PTR SS:[EBP+C]
PUSH ESI                                     PUSH 0
JE SHORT kernel32.77623FC1                   POP DWORD PTR DS:[EAX]
                                             XOR EAX,EAX
                                             POP EBP
                                             RETN 8
CALL DWORD PTR DS:[<&ntdll.NtQueryInformationProcess>]
TEST EAX,EAX
JL kernel32.77629176
XOR EAX,EAX
CMP DWORD PTR SS:[EBP+8],EAX
SETNE AL
MOV DWORD PTR DS:[ESI],EAX
XOR EAX,EAX
INC EAX
POP ESI
POP EBP
RETN 8
```

Figure 3: The original code (left) and modification code (right) of CheckRemoteDebuggerPresent()

3.3. CheckRemoteDebuggerPresent()

CheckRemoteDebuggerPrenset() function takes the Process Identifier (PID) of a particular process to detect whether it is debugged or not. It calls the NtQueryInformationProcess() function from inside to perform anti-debugging not only on malware but also on general software, and some packers use it to prevent debugging. The code on the left of Figure 3 is the original code of CheckRemoteDebuggerPrenset() and as shown on the left side of Figure 3, it calls NtQueryInformationProcess() from inside. There are two ways to defeat this function. Firstly, defeat the function called NtQueryInformationProcess() internally. Secondly, store 0x0 in the EAX register to make it always return 0x0.

This paper stores 0x0 in the stack using the PUSH instruction, as shown on the right side of Figure 3. It then stores 0x0 in the EAX register with POP instruction and terminates the function using RETN. There are three essential things to defeat this function. First, three existing instructions from the top of this code must be maintained. The second is to store 0x0 in the EAX register and return it. Last is to save the return address stored in the stack to the EBP register with POP instruction, then end with RETN. Otherwise, the flow of the debuggee gets tangled, and the analysis cannot proceed further.

3.4. ZwQueryInformationProcess()

The developer uses ZwQueryInformationProcess() functions for various reasons, such as finding a process path not only for anti-debugging. However, this section, only describes anti-debugging based on ZwQueryInformationProcess(). Figure 4 is a parameter for ZwQueryInformationProcess(). Among these parameters, ProcessInformationClass is the type of process information to be retrieved. It performs by set a specific value to the ProcessInformationClass parameter. If ProcessInformationClass is set to 0x7 which means ProcessDebugPort, it can determine whether it is debugged or not. In case of debugging, it returns 0xFFFFFFFF (-1). If not, it returns 0x0. Not only ProcessDebugPort, this function can executes anti-debugging bypassing 0x1E, which means ProcessDebugObjectHandle or 0x1F, which means ProcessDebugFlags to the parameter ProcessInformationClass. In the case of ProcessDebugObjectHandle, this function returns 0x000000B4 if debugged, and it returns 0x0 if not debugged. In the case of ProcessDebugFlags, this function returns 0x0 if debugged, and it returns 0x1 if not debugged. The code on the left of Figure 5 is the original code of ZwQueryInformationProcess().

```
NTSTATUS WINAPI ZwQueryInformationProcess(
    _In_      HANDLE              GetProcessHandle,
    _In_      PROCESSINFOCLASS    ProcessInformationClass,
    _Out_     PVOID               ProcessInformation,
    _In_      ULONG               ProcessInformationLength,
    _Out_opt_ PULONG              ReturnLength
);
```

Figure 4: Parameter of ZwQueryInformationProcess()

```
MOV EAX,0EA                        PUSH 100000
MOV EDX,7FFE0300                   RETN
CALL DWORD PTR DS:[EDX]
RETN 14                            MOV EAX,0EA
                                   MOV EDX,7FFE0300
                                   CMP DWORD PTR SS:[ESP+8],7
                                   JE SHORT 00100037
                                   CMP DWORD PTR SS:[ESP+8],1E
                                   JE SHORT 0010004A
                                   CMP DWORD PTR SS:[ESP+8],1F
                                   JE SHORT 0010005D
                                   PUSH 779A6052
                                   RETN
                                   MOV EAX,DWORD PTR SS:[ESP+C]
                                   PUSH 0
                                   POP DWORD PTR DS:[EAX]
                                   XOR EAX,EAX
                                   RETN 14
                                   MOV EAX,DWORD PTR SS:[ESP+C]
                                   PUSH 0
                                   POP DWORD PTR DS:[EAX]
                                   XOR EAX,EAX
                                   RETN 14
                                   MOV EAX,DWORD PTR SS:[ESP+C]
                                   PUSH 1
                                   POP DWORD PTR DS:[EAX]
                                   RETN 14
```

Figure 5: The original code (left) and modification code (right) of ZwQueryInformationProcess()

In order to bypass ZwQueryInformationProcess(), three things are crucial: (1) the first two instructions from the top of the original code must be executed; (2) it must return an appropriate

value; (3) it should not interfere with the flow of the debuggee. It needs a new virtual memory that must be allocated to handle the number of cases and consider non-anti-debugging cases. Also, `ZwQueryInformationProcess()` enters the kernel area through `KiFastSystemCall` and `SYSENTER` instructions. Since the debugger cannot enter the kernel, one can bypass these powerful instructions by writing new instructions in allocated virtual memory, as shown on the right of Figure 5. It determines whether the received `ProcessInformationClass` parameter is 0x7 (`ProcessDebugPort`), 0x1E (`ProcessDebugObjectHandle`), or 0x1F (`ProcessDebugFlags`). In each case, it jumps to the address with the specific instructions. Otherwise, it returns to the original instruction with `RETN`. The reason for returning to the original function is not to disturb the flow of the debuggee.

```
HWND FindWindowA(
  LPCSTR lpClassName,
  LPCSTR lpWindowName
);
```

```
HWND FindWindowW(
  LPCSTR lpClassName,
  LPCSTR lpWindowName
);
```

Figure 6: Parameters of `FindWindowA()` and `FindWindowW()`

```
MOV EDI,EDI
PUSH EBP
MOV EBP,ESP
XOR EAX,EAX
PUSH EAX
PUSH DWORD PTR SS:[EBP+C]
PUSH DWORD PTR SS:[EBP+8]
PUSH EAX
PUSH EAX
CALL USER32.76F8B865
POP EBP
RETN 8
```

```
MOV EDI,EDI
PUSH EBP
MOV EBP,ESP
XOR EAX,EAX
PUSH EAX
PUSH DWORD PTR SS:[EBP+C]
PUSH DWORD PTR SS:[EBP+8]
PUSH EAX
PUSH EAX
PUSH 220000
RETN
RETN 8

CALL USER32.76F8B865
CMP EAX,7018A
JE SHORT 000B000F
JMP SHORT 000B0011
XOR EAX,EAX
POP EBP
RETN 8
```

Figure 7: The original code (left) and modification code (right) of `FindWindowW()`

3.5. *FindWindow()*

The family of `FindWindow` functions can find the name of the specific window. It consists of `FindWindowW()`, `FindWindowA()`, `FindWindowExW()`, and `FindWindowExA()`. These are anti-debugging techniques that confirm the debugger's window name to see if the debuggers are running on the operating system. Figure 6 shows the parameters of `FindWindowA()` and `FindWindowW()`. The `lpClassName` gets a specific class name, and the `lpWindowName` gets a specific window name when calling these functions. If it finds the name of a specific window, it returns the handle of the window. If not, it returns 0x0. If malware calls these functions, it passes a particular debugger's window name to the `lpWindowName` parameter for anti-debugging. However, the analysts can defeat it always to returns 0x0.

The code on the left of Figure 7 is the original code of `FindWindowW()`, and the code on the right of Figure 7 is the code to defeat it. This study allocates a virtual memory and writes new instructions to disable anti-debugging techniques. After moving to the allocated address through the `PUSH` and `RETN` instructions, one must execute the existing `CALL` instruction. Next, it needs to get the debugger's window handle and compare it with the `EAX` register using `CMP` instruction. If it is the same, it jumps by `JE` instruction and stores 0x0 in the `EAX` register. Otherwise, it jumps by `JMP` instruction and restores the return address to the `EBP` register. The analysts can defeat this anti-debugging technique by writing new instructions in virtual memory. In Figure 7, the address of virtual memory is 0x00220000. However, this address must point to empty virtual memory.

3.6. *GetCurrentProcessId(), BlockInput()*

The commercial packer such as `Yoda's Protector` uses the two functions described in this section. The malware packed with `Yoda's Protector`, it calls `CreateToolHelp32Snapshot()` to obtain the PIDs and process all running processes' information. Next, it uses the `GetCurrentProcessId()` to get its process information and then compares the two results. If the parent process is a debugger, it interrupts analysis by terminating the process. Therefore, to defeat this anti-debugging technique, `GetCurrentProcessId()` always returns the PID of a debugger or analysis tool by modifying the PID of the debugger and debuggee to be the same. Analysts can defeat this technique. The code on the left side of Figure 8 is the original code of `GetCurrentProcessId()`, and the code is modified the same as on the right side of Figure 8 to defeat this function. To match the PID of the debugger and debuggee, it stores the PID of the debugger (0x0EA0 in Figure 8) in the `EAX` register and returns it. At this point, `NOP` instruction fills the remaining space of memory.

```
MOV EAX,DWORD PTR FS:[18]
MOV EAX,DWORD PTR DS:[EAX+20]
RETN
```

```
MOV EAX,0EA0
NOP
NOP
NOP
RETN
```

Figure 8: The original code (left) and modification code (right) of the GetCurrentProcessId()

```
MOV EAX,1141
MOV EDX,7FFE0300
CALL DWORD PTR DS:[EDX]
RETN 4
```

```
NOP
NOP
⋮
NOP
NOP
RETN 4
```

Figure 9: The original code (left) and modification code (right) of `BlockInput()`

The `BlockInput()` blocks the input events such as keyboard and mouse. When software calls this function, the computer blocks all events from input devices. By blocking all input devices' events, signals from devices do not affect the input queue's synchronous key state and asynchronous key state. If the software calls this function while running, the computer blocks the input devices' events, and there is no solution until the software calls again. The malware exploits this. If malware calls this function while debugging, the computer blocks all events from input devices, and

analysis stops. The only solution is to reboot the computer and analyze the malware again from the beginning. This function has a BOOL value as a parameter. If the parameter is TRUE, the computer blocks the events. If FALSE, the computer unblocks the events. This function returns to non-zero if it ends with no errors. The corresponding code is the left code of Figure 9. The way to defeat this technique is more straightforward than others. As shown on right side of Figure 9, this is defeated by modification with NOP instruction, which does nothing.

```
MOV ECX,DWORD PTR DS:[7FFE0324]
MOV EDX,DWORD PTR DS:[7FFE0320]
MOV EAX,DWORD PTR DS:[7FFE0328]
CMP ECX,EAX
JNZ SHORT KERNELBA.75D36770
MOV EAX,DWORD PTR DS:[7FFE0004]
MUL EDX
SHL ECX,8
IMUL ECX,DWORD PTR DS:[7FFE0004]
SHRD EAX,EDX,18
SHR EDX,18
ADD EAX,ECX
RETN
```

```
MOV EDX,7FFE0000
PUSH 5678ABCD
POP EAX
RETN
```

Figure 10: The original code (left) and modification code (right) of the GetTickCount()

```
CMP DWORD PTR DS:[73F78FD4],0
JNZ <JMP.&KERNEL32.GetTickCount>
CALL WINMM.73F526C0
SUB EAX,DWORD PTR DS:[73F78FD8]
PUSH 0
SBB EDX,DWORD PTR DS:[73F78FDC]
PUSH 2710
PUSH EDX
PUSH EAX
CALL <JMP.&ntdll._alldiv>
ADD EAX,DWORD PTR DS:[73F78FE0]
RETN
```

```
MOV EAX,0
RETN
```

Figure 11: The original code (left) and modification code (right) of timeGetTime ()

3.7. GetTickCount(), timeGetTime()

The anti-debugging techniques described so far are anti-debugging based on artifacts. Those are techniques in which malware detect analysis signs, analysis environment, or debugger. The anti-debugging based on artifacts is the most common and useful technique for malware to evade analysis. If a debugger runs the program, there are differences from normal execution. The anti-debugging based on artifacts confirms the debugger through these differences.

However, the anti-debugging described in this section is slightly different from the previous ones. The anti-debugging described in this section uses time elements. The anti-debugging based on time measures the time between a particular routine. It compares the execution in the real system and the execution analysis environment and indirectly detects the debugger. It measures the time stamp counter cycle before and after starting a particular routine in a program to confirm a debugger's presence because the analysis using a debugger takes a long time.

GetTickCount() starts counting after Windows boots. Through this function, malware or software can easily determine how long it has been since the computer booted up and got time value every time stamp counter cycle. The return value is the 32-bit form to keep the count for up to 49.7 days. Many developers use this function to measure elapsed time, and malware developers use it as well. The code on the left in Figure 10 is the original code of GetTickCount(). To defeat this, it must return the same value by hard coding. Hard coding is the way to develop software by embedding data directly into the source code or memory. In this way, it stores a meaningless value in the EAX register. Therefore, the return value of GetTickCount() is always meaningless value which hard coded. In other words, the measured time for a specific routine is always same.

timeGetTime() has less overhead than timeGetSystemTime(), because timeGetSystemTime() uses a MMTIME structure, which contains timing information for different multimedia data types. However, timeGetTime() uses a DWORD structure. Moreover, timeGetTime() reacts directly to timer interruptions and returns the time in milliseconds that have elapsed since the Windows booted, thus GetTickCount(). The debuggee can determine the presence of the debugger by using precise functions such as timeGetTime(). Disabling this is not much different with GetTickCount(). To defeat this function, one must to return the same value all the time. The code on the left of Figure 11 is the original code of timeGetTime(). However, if modified as the code on the right side of Figure 11, debuggee fails to detect the presence of debugger because it always returns the same value.

4. Experiments

This study implemented a sample program which includes anti-debugging techniques for the *(i) Sample-ex* experiment. The sample program has various anti-debugging techniques such as the PEB, and HEAP, among others. The sample program printed the return value to verify the results of the anti-anti-debugging intuitively. In the second experiment, *(ii) Packed-ex*, packed programs were used for packed malware or packed software. This study confirms the effectiveness of the proposed anti-anti-debugging mechanisms using these programs. Finally, real malware was employed in the in *(iii) Malware-ex* experiment. The anti-anti-debugging proposed in this paper works well with benign files, packed files, and actual malware through the step by step experiments.

Table 4: Anti-debugging techniques of each packer

Packer	Anti-Debugging Techniques
PECompact	IsDebuggerPresent()
Armadillo	IsDebuggerPresent()
ACProtect	IsDebuggerPresent()
	ZwQueryInformationProcess()
Themida	IsDebuggerPresent()
	ZwQueryInformationProcess()
Yoda's Protector	IsDebuggerPresent()
	GetCurrentProcessId()
	BlockInput()

Table 5: Information of real virus

File Type	Win.32.EXE
Virus Type	Worm/Win32.Abuse
MD5	068c3b80106b3300548bv6vv673a3db5
Anti-Debugging Techniques	`GetTickCount()`
	`FindWindowExA()`
	`FindWindowA()`

4.1. Experimental Data

This paper implemented a sample program including various anti-debugging techniques for *(i) Sample-ex* experiment, as shown in Figure 12. This program prints different results depending on the presence of a debugger. Next, sample files used in the *(ii) Packed-ex* experiment, packed the calculator program provided by Windows 7 (32bit) into five packers or protectors. The common feature of used packers or protectors is that they protect the original software by applying one or more anti-debugging techniques. Five packers called `Yoda's Protector`, `Themida`, `ACProtect`, `Armadillo`, and `PECompact` for *(ii) Packed-ex* experiment were used. These packers apply one or more anti-debugging techniques to protect original program against reverse engineering. Table 4 shows the anti-debugging techniques applied by each packer.

Among these, the most commonly used function is `IsDebuggerPresent()`, and the unusual cases are `GetCurrentProcessId()` and `BlockInput()` used by `Yoda's Protector`. Table 5 shows information about one of the malware samples collected from VirusShare[**], which is for the *(iii) Malware-ex* experiment. This malware sample uses `GetTickCount()`, `FindWindowExA()`, and `FindWindowA()` functions for anti-debugging. However, it is impossible to analyze some samples using different anti-debugging techniques such as `RDTSC` instruction and Self-Modifying technique.

`RDTSC` is an instruction that reads the current time-stamp counter variable. Unlike function such as `timeGetTime()`, `RDTSC` is executed as a single instruction. In order to bypass this instruction, it is necessary to modify the debuggee. However, modifying the debuggee requires careful attention. Self-Modifying is one of anti-analysis techniques. It is a technique to alter its own codes while the file is executing. Therefore, an analyst fails to identify the real instructions before executing Self-Modifying. Thus, `RDTSC` and Self-Modifying are out of our research scope and this paper does not handle these techniques.

4.2. Experimental Results

Figure 13 shows the results of analyzing sample data in the debugger in *(i) Sample-ex*. In the case of Figure 13 (a), all of the anti-debugging techniques used in the sample program detected a debugger. All of these anti-debugging techniques returned and printed artifact or evidence which indicates the presence of a debugger. However, an analysis was made again using anti-anti-debugging, and the results were confirmed, as shown in Figure 13 (b). It showed a completely different artifact or evidence and message against Figure 13 (a).

This research primarily confirmed that the anti-anti-debugging works successfully. Next, it conducted *(ii) Packed-ex* and analyzed using different data. The analysis from this point shows that the results of the analysis are limited. Therefore, the actual memory address and instructions were shown. Figure 14 (a) is the loaded file with `Yoda's Protector` in the debugger. The number series on the left correspond to each address, and instructions on the right correspond to the original code. Figure 14 (b) shows the new code to defeat `GetCurrentProcessId()`. It stored the PID of the debugger, which is 0x0B5C in the `EAX` register, to make it the same as shown in Figure 14 (b). This study also confirmed the results of `ZwQueryInformationProcess()` used by `ACProtect`, as shown in Figure 15. Figure 15 (a) is the original code and address of `ZwQueryInformationProcess()`, and Figure 15 (b) is the new instructions to defeat it. At this time, 0x003E0000 is the address of the virtual memory. Through this, whether the anti-anti-debugging is effective even in packed samples was secondarily ascertained.

```
void adbg_BeingDebuggedPEB(void)
{
    BOOL found = FALSE;
    _asm
    {
        xor eax, eax;           // clear eax
        mov eax, fs:[0x30];     // Reference start of the PEB
        movzx eax, [eax + 0x02];// PEB+2 points to BeingDebugged
        mov found, eax;         // Copy BeingDebugged into 'found'
    }

    if (found)
        printf("Caught by PEB.BeingDebugged check!\n");
    else
        printf("Not caught by PEB.BeingDebugged check!\n");

    printf("  Return value: 0x%08X\n", found);
}
```

(a) The code of `PEB.BeingDebugged` for sample program

```
void adbg_PEBNtGlobalFlag(void){
    BOOL found=FALSE;
    DWORD NtGlobalFlag;
    __asm
    {
        xor eax, eax // clear eax
        mov eax, fs:[0x30] // Address of PEB
        mov ecx, [eax+0x68] //Points to NtGlobalFlags
        mov NtGlobalFlag,ecx //FLG_HEAPENABLE_TAIL_CHECK(0x10)
                             //+ FLG_HEAP_ENABLE_FREE_CHECK(0x20)
                             //+ FLG_HEAP_VALIDATE_PARAMETERS(0x40)
    }
    if(NtGlobalFlag==0x70)
        printf("Caught by PEB.NtglobalFlag\n");
    else
        printf("Not caught by PEB.NtglobalFlag\n");
    printf("  Return value: 0x%08X\n", NtGlobalFlag);
}
```

(b) The code of `PEB.NtGlobalFlag` for sample program

```
void adbg_PEBHEAPFalgs(void){
    DWORD flags;
    __asm{
        mov eax, fs:[0x30]
        mov eax, dword ptr [eax+0x18]
        mov ecx, dword ptr [eax+0x40]
        mov flags, ecx
    }
    if(flags==0x40000062)
        printf("Caught by PEB.ProcessHeap.Flag\n");
    else
        printf("Not caught by PEB.ProcessHeap.Flag\n");
    printf("  Return Value: 0x%08X\n", flags);
}
```

(c) The code of `PEB.HEAP.Flags` for sample program

```
void adbg_PEBHEAPForceFalgs(void){
    DWORD forceflags;
    __asm{
        mov eax, fs:[0x30]
        mov eax, dword ptr [eax+0x18]
        mov ecx, dword ptr [eax+0x444]
        mov forceflags, ecx
    }
    if(forceflags==0x40000060)
        printf("Caught by PEB.ProcessHeap.ForceFlag\n");
    else
        printf("Not caught by PEB.ProcessHeap.ForceFlag\n");
    printf("  Return Value: 0x%08X\n", forceflags);
}
```

(d) The code of `PEB.HEAP.ForceFlags` for sample program

```
void adbg_IsDebuggerPresent(void)
{
    int found = 0;
    found = IsDebuggerPresent();
    if (found)
        printf("Caught by IsDebuggerPresent!\n");
    else
        printf("Not caught by IsDebuggerPresent\n");
    printf("  Return Value: 0x%08X\n",found);
}
```

(e) The code of `IsDebuggerPresent`()for sample program

```
void adbg_CheckRemoteDebuggerPresent(void)
{
    HANDLE hProcess = INVALID_HANDLE_VALUE;
    BOOL found = FALSE;
    hProcess = GetCurrentProcess();
    CheckRemoteDebuggerPresent(hProcess, &found);
    if (found)
        printf("Caught by CheckRemoteDebuggerPresent!\n");
    else
        printf("Not caught by CheckRemoteDebuggerPresent\n");
    printf("  Return Value: 0x%08X\n", found);
}
```

(f) The code of `CheckRemoteDebuggerPresent`()for sample program

```
void adbg_FindWindow(void){
    HWND result;
    result=FindWindow(NULL, L"Immunity Debugger - samplefile.exe");
    if(result==NULL)
        printf("Not caught by FindWindowW\n");
    else
        printf("Cautht by FindWindowW\n");
    printf("  Return value: 0x%08X\n", result);
}
```

(g) The code of `FindWindowW`() for sample program

```
void adbg_FindWindowA(void){
    HWND result;
    result=FindWindowA(NULL, "Immunity Debugger - samplefile.exe");
    if (result==NULL)
        printf("Not caught by FindwindowA\n");
    else
        printf("Cautht by FindWindowA\n");
    printf("  Return value: 0x%08X\n", result);
}
```

(h) The code of `FindWindowA`() for sample program

```
void adbg_gettickcount(void){
    DWORD beginTime=GetTickCount();
    DWORD endTime;
    __asm{PUSH ECX}
    __asm{MOV ECX, 3}
    __asm{PUSH EAX}
    __asm{XOR EAX, EAX}
    __asm{ADD EAX, 10}
    __asm{XCHG EAX, ECX}
    __asm{POP EAX}
    __asm{OR EAX, ECX}
    __asm{POP ECX}
    endTime=GetTickCount();
    if (endTime!=beginTime)
        printf("Caught by GetTickCount\n");
    else
        printf("Not caught by GetTickCount\n");
    printf("  Time difference: %d msec\n", endTime-beginTime);
}
```

(i) The code of `GetTickCount`() for sample program

```
void adbg_timegettime(void){
    DWORD beginTime=timeGetTime();
    DWORD endTime;
    __asm{PUSH ECX}
    __asm{MOV ECX, 3}
    __asm{PUSH EAX}
    __asm{XOR EAX, EAX}
    __asm{ADD EAX, 10}
    __asm{XCHG EAX, ECX}
    __asm{POP EAX}
    __asm{OR EAX, ECX}
    __asm{POP ECX}
    endTime=timeGetTime();
    if((endTime-beginTime)>0x10)
        printf("Caught by timeGetTime\n");
    else
        printf("Not caught by timeGetTime\n");
    printf("  Time difference: %d msec", endTime-beginTime);
}
```

(j) The code of `timeGetTime`() for sample program

```
void adbg_ZwQueryInformationProcess_ProcessDebugPort(void){
    DWORD ProcessDebugPort;
    PFZWQUERYINFORMATIONPROCESS pfZwQueryInformationProcess;
    HMODULE h_ntdll=GetModuleHandle(TEXT("ntdll.dll"));
    pfZwQueryInformationProcess=(PFZWQUERYINFORMATIONPROCESS)GetProcAddress(h_ntdll,"ZwQueryInformationProcess");
    pfZwQueryInformationProcess(GetCurrentProcess(),0x7,&ProcessDebugPort,4,0);

    if(ProcessDebugPort==-1)
        printf("Caught by ZwQueryInformationProcess(ProcessDebugPort)\n");

    else
        printf("Not caught by ZwQueryInformationProcess(ProcessDebugPort)\n");
    printf("  Return Value: 0x%08X\n", ProcessDebugPort);
}
```

(k) The code of `ZwQueryInformationProcess - ProcessDebugPort` for sample program

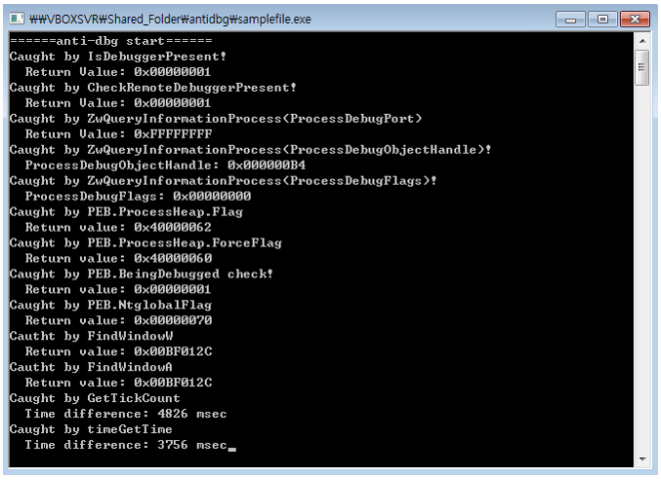
```
void adbg_ZwQueryInformationProcess_ProcessDebugObject(void){
    BOOL result=FALSE;
    DWORD fRet;
    DWORD ProcessDebugObjctHandle;
    PFZWQUERYINFORMATIONPROCESS pfZwQueryInformationProcess;
    HMODULE h_ntdll=GetModuleHandle(TEXT("ntdll.dll"));
    pfZwQueryInformationProcess = (PFZWQUERYINFORMATIONPROCESS)GetProcAddress(h_ntdll, "ZwQueryInformationProcess");
    fRet = pfZwQueryInformationProcess(GetCurrentProcess(), 0x1e, &ProcessDebugObjctHandle, 4, 0); // 0x1e : ProcessObjectHandle

    if(ProcessDebugObjctHandle)
        if(fRet==0) result=TRUE;
    if(result)
        printf("Caught by ZwQueryInformationProcess(ProcessDebugObjectHandle)!\n");
    else
        printf("Not Caught by ZwQueryInformationProcess(ProcessDebugObjectHandle)\n");

    printf("  ProcessDebugObjectHandle: 0x%08X\n", ProcessDebugObjctHandle);
}
```

(l) The code of `ZwQueryInformationProcess – ProcessDebugObjectHandle` for sample program

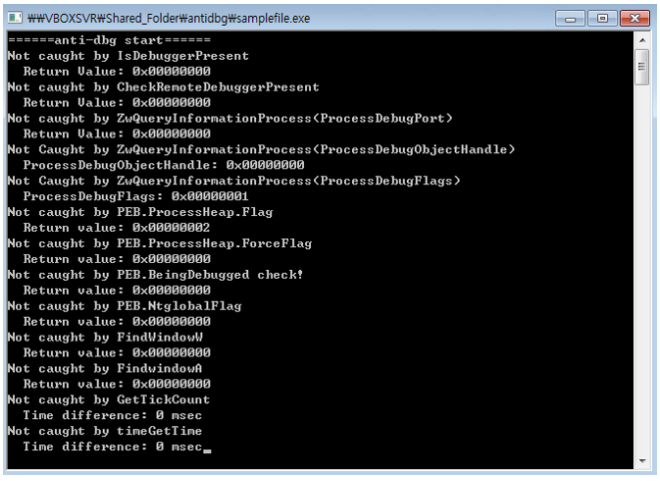
```
void adbg_ZwQueryInformationProcess_ProcessDebugFlags(void)
{
    BOOL result = FALSE;
    DWORD fRet;
    DWORD ProcessDebugFlags;
    PFZWQUERYINFORMATIONPROCESS pfZwQueryInformationProcess;
    HMODULE h_ntdll=GetModuleHandle(TEXT("ntdll.dll"));
    pfZwQueryInformationProcess = (PFZWQUERYINFORMATIONPROCESS)GetProcAddress(h_ntdll, "ZwQueryInformationProcess");
    fRet = pfZwQueryInformationProcess(GetCurrentProcess(), 0x1f, &ProcessDebugFlags, 4, 0); // 0x1f : ProcessDebugFlags

    if(!ProcessDebugFlags) result = TRUE;
    if(result)
        printf("Caught by ZwQueryInformationProcess(ProcessDebugFlags)!\n");
    else
        printf("Not Caught by ZwQueryInformationProcess(ProcessDebugFlags)\n");
    printf("  ProcessDebugFlags: 0x%08X\n", ProcessDebugFlags);
}
```

(m) The code of `ZwQueryInformationProcess – ProcessDebugFlags` for sample program

Figure 12: Anti-debugging functions for sample program



(a) Result without anti-anti-debugging



(b) Result with anti-anti-debugging

Figure 13: Results of (i) Sample-ex; (a) results without anti-anti-debugging; (b) results with anti-anti-debugging

```
75D368D0 MOV EAX,DWORD PTR FS:[18]
75D368D6 MOV EAX,DWORD PTR DS:[EAX+20]
75D368D9 RETN
```

(a) The original code and address of `GetCurrentProcessId()`

```
75D368D0 MOV EAX,0B5C
75D368D6 NOP
75D368D7 NOP
75D368D8 NOP
75D368D9 RETN
```

(b) The new code and address to defeat `GetCurrentProcessId()`

Figure 14: Results of (ii) Packed-ex in `Yoda's Protector`; (a) the original code and address of `GetCurrentProcessId()`; (b) the new code and address to defeat `GetCurrentProcessId()`

```
779A6048 MOV EAX,0EA
779A604D MOV EDX,7FFE0300
779A6052 CALL DWORD PTR DS:[EDX]
779A6054 RETN 14
```

(a) The original code and address of `ZwQueryInformationProcess()`

```
779A6048 PUSH 3E0000
779A604D RETN
003E0000 MOV EAX,0EA
003E0005 MOV EDX,7FFE0300
003E000A CMP DWORD PTR SS:[ESP+8],7
003E0015 JE SHORT 003E0037
003E0017 CMP DWORD PTR SS:[ESP+8],1E
003E0022 JE SHORT 003E004A
003E0024 CMP DWORD PTR SS:[ESP+8],1F
003E002F JE SHORT 003E005D
003E0031 PUSH 779A6052
003E0036 RETN
003E0037 MOV EAX,DWORD PTR SS:[ESP+C]
003E003E PUSH 0
003E0043 POP DWORD PTR DS:[EAX]
003E0045 XOR EAX,EAX
003E0047 RETN 14
003E004A MOV EAX,DWORD PTR SS:[ESP+C]
003E0051 PUSH 0
003E0056 POP DWORD PTR DS:[EAX]
003E0058 XOR EAX,EAX
003E005A RETN 14
003E005D MOV EAX,DWORD PTR SS:[ESP+C]
003E0064 PUSH 1
003E0069 POP DWORD PTR DS:[EAX]
003E006B RETN 14
```

(b) The new code and address to defeat `ZwQueryInformationProcess()`

Figure 15: Results of (ii) Packed-ex in ACProtect; (a) the original code and address of ZwQueryInformationProcess (); (b) the new code and address to defeat ZwQueryInformationProcess ()

```
76F86F69 MOV EDI,EDI
76F86F6B PUSH EBP
76F86F6C MOV EBP,ESP
76F86F6E PUSH 0
76F86F70 PUSH DWORD PTR SS:[EBP+14]
76F86F73 PUSH DWORD PTR SS:[EBP+10]
76F86F76 PUSH DWORD PTR SS:[EBP+C]
76F86F79 PUSH DWORD PTR SS:[EBP+8]
76F86F7C CALL USER32.76F88F4F
76F86F81 POP EBP
76F86F82 RETN 10
```

(a) The original code of FindWindowExA()

```
76F86F69 MOV EDI,EDI
76F86F6B PUSH EBP
76F86F6C MOV EBP,ESP
76F86F6E PUSH 0
76F86F70 PUSH DWORD PTR SS:[EBP+14]
76F86F73 PUSH DWORD PTR SS:[EBP+10]
76F86F76 PUSH DWORD PTR SS:[EBP+C]
76F86F79 PUSH DWORD PTR SS:[EBP+8]
76F86F7C PUSH 190000
76F86F81 RETN
76F86F82 RETN 10
00190000 CALL USER32.76F88F4F
00190005 CMP EAX,340722
0019000B JE SHORT 0019000F
0019000D JMP SHORT 00190011
0019000F XOR EAX,EAX
00190011 POP EBP
00190012 RETN 10
```

(b) The modification code of FindWindowExA()

```
76F88FF3 MOV EDI,EDI
76F88FF5 PUSH EBP
76F88FF6 MOV EBP,ESP
76F88FF8 XOR EAX,EAX
76F88FFA PUSH EAX
76F88FFB PUSH DWORD PTR SS:[EBP+C]
76F88FFE PUSH DWORD PTR SS:[EBP+8]
76F89001 PUSH EAX
76F89002 PUSH EAX
76F89003 CALL USER32.76F88F4F
76F89008 POP EBP
76F89009 RETN 8
```

(c) The original code of FindWindowA()

```
76F88FF3 MOV EDI,EDI
76F88FF5 PUSH EBP
76F88FF6 MOV EBP,ESP
76F88FF8 XOR EAX,EAX
76F88FFA PUSH EAX
76F88FFB PUSH DWORD PTR SS:[EBP+C]
76F88FFE PUSH DWORD PTR SS:[EBP+8]
76F89001 PUSH EAX
76F89002 PUSH EAX
76F89003 PUSH 170000
76F89008 RETN
76F89009 RETN 8
00170000 CALL USER32.76F88F4F
00170005 CMP EAX,340722
0017000B JE SHORT 0017000F
0017000D JMP SHORT 00170011
0017000F XOR EAX,EAX
00170011 POP EBP
00170012 RETN 8
```

(d) The modification code of FindWindowA()

```
75D36740 MOV ECX,DWORD PTR DS:[7FFE0324]
75D36746 MOV EDX,DWORD PTR DS:[7FFE0320]
75D3674C MOV EAX,DWORD PTR DS:[7FFE0328]
75D36751 CMP ECX,EAX
75D36753 JNZ SHORT KERNELBA.75D36770
75D36755 MOV EAX,DWORD PTR DS:[7FFE0004]
75D3675A MUL EDX
75D3675C SHL ECX,8
75D3675F IMUL ECX,DWORD PTR DS:[7FFE0004]
75D36766 SHRD EAX,EDX,18
75D3676A SHR EDX,18
75D3676D ADD EAX,ECX
75D3676F RETN
```

(e) The original code of GetTickCount()

```
75D36740 MOV EDX,7FFE0000
75D36746 XCHG EAX,EDI
75D36747 MOV EDI,5678ABCD
75D3674C XCHG EAX,EDI
75D3674D RETN
```

(f) The modification code of GetTickCount()

Figure 16: Results of *(iii) Malware-ex;* (a), (c) and (e) are the original code of each anti-debugging; (b), (d), and (f) are modification code to defeat each anti-debugging techniques

Table 6: Comparison results of anti-anti-debugging experiments with other works (O: successful, X: failed)

Technique	Proposed Work	Pin	Apate	Binary Substitution
`IsDebuggerPresent()`	O	O	O	O
`CheckRemoteDebuggerPresent()`	O	O	X	O
`FindWindow()`	O	O	X	O
`QueryInformationProcess()` (`ProcessDebugPort`)	O	O	X	O
`QueryInformationProcess()` (`ProcessDebugFlags`)	O	X	X	O
`QueryInformationProcess()` (`ProcessDebugObjectHandle`)	O	O	O	O
`GetTickCount()`	O	O	O	X
`timeGetTime()`	O	O	O	X
`GetCurrentProcessId()`	O	X	X	X
`BlockInput`	O	X	O	X
`PEB.BeingDebugged`	O	O	O	O
`PEB.NtGlobalFlag`	O	O	O	O
`HEAP.Flags`	O	O	O	X
`HEAP.ForceFlags`	O	O	O	X
`RDTSC` Instruction	X	X	O	O
`Self-Modifying`	X	X	X	X
Success Rate	87.5%	68.8%	62.5%	56.3%

Finally, this paper collected and experimented with malware, which has anti-debugging techniques for *(iii) Malware-ex*. As a result of the *(ii) Packed-ex* experiment, this paper listed the address and instructions in Figure 16, which shows the loaded malware sample results used in *(iii) Malware-ex* on the debugger and the results of defeating its anti-debugging techniques. Figure 16 (a) is the original code of `FindWindowExA()`, and Figure 16 (b) is the new code that defeats it. This study wrote a new code that was occupied in 0x00190000 virtual memory to defeat the function. Figure 16 (c) is the original code of `FindWindowA()`, and the code that defeats it is Figure 16 (d). This study occupied and wrote the new code on the memory address 0x00170000 to defeat this. As mentioned before, virtual memory addresses cannot always be the same and must occupy the memory with empty space. The last anti-debugging technique is `GetTickCount()`. Figure 16 (e) is the original code of this function, modified as in Figure 16 (f), in which malware failed to detect the presence of a debugger, and the malware analysis could proceed. Thus, this study experimented and confirmed step by step that the anti-anti-debugging proposed in this paper is effective.

Table 6 shows the experimental results for 16 anti-debugging techniques. The Binary Substitution [25], which shows the lowest success rate among each work, changes debuggee's binary. It requires special attention. Pin [4] is an excellent analysis tool, but it cannot defeat the anti-debugging techniques used by `Yoda's Protector`. Apate [23], which has a similar success rate to Pin, has also been unable to defeat the anti-debugging techniques used by `Yoda's Protector`, and it can defeat fewer anti-debugging techniques than Pin. As shown in Table 6, most anti-debugging techniques are based on APIs, PEB, and HEAP structures. However, there are far more anti-debugging techniques, and their research is needed for future works.

Among the many samples, there were some that this paper failed to analyze. Because this work can defeat some APIs and PEB structure-based anti-debugging techniques, the anti-debugging using different artifacts could not be defeated, such as `RDTSC` instruction, Memory Breakpoint, Self-Modifying, and Single-Step Detection [22].

5. Conclusion

Malware as well as commercial packers use various anti-debugging techniques to protect themselves from the analysis by reverse engineering. Anti-debugging techniques can bypass or neutralize debugging analysis. An analyst needs an advanced debugger and related knowledge in order to analyze malware with anti-debugging techniques. Thus, this study proposed an anti-anti-debugging against the anti-debugging techniques. In order to explain the anti-anti-debugging, we expounded the anti-debugging techniques used by malware and packers. Also, it set up a step by step experiment for verification. It implemented, experimented, and analyzed sample files, and packed files with anti-debugging techniques. Finally, it experimented with the actual malware with many anti-debugging techniques. As a result of a step by step experiment, it found that proposed anti-anti-debugging can defeat the anti-debugging techniques and deal with actual malware. It did not interfere with program execution flow as less as possible by minimizing the debuggee's direct modification. Of course, the method proposed in this paper cannot defeat all anti-debugging techniques. Nevertheless, the method serves an essential part in analyzing malware with a debugger. Furthermore, it improves the ability to handle sophisticated debugging evasion techniques.

Also, debuggees, such as `Yoda's Protector`, can call the anti-debugging techniques more than once to interfere with the analysis. However, the anti-anti-debugging proposed in this paper has the advantage of being able to defeat the anti-debugging

technique that is called multiple times by only modifying it once. There are several tools and methods to disable anti-debugging techniques, however, none elaborate on how to disable it or explain with assembly code. Thus, this paper aims to discover for analysts new ways to defeat anti-debugging techniques and help researchers develop their scripts for research and analyze malware.

Data Availability

The script file used in the experiments is available at https://github.com/goldbear564/antiantidebugging.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. NRF-2020R1A2C1012117, Development of machine learning based intrusion detection platform to secure end device in edge computing environment) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant by the Korea government (MSIT) (No.2017-0-00158, Development of Cyber Threat Intelligence (CTI) analysis and information sharing technology for national cyber incident).

References

[1] R. Goyda, S. Sharma, and S. Bevinakoppa, "Obfuscation of stuxnet and flame malware," In. Proc. of the 3rd international conference on applied informatics and computing theory(AICT'12), 150--154, 2012.

[2] J. W. Kim, J. Bang, Y.S. Moon, and M. J. Choi, "Disabling anti-debugging techniques for unpacking system in user-level debugger," in 2019 International Conference on Information and Communication Technology Convergence(ICTC), Jeju Island, South Korea, 2019, doi:10.1109/ICTC46691.2019.8939719.

[3] M. N. Gagnon, S. Taylor, and A. K. Ghosh, "Software protection through anti-debugging," IEEE Security & Privacy, **5**(3), 82--84, 2007, doi:10.1109/MSP.2007.71.

[4] C.K. Luk, R. Chon, R. Muth, H. Patil, and A. Klauser, "Pin: building customized program analysis tools with dynamic instrumentation, " ACM SIGPLAN Notices, **40**(6), 190--200, 2005, doi:10.1145/1064978.1065034.

[5] B. B. Rad, M. Masrom, and S. Ibrahim, "Camouflage in malware: From encryption to metamorphism, " International Journal of Computer Science and Network Security, **12**(8), 74--83, 2012. doi:10.1007/s11416-017-0291-9

[6] S. Bardin, R. David, and J. Y. Marion, "Backward-bounded DSE: targeting infeasibility questions on obfuscated codes," 2017 IEEE Symposium on Security and Privacy, 633--651, 2017, doi:10.1109/SP.2017.36.

[7] T. Blazytko, M. Contag, and C. Aschermann, "Syntia: Synthesizing the Semantics of Obfuscated Code," In Proc. of USENIX Security Symposium, 643--659, 2017.

[8] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandram, and S. Venkatran man, "Robust intelligent malware detection using deep learning," IEEE Access, **7**, 46717--46738, 2019, doi:10.1109/ACCESS.2019.2906934.

[9] X. Chen, J. Andersen, Z. M. Mao, and M. Bailey, "Towards an understanding of anti-virtualization and anti-debugging behavior in modern malware," in 2008 IEEE international conference on dependable systems and networks with FTCS and DCC(DSN), 177--186, 2008, doi:10.1109/DSN.2008.4630086.

[10] M. J. Choi, J. Bang, J. W. Kim, H. Kim, and Y. S. Moon, "All-in-one framework for detection, unpacking, and verification for malware analysis," Security and Communication Network 2019, **2019**(5278137), 1--16, 2019, doi:10.1155/2019/5278137.

[11] S. Cesare, Y. Xiang, and W. Zhou, "Malwise – an effective and efficient classification system for packed and polymorphic malware," IEEE Transaction on Computers, **62**(6), 1193--1206, 2013, doi:10.1109/TC.2012.65.

[12] F. Guo, P. Ferrie, and T. C. Chiueh, "A study of the packer problem and its solutions," in International Workshop on Recent Advances in Intrusion Detection, 98--115, 2008, doi:10.1007/978-3-540-87403-4_6.

[13] J. M. Borello, and L. Mé, "Code obfuscation techniques for metamorphic viruses," Journal in Computer Virology, **4**(3), 211--200, 2008, doi:10.1007/s11416-008-0084-2.

[14] I. You, and K. Yim, "Malware obfuscation techniques: a brief survey," in 2010 International conference on broadband, wireless computing, communication and applications, 297--300, 2010, doi:10.1109/BWCCA.2010.85.

[15] A. Moser, C. Kruegel, and E. Kirda, "Limits of static analysis for malware detection," in 23rd Annual Computer Security Applications Conference(ACSAC'07), 421--430, 2007, doi:10.1109/ACSAC.2007.21.

[16] R. R. Branco, G. N. Barbosa, and P. D. Neo, "Scientific but not academical overview of malware anti-debugging, anti-disassembly and anti-vm technologies," in Blackhat, USA, 2012.

[17] S. Gao, and Q. Lin, "Debugging classification and anti-debugging strategies," in 4th International Conference on Machine Vision(ICMV): Computer Vision and Image Analysis; Pattern Recognition and Basic Technologies, 729--734, 2012, doi:10.1117/12.924835.

[18] P. Chen, C. Huygens, L.Desmet, and W. Joosen, "Advanced or not? a comparative study of the use of anti-debugging and anti-vm techniques in generic and targeted malware," in IFIP International Conference on ICT Systems Security and Privacy Protection, 323--336, 2016, doi:10.1007/978-3-319-33630-5_22.

[19] S. Bhansali, W. K. Chen, S. D. Jong, A. Edwards, "Framework for instruction-level tracing and analysis of program executions," In Proc. of the 2nd International Conference on Virtual Execution Environments, 154--163, 2013, doi:10.1145/1134760.1220164.

[20] P. Feiner, A. D. Brown, and A. Goel, "Comprehensive kernel instrumentation via dynamic binary translation," In Proc. of the 17th international conference on Architectural Support for Programming Languages and Operating Systems, 135--146, 2012, doi:10.1145/2150976.2150992.

[21] N. Nethercote, and J. Seward, "Valgrind: a framework for heavyweight dynamic binary instrumentation," ACM SIGPLAN Notices, **42**(6), 89--100, 2007, doi:10.1145/1273442.1250746.

[22] S. Choi, T. Chang, S. Yoon, and Y. Park, "Hybrid emulation for bypassing anti-reversing techniques and analyzing malware," The Journal of Supercomputing, 1--27, 2020, doi:10.1007/s11227-020-03270-6

[23] H. Shi, and J. Mirkovic, "Hiding debuggers from malware with apate," In Proc. of the Symposium on Applied Computing, 1703--1710, 2017, doi:10.1145/3019612.3019791.

[24] J. K. Lee, B. J. Kang, and E. G. Im, "Rule-based anti-anti-debugging system," In Proc. of the 2013 Research in Adaptive and Convergent Systems, 353--354, 2013, doi:10.1145/2513228.2513301.

[25] J. K. Lee, B. J. Kang, and E. G. Im, "Evading anti-debugging techniques with binary substitution," International Journal of Security and its Applications, **8**(1), 183--192, 2014, doi:10.14257/ijsia.2014.8.1.17.

[26] P. Feng, J. Sun, S. Liu, and K. Sun, "UBER: combating sandbox evasion via user behavior emulators," International Conference on Information and Communications Security, 34--50, 2020, doi:10.1007/978-3-030-41579-2_3

[27] D. P. Pham, D. L. Vu, and F. Massacci, "Mac-A-Mal: macOS malware analysis framework resistant to anti evasion techniques," Journal of Computer Virology and Hacking Techniques, **15**(4), 249--257, 2019, doi:10.1007/s11416-019-00335-w.

[28] O. Or-Meir, N. Nissim, Y. Elovici, and L. Rokach, "Dynamic malware analysis in the modern era- a state of the art survey," ACM Computing Surveys (CSUR), **52**(5), 1-48, 2019, doi: 10.1145/3329786.

[29] C. V. Liţă, D. Cosovan, and D. Gavriluţ. "Anti-emulation trends in modern packers: a survey on the evolution of anti-emulation techniques in UPA packers," Journal of Computer Virology and Hacking Techniques, **14**(2), 107--126, 2018, doi:10.1007/s11416-017-0291-9.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 1189-1195 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Synthesis of SQL Queries from South African Local Language Narrations

George Obaido[*,1,2], Abejide Ade-Ibijola[2], Hima Vadapalli[1]

[1]*School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, 2001, South Africa.*

[2]*School of Consumer Intelligence and Information Systems, University of Johannesburg, Johannesburg, 2001, South Africa.*

ARTICLE INFO

Article history:
Received: 25 August 2020
Accepted: 18 October 2020
Online: 24 October 2020

Keywords:
Language translation
Synthesis of things
JFA applications
SQL queries

ABSTRACT

English remains the language of choice for database courses and widely used for instruction in nearly all South African universities, and also in many other countries. Novice programmers of native origins are mostly taught Structured Query Language (SQL) through English as the medium of instruction. Consequently, this creates a myriad of problems in understanding the syntax of SQL as most native learners are not too proficient in English. This could affect a learner's ability in comprehending SQL syntaxes. To resolve this problem, this work proposes a tool called local language narrations to SQL (`Local-Nar-SQL`) that uses a type of Finite Machine, such as a Jumping Finite Automaton to translate local language narratives into SQL queries. Further, the generated query extracts information from a sample database and presents an output to the learner. This paper is an extension of work originally presented in a previous study in this field. A survey involving 145 participants concluded that the majority found `Local-Nar-SQL` to be helpful in understanding SQL queries from local languages. If the proposed tool is used as a learning aid, native learners will find it easier to work with SQL, which will eliminate many of the barriers faced with English proficiencies in programming pedagogies.

1 Introduction

According to Linguists, English has much of its vocabulary borrowed from French and Latin, but it remains a Germanic language in terms of structure and sounds [1]–[2]. English is the predominant language of instruction used at most universities and remains the *de facto* language of communication for many industry sectors [3, 4]. It is interesting to note that many of the world's population do not speak English as their first language, but it is vastly used for communication and dissemination of knowledge [5, 6]. Most programming and scripting languages, such as C, Java, Python and Structured Query Language (SQL) have keywords, declarations and documentations based in English [6]–[7]. Even online resources, such as forums, blogs and developers communities that aid the learning of programming are primarily in English [8, 9]. For example, Stack Overflow, which is one of the largest online communities for developers, enforces that all questions and answers are posted in English. Hence, for a novice programmer to be proficient in programming, English is a necessity [10].

SQL is the most popular query declarative language used for performing operations on relational databases [11, 12]. As a standardised language, many of its syntaxes and statements are English-like, and SQL has found numerous applications in both industry and academia [13, 14]. Listing 1 shows examples of the SQL SELECT query statements that depict English-like formation. These set of queries would retrieve one or more records from a targeted database's table.

Listing 1: SQL SELECT statement query

```
SELECT * FROM myrecordinfo;

SELECT DISTINCT firstname, lastname
FROM myrecordinfo;

SELECT * FROM myrecordinfo
WHERE firstname='karabo' AND lastname='hlophe';

SELECT * FROM Customers
WHERE country='South Africa'
OR city = 'Pietermaritzburg';
```

[*]Corresponding Author: George Obaido, University of the Witwatersrand, School of Computer Science and Applied Mathematics, Johannesburg, 2001, South Africa. Email: rabeshi.george@gmail.com & Contact: +27832227427

South Africa is a multilingual country with eleven official languages [15]–[16]. Most South African learners prefer to communicate in their local dialects, rather than English [17, 16]. Educational instruction for learning programming requires a good proficiency in English, which creates numerous challenges for a local student whose native language is not English [6, 18, 19]. Just like any other programming language, anecdotal evidence has shown that SQL is hard for students [11]–[13], [20]. Even the straightforward, English-like syntax of SQL is misleading [21, 22]. Other studies have identified that the burden of remembering SQL keywords and database schemas are often challenging [20, 23]. These challenges have made it crucial for researchers and developers to create *culturally-agnostic* aids that would assist learners to improve their understanding of SQL. Such systems could be developed to enable learners express their requests in *free-forms* using local language narrations. To our knowledge, this is the first study that attempts to translate South African local language narrations into SQL queries.

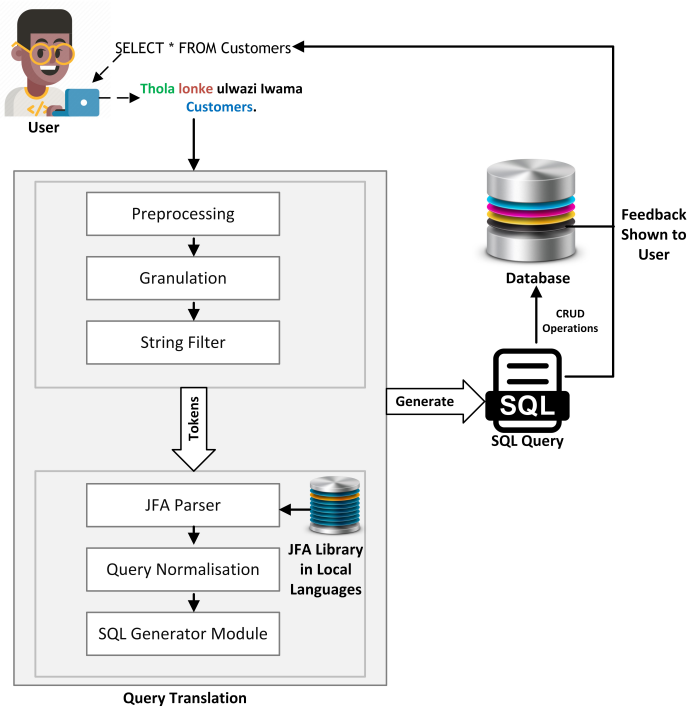


Figure 1: The translation process (adapted from [11])

This paper is an extension of work originally presented in [11], where natural language specifications were translated into SQL queries. We extended the use of a Jumping Finite Automaton, a type of Finite Machine for the translation of local language specifications into SQL queries. This idea was further implemented into a desktop-based application called `Local-Nar-SQL`. The process of translating the local language specifications into SQL queries is presented in Figure 1. Here, a user provides narrations in a local language, that is then processed by `Local-Nar-SQL`. The system preprocesses and passes the provided narrations to the JFA parser for matches. Further, the tokens are matched and a query is generated, which is made available to the user. In this paper, the following contributions were made. This study has:

1. developed a method for the recognition of local language specifications using a JFA,
2. extended this approach into a tool, called `Local-Nar-SQL`, for the translation of local language specifications into SQL queries, and
3. evaluated this approach using human subjects and presented initial results.

The rest of this paper is organised as follows: Section 2 highlights the background and related works, Section 3 presents the methodology used for this study, Section 4 shows the implementation and results details, the evaluation information is provided in Section 5 and Section 6 concludes the paper and highlights discussions for future studies.

2 Background

This section presents the problems as research questions, highlights the motivation behind this work and discusses related work.

2.1 Research Questions

This study attempts to solve the following problems:

1. how can we successfully translate local language narrations into SQL queries?
2. can we describe a JFA and implement this into a tool that addresses the above problem?
3. what are human perceptions of the developed tool?

These problems were addressed in Section 3, 4 and 5.

2.2 Motivation

English proficiency is a problem in South Africa English proficiency remains a major concern for local students learning programming [10]. Many programming language constructs, keywords, API methods and libraries are mostly written in English, which pose a great deal of difficulty for a local learner [6, 18, 19].

Decolonisation of science Issues regarding the decolonisation of the science curriculum is currently being debated in South Africa and the rest of the world [24]. Many local learners agree that the legacies of apartheid (entrenched colonialism) are still evident in education sectors, built by western perspectives, which might impinge access to learning [25, 26].

Comprehension problems in SQL Generally, the English-like syntax of SQL is misleading for learners [12, 14]. Similarly, most query interfaces are built to report errors in English, which might be confusing for learners to decipher [27, 28].

Abstract nature of programming From a novice learners' perspective, the process of learning a new programming language remains a daunting task [29, 30]. Even so, the abstract nature of programming has resulted in high failure rates for Computer Science (CS) disciplines [31, 32].

2.3 Why Local Language Narrations

In Section 2.2, it was seen that English language poses numerous difficulties for local language learners. An ideal approach would be to enable such learners to express their intention, free from natural language ambiguities – *in narrations*. Narrations have been shown to provide textual explanations for programming and scripting language paradigms [11, 13, 32]. This practice allows learners to express their requests in *free-forms*, ignoring programming language syntax. Such *free-forms* allow learners to express their requests in their local language, free from English language ambiguities. For programming languages, especially in low-resource settings, such an approach has been tried in [7, 33, 34].

2.4 Related Work and Terms

This section presents similar work in the area of language translations and the applications of JFA.

Language Translation An evolutionary method for translating Portuguese language into SQL query [35], a pattern matching approach for translating Hindi language into SQL query [36], Telugu language translation into SQL query [37], translating Italian language to SQL query [38], translating Chinese language into SQL query [39] and Arabic translation into SQL queries [40]–[41].

JFA Applications Abstraction of frequently asked questions (FAQs) for comprehension purposes [42], synthesising SQL queries from narratives [11] and the automatic comprehension of tweets [43].

SQL Comprehension Generating narratives of SQL queries [12], generating narratives of nested SQL queries [14], generating SQL queries from visual specifications [13] and synthesising SQL queries from verbal specifications [44].

The method used in this paper is a JFA, and definitions of this term are as follows.

Definition 1 (A Jumping Finite Automaton (JFA) [45]) *A JFA is a five-tuple, M = (Q, Σ, R, s, F) where:*

1. *Q is a bounded set of states,*
2. *Σ is an input alphabet,*
3. *R is the bounded set of rules, where py → q (p,q ∈ Q, y∈Σ),*
4. *s∈Q is the start state, and*
5. *F⊆Q is the final state.*

3 Methodology

This section presents the methodology for this study. To begin, local languages were abstracted into a JFA, then queries were normalised, where irrelevant details were removed. Next, a query was generated.

3.1 Abstracting Local Languages to SQL

For local language abstraction into a JFA, entities were identified with colours, such as query types ($\sum_{QT} = a_x$) in green, attributes ($\sum_{CT} = b_y$) in red and relations ($\sum_{ET} = c_z$) in blue. This information is presented in Table 1. The Chinook DB [46] was selected for our use-case with a corpus containing a total of 120 different local languages conversions. This database contains a typical digital music store, consisting of eleven tables and multiple records. Some examples of the JFA abstractions are presented.

Table 1: JFA symbols 1 - 30

a_x	b_y	c_z
a_0 Khombisa	b_0 EmployeeID	c_0 Employee
a_1 Ungitholele	b_1 LastName	c_1 Genre
a_2 Ngitholela	b_2 FirstName	c_2 Customer
a_3 Ukuthola	b_3 Title	c_3 Mediatype
a_4 Usika	b_4 ReportTo	c_4 Track
a_5 Fumana	b_5 Address	c_5 Invoice_item
a_6 Khetha	b_6 State	c_6 Invoice
a_7 Enta	b_7 City	c_7 Phone
a_8 Faka	b_8 PostalCode	c_8 Playlist_track
a_9 Susa	b_9 Fax	c_9 Album
a_{10} Funda	b_{10} Country	c_{10} Artist
a_{11} Wys	b_{11} Email	–
a_{12} Vind	b_{12} CustomerID	–
a_{13} Lys	b_{13} SupportRepID	–
a_{14} Kry	b_{14} TrackID	–
a_{15} Plaas	b_{15} ArtistID	–
a_{16} Opdateer	b_{16} InvoiceID	–
a_{17} Verwyder	b_{17} MediaTypeID	–
a_{18} Skep	b_{18} InvoiceLineID	–
a_{19} Lees	b_{19} Name	–
a_{20} Nghenisa	b_{20} UnitPrice	–
a_{21} Sula	b_{21} Composer	–
a_{22} Susa	b_{22} Company	–
a_{23} Engetela	b_{23} Al | Alle | Alles	–
a_{24} Tumbuluxa	b_{24} Konke | Yonke	–
a_{25} Tlakusisa	b_{25} Yothe | Vhothe	–
a_{26} Ulonga	b_{26} –	–
a_{27} Ubvisa	b_{27} –	–
a_{28} Udadzisa	b_{28} –	–
a_{29} Utomola	b_{29} –	–
a_{30} Usika	b_{30} –	–

- Ek will al die customer besonderhede vind

Example 1 above shows a query request sentence in Afrikaans. The English translation for this sentence is: "I want to find all the customer's details". The equivalent JFA shows:

$$\mathbf{M} = (\{I, J, K, L\}, \{b_{23}, c_2, a_{12}\}, R, I; \{L\})$$

where $\{I, J, K, L\}$ are the states,
$\{b_{23}, c_2, a_{12}\}$, are the input alphabets,
R is the set of rules,
I is a start state, and
$\{L\}$ is a final state.

with

$$\mathbf{R} = \{Ib_{23} \rightarrow J, Jc_2 \rightarrow K, Ka_{12} \rightarrow L\}$$

accepts

$$\text{L(M)} = \{w \in \{b_{23}, c_2, a_{12}\}^*: |b_{23}| = |c_2| = |a_{23}|\}$$
$$\text{i.e. } b_{23} = al; c_2 = customers; a_{12} = vind$$

$$\begin{aligned} b_{23}c_2a_{12}a_{12}c_2\underline{I}b_{23} &\curvearrowright b_{23}c_2a_{12}\underline{J}a_{12}c_2 && [Ib_{23} \rightarrow J] \\ &\curvearrowright b_{23}c_2a_{12}\underline{K}a_{12} && [Jc_2 \rightarrow K] \\ &\curvearrowright \underline{L}b_{23}c_2a_{12} && [Ka_{12} \rightarrow L] \end{aligned}$$

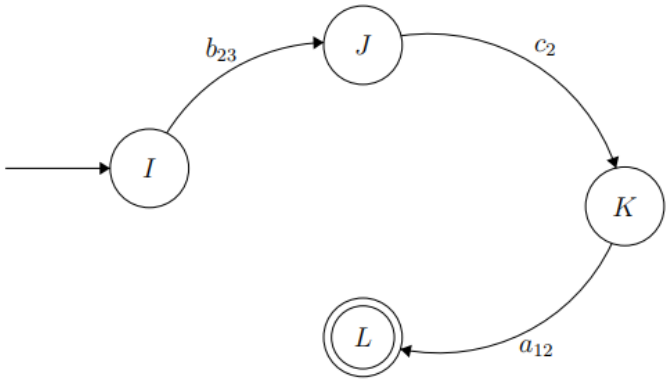


Figure 2: An example of a sample JFA

Figure 2 shows an example of a sample JFA with three transitions. The diagram consists of four states, labelled as $\{I, J, K, L\}$, where the start state is I and accepting state is L. Ib_{21} begins the movement with the *al* keyword, then Jc_2 shows the movement with the *customer* keyword, and the transition, Ka_{12} ends with the *vind* keyword. Since local languages are rich in vocabulary, the input alphabet may differ in another language, which might result in a change in the states and transitions.

- Ngifuna ukuthola yonke imininingwane ya ma customer

The above sentence is the Zulu translation of: "I want to find all the customer's details". The equivalent JFA shows:

$$\mathbf{M} = (\{I, J, K, L\}, \{a_3, b_{24}, c_2\}, R, I; \{L\})$$

where $\{I, J, K, L\}$ are the states,
$\{a_3, b_{24}, c_2\}$, are the input alphabets,
R is the set of rules,
I is a start state, and
$\{L\}$ is a final state.

with

$$\mathbf{R} = \{Ia_3 \rightarrow J, Jb_{24} \rightarrow K, Kc_2 \rightarrow L\}$$

accepts

$$\text{L(M)} = \{w \in \{a_3, b_{24}, c_2\}^*: |a_3| = |b_{24}| = |c_2|\}$$
$$\text{i.e. } a_3 = ukuthola; b_{24} = yonke; c_2 = customers$$

$$\begin{aligned} b_{24}a_3c_2b_{24}c_2\underline{I}a_3 &\curvearrowright b_{24}a_3c_2\underline{J}b_{24}c_2 && [Ia_3 \rightarrow J] \\ &\curvearrowright b_{24}a_3c_2\underline{K}c_2 && [Jb_{24} \rightarrow K] \\ &\curvearrowright \underline{K}b_{24}a_3c_2 && [Kc_2 \rightarrow L] \end{aligned}$$

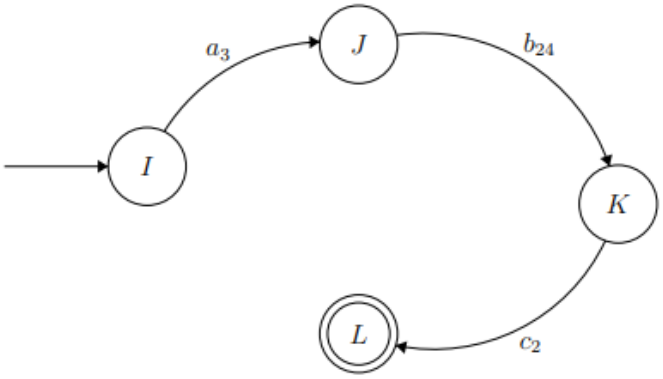


Figure 3: An example of a sample JFA

Figure 3 shows an example of a JFA with three transitions. The four states as shown in the diagram are $\{I, J, K, L\}$, with I representing the start state and L denoting the accepting state. Ia_3 begins the movement with the *ukuthola* keyword, then Jb_{24} shows the next input label with the *yonke* keyword, Kc_2 ends with the *customer* keyword. It is interesting to note that these examples only cater for a single relation. We may have situations of many states and transitions, taking into account the input statement provided by the learner.

3.2 *Normalising Queries*

At the normalisation phase, irrelevant details were removed and keywords that appears in our JFA design were used to semantically form a query. The normalisation process is passed to the SQL query generator, which is then used to create and retrieve records from the Chinook DB. As illustrated in Table 2, each local language word is mapped to a query operation. This stage is useful for the SQL generation phase.

Table 2: Mapping Local Language Keywords

Keywords	Query Operation
Ukuthola, Thola, Ngtholele, Fumana, Ngitholela, Vind, Kies	SELECT
Al, Alle, Alles, Konke, Yonke, Yothe, Vhothe	ALL
Skep, Usika, Dala	CREATE
Faka, Plaas, Ulonga	INSERT
Susa, Lees, Verywyder, Utomola, Ubvisa Vhothe	DELETE

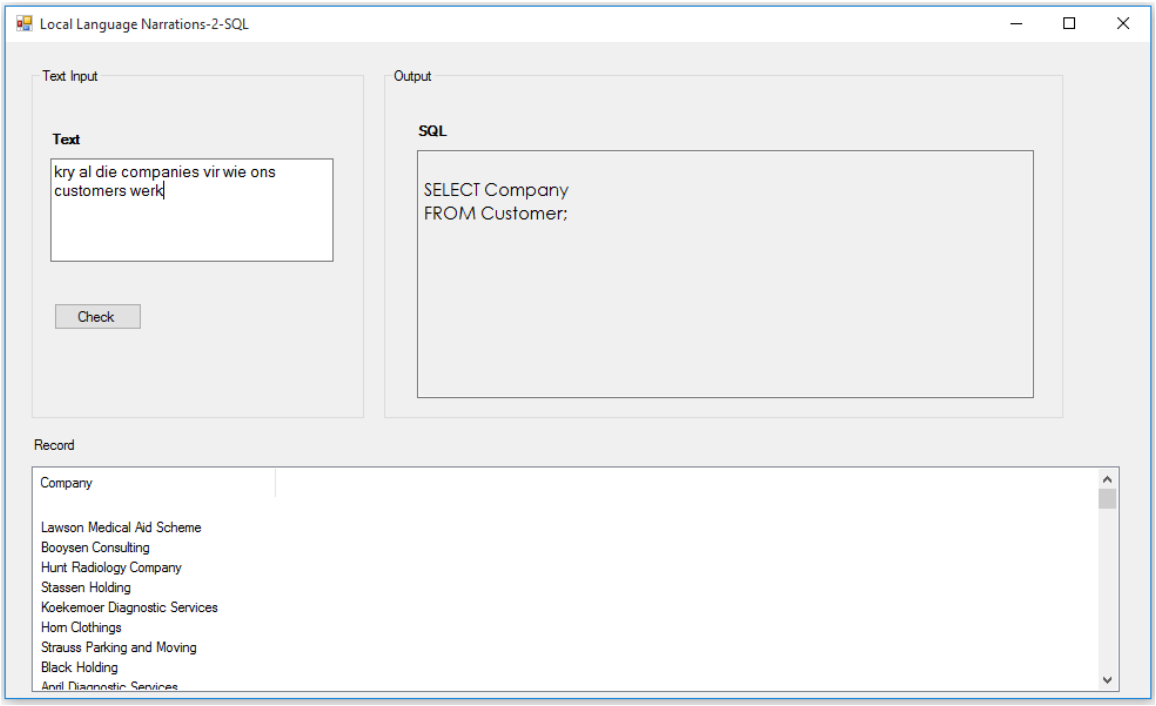


Figure 4: A sample output of `Local-Nar-SQL`

3.3 Generating SQL Queries

This phase takes information from the normalisation phase to semantically generate a query. The generated query is used against the Chinook DB, which then produces an output. Since this task is limited to a single relation, the JFA algorithm was useful for this purpose. Similarly, the task was quite straightforward.

4 Implementation and Results

In Section 3, JFA technique was described which showed how a query is generated. Here, the implementation and results of the local language narration process are presented. The JFA approach was implemented using the Microsoft C# .NET framework[1], which was developed into a software tool, called `Local-Nar-SQL`.

`Local-Nar-SQL` was created to allow a native learner to specify a query request in any of the eleven local languages spoken in SA. The software tool uses the learner's input to generate a query and retrieve records from the sample database. Further, the result is made available to the learner. The sample database used for our use-case was the ChinookDB[2], which contains multiple records. A sample output of this request is provided in Figure 4.

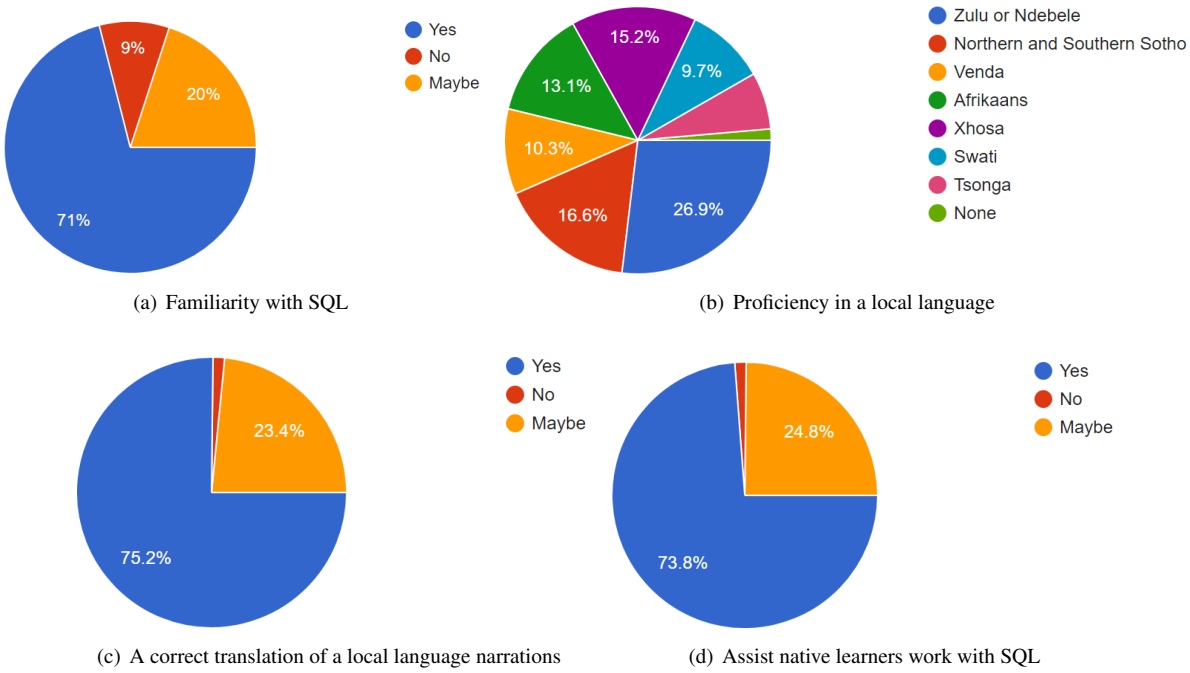
5 Evaluation

A total of 145 participants participated in a survey that was conducted online. The majority of the participants are native language learners who are familiar with programming and have taken numerous computer science courses. The questions for the survey is available via `https://forms.gle/uSXn8TZpBEWmnGk66`. The results are presented in Figure 5.

Of the responses received, 71% indicated familiarity with SQL, whilst 9% of respondents claimed no familiarity with SQL and 20% were unsure of their answers (Figure 5(a)). The participants were asked about their native language proficiencies. Approximately 29% of learners were proficient in Zulu, 16.6% were proficient in Northern and Southern Sotho, 15.2% claimed to be proficient in Xhosa and 13.1% admitted to being proficient in Afrikaans. The remaining respondents were proficient in Venda (10.3%), Swati (9.7%) or Tsonga (6.9%), and 1.4% of the participants did not list any of these native languages (Figure 5(b)). Furthermore, the participants were asked if the `Local-Nar-SQL` tool provided a correct translation of local language narrative to SQL: 75.2% affirmed that the tool was accurate, 23.4% were unsure and 1.4% indicated no (Figure 5(c)). When the participants were asked if they think that the tool will assist native learners work with SQL, about 73.8% affirmed that the tool would assist learners, 24.8% were unsure and 1.4% indicated no (Figure 5(d)).

The survey showed that the majority of the learners found the `Local-Nar-SQL` tool to be useful and helpful towards their interest of learning SQL. If our proposed tool is used as a learning aid, native learners will find it easier to work with SQL, which will eliminate many of the barriers faced with English proficiencies in programming pedagogies.

[1] https://docs.microsoft.com/en-us/dotnet/framework/
[2] https://github.com/lerocha/chinook-database



(a) Familiarity with SQL

(b) Proficiency in a local language

(c) A correct translation of a local language narrations

(d) Assist native learners work with SQL

Figure 5: The survey results

6 Conclusion and Future Work

This paper presented an extended version of a JFA for the automatic synthesis of SQL queries from local language narrations. This automata-based algorithm was further implemented into a software tool called `Local-Nar-SQL`, designed for native learners to specify query request in any of the eleven official South African languages of their choice. `Local-Nar-SQL` takes these requests, generates a query and produces an output that is presented to the learner. If implemented on a large scale, this tool will serve as:

1. a comprehension aid: With `Local-Nar-SQL`'s textual and translation approaches, it can improve the cognitive workload of a native learner towards understanding the SQL concept.
2. a learning aid: The *free-form* approach offered by the tool will enable native learners improve and enhance their SQL query skills.
3. a solution to English language barriers: Native learners are not required to attain a high English proficiency to use `Local-Nar-SQL`. Since local languages are widely spoken in South Africa, it is seen that this tool will stimulate a native learner's interests in becoming proficient in SQL.

To this end, our evaluation results showed that the majority of learners agreed that this tool will be helpful to native learners yearning to be proficient in SQL. Our evaluation results are consistent with the works of [38] where majority of the native participants agreed that the developed tool helped serve their information needs . In the future, we aim to improve this study to accommodate nested SQL query and other query operation tasks. Also, it would be interesting to conduct a performance evaluation on our proposed tool. Such information would provide us with the accuracy of `Local-Nar-SQL`.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work is based on research supported by the National Research Foundation (NRF) of South Africa (Grant Number: 119041). Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and therefore the NRF does not accept liability in regard thereto.

References

[1] R. D. Fulk, "English as a Germanic language," Companion to the history of the English language, 1–8, 2008.

[2] D. Borodenko, "Formation of the English Language," Science progress in European countries: New concepts and modern solutions, 304 – 427, 2019.

[3] F. Ndhlovu, L. Siziba, "English in Southern Africa," in The Social and Political History of Southern Africa's Languages, 65–92, Springer, 2018.

[4] S. Passera, A. Kankaanranta, L. Louhiala-Salminen, "Diagrams in contracts: Fostering understanding in global business communication," IEEE Transactions on Professional Communication, **60**(2), 118–146, 2017, doi: DOI:10.1109/TPC.2017.2656678.

[5] O. Inbar-Lourie, "English only? The linguistic choices of teachers of young EFL learners," International Journal of Bilingualism, **14**(3), 351–367, 2010, doi:https://doi.org/10.1177%2F1367006910367849.

[6] P. J. Guo, "Non-native English speakers learning computer programming: Barriers, desires, and design opportunities," in Proceedings of the 2018 CHI conference on human factors in computing systems, 1–14, 2018.

[7] A. G. Soosai Raj, E. Zhang, S. Mukherjee, J. Williams, R. Halverson, J. M. Patel, "Effect of Native Language on Student Learning and Classroom Interaction in an Operating Systems Course," in Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education, 499–505, 2019.

[8] A. Nguyen, P. Rigby, T. Nguyen, D. Palani, M. Karanfil, T. Nguyen, "Statistical translation of English texts to API code templates," in 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME), 194–205, IEEE, 2018.

[9] A. Barua, S. W. Thomas, A. E. Hassan, "What are developers talking about? an analysis of topics and trends in Stack Overflow," Empirical Software Engineering, **19**(3), 619–654, 2014.

[10] K. Reestman, B. Dorn, "Native Language's Effect on Java Compiler Errors," in Proceedings of the 2019 ACM conference on international computing education research, 249–257, 2019.

[11] G. Obaido, A. Ade-Ibijola, H. Vadapalli, "Synthesis of SQL Queries from Narrations," in 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), 195–201, IEEE, 2019.

[12] A. Ade-Ibijola, G. Obaido, "S-NAR: generating narrations of SQL queries using regular expressions," in Proceedings of the South African Institute of Computer Scientists and Information Technologists, 1–8, 2017.

[13] G. Obaido, A. Ade-Ibijola, H. Vadapalli, "Generating SQL queries from Visual Specifications," in Annual Conference of the Southern African Computer Lecturers' Association, 315–330, Springer, 2018.

[14] G. Obaido, A. Ade-Ibijola and H. Vadapalli, "Generating narrations of nested SQL queries using context-free grammars," in 2019 Conference on Information Communications Technology and Society (ICTAS), 1–6, IEEE, 2019.

[15] H. Brookes, "Youth Language in South Africa: The Role of English in South African Tsotsitaals," English in multilingual South Africa: The linguistics of contact and change, 176–195, 2020.

[16] F. Banda, "The dilemma of the mother tongue: Prospects for bilingual education in South Africa," Language culture and curriculum, **13**(1), 51–66, 2000, doi:https://doi.org/10.1080/07908310008666589.

[17] S. Coetzee-Van Rooy, "Being English in multilingual South Africa," World Englishes, 2020.

[18] Y. Pal, S. Iyer, "Classroom versus screencast for native language learners: Effect of medium of instruction on knowledge of programming," in Proceedings of the 2015 ACM conference on innovation and technology in Computer Science Education, 290–295, 2015.

[19] A. K. Veerasamy, A. Shillabeer, "Teaching English based programming courses to English language learners/non-native speakers of English," International Proceedings of Economics Development and Research, **70**, 1–6, 2014.

[20] A. Migler, A. Dekhtyar, "Mapping the SQL Learning Process in Introductory Database Courses," in Proceedings of the 51st ACM Technical Symposium on Computer Science Education, 619–625, 2020.

[21] S. Prabhu, S. Jaidka, "SQL and PL-SQL: Analysing teaching methods," in The Proceedings of the Computing and Information Technology Research and Education New Zealand (CITRENZ), 105–109, 2019.

[22] S. Ardeleanu, "SQL: Beauty and the Beast," in Relational Database Programming, 17–30, Springer, 2016.

[23] J. Heller, "Use SQL More Often with Basic Dynamic SQL," in Pro Oracle SQL Development, 339–353, Springer, 2019.

[24] J. de Beer, N. Petersen, "Decolonisation of the science curriculum: a different perspective (# cookbook-labs-must-fall)," 2016.

[25] V. Mudaly, "Decolonising the mind: Mathematics teachers explore possibilities for indigenising the school curriculum," Journal of Education (University of KwaZulu-Natal), **74**, 67–84, 2018, doi:http://dx.doi.org/10.17159/2520-9868/i74a05.

[26] L. Le Grange, "Decolonising the university curriculum: Leading article," South African Journal of Higher Education, **30**(2), 1–12, 2016.

[27] M. Llopis, A. Ferrández, "How to make a natural language interface to query databases accessible to everyone: An example," Computer Standards & Interfaces, **35**(5), 470–481, 2013.

[28] J. Barzdins, M. Grasmanis, E. Rencis, A. Sostaks, A. Steinsbekk, "Towards a More Effective Hospital: Helping Health Professionals to Learn from their Own Practice by Developing an Easy to use Clinical Processes Querying Language," Procedia Computer Science, **100**, 498–506, 2016.

[29] A. Ade-Ibijola, S. Ewert, I. Sanders, "Introducing Code Adviser: A DFA-driven electronic programming tutor," in International Conference on Implementation and Application of Automata, 307–312, Springer, 2015.

[30] A. O. Ade-Ibijola, Automatic novice program comprehension for semantic bug detection, Ph.D. thesis, 2016.

[31] M. Konecki, "Problems in programming education and means of their improvement," DAAAM international scientific book, **2014**, 459–470, 2014.

[32] A. Ade-Ibijola, S. Ewert, I. Sanders, "Abstracting and narrating novice programs using regular expressions," in Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology, 19–28, 2014.

[33] A. G. Soosai Raj, K. Ketsuriyonk, J. M. Patel, R. Halverson, "Does Native Language Play a Role in Learning a Programming Language?" in Proceedings of the 49th ACM technical symposium on computer science education, 417–422, 2018.

[34] K. Varadarajan, M. L. Chu, "Declarative programming model with a native programming language," 2020, US Patent 10,585,653.

[35] A. Afonso, L. Brito, O. Vale, "An evolutionary method for natural language to SQL translation," in Asia-Pacific Conference on Simulated Evolution and Learning, 432–441, Springer, 2008.

[36] R. Kumar, M. Dua, "Translating controlled natural language query into SQL query using pattern matching technique," in International Conference for Convergence for Technology, 1–5, IEEE, 2014.

[37] R. Reddy, N. Reddy, S. Bandyopadhyay, "Dialogue based question answering system in Telugu," in Proceedings of the Workshop on Multilingual Question Answering-MLQA, 1–8, 2006.

[38] L. Siciliani, P. Basile, G. Semeraro, M. Mennitti, "An Italian Question Answering System for Structured Data based on Controlled Natural Languages." in CLiC-it, 1–6, 2019.

[39] X. Meng, S. Wang, "Nchiql: The chinese natural language interface to databases," in International Conference on Database and Expert Systems Applications, 145–154, Springer, 2001.

[40] B. Al-Johar, A portable natural language interface from Arabic to SQL, Ph.D. thesis, University of Sheffield, 1999.

[41] H. Bais, M. Machkour, L. Koutti, "An independent-domain natural language interface for relational database: Case Arabic language," in 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), 1–7, IEEE, 2016.

[42] N. E. Okwunma, "Automatic comprehension of customer queries for feedback generation," 2018, MSc thesis.

[43] S. Obare, A. Ade-Ibijola, G. Okeyo, "Jumping Finite Automata for Tweet Comprehension," in 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), 1–7, IEEE, 2019.

[44] G. Obaido, A. Ade-Ibijola, H. Vadapalli, "TalkSQL: A Tool for the synthesis of SQL Queries from Verbal Specifications," in International Multidisciplinary Information Technology and Engineering Conference (IMITEC) Vanderbijlpark, South Africa, 1–9, 2019.

[45] A. Meduna, P. Zemek, "Jumping finite automata," International Journal of Foundations of Computer Science, **23**(07), 1555–1578, 2012.

[46] B. Schultheiss, A. Vollebregt, C. Hummelink, "Storing heterogeneous helicopter signal data: Advantages of using an XML database," in 2007 IEEE Autotestcon, 469–475, IEEE, 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1190-1206 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Development of a Technology and Digital Transformation Adoption Framework of the Postal Industry in Southern Africa: From Critical Literature Review to a Theoretical Framework

Kgabo Mokgohloa*, Grace Kanakana-Katumba, Rendani Maladzhi

Department of Mechanical and Industrial Engineering, University of South Africa, Johannesburg, 0002, South Africa

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 23 November, 2020
Online: 14 December, 2020

Keywords:
Technology adoption
Systems approach
Digital transformation

ABSTRACT

The pressing and most urgent challenge for most of the Posts in Southern Africa is to transit from its historical reputation of being a snail-paced, inefficient, loss-making and ineffective service provider to an agile, innovative, solution-driven and highly competitive service provider that strive for excellence. In this context, adoption, and diffusion of technology and digital transformation by Posts should deliver the necessary traction to steer the digital transformation journey of the postal industry in Southern Africa with the necessary velocity while cognizant and mindful of the appropriateness of the technology in the African context in the light of a VUCA (Volatile, Uncertain, Complex and Ambiguous) world. The traditional technology adoption models are characterized by "linearity" which is the opposite of a dynamic setting characterized by "causality" which is the fundamental principle of a system thinking approach. The multi-stakeholders in the postal sector with often competing interests renders a linear approach defunct and requires a system approach to digital transformation and technology adoption. The conceptual framework developed from the critical review of literature offers considerable leverage to postal sector in developing countries and to Southern Africa in particular. The proposed conceptual framework integrates dimensions such as Postal Industry 4.0 Envisioning, Strategy, Institutional factors, Organizational factors, Individual factors, Industry 4.0 environment and outcomes.

1. Introduction

This paper is an extension of "Postal Development: Literature Review into Adoption Models" presented at IEEM 2019 Conference in Macau by authors in [1]. The rise of digitalization powered by a digital revolution of the 21st century ushered by the Fourth Industrial Revolution and personified by Industry 4.0, has prompted the postal sector across the world to expand their services well beyond the original service of postal operators (POs) which was delivery of physical mail in the form of letters and post cards.

In [2], the authors propose that although there are Posts in different nations that are struggling to keep afloat due to economic and political turbulence there are other Posts that are trailblazers in postal digital transformation which is underpinned by digital innovation which leads to digital disruption. It comes as no surprise that Swiss Post, The Netherland Post, Deutsche Post, La Poste, and Japan Post are trailblazers in digital innovation.

In [3], the authors postulates that societal anatomy is rapidly changing, and that the digital age shepherded changes in consumption patterns which has resulted in the evolution of the customer of the future with unique needs and expectations that industries should meet. A systems (holistic) approach is required to comprehend a new world that is driven by volatility, uncertainty, complexity, and ambiguity (VUCA). VUCA world is categorised by the traits of volatility, uncertainty, complexity and ambiguity [4]. In [5], the authors argue that Volatility, Uncertainty, Complexity and Ambiguity (VUCA) explains the competitive ecosystem of the digital economy in which corporations ought to acclimate business process to match up briskly shifting and more complex dynamic settings.

In [6], the authors propose that in the current Volatile, Uncertain, Complex and Ambiguous (VUCA) world, organizations and employees are faced with a plethora of challenges that overly dynamic and totally different from the challenges of the past. In [6], the authors further argue that there are several of drivers that propel these changes, and that

*Corresponding Author: Kgabo Mokgohloa, mokgok@unisa.ac.za

technology, accessibility to information, a fast shifting global economy, the establishment of a global consumer society and a very insistent customer with unique and diverse requirements are only some influences that have shaped an extremely complex and ambiguous world.

In [7], the authors advance that organizational and societal systems in the digital era encompass characteristics of volatility, uncertainty, complexity, and ambiguity; and as a result a systems approach to digital transformation and technology adoption is a key success factor for Posts in Southern Africa.

1.1. Problem statement

The crises that has befell on the Designated Postal Operators (DPOs) in Southern Africa has precipitated the near collapse of the postal logistics sector due poor performance on the four performance indicators of the 2IPD and consequently losses except for a few Designated Postal Operators making a profit, this is characterized by the scores respective Post in Southern Africa obtained on the 2IPD.

The underperformance by SADC DPOs is fuelled by a plethora of dynamics including inflexible or rather outdated business models that are not adapting to the technological changes that are disrupting current business models immensely, the archaic business models cannot compete in the 21st century and this result in majority of Postal Operators in Southern Africa posting losses year in year out. The unsustainability of the DPOs leave the State in respective countries with no option but to bail out these institutions. This performance is contrary to the first-tier countries who occupy pole positions on the integrated index on postal development (2IDP).

1.2. Research question

Can a technology and digital adoption framework in the context of postal industry 4.0 be developed for Posts in Southern Africa?

1.3. Concluding remarks to introduction

The paper is divided into (a) Introduction which introduces the phenomenon under study as well as articulating the problem statement and the research question which guided the research study. (b) Literature review which delved into the postal landscape, articulation of the Integrated Index on Postal Development (2IDP) which is a global measure of performance of the postal sector globally, and with specific reference to the top 5 performers on the 2IDP and digital readiness index. The literature review further delved into the phenomenon of industry 4.0, technology adoption, digital transformation, reductionist vs systems approach and capability maturity models. The report then detailed the research methods adopted, the discussion and findings that included the development of the theoretical framework, deductions and upcoming research concludes the study.

2. Literature review

2.1. The Postal landscape

In [2], the authors articulates an international view point of the Postal Industry, accentuating that the postal network encompasses a worldwide grid of over 677 000 retail offices, 5.3 million members and substantial infrastructure comprising 192 nations.

In [8], the authors hypothesizes that imperfect internet connectivity, inadequate investment and/or recapitalization of the network infrastructure, deprived value of service, restricted interoperability and connectivity among the systems of value chain participants, as well as uncertainty in postal industry characterisation in southern Africa are the foremost motives for the snail-paced advancement.

In [9], the authors argues that the regional scenery of the postal industry in southern Africa encompass 50 mail centres and a distribution infrastructure of 16, 064.394 deliverance points. Of these, South Africa has a lion share of 26 mail centres and 14, 106.896 delivery points. The postal industry in Southern Africa comprise of fourteen designated postal operators in South Africa, Botswana, Zambia, Eswatini, Seychelles, Democratic Republic of Congo, Malawi, Comoros, Madagascar, Mauritius, Tanzania, Lesotho, Mozambique, and Angola.

2.2. The Integrated Index on Postal Development

In [3], the authors note that most Posts are performing disappointingly on the Integrated Index for Postal Development (2IPD). In [3], the authors further proposes that measuring the multiple facets of postal development is no insignificant task, and hypothesize that in order to surmount this difficulty; the Universal Postal Union (UPU) has been tapping on abundance of postal big data to appraise the performance of Posts globally. One of the most important sequels of these endeavours was the development of the Integrated Index for Postal Development commonly known as the 2IPD. In [10], the author argues that 2IPD gauges the performance of Posts in the four vital aspects of postal advancement which are known as the 4Rs (reliability, reach, relevance and resilience).

In [11], the authors outlines the four aspects of postal advancement (development) as follows: (a) Reliability is a fusion of excellence of service delivery, including certainty across all categories of postal delivery service; it eventually gauges the level of postal operational efficiency (b) Reach is fusion of worldwide postal connectedness at intercontinental level across all categories of international postal delivery services, it ultimately gauges the level of globalization of the postal services (c) Relevance encompass the strength of demand for the full range of postal services in each postal section; it eventually gauges the level of attractiveness of the postal market (d) Resilience encompass ability to innovate, deliver wide-ranging postal services and assimilate sustainable development targets in postal business models; it ultimately gauges the level of adaptableness of postal business models.

In [3], the authors advances that a total of 172 countries were analysed for the development of the 2019 global ranking (based on complete data for 2018). Table 1 depicts the top 5 countries in alphabetical order with their corresponding scores on the 2IDP and are as follows: France, Germany, Japan, Netherlands, and Switzerland. In [12], the authors postulates that these top 5 DPO's are performing outstandingly on the 2IDP because they have steadily build dependable, well-connected, appropriate and agile postal services. The report further highlights that other DPOs

especially in developing countries are underperforming relative to the 2IDP indicators.

Table.1: Top 5 Designated Postal Operators performance on Integrated Index for Postal Development, adapted from [12]

Country	2IPD Score 2018	2IPD Score 2019
France	83.3	86.86
Germany	91.3	90.79
Japan	91.6	87.19
Netherlands	93.7	93.67
Switzerland	100	100

The analysis of performance on the 2IPD points to Posts in developing countries (Southern Africa included) as laggards' when compared to the top performing DPOs on the 2IPD. Table 2 lists the SADC (Southern African Development Community) Designated Postal Operators with their corresponding 2IPD scores.

Table.2: Southern African Development Community Designated Postal Operators performance on Integrated Index for Postal Development, adapted from [12]

Country	2IPD Scores 2018	2IPD Scores 2019
Angola	21.22	30.54
Botswana	23.72	21.00
Comoros	11.29	11.76
Democratic Republic of Congo	15.87	17.01
Eswatini	24.72	15.80
Lesotho	14.09	13.44
Madagascar	30.96	30.63
Malawi	24.78	13.74
Mauritius	49.17	40.53
Mozambique	4.95	6.28
Namibia	31.28	27.52
Seychelles	30.68	24.00
South Africa	33.34	33.34
Tanzania (Republic)	31.81	39.12
Zambia	7.73	8.75
Zimbabwe	19.53	12.74

In [13], the authors explains that the industrial sector has always been critical to the economic development of countries. Since the end of the 18th century, businesses have gone through enormous modifications that revolutionized a way how goods are produced and has yielded numerous benefits, predominantly related to efficiencies and productivity. In [14], the authors argues that presently, after three preceding industrial revolutions, the combination of advanced technologies and the internet age is completely transfiguring the industrial landscape and it is labelled the Fourth Industrial Revolution or its personification known as the Industry 4.0. The postal sector is not immune to these fast-paced technological revolutions that continue to chance the postal landscape daily.

The rise of digitalization brought upon the digital age has prompted the postal industry globally to enlarge their services further than the traditional mail delivery service of Posts. In [15], the authors hypothesise that with the global adoption of digitalization and digitization at an exponential rate, stakeholders are progressively demanding to interact directly with the postal service via digital networks.

2.3. Industry 4.0

In [13], the authors elucidates that the industrial sector has always been critical to the economic development of countries. Since the end of the 18th century, businesses have gone through enormous modifications that revolutionized a way how goods are produced and has yielded numerous benefits, predominantly related to efficiencies and productivity. In [14], the authors argue that today, after three preceding industrial revolutions, the combination of advanced technologies and the internet age is completely transfiguring the industrial landscape.

In [16], the authors argue that the aim of Industry 4.0 is to digitize and incorporate processes vertically across the entire value chain of the organisation, from conceptualization all the way to sales and after-sales support. Data throughout the value chain is available instantaneously, backed by augmented reality and optimized in a unified network Industry 4.0 is a farsighted initiative of the German State that traditionally supports advancement of the German industrial sector. In [17], the authors support this view and propose that in this sense, Industry 4.0 can be viewed as an action towards nurturing Germany's posture as one of the most leading technologically advanced nation.

In [18], the authors postulate that Industry 4.0 is envisioned at the fixated assemblage and application of prompt data and information by means of linking all individual rudiments of a system, with a view of reducing the intricacy of processes, while raising the efficacy and value of operations.

Figure 1 presents the Industry 4.0 framework and associated technologies at which the core of Industry 4.0 is based on the concept of the three underlying fundamentals which are:

- Digitalization and amalgamation of vertical and horizontal value chains
- Digitalization of offerings
- Digital industry models and client access

In [16], the author notes that the underpinnings to industry 4.0 could be elaborated as follows:

- Digitalization and amalgamation of horizontal & vertical value chains comprise of two features; The initial feature comprises amalgamation of processes vertically across the entire firm and entails all business processes from conceptual

design all the way to delivery of the product or service to the customer or end-user.

- Digitalization of offerings incorporates the extension of current offerings which includes the addition of smart sensors that could be utilized with data analysis tools as well as development of new digitized offerings. The ability to analyze big data could benefit firms to better respond to changing customer demands.
- Digital industry models and client access encompasses enlargement of offering by supporting disruptive digital solutions. Disruptive digital industry models have potential to create new revenue streams by responding with speed to the changing customer demand changes in a dynamic setting.

In [16], the authors argue that big data and data analysis tools are the principal competences of Industry 4.0 and is enabled by technologies such as smart sensors, advanced algorithms, Internet of Things, location technologies, and many other technologies that support the digital revolution.

Figure 1 presents the Industry 4.0 setting in which an argument is advanced that the crucial significance to Industry 4.0 is its ability to communicate with other smart "things" such as the digital capabilities of smart transportation, smart electrical grid, smart supply chain management, and other smart infrastructure. In [18], the authors argue that the linkage between social web and business web is of critical importance and will see further radical transformation in the future and will further reduce the digital divide between business and society at large as technology develops and becomes accessible to society.

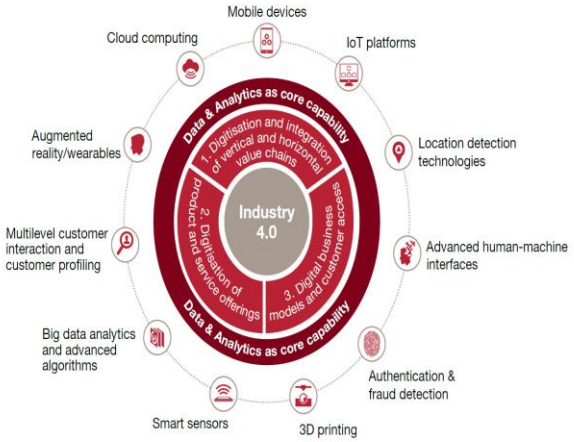


Figure 1: Industry 4.0 framework and contributing digital technologies [16]

In [19], the authors suggests that the important elements that outline the notion of Industry 4.0 are:

- Cyber-physical system (CPS), the term that describes the amalgamation of cyber with physical systems. CPS incorporates devices which can assemble and transmit data through the internet.
- Internet of things (IoT) facilitates the interaction with other systems and between systems and ultimately with the users.
- The ability to assemble data and analyze the data into useful information instantaneously is an important feature of a CPS and enables decision makers to make informed decisions.

In [20], the authors advocates that the concept of Industry 4.0 is underpinned by six design principles that are pertinent to its efficacious implementation in respective industries which are interoperability, virtualization, decentralization, real-time capability, service orientation and modularity depicted in Figure 2.

	Cyber-Physical Systems	Internet of Things	Internet of Services	Smart Factory
Interoperability	X	X	X	X
Virtualization	X	-	-	X
Decentralization	X	-	-	X
Real-Time Capability	-	-	-	X
Service Orientation	-	-	X	-
Modularity	-	-	X	-

Figure 2: Design principle for each of the main 4 Industry 4.0 components [20]

The Fourth Industrial Revolution and its application as espoused by Industry 4.0 offers both challenges and opportunities for the postal sector, the major challenge is that Posts can no longer operate using business models of the past because digitalization of the Post is something that is not optional but mandatory if the Post still must survive another 100 years.

In [21], the authors argue that today, the concept of the internet is often reflected as "given" due to its omnipresent manifestation and quickening effect on the ways in which humanity relates. The world is changing rapidly because of the digital revolution. In [22], the authors argue that one such manifestation of the digital revolution is the IOT (Internet of things) with which digitization is no longer a simulated reality. IOT links several "things" through the internet and allows them to communicate information amongst each other. In [23], the authors propose that the Internet of Things (IoT) is sensor technologies that enable physical objects to gather and transmit data through the Internet in instantaneously. It is one of the latest advances taking hold in many industries and most appropriate for the postal sector due to its possible applications in the postal value chain.

In [24], the authors propose that the IoT architype is envisioned at devising a sophisticated system with the combination sensors and associated technologies with ability to communicate in real-time. Contrarywise, assembling data and preserving the privacy of users, and then taking cognizance of confidentiality and internet security specification in IoT remains a significant concern amongst stakeholders in the IoT space.

In [23], the authors suggests that the current rise of IoT applications is fuelled by a conjunction of issues such as omnipresent connectivity and the diminishing expenditure and enhanced performance of measuring devices (sensors) and analytics. These technology developments relate to expanding consumer requirement for more data on the products and services they purchase. The flow of information is crucial to the "flow of things".

The United States Postal Service has embodied the Internet of Things (IoT) and coined the Internet of Postal Things (IoPT). In [23], the authors articulate a postal standpoint and reason that IoT can be better pronounced as "Things on the Internet" which are

unified webs of distinctively recognized physical "things" embedded with sensors that amass, converse, and act upon a widespread diversity of data such as locality pin-pointing, temperature, and movement. These "smart objects" amass and transmit data that can be scrutinized for improvement of business processes across organizations and for improved solutions to customer demands. Figure 3 depicts the chief elements of the IoT value-chain.

In [23], the authors argue that IoT unwraps new opportunities for various organizations and in particular entities like the postal service which has a geographic reach that surpass many industries. In [25], in the authors propose that the postal service world-wide has a widespread network of over 677,000 post offices, 5.32 million staff and physical network of buildings encompassing 192 countries and asserts that the postal sector performs a significant socio-economic role in society at large.

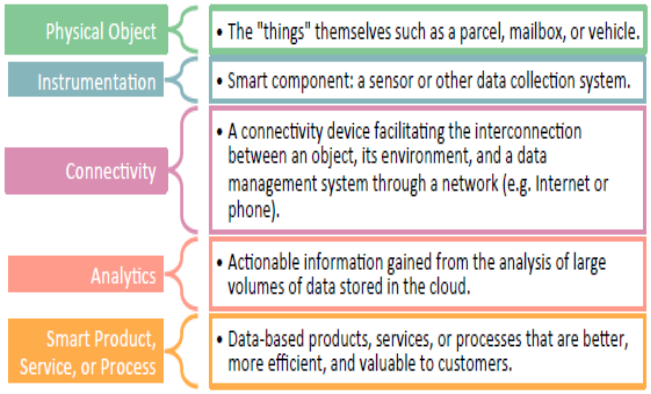


Figure 3: IoT value chain [23]

Posts could significantly profit in implementing Industry 4.0 fundamentals in the quest to adapt to challenges and opportunities presented by the digital era, and could greatly improve service delivery to new market demands and improve their sustainability (both economic, social and environmental). Applications of Industry 4.0 fundamentals can significantly improve sustainability of Posts in Southern Africa.

2.4. Innovation Diffusion Theory

In [26], the authors postulates that the most noticeable and dominant attempt to dissect the dynamics influencing the adoption and diffusion practices was pioneered by [27], however the authors argue that the framework is appropriate for conceptualization of adoption and diffusion from a qualitative point of view as opposed to quantitative prediction of adoption of innovative practices. In [28], the author argues that diffusion is the process by which an invention is conversed through networks over time amongst the participants of a network. In [29], the authors argue that technology adoption is a developed area of research in information systems (IS).

In [30], the authors articulates the difference between innovation and technology which is often confused and argues advance that technology is concerned in resolving real social problems to advance humanity. In [31], the authors argue that innovation on the other hand comprises of the creation of an innovative concepts and development of the concept into new offerings resulting in economic development. In [32], the authors argue that innovation is not merely creativity, research and development, invention, bright ideas or fancy gadgets. In [33], the authors argue that innovation is rather a development of new resources or new resource capacity that are able to create wealth. In [34], the authors argue that innovation is a complex, organization-wide effort, that demands a set of procedures and activities to shape, coordinate, and promote it.

In [28], the authors investigates drivers that define the speed of acceptance (adoption) of innovations and argues that the following factors regulate the degree of adoption. (a) Apparent characteristic of innovation affected by considerations such as relativity of the advantage to adopt, compatibility of adoption to current systems, complexity of the innovation, trialability of the innovation, and observability of the results. (b) Innovation choice, which is driven by issues such as is the innovation optional or mandatory (c) Type of societal structures including a component such as the degree of linkages of the societal structures; and (d) The effectiveness and efficiency of transformation (change) champions in advocating transformational change.

In [11], the authors suggests that each invention has distinctive attributes that affect the dispersion (diffusion) process. These attributes can be broken-down into five elements:

(a) Relative advantage is the extent to which an invention is understood as being superior to the concept it succeeds, typically stated as cost-effectiveness, societal stature, or other advantages.
(b) Compatibility as recognized by participants in society is constructively related to the level of acceptance (adoption) and influences the way customers act towards an invention.
(c) Complexity is the extent to which an invention is understood as comparatively complicated to comprehend and utilize.
(d) Trialability is the extent to which an invention may be tested on a restricted basis. The trialability of an invention is certainly related to the pace of acceptance (adoption).
(e) Observability is the extent to which outcomes are observable to the user and to others.

2.5. Technology Adoption Theory

In [35], the authors argues that Theory of Reasoned Action (TRA) is the early notion that attempted to explain user adoption of technology. TRA describes user behaviour from a social point, and with a specific aim of pin-pointing the causes of conscious behaviour. In [36], the authors propose that the model (Technology Acceptance Model- TAM) was developed from TRA as its basis but further refined.

In [37], the authors argues that the basis of TRA is that behaviour is predicted by intention (I) and further highlights that intentions are conjointly driven by two factors (a) Attitude (A) towards the act which is a function of attitudes about effects of this conduct and (b) subjective norms (SN) defined as a person's viewpoint of whether stakeholders key to the person consider the actions should be implemented. Attitude towards the behaviour is defined as the individual's negative or positive outlook about executing the behaviour.

In [38], the authors propose that TRA contends that views influence mindsets, which lead to intent and as a result engender behaviour. The TRA is graphically illustrated in Figure 4.

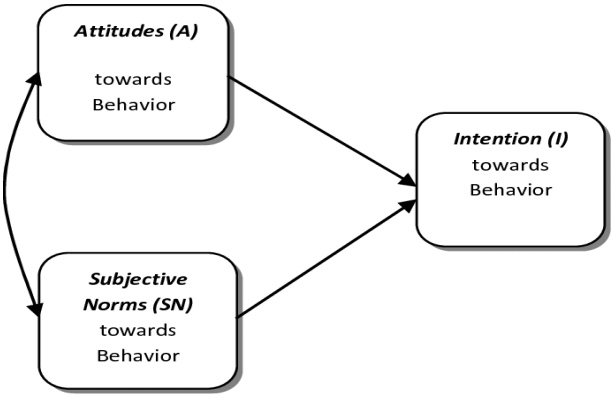


Figure 4: Theory of reasonable action. Source [37].

In [39], the authors state that consumer rationale to accept and adopt a technology could be explained by perceived usefulness (PU), perceived ease of use (PEU) and attitude towards the system. In [39], the authors hypothesize that attitude of a user towards the system was the core factor to determine if the user will accept or reject the innovation, and that attitude is directly influenced by (a) Perceived Usefulness (PU) and (b) Perceived Ease of Use (PEU); and concludes that apparent ease of use unswervingly influence perceived usefulness.

In [38], in the authors argue the concepts in the earliest TAM (see Figure 5) as follows: Perceived Usefulness (PU), Perceived Ease Of Use (PEOU), Attitude, and Behavioural Intention to Use. Among the constructs, PU and PEOU establish a user's confidence on a technology and consequently predict his or her stance on the technology, which in turn foresees its approval.

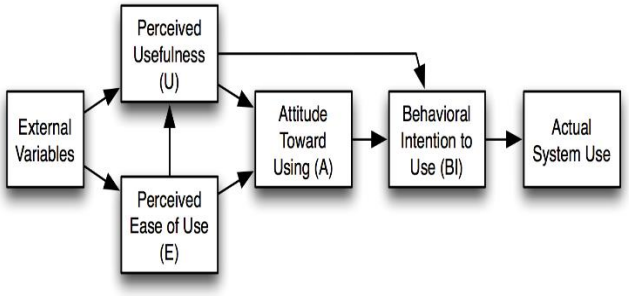


Figure 5: Technology Adoption Model. Source [38]

In [40], the authors maintains that while the TRA was intended to illuminate essentially any human behaviour, the TAM on the other hand was intended to offer an clarification of the factors of computer acceptance throughout a comprehensive range of end-user computing technologies and user populations. In [37], the authors subsequently contends that according to the TAM, actual system usage is directly defined by behavioural intention (BI) to use. BI is in turn swayed by the user's mindset towards utilizing the system and perceived usefulness. Attitude is propelled by perceived usefulness (PU) and perceived ease of use (PEOU) of the system. Both PU and PEOU are influenced by external variables, such as individual variations, situational constrictions, organizational traits, and system attributes.

As the TAM progressed past its initial positioning, several researchers established shortcomings in the model which required to be tackled. The outcome of research that was conducted by several researchers ensued in adjustments to TAM. In [37], the authors suggest that the adjustments entailed the establishment of additional variables and concepts resulting in TAM2 as depicted in Figure 6.

In [37], the authors argue that the major additions to the TAM that were introduced by TAM2 were the inclusion of societal influence activities (subjective norm, voluntariness, and image) and mental method procedures (job relevance, output quality and results demonstrability).

TAM2 theorizes that users' mental assessment of the match between considerable goals at work and the subsequent effects of executing tasks using the system provides a basis for establishing views concerning the usefulness of the system. The results of the research by uncovered that TAM 2 performed well in both voluntary and obligatory environments.

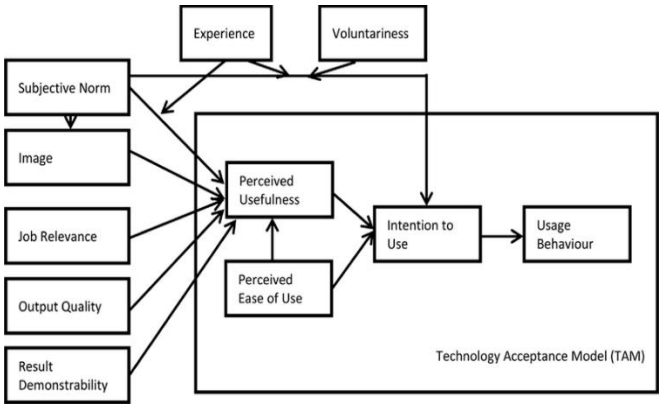


Figure 6: Technology Adoption Model 2. Source [42].

In [37], the authors argues that the succeeding foremost addition to TAM was the Unified Theory of Acceptance and Use of Technology (UTAUT), the author further proposes UTAUT was framed, with four fundamental determining factor of intention and usage and up to four moderators of crucial relations. The theory proposed that four important concepts perform a significant part as direct factors of user acceptance and usage behaviour: (a) performance expectancy (b) effort expectancy (c) social influence and (d) facilitating conditions. The UTAUT is presented in Figure 7.

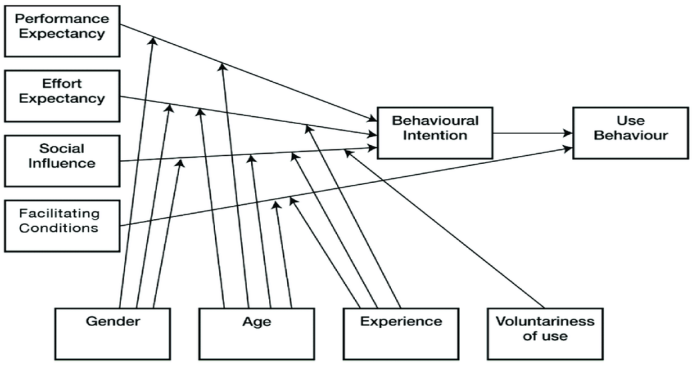


Figure 7: Unified Theory of Acceptance and Use of Technology (UTAUT. Source [43].

In [37], the authors propose that UTAUT advocates that facilitating conditions would unswervingly impact usage, while the remaining three concepts would influence Behavioural Intention (BI). In [44] the authors argue that the model describes facilitating conditions as the extent to which a person deems that organizational and technical infrastructure exists to support the utilization of a system they are accepting (adopting). In [45], the authors argue that the UTAUT further extended the concepts in the earliest TAM through the recognition of the vital role of endeavour in performing the new task and the role support structures determines adoption in organizations.

In [46], the authors argues that TOE (Technology-Organization-Environment) framework entails three categories of settings that may influence technological innovation adoption and implementation process. TOE is premised on the three context which are technological, organizational, and environmental. In [47], as represented in Figure 8, the authors explains that technological setting defines the internal and external technologies suitable to the organization, while organizational setting denotes descriptive actions about the organizational scope, size and structure; finally, environmental setting is the arena in which an organization conducts its business.

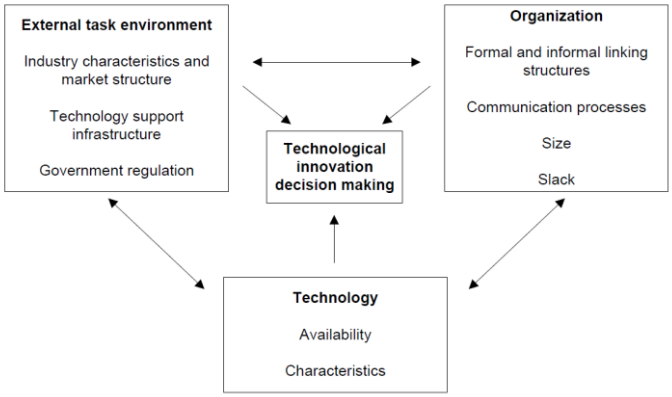


Figure 8: Technology, organization, and environment framework

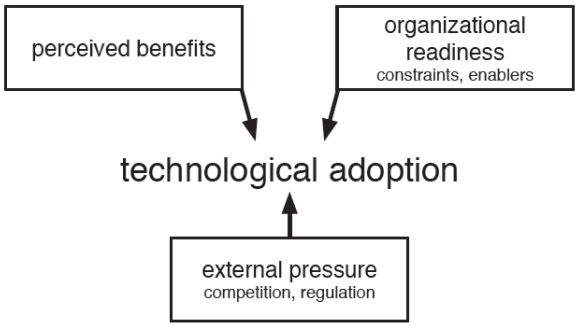


Figure 9: BOE Model [48]

In [48], the authors postulates that the TOE model defines the dynamics for technological adoption, focusing on the external stresses (e.g. market forces and governmental regulatory requirements), organizational structures, and technological availability; in addition the BOE model was originally developed to understand the adoption of EDI (electronic data interchange) technology, but has since been used as a general technology adoption model. The BOE model comprises three factors: external pressure, organizational readiness, and perceived benefits.

In [48], the authors advance that by distinction, the BOE model as illustrated in Figure 9 (Benefits, Organizational Readiness and External Pressures) integrates the organization and technology context of TOE into organizational readiness and augments a perceived benefit aspect for viewing at adoption of technology.

In [26], the authors argue that the Adoption and Diffusion Outcome Prediction Tool (ADOPT) was developed as a tool to encourage application in project planning and implementation of sound insight of the social and economic factors affecting acceptance (adoption) of agricultural innovations. In [49], the authors proposes that ADOPT focuses on factors that have demonstrated to impact the pace and/or highest point of adoption within a population. These are (a) attributes of the innovation, (b) attributes of the population, (c) tangible relative advantage of employing the innovation, and (d) learning of the actual relative advantage of the technology. In [26] the authors proposes that the variables of the theoretical context can be divided into two main categories: (a) those that relate to traits of the target population, and (b) those that relate to traits of the practice. These elements are presented in Figure 10.

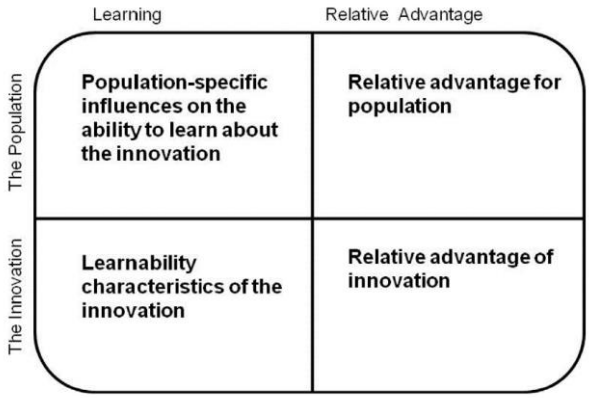


Figure 10: Conceptual characterization of factors included in ADOPT

2.6. Capability Maturity Models

In [50], the authors states that the original Capability Maturity Model (CMM) was developed in the software engineering space to support advances in the dependability of software development organizations, that is, in their aptitude to deliver software within time, performance, cost and quality. More specifically, it was "intended to assist developers to choose process-improvement strategies by defining their existing process maturity and finding the most critical issues to improving their software quality and process".

In [51], the authors describes the five maturity levels and further emphasize the basic process changes made at each level:

- Initial: The process is described as ad hoc, and sporadically even anarchic. Limited processes are defined, and accomplishment depends on individual effort.
- Repeatable: Fundamental business processes are established to track expenditure, plans, and functionality.

- Defined: The business processes are documented, standardized, and integrated.
- Managed: Detailed measures of the business processes are collected. Business processes are quantitatively understood and controlled.
- Optimizing: Continuous process improvement is enabled by quantifiable feedback from the business process.

The CMM provides a framework for shaping these incremental steps into five maturity levels that lie down consecutive grounds for continuous process improvement. These five maturity levels characterize an ordinal scale for evaluating the maturity of an organization's processes and for appraising its process capability. The levels also help an organization highlight its improvement endeavours. The steps above are graphically presented in Figure 11.

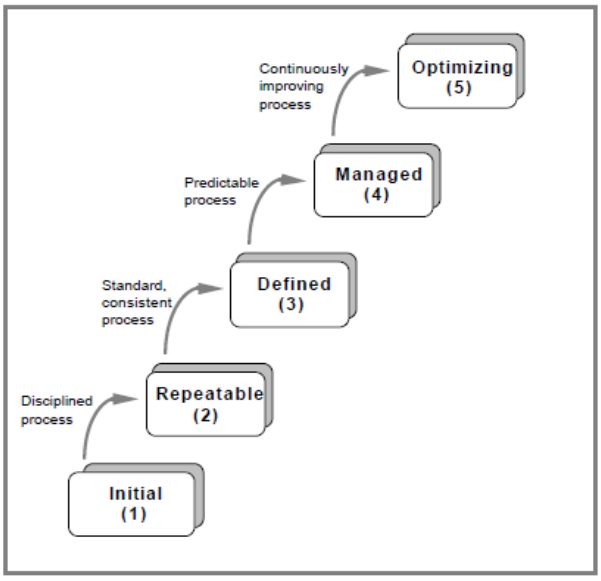


Figure 11: The five levels of software maturity [51]

In [52], the authors argue that capability maturity models are used in three ways; (a) Descriptive use which focuses on (1) To assess the 'as-is' state of specific organizational capabilities (2) As an analytical diagnostic tool (3) To communicate maturity levels to internal and external stakeholders (b) Prescriptive use which focus on (1) To ascertain desired maturity level and receive guidance on how to continuously improve from current to desired level (2) To follow specific and detailed sequences of action (3) Comparative use which emphasizes on comparing the performance of the specific organisation through internal or external benchmarking (best practices).

In [53], the authors affirms that one of the barriers to industry's wider headway in digital transformation is the absence of a vivid industry focused blueprint. The report proposes that the Digital Maturity Model presented in Figure 12 is a valuable instrument to offer guidelines for a vivid pathway throughout the transformation journey.

The five core dimensions underpinning the model are:

- Customer, this facet is focused on offering an understanding where consumers perceive the organization as their digital collaborator utilizing their desired networks of interaction to control their connected future on and offline
- Strategy, this facet stresses on how organizations transform or operate to enhance its competitive advantage through digital projects; it is entrenched within the whole organizational strategy
- Technology, this facet emphasizes the success of digital strategy by enabling the creation, processing, storage, securing and transmission of data to meet the needs of customers effectively and efficiently.
- Operations, this facet is premised on executing and evolving processes and tasks by leveraging on digital technologies to drive strategic management and enhance business efficiency and effectiveness
- Organization and culture, this facet focuses on identifying and cultivating an organizational culture with governance and talent management processes to boost advancement along the digital maturity curve, and the ability to realize expansion and invention (innovation) goals.

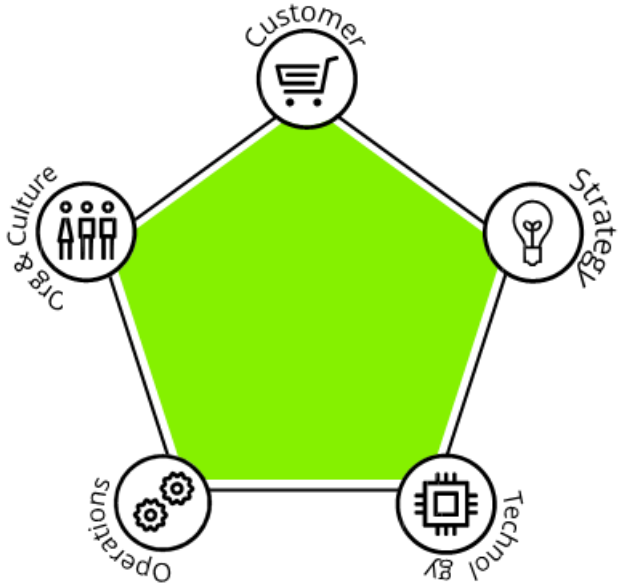


Figure 12: Deloitte Digital Maturity Model [53]

Figure 13 crystalizes the sub-dimensions that are associated with the five core dimensions presented in Figure 12. The sub-dimensions elucidate and unpacks each of the five core dimensions that are critical to digital transformation, these dynamics are such as customer engagement, customer insight and behaviour, brand management, finance & investment, connected things, data & analytics, real-time insights & analytics, agile change management, organizational culture, leadership & governance, organizational design & talent management, workforce enablement, automated resource management, stakeholder management, ecosystem management, strategic management, customer trust & perception, market & customer, network, security technology architecture and innovation management are key towards a successful digital transformation journey.

In [52], the authors propose a Digital Maturity Model which consists of four stages. The stages proposed in their model are (a) Initiate (b) Competent (c) Purposeful and lastly (d)

Transformative. The key attributes of businesses related to each of these maturity stages are presented in Figure 14.

Customer	Strategy	Technology	Operations	Organisation & Culture
Customer Engagement	Brand Management	Applications	Agile Change Management	Culture
Customer Experience	Ecosystem Management	Connected Things	Automated Resource Management	Leadership and Governance
Customer Insights & Behaviour	Finance & Investment	Data & Analytics	Integrated Service Management	Organizational Design & Talent Management
Customer Trust & Perception	Market & Customer	Delivery Governance	Real-time Insights & Analytics	Workforce Enablement
	Portfolio Ideation & Innovation	Network	Smart and Adaptive Process Management	
	Stakeholder Management	Security	Standards & Governance Automation	
	Strategic Management	Technology Architecture		

Figure 13: The 5 core dimensions divided into 28 sub-dimensions [53]

In [54], the authors suggests that the use of maturity models to the Industry 4.0 phenomenon could potentially aid corporations to assimilate this practice into their ethos and argue that the Industry 4.0 vision portrays a completely new approach to business process and a completely new business model. They propose a three-step systematic process to guide organizations in their respective Industry 4.0 journey to find footing and determine their vision which is precursor for strategy.

In [54], the authors proposes that the three-stage process comprise of (a) Envision 4.0 which relates to defining a tailor-made Industry 4.0 vision and it requires companies to develop their comprehension of the generic Industry 4.0 concepts with company-specific proficiencies and resources (b) Enable 4.0 starts with the significant Industry 4.0 vision developed in the previous phase, and based on this vision, the organization defines the digital portfolio and proficiencies required to provide support to the new products or services identified. (c) The enact phase is characterized by a well-timed, directed and multi viewpoint plan of the overall strategy towards the Industry 4.0 vision that fosters the strategic setting for tangible measures and outcomes.

In [54], the authors further argue that in order to employ maturity models to execute the three-staged process described earlier in the context of Industry 4.0, the maturity gauges (scales) for the three-stage model are described as follows: (a) Level 1 represents "initial" level where there is no industry 4.0 vision in place, (b) Level 2 represents "managed" level whereby a Roadmap of industry 4.0 strategy is in place, (c) Level 3 represents "defined" level whereby customer sections (segments), value intention (proposition), and significant (key) resources and competencies are clearly defined; (d) Level 4 which represents "transform" level whereby strategic imperatives are transformed into tangible projects, and lastly (e) Level 5 which represents "Detailed Business Model" level whereby the business model is completely transmuted in line with Industry 4.0 objectives.

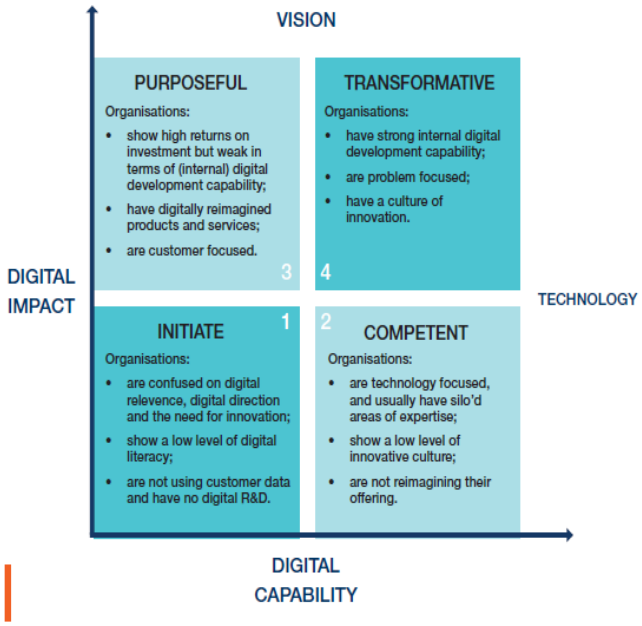


Figure 14: Stages of Digital Maturity and Key Attributes of Businesses in Each Maturity Levels [52]

2.7. Barriers towards movement to digital Post

In [15], the authors articulates the top five barriers (obstacles) towards movement into digital Posts by Region and for purposes of this study only the top performers (Industrialized countries) and Africa will be considered. These barriers are tabled in Table 3.

Table 3: Barriers for digital adoption and diffusion in the postal sector [15]

Geographic area	**Barriers**
Globally	Resource constraints
	Transition towards a digital culture
	Limitations of IT infrastructure
	Lack of adequate inhouse expertise required for development of digital services
	On-time custom clearance capability and capacity
Industrialized countries	The time taken for the corporation to shift in the direction of a digital culture
	Slow overall consumer embracing of digital postal services
Africa	Limited resources
	Lack of IT infrastructure
	Poor digital culture
	Lack the internal resources to develop digital services
	Slow overall consumer embracing of digital postal services

Clearly, Africa is faced with a plethora of challenges as depicted in Table 3, and Southern Africa as a Region in Africa suffers from the same challenges. In [8], the authors hypothesises that inadequate internet connectivity, inadequate capital outlay and

or recapitalization of the network infrastructure, meagre excellence of service, inadequate interoperability and connectivity among the business value chain and vagueness in postal sector definition in Southern Africa are proposed as foremost motives for the low advancement of the Region in embracing technologies associated with the digital revolution.

In [15], the authors affirms that the best ranking designated operator when it comes to digital readiness achieves a normalized score of 1, while the lowest possible performer obtains a normalized minimum score of 0. All normalized scores can be understood as the gap of any given designated operator with respect to the best (score 1), the midway (score above 0.5) or the worst (zero score) global performer. Table 4 represents the postal digital readiness of the industrialized countries with Switzerland leading with a full score of 1 while Table 5 portrays Southern Africa as a region and South Africa leads with a score of 0.46 which is just below the intermediate score of 0.5 while the worst performing countries are Madagascar and Democratic Republic of Congo which are both at 0.05. The distance between the worst performer and the best performer stands at a mammoth gap of 0.95. Clearly Southern Africa is faced with enormous challenges to realize a digital Post.

Table 4: Digital readiness score of Industrialized countries [15]

Country	Normalized score 2019
France	0.81
Germany	0.81
Japan	-
Netherlands	0.62
Switzerland	1.00

Table 5: Digital readiness score of Southern Africa countries [15]

Country	Normalized score 2019
Angola	-
Botswana	0.32
Comoros	0.14
Democratic Republic of Congo	0.05
Eswatini	-
Lesotho	0.19
Madagascar	0.05
Malawi	0.08
Mauritius	-
Mozambique	0.11
Namibia	0.27
Seychelles	-
South Africa	0.46
Tanzania (Republic)	0.22
Zambia	-
Zimbabwe	0.14

The literature has pointed out that adoption of technology of organizations in general or digitalization (Adoption of digital technologies) of the Post is a complex non-linear phenomenon with a variety of drivers and barriers. Adoption and diffusion barriers and drivers reveal a systemic character, which demands an all-inclusive approach to managing the complex nature of adoption and diffusion barriers and drivers that involve exploration of interactions between adoption and diffusion barriers and drivers as well as management of the casual relationship between the drivers and barriers of adoption and diffusion of technology/digitalization.

2.8. From reductionist to system thinking approach

In [55], the authors explains that reductionism is understood as the conventional means of tackling problem solving and is widespread throughout numerous fields. Reductionism also goes under many names including 'stepwise enhancement', 'disaggregation' and merely 'breaking problem into sub-problems'. Reductionism inclines to refer to comprehension instead of problem-solving but the latter seems more appropriate in this case. In [56], the authors argue that reductionism as the opposite of holism and systems thinking agrees with the view that all objects or systems are breakable to lower levels in the hierarchy of their composition.

In [57], the authors contends that in accordance with reductionist thinking "The whole can be broken down into its parts and put back together from its parts. Parts are related through a simple cause-effect relationship; Thus, its defining characteristics exist in its parts." However, there are limitations to a reductionist thinking approach because systems most often defy the reductionist analysis because by focusing on the system's parts, researchers miss the whole, which frequently took on a form that was not distinguishable from the reconstructed parts. This view can be interpreted that it assumes that for a system to perform effectively and efficiently, its components (parts) should perform effectively and efficiently.

Conversely, systems are complex and as result the interaction of the components (parts) in a dynamic setting invariably affect performance of the entire (whole) system. Therefore, effective, and efficient technology adoption must focus on the dynamic setting (interaction) of drivers and barriers rather than on their separate actions. In [58], the authors argue that system is a unified set of components that are logically organized in a way that achieves a particular objective. In [59], the authors proposes that a system is a sequence of intermingling components organised to achieve stated goals.

In [60], the authors suggests that there are key dissimilarities in behaviours of reductionist thinkers when contrasted with systems thinkers and the crux of the differences is that while a reductionist approach simplifies the problem to its smallest essence to enable resolution of the problem through analytical methods; a system thinking approach on the other hand seek to understand the bigger picture and how elements of the problem interact in a dynamic setting. Because organizations are complex due to many interacting dynamics and forces, it is probable to propose that postal sector in Southern Africa lags in adoption of innovative technologies due to its reductionist thinking approach rather than a system thinking approach.

In [61], the authors propose that an event-oriented viewpoint is logical, action-oriented, irresistibly uncomplicated and frequently narrow-minded. Figure 15 portrays this thinking as a concept. In [62], the authors further argues that this event-oriented mind-set

signifies a conviction that challenges are irregular, emanating from overwhelming events in the outside world and that life is unpredictable. He argues that this mind-set has a conviction that events pop out from nowhere or at least there is no time to worry about their causes, what is important he argues, is to fix the problem as soon as possible. The typical thinking is characterized by linearity whereby problems are events and solutions as fixes.

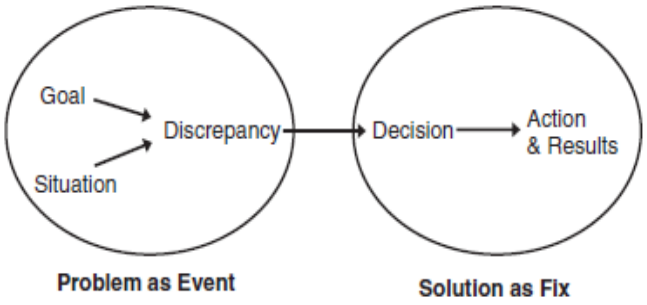


Figure 15: Event oriented worldview [62]

In [62], the author contends that feedback systems thinking is paradigm shift, a new approach of understanding the world, and a remedy to narrow thinking; and further advances that problems and solutions are viewed as intertwined as depicted in Figure 16.

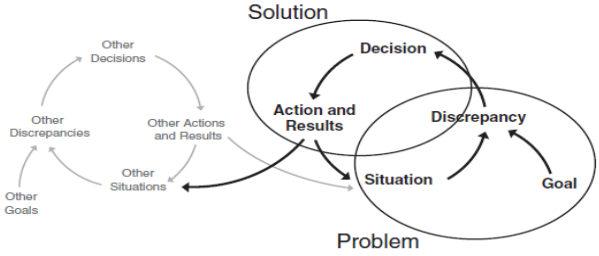


Figure 16: A shift of mind: A feedback perspective [62]

In [61], the authors contends that the change is the ultimate constant, and that change confronts conventional traditions, methods, and values. Important, most of the changes we battle to grasp occur as after-effects, deliberate, and not deliberate. In many cases, well-intentioned exertions to resolve unrelenting problems often lead to policy resistance, where policies are overdue, weakened, or thrashed by the unanticipated reactions of other people or of nature. This is a frequent challenge in the postal sector globally, in Africa and for purposes of this study in Southern Africa where for instance the digital age has brought unintentional consequences of the archaic postal business models that are putting in peril the very survival of postal operators.

In [63], the authors affirm that a system thinking approach is different from a reductionist approach, the reductionist approach focuses on analysis which is about dissection of complexity into manageable components or an event-oriented approach which has the conviction that problems emanate from nowhere. This approach clearly fits into the linear and mechanistic worldview, where the world and problems are fragmented into parts. Diametrically opposed to a reductionist approach of "analysis" is systems thinking approach which relies on synthesis rather than analysis. Fusion (synthesis) relates to comprehending the whole and the fragments (parts) simultaneously, alongside the relationships and links that constitutes the dynamic setting of the whole.

In [57], the authors supports this notion and argues that the whole emanates because of the interfaces between its components or fragments and proposes that fragments are related through intricate multiple effects. Thus, its defining attributes do not live in its fragments but in the whole. In [60] the authors suggest that systems thinking means the aptitude to distinguish and analyse the inter-connections within and between systems.

In [64], the authors contend that system thinking is a strong approach to deal with the system of innovation. In [64], the authors defines systems thinking as a field for observing aggregates and a context for spotting interrelatedness rather than things, for spotting patterns of change instead of stagnant snapshots. In [65], the authors define system thinking as the capability to epitomize and measure dynamic intricacy (e.g., behaviour that develops from the interface of a system's agents over a period), of systems.

In [66], the authors maintain that systems thinking is a set of harmonious diagnostic competencies used to enhance the ability to discover and comprehend systems, forecasting their conducts, and developing adjustments to them to create the anticipated effects. These competencies act as a system. In [63], the authors suggest that systems thinking is a process of examining and improving mental models and its core is based on four fundamental concepts (a) connectedness (b) Synthesis (c) Feedback loops, and lastly (d) Causality

In [60], the authors suggest that Systems Thinking requires recognition that systems designed by humans, recurrent occurrences or relations originate from systemic constructs which emanate from mental models. In [60], authors further propose that the Iceberg Model (Depicted in Figure 17), is a fundamental aspect of systems thinking. The Iceberg Model contends that actions (events) and patterns (which can be observed) are triggered by universal (systemic) structures and mental models, which are usually not observable.

In [60], the authors further propose that systemic constructs as in structures are the structural pyramid; societal order; interrelatedness; instructions and techniques; establishments and endorsement levels; process flows and methods; enticements (incentive programme), reward (compensation), aims, and system of measurement (metrics); manners (attitudes); responses and the incentives and qualms that cause them; organizational culture; feedbacks and interruptions (delays) in a dynamic setting; and fundamental forces that occur in organizations. Behaviours develop from these complex interactions, which exist because of mental models.

In [67], the authors argue that system dynamics, a subset of systems thinking signifies variations in a complicated system, and is beneficial for mimicking the stock-flow-feedback processes that usually trigger the behaviour of intricate systems. In [68] the authors argues that the modelling approach depend on the following crucial elements. "Stocks" exemplify the rise of something; "flows" exemplify the rate of flow into or out of a stock.

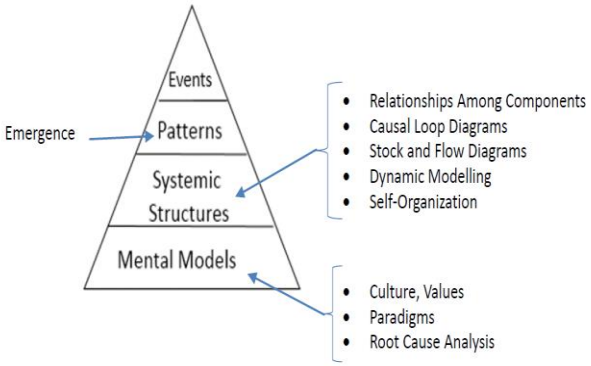


Figure 17: Integrated Model of Systems Thinking [60]

Stocks and flows, and variables characterises a system of linked non-linear differential equations, which can be modelled over a period to ascertain how stocks shift over time. In [69] the authors argues that one benefit of application of system dynamics approach over more conventional models is the addition of feedback consequences. Feedbacks ensure that a model's output is precise and uncovers intricacy that could be unobserved in other modelling approaches. Therefore, decision-makers can more definitely distinguish the interrelatedness of the model in a dynamic environment.

In [69], the authors argue that learning is a response (feedback) process. In [67], the author argues that as decisions are made these action modify the real world, and information feedback about the real world is received; and applying new insights, interpretation of the world is adjusted and becomes nearer to the goals.

This observation is principally valuable in the context of postal sector in Southern Africa which is overwhelmed by multiple stakeholders with multiple interests including policy directions internationally (Universal Postal Union), continentally (Pan African Postal Union), regionally (Southern African Postal Operators Association as implementing Agency of Southern Africa Development Community) and lastly nationally where individual postal operators operate. In this research the enablers and impediments that play in the adoption and diffusion of technology are identified including feedback loops that associate the several factors involved in the adoption and diffusion of technology process.

3. Methodology

In [70], the authors advances that research design can be classified in three categories which are (a) Exploratory research studies; (b) Descriptive and diagnostic research studies, and (c) Hypothesis-testing research studies. This study adopted the qualitative research design approach which is exploratory in nature.

In [71], the authors suggest that the idea behind qualitative research is to tenaciously hand-pick partakers or locations (or records or graphic materials) that enables the researchers to comprehend the conundrum (problem) and the research question. This fundamental feature of qualitative research design resulted in the in collection of specific secondary data that provides insight to the researcher to comprehend the phenomenon under study.

In [72], the author argues there are diverse forms of review methodologies; systematic, semi-systematic and integrative approaches and argues that each approach is appropriate depending on the purpose of analysis. This study adopted an integrative approach to literature reviews. An integrative approach is characterized by tenets which are articulated in Table 6.

Table 6: Tenets of an integrative approach to literature reviews

Tenet	Integrative Approach
Typical rationale	Critique and integrate
Research questions	Narrow or broad
Exploration approach	Usually not methodical
Sample attributes	Peer reviewed articles
Evaluation	Qualitative
Examples of contribution	Classification (Theoretical models or frameworks)

The data was collected from archival data from literature and industry reports, this data focused on a variety of drivers and barriers common in adoption of technology in organizations as well as characteristics of factors that drive or impede technology adoption in organizations and as well as in the postal sector.

This exploratory, qualitative research design using integrative approach was aimed at answering at great length the first research question which is, "Can a generic framework for technology and digital transformation adoption be developed for Posts in Southern Africa?

In [71], the author avows that "The data collection steps include setting the boundaries for the study, collecting information through unstructured or semi structured observations and interviews, documents, and visual materials, as well as establishing the protocol for recording information". In [71], the author further argues that the objective of qualitative research is to tenaciously choose members or sites (or documents or visual material) that will assist the investigator to fully grasp the challenge and the research question. The study purposefully selected relevant literature from Universal Postal Union, the different postal operators, specific literature that deals with technology adoption, ICT adoption and digital transformation with the view of comprehending the problem and the associated research question.

4. Discussion and findings

The literature review accentuates the significance of not only dealing with barriers and drivers to adoption of technology and Industry 4.0 technologies but highlights the fact that these barriers and drivers are not linear but intermingles with each other in a dynamic setting. Literature divulges that adoption and technology diffusion processes are very complex and entails multiple interdependent drivers and barriers which are very dynamic involving numerous feedback processes that demonstrate nonlinearity characteristics.

Literature details the postal sector landscape, and reflects on the digital panorama globally, continentally and in the region

(Southern Africa) and points to serious challenges faced by developing and less developed countries from a digital ecosystem point of view. These challenges are reflected in several barriers that need to be overcome and drivers that needs to be leveraged on in a dynamic setting.

Literature points to vast variances in digital adoption readiness between developed countries and developing countries, with specific reference to Southern Africa as depicted in Table 4 and Table 5 respectively and graphically depicted Figure 18 and Figure 19, respectively.

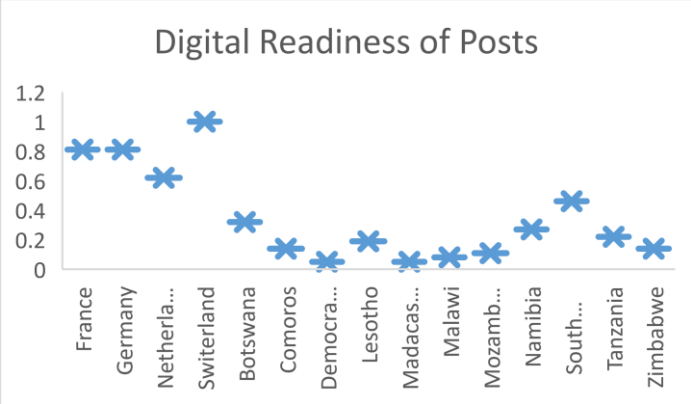


Figure 18: Digital readiness score of Posts

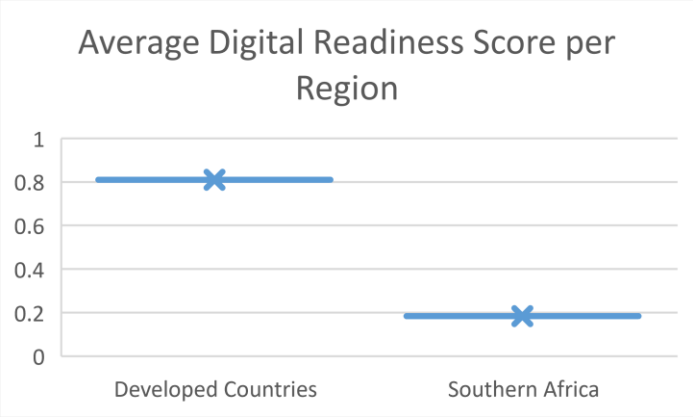


Figure 19: Average Digital readiness score per Region

The digital readiness score illustrates that Switzerland scores 100% and this is further supported by the performance of Switzerland on the Integrated Index on Postal Development in which it scores 100%. Switzerland is followed by both Germany and France scoring 81% and The Netherlands scoring 62%. The average of the four leading countries stands at 81%.

Southern Africa ranges from maximum of 46% which represent the performance of South Africa to a minimum of 5%, the variation within Southern Africa cumulatively results in an average of 18%. The performance of Southern Africa should be seen in the context of a plethora of challenges facing the continent and the Southern Africa in the context of this study. These challenges require a concerted effort by Posts and respective national governments and regional bodies to address the legacy of colonialism but also the problems that occurred post-independence that continue to halt prospects of development.

Literature also points to Technology Acceptance Models TAM, TAM2, TRA, UTAUT, TOE, BOE, and ADOPT are among the most popular which highlight factors related to system users; however, there is a literature knowledge gap in these models handling other stakeholders who are directly or indirectly affected by the adoption process and the interaction of drivers and barriers in a dynamic setting. These technology acceptance models lack the important aspect of feedback systems which are at the core of systems thinking approach.

TRA, TAM, TAM2, and UTAUT are significant pioneers in the technology adoption space by elucidating the aspects that perform a prominent role in the process of technology adoption. The evolution from TRA to UTAUT is exemplified in Figure 20.

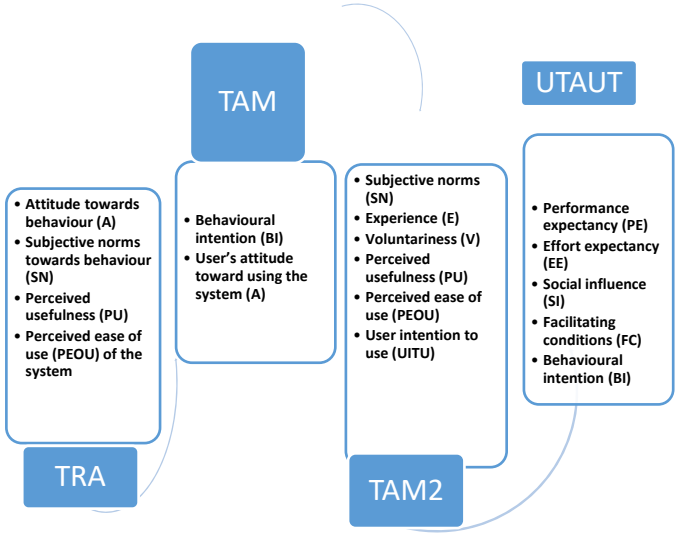


Figure 20: Evolution of technology adoption models

The factors elucidated in Figure 20 directly determine the user adoptability potential of technology, however they are applicable to individuals rather than organizations. The BOE and TOE attempts to address the limitations of TRA, TAM, TAM2 and UTAUT which are only applicable to individual's adoption of technology through articulation of constructs that are applicable to organization-wide adoption of technology. The constructs of BOE (Perceived Benefits, Organization Readiness and External Pressures) are depicted in Figure 21 and TOE constructs (Technology, Organization and External Environment) are depicted in Figure 22.

These factors associated with both TOE and BOE are an attempt to integrate organization-wide factors and external factors that are at play in the context of organization-wide adoption of technology adoption as opposed to the traditional technology adoption models in the form of TRA, TAM, TAM2 and UTAUT which are relevant to individual adoption of technology or new innovation. While TOE and BOE attempts to integrate factors associated with organization-wide adoption of technology and new innovation, they both lacked "causality" which is at play in a dynamic setting resulting in new emergence of patterns and mental models due to interacting factors that inhibit or enable adoption

and diffusion of technology in a dynamic setting which result in even greater complexity. BOE and TOE are therefore not sufficient to explain nor explore the complex nature of interacting factors in the dynamic setting of the postal sector in Southern Africa.

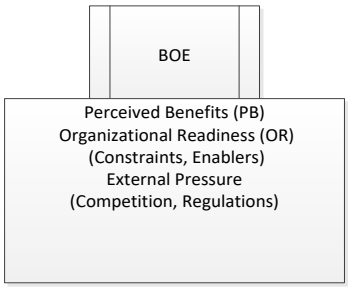


Figure 21: Constructs of the BOE model

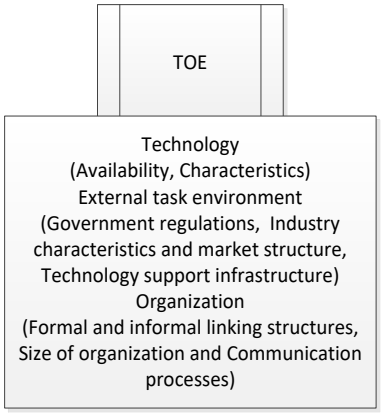


Figure 22: Constructs of the TOE model

The purpose of this research is to explore how best to close the limitations of the "linear" technology adoption models through the application of a holistic or systems thinking approach to identify systemic issues in technology adoption within Southern African context by considering the complete stakeholder system a "system of stakeholders" with their corresponding interests giving rise to various dynamic settings that could impede or enable technology adoption and digital transformation in the postal sector in Southern Africa.

The System Thinking approach which inherently integrates feedback systems is suitable for complex phenomenon like adoption of technology and digital transformation in the postal sector which is underpinned by multiple stakeholders with often opposing interests. A systems approach in comparison to traditional "linear" technology adoption models (TRA, TAM, TAM 2, UTUAT, TOE, BOE and ADOPT) was found to be the most appropriate and suitable for managing the interaction of barriers and drivers that are at play in a dynamic setting in the context of the postal sector in Southern Africa.

The literature suggests that Industry 4.0 is centred on digital transformation of business models. At its core is digitalization and integration of vertical and horizontal value chains, digitalization of offerings (products and services) and digitalization of business models and customer access. Of which, data and data analytics are core competencies interacting with digital technologies such as wearables and augmented reality, IoT platforms, mobile devices, big data and advanced algorithms, location detection technologies, 3D printing and other related Industry 4.0 technologies.

The critical review of literature explicates that the foundation to digital transformation is a systematic approach that is directed from a digital vision envisaged by the organization, the model proposed in [52] where the authors argues that the transformation requires organizations to move from level 1 (Initiate), to level 2 (Competent) to level 3 (Purposeful) and finally to level 4 (Transformative). The levels start from chaotic and no ambition or culture for innovation (level 1) to structured, systematic and innovative (level 4).

In [54], the authors articulates the three-stage model of envisioning of Industry 4.0, enablement of Industry 4.0 and lastly enacting of Industry 4.0. The three-staged model proposes 5 levels of digital maturity ranging from the lowest level (Level 1) in which the Industry 4.0 vision is not defined to the highest level (Level 5) in which there is complete business model transformation. The insights and inferences that could be drawn that without a solid vision and an underpinning strategy which acts as a roadmap for organizations with demonstrable projects; the adoption of technological innovation and digital transformation is set to fail.

In [53], the authors proposes the five dimensions that underpins digital transformation journey and are (a) Customers, (b) Strategy, (c) Technology, (d) Operations and (e) Organization and culture are key towards digital transformation journey that is structured to ensure success.

The proposed technology and digital adoption conceptual framework as presented in Figure 23 espouses the insights drawn from the critical review of selected literature.

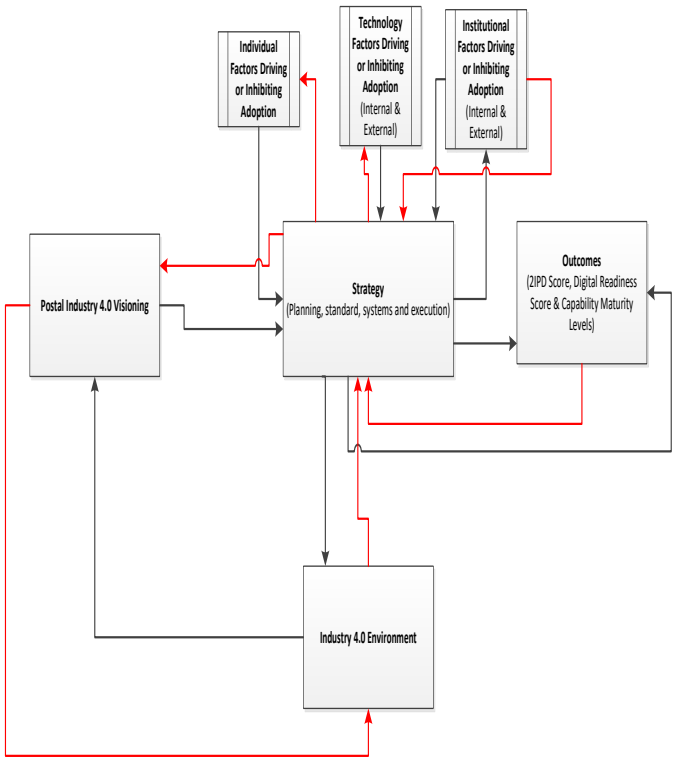


Figure 23: Proposed postal industry 4.0 adoption framework

The framework proposed commences with the visioning dimension which is a key critical success factor in adoption of digital technology and digital transformation. It sets the agenda for digital transformation of the Postal Sector in Southern Africa and without a vision for postal industry 4.0, the digital transformation process is bound to fail. The visioning dimension is supported by the strategy which integrates aspects of planning, systems relating to Industry 4.0, standards relating to Industry 4.0 and execution of strategic imperatives into concrete programmes and projects.

This dimension of strategy interacts with factors that inhibit or enable digital technology adoption and technology adoption broadly. These factors that either inhibit or enable adoption of digital technology and technology broadly are technological, institutional, and individual focused. These factors should be factored in during the strategy development phase and take cognizance of Industry 4.0 standards and systems prevailing. The strategy dimension also feedforwards to the Industry 4.0 environment which is meant to facilitate the core capabilities of Industry 4.0 which are data and analytics, aimed at supporting the core fundamental goal of Industry 4.0 which are: Digitalization and integration of horizontal & vertical value chains, digitalization of products and service offerings includes the expansion of existing products; and lastly, digital business models and customer access.

The execution of the vision for postal industry 4.0 through execution of strategy is envisaged to improve outcomes in the form of an improved 2IPD score, improved digital readiness score and improved capability maturity for the postal industry in Southern Africa.

The technology and digital objectives are underpinned by 5 strategic objectives which are:

1. Adoption of Industry 4.0 in the postal sector (Postal Industry 4.0)
2. Digital transformation readiness
3. Improvement of digital readiness by Posts in Southern Africa through capability maturity modelling
4. Adoption and diffusion of technological innovation
5. Improvement on the 2IPD score

A system approach is represented by feedback systems that are represented as red lines and arrows. Feedback systems are the foundation of the proposed approach as they mimic a dynamic setting due to interaction of a plethora of factors and outcomes.

The critical findings of this study can be summarized as:

(a) The traditional technology adoption models are particularly relevant to individual level rather than organizational level

(b) The TOE (Technology, Organization and Environment) model is the first attempt to incorporate organizational-wide adoptability of technology. However, the approach is linear in nature rather than systemic as feedback systems which are at play in a dynamic setting are lacking.

(c) That there is a big variation in performance on the 2IPD and on digital readiness index between the first-tier countries and Southern Africa. This variation is driven by digital culture which is lower in Southern Africa as compared to top-tier countries (Switzerland, The Netherlands, Germany, France, and Japan).

(d) Visioning and strategy drive digital transformation, without a robust visioning and strategy (including digital culture), the outcomes in terms of performance on both 2IDP and digital readiness index will always be below expectation.

(e) A linear approach to digital transformation and technology adoption fails to capture the complexities associated with interaction of factors in a dynamic setting resulting with even greater complexity.

5. Conclusions

Universally, the postal sector is faced with a plethora of challenges regarding the ever-changing world that is driven by volatility, uncertainty, complexity, and ambiguity (VUCA). This research presented the postal landscape with specific reference to performance of Southern Africa as a Region against the top 5 postal operators on the Universal Postal Union's Integrated Index on Postal Development.

The limitations of the technology adoption models were presented and discussed in context of their "linearity" or rather, a reductionist approaches to technology adoption as opposed to a systems approach which was presented and discussed in detail together with its respective tenets of which feedback systems are fundamental cornerstone to a system approach where the paradigm is "causality" rather than "linearity".

The conceptual framework that incorporates a system approach was presented and discussed in detail, with its respective tenets derived through insights obtained from critical review of literature. In [1], the authors presented a theoretical model at IEEM 2019. The proposed conceptual model as depicted in Figure 23; it is an enrichment of the earlier theoretical framework presented. It is a build-up towards the final goal of developing a dynamic model towards technology and digital adoption in the postal sector in Southern Africa. The conceptual framework answers the research question which was the guiding ingot to the critical literature review.

The research gap identified in this study is the failure of traditional technology adoption models in driving digital transformation agenda. The linearity fails to capture adoption dynamics in a dynamic setting as opposed to the proposed adoption model with build-in feedback loops to better understand interaction of drivers and barriers in an evolving dynamic setting as in the real world.

Future research will be the development of a conceptual framework using system dynamics where the drivers and barriers identified in the research are modelled in a dynamic setting to understand patterns that emerge from their interaction. The developed system dynamic conceptual model will be rigorously tested by selected participating Posts to further refine the model until it closely represents reality. The validated model could be used by Posts to model different drivers and barriers with a view of mitigating against barriers that would enable Posts in Southern Africa to fully adopt a digital transformation journey in order to be sustainable now and in the future.

6. Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. Mokgohloa, M.G. Kanakana-Katumba, R.W. Maladzhi, J.A. Trimble, "Postal Development: Literature Review into Adoption Models," IEEE International Conference on Industrial Engineering and Engineering Management, **2**, 764–768, 2019, doi:10.1109/IEEM44572.2019.8978531.

[2] UPU, Annual Strategic Review 2017, Bern, 2017.

[3] UPU, "Postal Development Report 2018 - Benchmarking a critical infrastructure for sustainable development," (April), 1–32, 2018.

[4] C. Barman, A., Potsangbam, "Shifts of Strategic Paradigms in the VUCA World- Does 'outside the box thinking' a meaningful cliché for the business world?," (July 2018), 2017.

[5] B. Cousins, "Design thinking: Organizational learning in VUCA environments," Academy of Strategic Management Journal, **17**(2), 1–18, 2018.

[6] N. Balasubramanian, "Innovative Operation Strategies in VUCA world," 2319–7668, 2018.

[7] P. Hallo, Leonie; Nguyen, Tiep; Gorod, Alex; Tran, "Effectiveness of Leadership Decision-Making in Complex Systems," Systems, **8**(5), 2–21, 2020.

[8] SADC, SADC postal strategy 2017-2020, 2017.

[9] SAPOA, SAPOA Strategy Implementation Plan 2017 - 2020 Version 4, SAPOA, Maseru, 2016.

[10] UPU, postal Development Report 2018, Universal Postal Union, Bern, 2018.

[11] Universal Postal Union, The digital economy and digital postal activities – a global panorama, Universal Postal Union, Bern, 2019.

[12] UPU, Postal Development Report 2019, Berne, 2019.

[13] T. Santos, Beatrice; Chaura-Santos, F; Lima, "Industry 4.0: an overview," Research G, (July), 2018.

[14] H. Lasi, P. Fettke, H.G. Kemper, T. Feld, M. Hoffmann, "Industry 4.0," Business and Information Systems Engineering, **6**(4), 239–242, 2014, doi:10.1007/s12599-014-0334-4.

[15] Universal Postal Union, The digital economy and digital postal activities – a global panorama, 2019.

[16] G. Reinhard, V. Jesper, S. Stefan, Industry 4.0: Building the digital enterprise, 2016, doi:10.1080/01969722.2015.1007734.

[17] A. Rojko, "Industry 4.0 Concept: Background and Overview," International Journal of Interactive Mobile Technologies (IJIM), **11**(5), 2017, doi:10.3991/ijim.v11i5.7072.

[18] Deloitte, Industry 4.0 Challenges and solutions for the digital transformation and use of exponential technologies, 2015, doi:10.1057/9780230514027_2.

[19] L.D. Galindo, The Challenges of Logistics 4.0 for the Supply Chain Management and the Information Technology, 2016, doi:10.1038/npp.2015.289.

[20] B. Hermann, Mario; Pentek, Tobias; Otto, Ultrafast synthesis of bifunctional Er 3+ /Yb 3+ -codoped NaBiF 4 upconverting nanoparticles for nanothermometer and optical heater, Gallen, 2015, doi:10.1016/j.jcis.2017.12.027.

[21] J. Macaulay, L. Buckalew, G. Chung, Internet of Things in Logistics, 2015.

[22] F. Jindal, R. Jamar, P. Churi, "Future and Challenges of Internet of Things," International Journal of Computer Science and Information Technology, **10**(2), 13–25, 2018, doi:10.5121/ijcsit.2018.10202.

[23] USPS, The Internet of Postal Things Report Number RARC-WP-15-013, Virginia, 2015.

[24] M.A. Rahman, A.T. Asyhari, "The emergence of internet of things (Iot): Connecting anything, anywhere," Computers, **8**(2), 8–11, 2019, doi:10.3390/computers8020040.

[25] UPU, Postal development report, Berne, 2018.

[26] G. Kuehne, R. Llewellyn, D.J. Pannell, R. Wilkinson, P. Dolling, J. Ouzman, M. Ewing, "Predicting farmer uptake of new agricultural practices: A tool for research, extension and policy," Agricultural Systems, **156**(August 2016), 115–125, 2017, doi:10.1016/j.agsy.2017.06.007.

[27] E.M. Rogers, Diffusions of Innovations, 1995.

[28] E.M. Rogers, M. Everett, Diffusion of Innovatons, 1983.

[29] R. Sharma, R. Mishra, "A Review of Evolution of Theories and Models of Technology Adoption," Indore Management Journal (IMJ), **6**(2), 17–29, 2014.

[30] C. Dube, V. Gumbo, "Diffusion of Innovation and the Technology Adoption Curve: Where Are We? The Zimbabwean Experience," Business and Management Studies, **3**(3), 34, 2017, doi:10.11114/bms.v3i3.2500.

[31] A. Kogabayev, Timur; Maziliauskas, "The definition and classification of innovation," Holistica, **8**(1), 59–72, 2017, doi:10.1515/hjbpa-2017-0005.

[32] S. Xaba, Evaluation of South African Post Office SOC Ltd Innovation Efficacy, University of Pretoria, 2012.

[33] P.F. Drucker, Innovation and Entrepreneurship: Practice and Principles, HarperCollins Publishers, London, 1985.

[34] M. De Jong, N. Marston, E. Roth, "The eight essentials of innovation," McKinsey Quarterly, (2), 36–47, 2015.

[35] J.P. Wentzel, K.S. Diatha, V.S.S. Yadavalli, "An application of the extended Technology Acceptance Model in understanding technology-enabled financial service adoption in South Africa," Development Southern Africa, **30**(4–5), 659–673, 2013, doi:10.1080/0376835X.2013.830963.

[36] D. Mugo, K. Njagi, B. Chemwei, J. Motanya, "The Technology Acceptance Model (TAM) and its Application to the Utilization of Mobile Learning Technologies," British Journal of Mathematics & Computer Science, **20**(4), 1–8, 2017, doi:10.9734/BJMCS/2017/29015.

[37] J.P. Wentzel, Modelling Financial Services Adoption Through an Intermediary in South Africa: TAM and SEM Approach, Pretoria, 2012.

[38] Q. Ma, L. Liu, "The Technology Acceptance Model," Advanced Topics in End User Computing, **4**, (January), 2011, doi:10.4018/9781591404743.ch006.ch000.

[39] M. Chuttur, "Technology Acceptance, Information System Deployment, TAM, Information System Theory," Sprouts, **9**(2009), 2009.

[40] P.. Davids, F.D; Bagozzi, R.P; Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," Management Science, **35**(8), 982–1003, 1989.

[41] P. Lai, "the Literature Review of Technology Adoption Models and Theories for the Novelty Technology," Journal of Information Systems and Technology Management, **14**(1), 21–38, 2017, doi:10.4301/s1807-17752017000100002.

[42] V. Venkatesh, F.D. Davis, "Theoretical extension of the Technology Acceptance Model: Four longitudinal field studies," Management Science, **46**(2), 186–204, 2000, doi:10.1287/mnsc.46.2.186.11926.

[43] F. Venkatesh, Viswanath; Morris, Michael; Davis, Gordon; Davids, "User Acceptace of Information Technology: Towards a Unified View," MIS Quarterly, **27**(3), 95–98, 2003.

[44] C.M. Chao, "Factors determining the behavioral intention to use mobile learning: An application and extension of the UTAUT model," Frontiers in Psychology, **10**(JULY), 1–14, 2019, doi:10.3389/fpsyg.2019.01652.

[45] Y.K. Dwivedi, N.P. Rana, A. Jeyaraj, M. Clement, M.D. Williams, "Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a Revised Theoretical Model," Information Systems Frontiers, **21**(3), 719–734, 2019, doi:10.1007/s10796-017-9774-y.

[46] H. Gangwar, H. Date, R. Ramaswamy, "Understanding determinants of cloud computing adoption using an integrated TAM-TOE model," Journal of Enterprise Information Management, **28**(1), 107–130, 2015, doi:10.1108/JEIM-08-2013-0065.

[47] T. Oliveira, M. Martins, "Literature Review of Information Technology Adoption Models at Firm Level.," Electronic Journal of Information …, 2011.

[48] A. Dasgupta, S. Wendler, "AI Adoption Strategies," (9), 1–13, 2019.

[49] G. Kuehne, C. Nicholson, M. Robertson, R. Llewellyn, C. McDonald, "Engaging project proponents in R&D evaluation using bio-economic and socio-economic tools," Agricultural Systems, **108**, 94–103, 2012, doi:10.1016/j.agsy.2012.01.009.

[50] K. Crowston, J. Qin, "A capability maturity model for scientific data management: Evidence from the literature," Proceedings of the ASIST Annual Meeting, **48**(February), 2011, doi:10.1002/meet.2011.14504801036.

[51] M.C. Paulk, B. Curtis, M.B. Chrissis, C. V Weber, "for Software , Version 1 . 1 Software Engineering Institute," Office, **40**(February), 1–26, 1993.

[52] M. Shahiduzzaman, M. Kowalkiewicz, R. Barrett, M. McNaughto, Digital Business: Towards a Value Centric Maturity Model, 2017.

[53] Deloitte, Deloitte Digital Maturity Model, 2018.

[54] J. Ganzarain, N. Errasti, "Three stage maturity model in SME's towards industry 4.0," Journal of Industrial Engineering and Management, **9**(5), 2016, doi:10.3926/jiem.2073.

[55] M. Rafferty, "Reductionism, Holism and System Dynamics," Proceedings of the 2007 International Conference of the System Dynamics Society, 2007.

[56] M. Dafermos, "Encyclopedia of Critical Psychology," Encyclopedia of Critical Psychology, (April), 2014, doi:10.1007/978-1-4614-5583-7.

[57] P. Mella, "Systems Thinking: The art of understanding the dynamics of systems," Journal of Learning, **15**(January), 2015.

[58] D.H. Meadows, Thinking in Syatems, Earthscan, London, 2009.

[59] H. Sillitto, P. Godfrey, "Defining " System " : A Comprehensive Approach," in 27th Annual INCOSE International Symposium, INCOSE, Adelaide,

2017.
[60] M. Jamshidi, "Systems thinking," IEEE Transactions on Neural Systems and Rehabilitation Engineering, **16**(4), 1–4, 2008, doi:10.1109/TNSRE.2008.929253.
[61] J.D. Sterman, Business Dynamics: Systems thinking and modeling for a complex world, McGraw-Hill, London, 2000, doi:10.4324/9781351152723-7.
[62] J.D.W. Morecroft, Strategic Modelling and Business Dynamics, 2015, doi:10.1002/9781119176831.
[63] E.. Schumacher, Introduction to Systems Thinking Principles and Analytical Tools, 2019.
[64] S. Peter, The Fifth Discipline:The art and practice of learning organisation, Doubleday, New York, 1990.
[65] J.D. Sterman, L.B. Sweeney, "Bathtub dynamics: initial results of a systems thinking inventory," System Dynamics Review, **16**(4), 249–286, 2000.
[66] R.D. Arnold, J.P. Wade, "A definition of systems thinking: A systems approach," Procedia Computer Science, **44**(C), 669–678, 2015, doi:10.1016/j.procs.2015.03.050.
[67] J.D. Sterman, "System dynamics modeling: Tools for learning in a complex world," IEEE Engineering Management Review, **30**(1), 42–52, 2002, doi:10.1109/EMR.2002.1022404.
[68] M. Reinker, E. Gralla, "A System Dynamics Model of the Adoption of Improved Agricultural Inputs in Uganda, with Insights for Systems Approaches to Development," Systems, **6**(3), 31, 2018, doi:10.3390/systems6030031.
[69] D.K. Fisher, J. Norvell, S. Sonka, M.J. Nelson, "Understanding technology adoption through system dynamics modeling: Implications for agribusiness management," International Food and Agribusiness Management Review, **3**(3), 281–296, 2000, doi:10.1016/s1096-7508(01)00048-9.
[70] P.D. Leedy, J.E. Ormrod, John W. Creswell, S. Greener, W. Askarzai, B. Unhelkar, D.J. Currie, C. Smith, P. Jagals, D. Williams, H. Snyder, H.K. Mohajan, H.K. Mohajan, J.M. Brayer, O. Skovsmose, M. Borba, M.A. Ragab, A. Arisha, S. Rajasekar, P. Philominathan, S. Macdonald, N. Headlam, U. Sekaran, C. Kothari, Research Design, Eleventh E, SAGE, Essex, 2004, doi:10.1186/s12889-018-5318-8.
[71] J.W. Creswell, Research Design, Fourth Edi, SAGE, London, 2014.
[72] H. Snyder, "Literature review as a research methodology: An overview and guidelines," Journal of Business Research, **104**(August), 333–339, 2019, doi:10.1016/j.jbusres.2019.07.039.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1204-1213 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Examination of a Skill Sampling Method of an Athlete Using the Athlete's Movement and Eye Movement for the Development of an AI Coach

Takuya Sarugaku[*,1], Jun Lee[2], Yasuaki Matsumoto[2], Mitsuho Yamada[1]

[1]*Graduate School of Information and Telecommunication Engineering, Tokai University, 108-0074, Japan*

[2]*Takei Scientific Instruments Co., Ltd., 956-0113, Japan*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 17 September, 2020
Online: 24 October, 2020

Keywords:
Sports
AI
4K
Gaze movement
Wireless eye movement measurement device

ABSTRACT

From amateur players who enjoy sports throughout their lives to top athletes who participate in international competitions, interest in improving sports skills is growing. Coaches and their coaching are indispensable for improving sports skills, but it is difficult for many athletes, especially amateur athletes, to secure coaching. However, we thought that anyone could easily receive coaching through the use of an artificial intelligence (AI) coach. In order to bring about AI coaching, learning is important. The set of learning data must include data such as players' skills as they correspond to their gaze and performance. In particular, it is thought that analyzing gaze movement during sports may provide insight into exceptional athletic skills. In this study, we propose a skill sampling method of collecting learning data for the specific purpose of creating an AI coach, using a wireless eye movement measurement device and 4K imaging.

1. Introduction

There is growing interest in sports, including the promotion of lifelong sports with an eye on the aging population, as well as the improvement of the skills of athletes aiming to participate in the Olympic Games and other international competitions. Many people from professionals to amateurs wish to improve their athletic ability and enjoy sports as a regular activity. Having a good coach is essential to improve sports skills, however, not everyone has access to a superior coach. We believe that by extracting the skills of both athletes and coaches, and applying learning data, it is possible to develop an AI robot that can coach anytime and anywhere.

An important task in realizing such AI is to include skill learning. It is necessary to collect learning data as a set of both players' and coaches' performances, gazes, and ideal skills. However, to the best of our knowledge, no such sets of learning data yet exist. In the present paper, we examine a skill sampling method of collecting learning data based on the performance and gaze of an athlete.

We focused on eye movement because we believe that it is essential for players during sports to watch the ball, to watch the opponent, and to observe the situation of the game. In fact, in an experiment that analyzed the eye movements of athletes performing on a gymnastics pommel horse, it was confirmed that posture fluctuations can be predicted from the stability of eye movements [1]. In martial arts such as kendo and fencing, it is said that it is important to maintaim a wide field of view by concentrating on a single point without moving one's eyes (as the Japanese proverb says, *enzan no metsuke*: "eyes on the far mountain" in a rough English translation)[2]. In a study by Natsuhara et al. [3], experts and semi-experts in football clarify whether there is a difference in eye movements during play depending on skill.

It is thought that an examination of eye movements during competition in various sports can reveal the points of attention and important gaze points for each sport.[4-8] For example, in relation to eye movements, experiments that block vision while a subject shoots a basketball have demonstrated that visual information at the final stage of the shot is important [9].

However, simply measuring eye movements without further context leaves it unclear what game situation brought about those eye movements. Therefore, combining athlete images taken from different directions with eye movement data can help match the movement of the athlete at any given point to the specific playing situation and eye movement. By connecting an actual athlete's action output with data obtained in this way, a set of learning data can be obtained. The required learning data vary by sport. In table

[*]Corresponding Author: Takuya Sarugaku, Email: tky.j10@fuji.tokai-u.jp

tennis, for example, useful data include the players' gaze, posture and movements, how to hit the ball, the score rate, etc. In dance, on the other hand, in addition to data on the dancer's gaze, posture and movements, it is also useful to know his or her velocity of rotation, etc. It is necessary to construct a database that associates the characteristic play of skilled athletes with results such as score rate and rotation speed.

In this paper, we examine a skill sampling method based on an athlete's performance and gaze in order to acquire a data set to be used in creating an AI coach. First, we consider the configuration and accuracy evaluation of a wireless eye movement measurement device developed to easily measure the eye movements of athletes who exercise intensely. Next, we present the results of experiments conducted during table tennis and dance rotation as examples of data acquisition.

2. Configuration of a wireless eye movement measurement device

Many sports science experiments using eye movements have been conducted [10]. In the work [11], experiments have been conducted with the aim of elucidating the role of head, eye and arm movements during table tennis competitions, using eye movement measuring devices and other equipment. However, the eye movement measurement device used in this experiment requires a wired connection to an interface PC and control unit etc, which is difficult to measure in competitions with a lot of distance to travel or competitions involving rotation. In addition, the study [12] describes the eye movements and shooting strategy in soccer penalty kicks. The eye movement measurement device used in this study was linked to a PC via a 10m fire wire cable from a recording device in a pouch wrapped around the waist, so it was not possible to measure in a completely free state. One study by [4] measured the eye movements of athletes watching competition videos. As in these examples, few studies have measured eye movements in a state where athletes are free to compete.

There are several reasons why it is difficult to measure the eye movements of athletes during competition. First, wired devices can hinder the athlete's performance. Also, calibrated sensors may be displaced by the intense movement of the athlete and, when a sensor shifts, the experimenter may not notice the shift unless measurement status can be confirmed in real time. We solved these problems using wireless measurement devices and by enabling real-time monitoring and recalibration of the measurement situation. Figure 1 shows the configuration of the developed eye movement measurement device, and a person wearing the device is shown in Figure 2. A video transmitter and external control system with a mouse were added to a TalkEye Lite system (Takei Science Instruments Co., Ltd., Niigata, Japan), which uses a corneal reflection method. The experimenter can control recording start/stop, offset of the line of sight, etc. with a Bluetooth mouse by watching a display transmitted wirelessly. Two devices can be used at the same time. Because experimental data is recorded in the TalkEye Lite, there are no direct effects of the wireless delay. The device can be fixed to the head with a hook and loop fastener so that the eye movement sensors do not slip off during exercise.

Figure 3 shows the operation screens shown on the external display: (a) is the eye movement calibration screen, (b) is an offset button for correcting the displacement of gaze movement when the subject gazes at the center of the field of view before the experiment, and (c) is the experimental screen. Basic operation is possible wirelessly when the athlete carries the eye movement measuring equipment in a backpack. It is possible to measure the gaze movement of two athletes simultaneously by radio.

We evaluated the accuracy of this eye movement recording system using a dancer who agreed to serve as a subject for the experiments discussed in Section 4. We had the dancer perform two high-speed turns and compared eye movement accuracy before and after the turns. Five-point calibration was applied, as shown in Figure 4. The distributions of the gaze points when the subject gazed at the calibration points before and after two high-speed turns are shown in Figure 5. The subject's accuracy before turning was x: -0.82±0.64 deg., y:-0.72±0.69 deg., and that after turning was x: 0.92±0.87 deg., y:-1.71±1.75 deg. Considering that the diameter of the fovea, which has the highest resolution on the retina, is about 1 degree, it was confirmed that eye movement accuracy barely changes at all with high-speed turns. However, it was confirmed that the dancer's gaze shifted upward when viewing the lower index point after rotation.

The maximum transmission distance of video in the room is 19.7m, and the maximum distance at which the TalkEye Lite can be controlled is 43m. The device weighs 1.3 kg. The size of the goggles with the eye movement sensors and a field of view camera is 19 cm in length, 21 cm in width, and 9 cm in height.

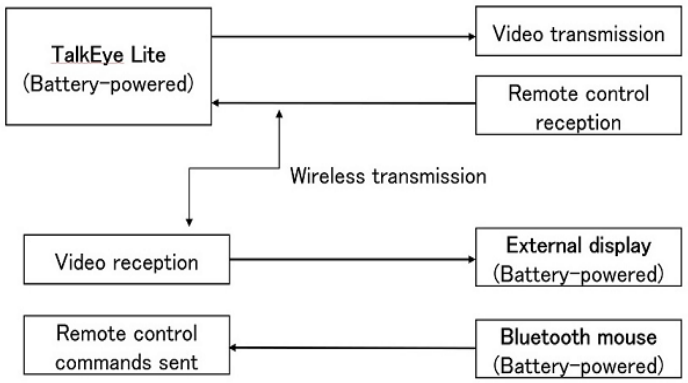


Figure 1: Configuration of the developed device



Figure 2: A person wearing the device (the eye movement measurement device is in the backpack)

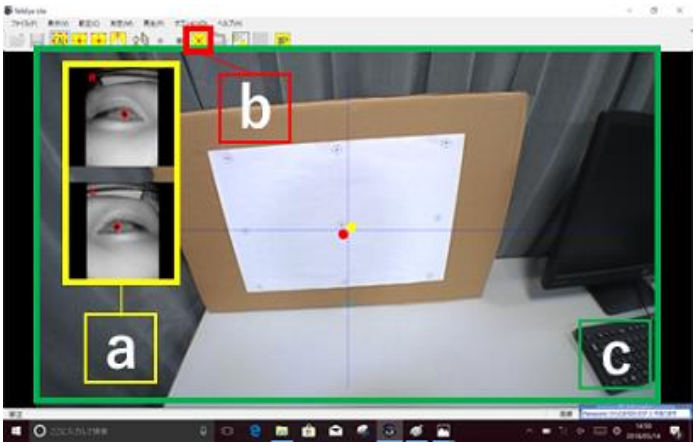


Figure 3: The operation screens shown on the external display

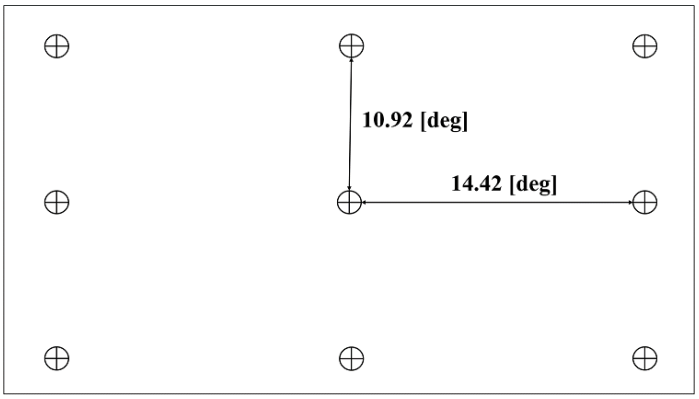


Figure 4: Fixation points for five-point calibration

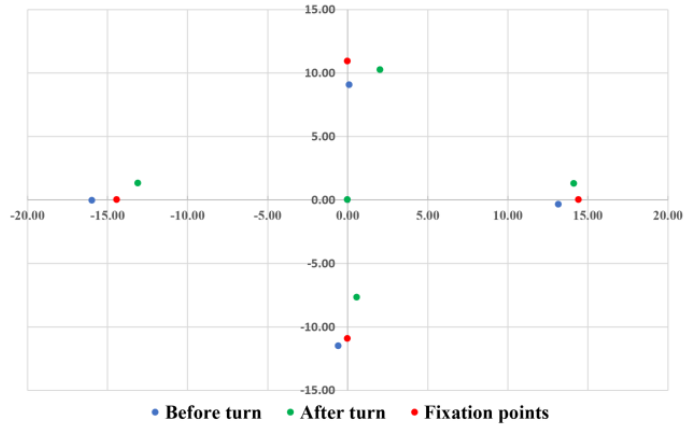


Figure 5: Distributions of the gaze points when the subjects gazed at the calibration points before and after two high-speed turns

3. Measurement of gaze movement during table tennis play

We first conducted experiments on table tennis as an example to demonstrate the wireless function of our device. Our subjects were two players on the table tennis team at Tokai University Takanawadai Junior & Senior High School. A table tennis table was brought to our 4K studio, each athlete wore a backpack device, and the eye movement of each was acquired. In order to record the actions of the athletes, video was taken with a 4K camera from the front and the side of the table tennis table and were recorded synchronously with recording by the eye movement measurement device. Figure 6 shows the layout of the equipment during the experiment, and Figure 7 shows an image taken during play.

An example of an image edited after measurement is shown in Fig8. "Cam 1" in Figure 8 shows the performance of Player A, "Cam 2" shows the movement of the ball during the game, and "Cam 3" shows the performance of Player B. The images marked "Player A" and "Player B" in the figure show their respective lines of sight. In addition to the images of the two athletes' gaze movement superimposed on the image from each camera, the expression and movement of the athletes during the game are displayed. Our 4K system makes it possible for individual images to maintain a resolution quality nearly as good as that of high-definition television even when we divided an image, as in this figure; high-precision analysis was thus possible.

The purpose of this experiment was to sample the athlete's characteristic eye movements and performance during table tennis by using the above analysis method. The specific data to be collected include information on the athlete's gaze and performance during rallies, for example. These data will be useful in the creation of an AI coach. An example of analysis using this apparatus is shown in Figure 9, which shows the gaze movement of both players at the moment Player A hits the ball. The left and right lines of sight are indicated by the yellow and green numbers, respectively, where their number indicates their order.

Furthermore, we propose a method to compare the body and eye movements of Players A and B in the time series shown in Figure 10. Figure 10 compares the actions and lines of sight of Players A and B. Here, for greater clarity, the ball is circled in white and the line of sight is circled with yellow dots. The area over which the gazing points were distributed is expressed by the size of the circle.

In this sequence, Player B hits the ball at Time (a) [Column a, Row 3], and Player A hits the ball at Time (f) [Column f, Row 1]. Let us first focus on Player A. Looking at Player A's line of sight [Row 2] at Time (a) [Column a], Player A looks at Player B's racket at the moment Player B hits the ball. At Times (b-f) [Columns b-f, Row 2], it can be confirmed that Player A continues watching the flying ball. Player A's movement at Time (d) [Column d, Row 1] confirms that he starts to drop his waist to return the ball. Comparing Times (e), (f) and (g) [Columns e-g, Row 1], we see that Player A's waist has dropped most at Time (g) [Column g, Row 1], after he has finished hitting the ball. In addition, we can confirm that he follows the ball and gazes at his partner's racket at Times (f-h) [Columns f-h, Row 2].

Next, when we examine Player B's line of sight where he follows the trajectory of the flying ball at Time (a) [Column a, Row 4], and after hitting the ball toward Player A at Time (b) [Column b, Row 4], it can be seen that his eye follows the trajectory of the ball as it flies to Player A. Also, Player B's movement at Time (a) [Column a, Row 3] confirms that he is leaning forward to hit the ball. The sequence at Times (f-h) shows that Player A hit the ball higher than at Times (a-c), and this is confirmed by comparing Player B's movement [Column b, Row 3] and [Column g, Row 3] and continuing into Time (h) as he follows the downward motion of the ball [Column h, Row 4].

Judging from the relative sizes of the circles of yellow dots, Player B's eye movement area is larger than Player A's. Head movement was judged from images of the table tennis table taken by each player's field-of-view camera. Images from Player A's

field-of-view camera [Row 2] moved largely along the vertical direction, and changes in the image of the table are therefore large. In contrast, the images from Player B's field-of-view camera [Row 4] show little change, suggesting that his head moved only a little. These results indicate that Player A's head movement supplemented his eye movement since his eye movement area was small, while Player B did not need to supplement his eye movement with head movement. In this regard, the coach of the high school table tennis team pointed out that small head movement during a game, like Player B's, is better.

Time (g) in Figure 10 also shows another characteristic result. When the ball hit by Player A reaches the top of its trajectory after bouncing off the table, Player B looks at the highest point of the ball. It is known that the rotation direction of the ball is easy to see when the vertical velocity of the ball is nearly at the top of its trajectory. In other words, we were able to capture the moment when Player B looked at the top of the trajectory to determine the rotation direction of the ball.

Two important features of our system are that two wireless eye movement analysis devices can be employed simultaneously and that it uses high-definition 4K imaging, making it possible to extract such moments from rallies between two players. By using two devices, it is possible to know the athletes' tactics using gaze. We believe that this can be useful as a sports video analysis method or as a new training method. By using this method, the athlete's body movement, situation, and eye movement can be obtained simultaneously at any point during competition. At present, coaching emphasizes movement but, by using gaze data, it is possible to provide coaching based on the information skilled athletes obtain from their own gaze and from their own plays as well as those of their opponents. Furthermore, a database can be created in which these measured data are combined with the score rate, smash success rate, etc. By then using this information as learning data, an AI coach that can teach expert skills can be created.

4. Measurement of gaze movement during dance

Next, let us consider an experimental example of dance. Considerable research on figure skating is currently being conducted. Resistance to vestibular system input is typically weak in novices who have just started figure skating, and it has been noted that they are remarkably susceptible to physiological and mental disorders. Rotation experiments using a Barany chair have shown heartbeat irregularities and malfunction of the coordination of movements of the hands and feet [13].

It is difficult to measure eye movements while a subject is rotating during competition. Therefore, experiments have been conducted in which circulation vection is generated using virtual reality in order to examine eye movements closely related to the vestibulo-ocular reflex (VOR). In a comparison of professional skaters and beginners in which the angle of the view to be displayed was changed, significant differences were observed all components of eye movement (decrease in fixation time, decrease in saccade amplitude, increase in blinking, etc.) in results corresponding to a visual angle exceeding 180°[14].

These studies show that training of the vestibulo is an important aspect of figure skating training, and eye movement is useful for evaluation. From the point of view of training athletes for sports that require high speed rotation, such as figure skating and dancing, the athlete should set the gaze position at the time of rotation at a single point and suppress the VOR, which is the function of the semicircular canal at the back of the ear. This is said to prevent eye movement in the direction opposite to the rotation of the head, which occurs to compensate for head movement during head rotation [15-17].

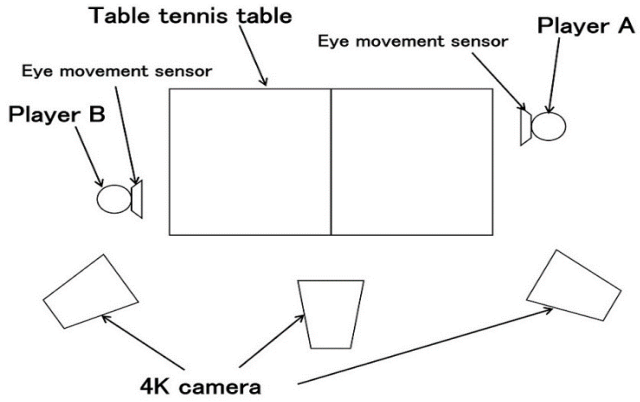


Figure 6: Arrangement of the developed equipment



Figure 7: Photo taken during measurement



Figure 8: Edited image

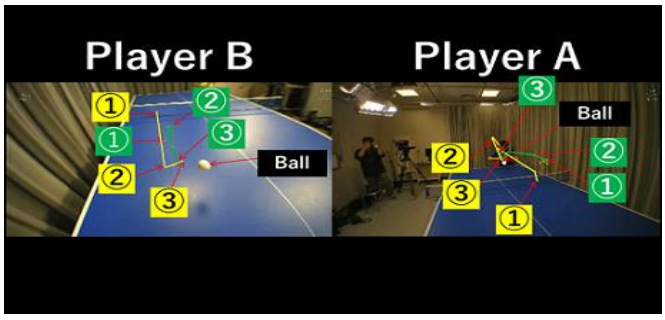


Figure 9: Analysis example

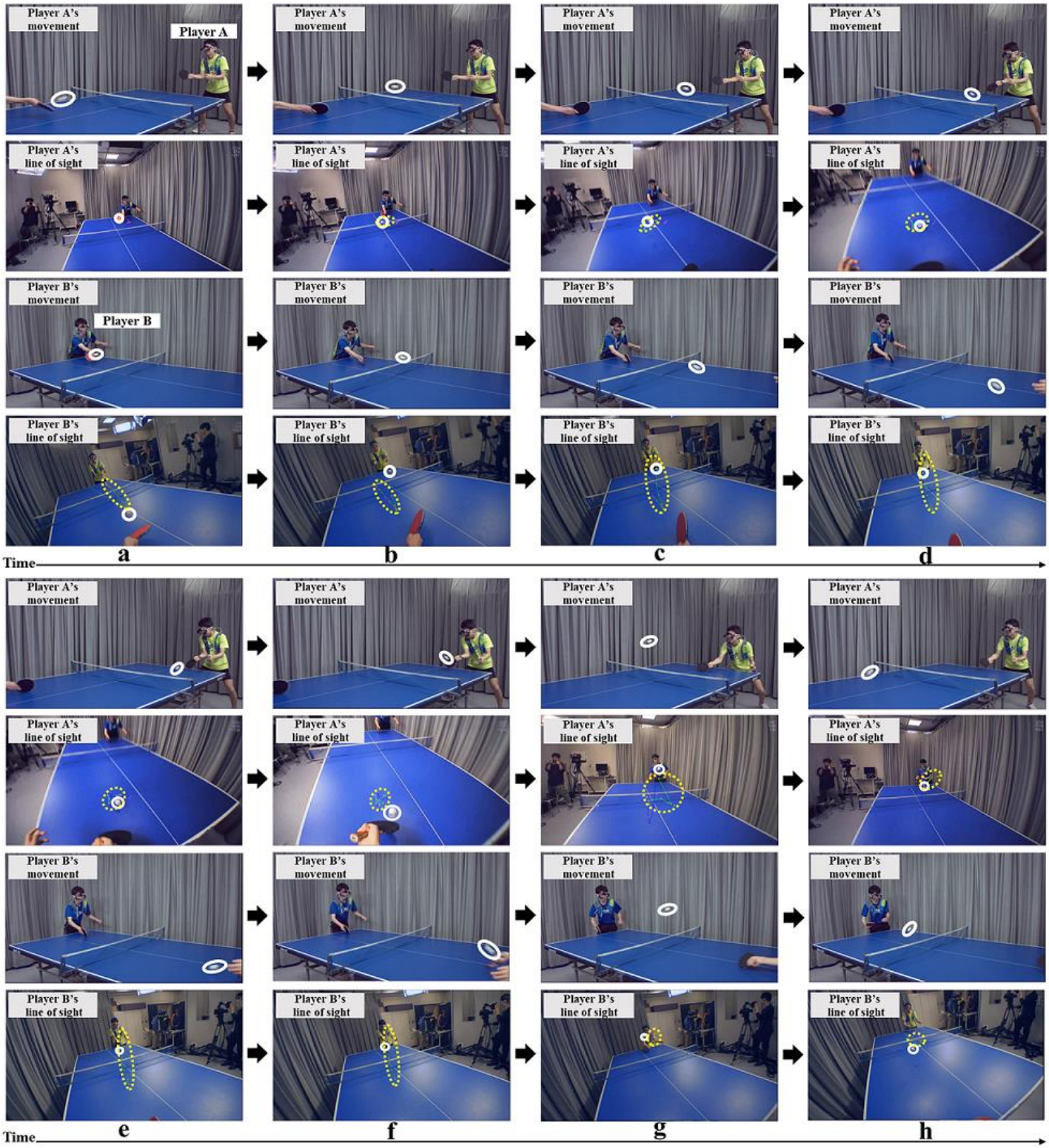


Figure 10: Body and eye movements of Players A and B in a time series

Previous studies have used rotating chairs and images such as VR environments but to the best of our knowledge, none have actually measured eye movements during performance. Since the athlete rotates, a wired system is impossible and a wireless system is indispensable. Furthermore, since the system undergoes heavy centrifugal force during rotation, it is necessary to ensure that the goggles for eye movement measurement do not shift, and robustness is also required for the eye movement measurement device itself. Additionally, due to the intense motion during performance, it is possible that the calibration of the eye movement device may shift. Thus, it is necessary for the experimenter to be able to recalibrate the eye movement measuring apparatus during performance of subject.

We believe that the device introduced in the present study meets the requirements for eye movement measurement during dance. In the present dance experiment, a dance expert with a dance history of 17 years beginning at the age of 4 wore a wireless eye movement measurement device and rotated in place. To the best of our knowledge, there have been no reported examples of eye movement measurement during dance using a wireless-type eye movement measurement device. All experiments were conducted at the 4K studio of our university, and videos were taken with GoPro from the top, and with a 4K camera from the front, right and left; all videos were recorded synchronously with the wireless eye movement measurement device. The sampling rate was 120 Hz for the GoPro device, 60 Hz for the 4K camera, and 30 Hz for the wireless eye movement measurement device. The experimental set-up is shown in Figure 11 and 12.

After the experiment, we edited the video from the three recording devices, and found the rotation angles of the head and body. An example of the edited video is shown in Figure 13. Images of the dancer's eye movement and of the dancer from various directions were recorded in high definition using a 4K production system. The purpose of this experiment was to sample a dancer's characteristic eye movements and body posture and movement during a dance competition. As mentioned above, to the best of our knowledge, there has been no actual measurement of gaze during a dance turn. Our proposed skill sampling method can clarify not only the gaze of the spinning dancer, but also his or her movement. A sequence of images from a dance turn is shown in Figs.14-17. Figure 14 shows the series of images taken from overhead, and the same sequence is then shown from the front (Figure 15), from the right (Figure 16), and from the left (Figure 17). A picture was taken every 45°, and the numbers correspond across the different views. The present method makes it possible to acquire the relationships among the body, head, and eye movements of a dancer during a dance.

Figs.18 and 19 are graphs of the subject's lines of sight at the time of rotation. The solid line shows the rotation angle of the head, the broken line shows that of the body, and the dotted line shows that of the right eye (gaze). Where there is no dotted line, the subject blinked and eye movement could not be acquired. The color of each line corresponds to the number of the trial. The left vertical axis in Figure 18 shows the angle between the head and the body, the right vertical axis shows the angle of the line of sight, and the horizontal axis shows the corresponding numbers of the frames on the wireless eye movement measuring device. In Figure 19, the vertical axis shows the velocity of the corresponding rotation angle, and the horizontal axis shows the numbers of the frames on the wireless eye movement measuring device. First, let us focus on head movement. The head does not move significantly until Frames 7-10 in Figure 18. This is because the athlete is trying to look at his chosen front fixation point. After that, the head rotates rapidly from Frame 8 onward, overtaking the body, which rotated previously. The velocity of the head rotation angle reaches its maximum in Frames 10-14 in Figure 19, and the average of the velocity of the head rotation angle of 11 trials is 1712±112 deg/sec. Next, let us examine the motion of the body. The velocity of the body rotation angle (Figure 19) reached its maximum with the same timing as the head, at Frames 8-12. However, it can be seen that there is no large change in speed compared to the head; the body rotates at an approximately constant speed. The dancer's eye moves to the left until Frame 11, at which point the body and the head are both rotating to the right, indicating that the dancer is gazing at his chosen front fixation point on the left. Some data were lost due to blinking but, after 18 frames, the body and head passed right behind and turned to the right, and the dancer was gazing at his chosen front fixation point on the right. Eventually, the body and head rotate 360 degrees and again face the front and the eye movement is also 0 degrees, but before that time, a number of sawtooth-like eye movements like OKN (optokinetic nystagmus) that pass through the predetermined frontal fixation point and back again have occurred.

Next, Figs. 20 and 21 show the rotation angle and rotation angle velocity of a typical trial. The axes in Figs. 20 and 21 are the same as those in Figs. 18 and 19. The forms at the bottom of Figs.20 and 21 show the direction of the body, head, and eyes (arrow) at points corresponding to the images in Figs.15-17. These figures are intended to make it easier to match the athlete's body movements with other data obtained at the same point. From ⑤ to ⑥ , the subject blinked and we were unable to acquire eye movement data. Figure 20 shows that the body rotates first, followed by the head, which turns rapidly around a body rotation angle of about 90° (③ in Figs. 14-17, 20 and 21) until the body rotation angle reaches about 140 deg (④ in Figure 14-17, 20 and 21), when the head begins to lead. Figure 21 confirms that the maximum velocity reached 1920 deg/sec when the head was sharply turned. The velocity of the body rotation angle did not change as rapidly as that of the head.

Generally, the eye rotates in the direction opposite to that of the rotation of the head due to the VOR, but when the head is rotating at a high speed, the VOR is prevented, and we confirmed that the eye did not move in Figure 21 (④ and ⑤). Furthermore, the video from the field-of-view camera of the wireless eye movement measuring device confirmed that the subject continued to gaze at the front camera between 0° and 90° (① - ③ in Figure 14-17, 20 and 21) and between 270° and 360° (⑦ - ⑨ in Figure 14-17, 20 and 21). From Figure 20, it was confirmed that the eye rotated to the left from the front as it rotated (① - ④), and from the right to the front after turning (⑦ - ⑨), in order for the dancer to gaze at a point that he had chosen in advance (the front camera, in this example) even as the body rotated. After that, eye movement showed a large reverse in direction at ⑦. Regarding gaze position, it was confirmed that the dancer gazed at a point that he had chosen in advance by asking "Where do you look when you are turning?" after the experiment.

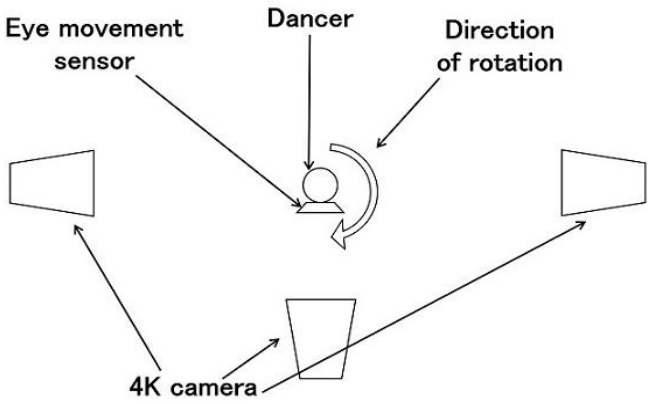


Figure 11: Experimental set-up from the top

Generally speaking, when one sits on a rotating chair and rotates, the VOR causes dizziness. However, a dancer keeps his or her posture steady without dizziness by watching a certain point while rotating [11-13]. To the best of our knowledge, the present study is the first time that this could be confirmed through the use of an eye movement measuring device. The present method makes it possible to acquire the relationships among the body, head, and eye movements of a dancer while dancing. By using this skill sampling method to create a database of the head, body, and eye movement data obtained from many skilled athletes, an AI coach that can teach beginners the optimal rotation method can be created.

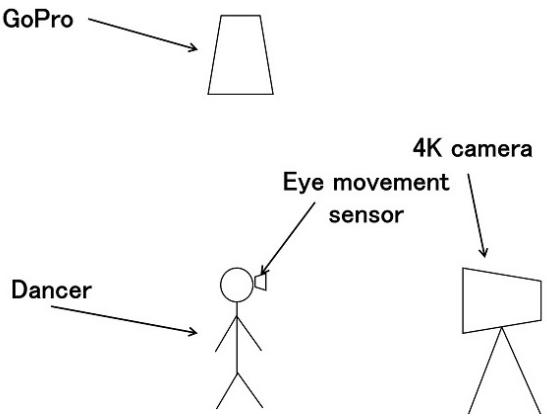


Figure 12: Experimental set-up from the side

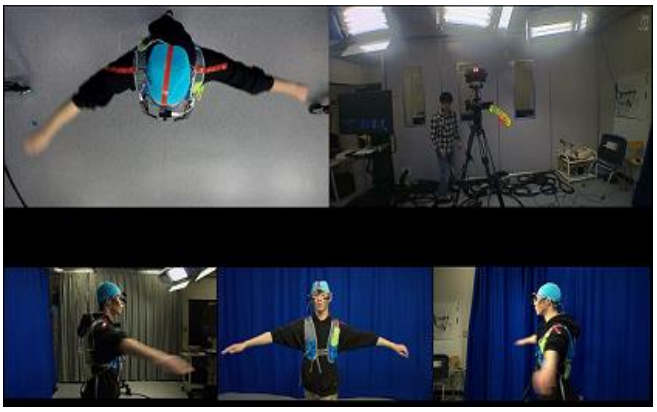


Figure 13: Video output after editing

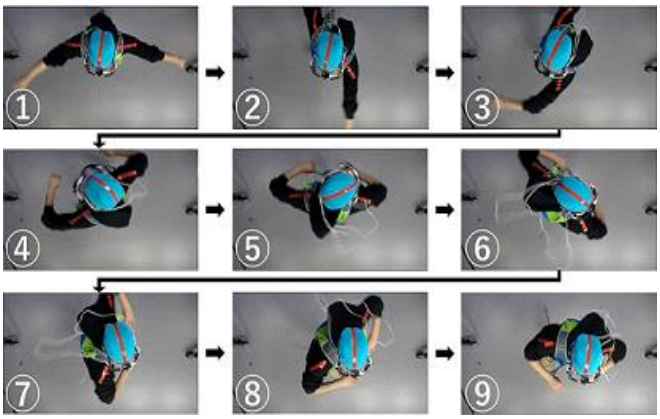


Figure 14: View from the top

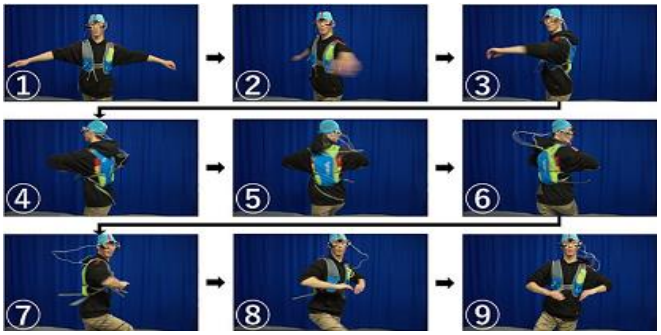


Figure 15: View from the front

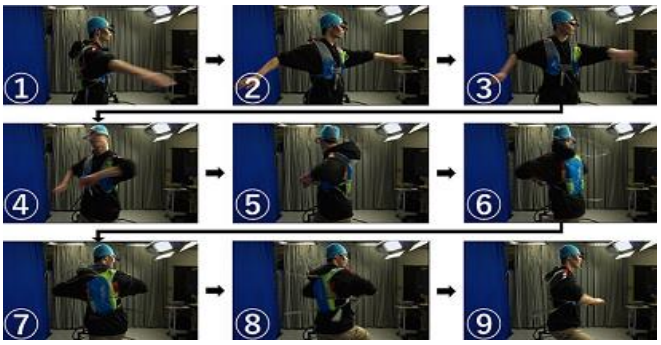


Figure 16: View from the right side

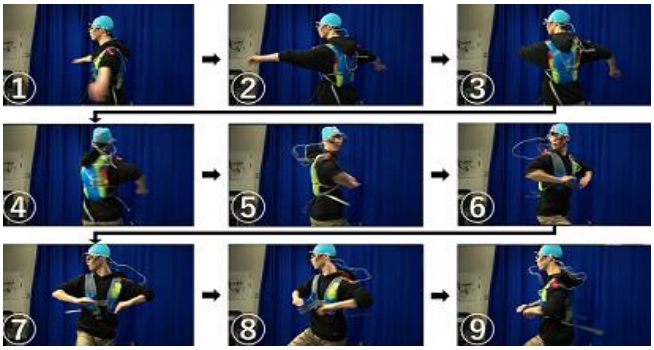


Figure 17: View from the left side

5. Conclusion

There is a big difference in the level of sports ability to aim for, from the promotion of lifelong sports aimed at healthy longevity to the training of athletes who are able to compete in the Olympic Games and other international competitions. For this reason, coaches for many different levels of athletes are required. However, it cannot be said that there are enough coaches, except for athletes who are aiming for professional sports and the Olympics. If an individual learns tips for improving and enjoying various sports, he or she will be able to start sports at an early age and continue throughout life; such tips may also help the handicapped and the elderly. Therefore, if we can develop an AI coach that can teach these tips, we believe that the shortage of coaches can be solved at once.

However, learning data are required in order to develop AI robots. For example, if the goal is a successful shot in soccer, the performance of the athlete up to that point and during the immediate situation becomes learning data. The situation during a game changes every second, and the exact same game situation is rarely, if ever, reproduced. Additionally, the athlete does not experience all possible game situations and does not calculate the

next play based on all past experiences during games. It is possible to develop an AI coach that can make judgments similar to those of the skilled athlete by applying as much learning data and as many outcomes as possible. To the best of our knowledge, no analysis of the behavior of athletes during a match or performance has yet been carried out for this purpose. Therefore, given the fact that the eye movements of advanced and beginner athletes differ significantly in various skills, and taking advantage of the superb resolution of 4K imaging, which allows images taken during play to be synthesized with high resolution, we worked on building databases that make it possible to learn from the eye movements and performance of athletes during a game or performance. In this paper, we present measurement examples of athletes both playing table tennis and performing dance rotation, and analysis examples using 4K video; these data can be used to create a database.

In table tennis, the analysis included data collected when two players were competing, specifically, data regarding eye movement, the position of the racket, and the position of the ball. In the case of dance rotation, eye movement, head movement, and body movement during rotation were examined. The data introduced here are those of non-professional athletes and are not necessarily sufficient as input for learning data. However, by applying our skill sampling method with cooperation from top athletes, it will be possible to collect the necessary data. Using our method, we would like to develop AI coaching through the use of learning data that match eye movements, body movements, game situations and outcomes during play.

Acknowledgment

This work was supported in part by the Hoso Bunka Foundation (2018/2019 Grant Cycle). We would like to express our sincere gratitude to the table tennis team at Tokai University Takanawadai Junior & Senior High School for kindly allowing us to take movies of them playing. We also thank Mr. Hiroki Kudo in our laboratory who danced for our dance experiment.

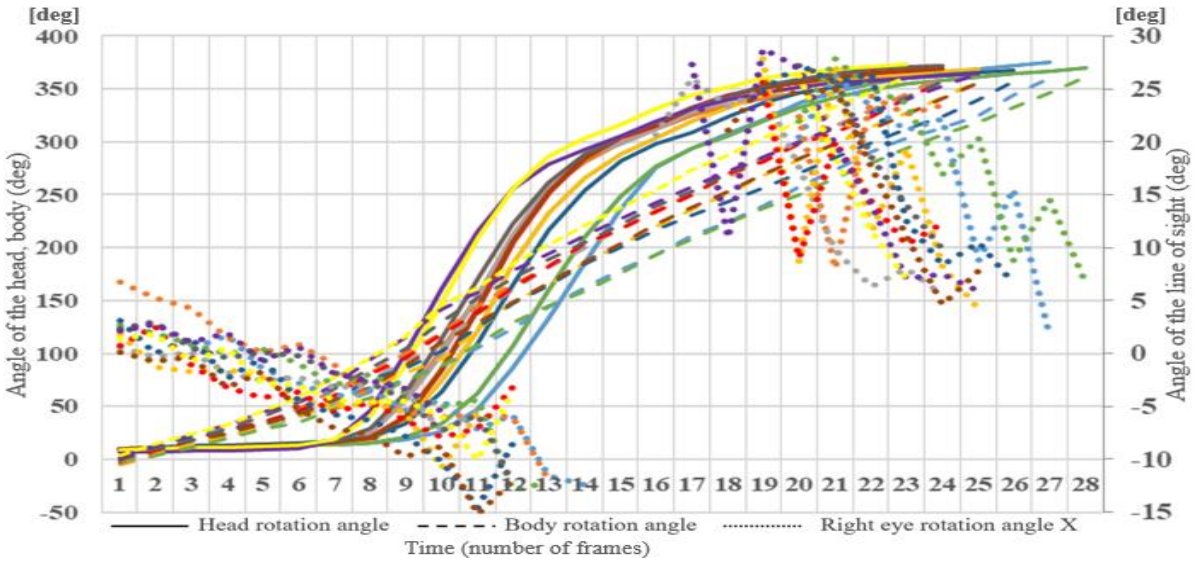


Figure 18 : Angle of gaze, head, body (11 trials)

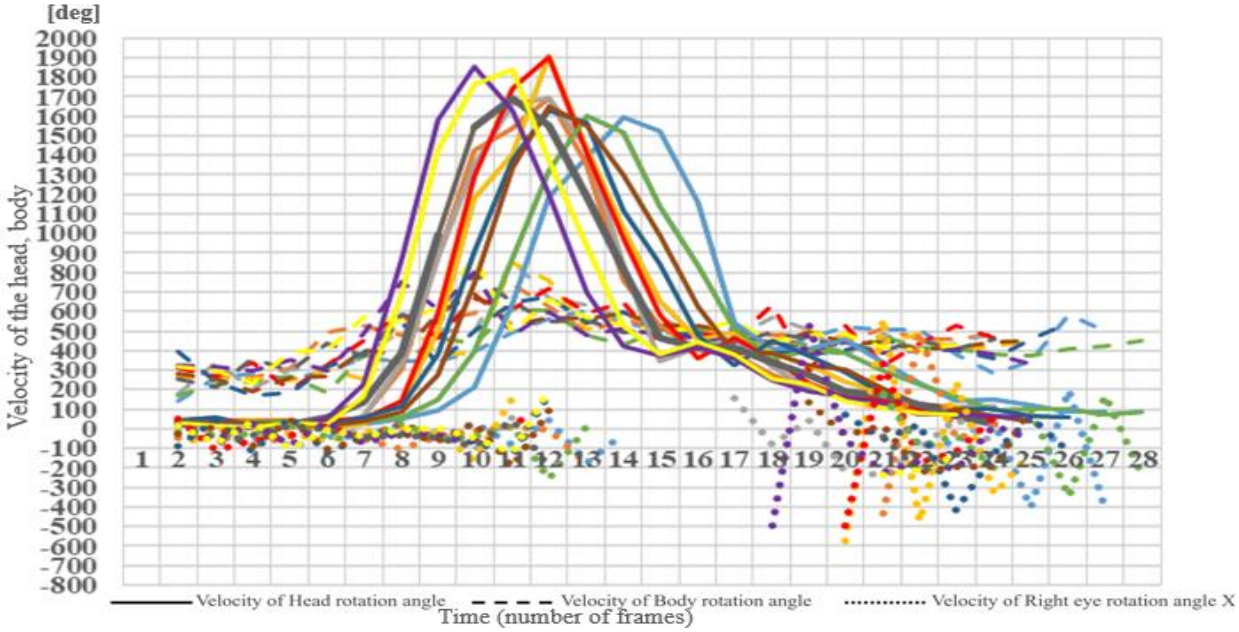


Figure 19 : Velocity of gaze, head, and body rotation angle (11 trials)

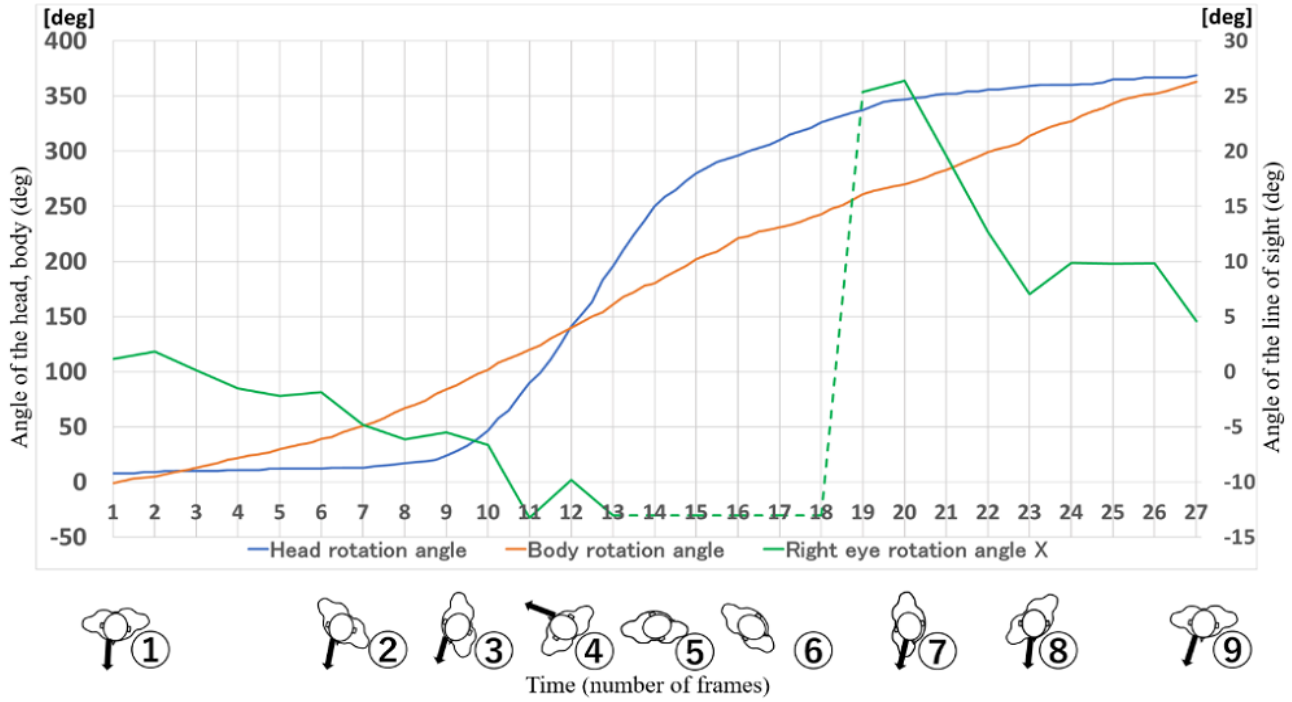


Figure 20 : Angle of gaze, head, body

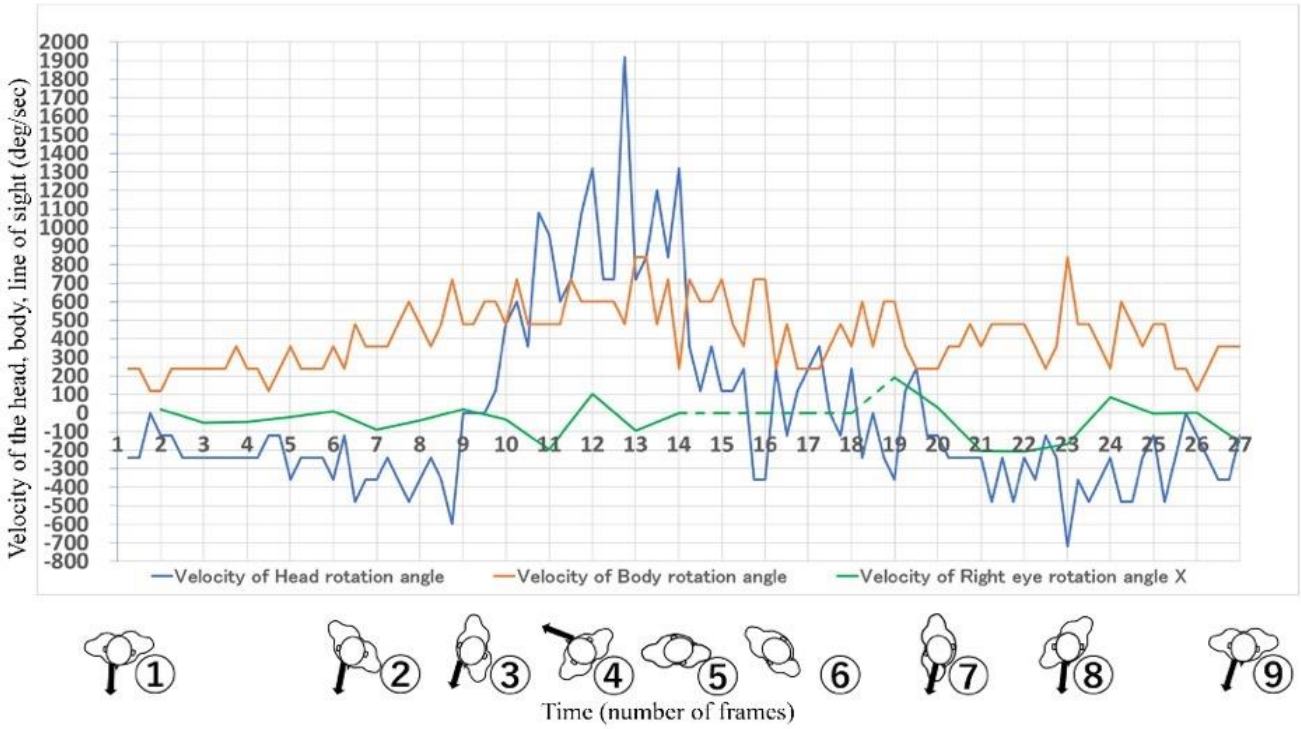


Figure 21 : Velocity of gaze, head, and body rotation angle

References

[1] Y. Sato, "Gaze Behavior during Circles on Pommel Horse", Bulletin of College of Commerce (comprehensive cultural studies), Nihon University, **24**(1,2,3), 221-236, 2019.

[2] K. Hasegawa et al., "Study on the Point of Observation the Moment of Stroke : With Special Reference to the Point of Observation," Metsuke", in Kendo", The bulletin of School of Physical Education, Tokai University, **16**, 55-60, 1987.

[3] T. Natsuhara et al., "Soccer ni okeru senjyutsuteki handan wo tomonau pasu no suikou wo sasaeru ninchi purosesu (Cognitive process that supports the execution of a path with tactical judgment in football)", Japan J. Phys. Educ. Hlth. Sport Sci., **60**(1), 71-85, 2015, doi:10.5432/jjpehss.14001.

[4] E. Hiroshi Edagawa, Monthly Book OCULISTA, No.58, 1, 2018.

[5] Z. Jian, W. Kazuhiko, M. Mai, " A study on anticipatory accuracy and visual search strategy of skilled and novice soccer players:Compared between 1 vs. 1 and 3 vs. 3 situations", Jpn J. Phys. Educ. Health Sport Sci., **53**(1), 29-37, 2008, doi: 10.5432/jjpehss.0478.

[6] Y. Mizusaki, H. Nakamoto, S. Mori, " The Effect of Visual Occlusion during Quiet Eye Training on Free-Throw Accuracy in Novice Basketball Player", Annals of Fitness and Sports Science, National Institute of Fitness and Sports in Kanoya, **47**, 21-28, November, 2013.
[7] M. Takahashi et al., " Visual Behavior of Expert Baton Twirlers during Catching Tasks", The Japanese journal of ergonomics, **46**(1), 31-36, February, 2010, doi: 10.5100/jje.46.31
[8] T. Kato, T. Fukuda, " Visual Search Strategies of baseball batters during the preparatory phase of batting", The Japanese journal of ergonomics, 38(6), 333-340, December, 2002, doi: 10.5100/jje.38.333.
[9] R. Oudejans, van de Langenberg RW, Hutter RI, "Aiming at a far target under different viewing conditions: visual control in basketball jump shooting.", Hum Mov Sci, **21**, 457-480, 2002, doi:10.1016/s0167-9457(02)00116-1.
[10] S. Hüttermann, B. Noël, D. Memmert, "Eye tracking in high-performance sports: Evaluation of its application in expert athletes", IJCSS – **17**(2), 2018, doi:10.2478/ijcss-2018-0011.
[11] S.T. Rodrigues, Joan N Vickers, Andrew Mark Williams, "Head, eye and arm coordination in table tennis", Journal of Sports Sciences, **20**,187-200, April, 2002, doi: 10.1080/026404102317284754.
[12] G. Wood, Mark R Wilson, "Gaze behaviour and shooting strategies in football penalty kicks: Implications of a 'keeper-dependent approach", International journal of sport psychology **41**(3), pp.293-312, November, 2010.
[13] J. Wetzig, M.Reiser, E.Martin, N.Bregenzer, R.J.von Baumgarten, "Unilateral centrifugation of the otoliths as a new method to determine bilateral asymmetries of the otolith apparatus in man", Acta Astronautica, **21**, 519-525, 1990, doi:10.1016/0094-5765(90)90070-2.
[14] G.Y. Menshikova, Aterm I. Kovalev, Oxana A. Klimova, Alexander M. Chernorizod, "Eye movements as indicated of vestibular dysfunction", Perception, **44**, 8-9: 1103-1109, First Published August 14, 2015, doi:10.1177/0301006615594916.
[15] R. Letzter, "The brutal neuroscience of figure skating: How spinning athletes overcome dizziness", https://www.livescience.com/61795-ice-skating-brain-spin-dizzy.html, February 16, 2018.
[16] A. Kheradmand, "Why don't figure skaters get dizzy when they spin", https://www.scientificamerican.com/article/why-don-t-figure-skaters-get-dizzy-when-they-spin/, July 24, 2013.
[17] C. Dehesdin, "Spin, Dance, Jump, Repeat!, Why don't figure skaters get dizzy?", http://www.slate.com/articles/news_and_politics/explainer/2010/02/spin_dance_jump_repeat.html, February 10 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1207-1216 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Evolution of Teaching Approaches for Science, Engineering and Technology within an Online Environment: A Review

Rendani Wilson Maladzhi*, Grace Mukondeleli Kanakana-Katumba

University of South Africa, Department of Mechanical and Industrial Engineering, Johannesburg, South Africa

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 25 November, 2020
Online: 16 December, 2020

Keywords:
Augmented reality
Online teaching and learning
Science
Engineering and technology
Teaching Approaches
Virtual reality
Virtual laboratory

ABSTRACT

The emergence of the COVID-19 pandemic earlier this year destabilised the operations in the education sector, particularly in institutions of higher learning. Most of these institutions are now expected to teach online, assess their students using non-venue examinations and offer remote laboratory practical and experiments. However, many of these institutions were not prepared for such dynamic change in a such short space of time. Consequently, most of the institutions moved their mid-year examinations to the October/November period. In addition, they are finding methods to conduct laboratory practical and experiments. Prominent researchers agree that institutions of higher learning are challenged by the current dispensation where academics are expected to implement new pedagogical approaches and take advantage of information and communications technology (ICT) for teaching and learning. Every module requires a specific teaching method, as a result, academics need to know about various methods available. Generally, the sciences and engineering modules are taught within the classroom environment and require experiential platforms to solidify the theoretical knowledge gained. The current study is aimed at assessing the evolution of teaching approaches and technologies in distance education regarding online delivery particularly in science, engineering and technology. Various search engines such as Google Scholar, Scopus, Sabinet, ProQuest and EBSCO were used to obtain relevant literature to depict popular teaching approaches and the relevant technologies. In order to access relevant literature, various key strings were used. The findings of the study revealed problem-based, apprenticeship and experimental, and competency-based learning as the most popular teaching approaches, particularly between 2010 and 2020. Project-based learning, case-based learning and communities of practice equally share the third position followed by integrative learning. Students and academics show that instant feedback provided by virtual labs yielded encouraging results. This shows that many students and their academics prefer the introduction of the virtual labs in their learning environment, especially from 2012 until recent. The literature further confirms the current findings that institutions of higher learning need to equip academics to be able to integrate technology within their teaching methods for students to continue to learn at ease.

1. Introduction

This paper is an extension of "Online Learning Approaches for Science, Engineering and Technology in Distance Education" presented at the IEEM 2019 Conference in Macau [1]. Higher education institutions are currently endorsing online courses as a business reality since their budgets are now linked to their enrolments [2]. The introduction of an online learning environment has become a long-term business strategy [3] for many institutions worldwide, to the extent that the proportion of students taking at least one online course has increased from 1 in 10 to nearly one-third in 2010 [3]. The accessibility of the internet of things (IOT) in universities worldwide caused them to invest in e-learning systems to enhance their teaching activities to improve their students' learning experiences and performance [4]. Online teaching and learning environments have brought some mixed feelings to students who believe that video lecturers slightly outperform face-

*Corresponding Author: Rendani Wilson Maladzhi, maladrw@unisa.ac.za

to-face lecturers [5] and some students feel interactive online videos are even better [6]. In addition, others feel that online homework is as effective as paper and pencil [7], [8] and intelligent tutoring is seen to be as effective as human tutors [9]. The dynamics above describe how teaching and learning have evolved over the years with reference to five generations of distance education.

The transition through the five generations of distance education has, over the years, witnessed a change in the use the delivery media of teaching and learning that ranged from text and images, sound and video, live television broadcasting to Web 2.0 interactive online technologies such as podcasts, video conferences, virtual spaces etc. [10]-[12]. Based on the above stated transition, the exchange of information between academics and students remained key, but the mode of delivery and the technological tools made all the difference. Hence, various disciplines require different teaching methods accompanied by various technological tools, as is the case with the current research that focusses on the science, engineering and technology sector. Teaching science, engineering and technology in an online environment has been considered an impossible mission for several years [13]. It is a fact that the sciences and engineering disciplines customarily require laboratory work where students perform their practicals to demonstrate their understanding of the theoretical knowledge.

In another instance, an Engineering Graphics Online Course was first introduced in Tianjin University, where 20 colleges and universities participated to cater for almost 1500 students. As a result, this online environment enabled educational resources to be shared in different areas while quality teaching of engineering graphics was made possible [15].

Laboratory classes are an essential component of most science and engineering courses with the potential to achieve several practical and theoretical objectives [16]. Arguably, students must not only learn manipulative techniques, but also link theory to practice, solve different kind of problems, interpret data, interact with staff and other students, and successfully navigate the lab [16]. Learning in this situation can be greatly assisted by academics who are able to guide students through this complex process. However, the effectiveness of laboratory classes is often not achieved to their full potential [16].

The emergence of the COVID-19 pandemic brough along uncertainties for the institutions of higher learning regarding teaching activities and practical and laboratory experiments. On the other hand, it presented distance education institutions with an opportunity to introduce non-venue-based assessments and remote practical and laboratory experiments. The traditional teaching and learning platform greatly benefited by the development of internet and ICT capabilities [17]. Virtual labs afforded students an opportunity to revive their curiosity and creativity while learning from their own respective spaces [18]. The mode of teaching such as synchronous and asynchronous modes made it possible for teaching and learning to take place at a distance. As such, students taught through the asynchronous mode displayed higher cognitive achievement while those taught through the synchronous e-learning mode displayed improved skills acquisition [19]. Hence, the importance of academics understanding the engineering, science and technology (SET) environment to better equip the students appropriately. A synchronous mode of delivery facilitates teaching and learning in real time, whereas an asynchronous mode requires self-paced teaching and learning activities that make use of email, blogs, wikis, discussion boards, web-supported textbooks, hypertext documents, audio or video courses, and social networking using Web 2.0. [19]. The emergence of Massive Open Online Courses (MOOCs) has made it possible for students all over the world to access free credible multimedia resources utilised by academics from recognised universities across the world [20].

Theories about teaching methods do not exist, particularly in post-secondary level, but apprenticeships that resemble learning by teaching, do exist [21]. However, other researchers viewed apprenticeship or experiential learning or learning by doing as an opportunity for SET learners to engage in active learning while reflecting and putting into practice what they learned in the classroom or outside of the classroom [22]. Traditionally, SET experiential learning methods such as laboratory work, workshops; apprenticeships; problem-based learning; case-based learning; project-based learning; inquiry-based learning; and cooperative learning [21], were mainly found within the SET environment. Additive manufacturing [23], augmented reality, virtual labs, online teaching modules and educational aids (Kit) [24]; modelling activities by 3D printers [25]; Engineering Graphics Online Course [15]; online tutorial system [30]; flipped classroom [23]; virtual simulation teaching [26]; virtual reality; the Digital Engineering module were all transformed it into an online teaching environment.

The purpose of this paper is to determine teaching approaches in SET that are gaining popularity due to the evolution of technology within the online environment to improve students' performance. The technological advances over the years contributed positively in turning traditional teaching approaches in science, engineering and technology into suitable online learning.

2. Literature Review

Teaching and learning have evolved over the years due to technological advancements and knowledge transfer between academics and students. It is a fact that teaching modalities also depend on various disciplines offered in various institutions of higher learning.

Experiential learning took a form of learning by doing and as such, is preferable in the SET environment [21]. However, its successful implementation particularly in an online environment requires virtual technologies. As a result, literature that covers teaching methods and respective technologies are viewed as follows: integrative, apprenticeship and experiential learning, problem-based learning, project-based learning, case-based learning, inquiry-based learning, competency-based learning and communities of practice. It has now become apparent that technologies that enhance teaching and learning be exposed to the world for the benefit of the SET sector.

2.1. Integrative approach

The integrative approach is regarded as an innovative practice due to its ability to introduce a concerted approach to mathematics

and science academics in the teaching and learning environment [27]. It permits students to learn about life's origin, then life's evolution, and thereafter about the future of life and humankind in the Earth and Life Sciences, Mathematical modelling concepts, such as the modelling cycle, solves problems regarding real-life situations [27]. Some of the characteristics of the interactive approach are like those in seminars and tutorial [27]. The emergence of distance learning through online environments calls for teaching activities to be conducted differently. In other instances, the teaching of mathematics education in engineering spaces is enriched through modelling activities by 3D printers [25]. As a result, academics should be provided with technological skills to function better and teach effectively.

Virtual learning systems (VLSs) in the academic fraternity play a similar role as enterprise resource planning systems (ERPSs) in the administration arena [28]. Virtual learning systems are known for their role in the teaching and learning avenues that coordinate activities such as learning objectives, class exercises, quizzes and tests in both synchronous and asynchronous communication [29]. The introduction of online tutorial systems within the Australian universities has seen a tremendous increase in the number of students entering the universities [30]. It is a fact that the new developed online teaching approaches introduced a diverse background of skills, motivations and prior knowledge. In addition, the online tutorial system was found to be instrumental in modules such as engineering physics and mathematics, and as such, students did not only pass their examinations, but it also improved the retention rate of first-year engineering students with the financial benefit to the institution [30].

2.2. Apprenticeship and experiential learning

The teaching approaches such as apprenticeship, experiential learning, cooperative learning and adventure learning are commonly used in the fields of SET. They are preferred for their capability to provide students with prospects to experimentally practice their learned experiences [21]. Learning by doing was traditionally instrumental in the mechanical, civil, electrical and electronics vocational fields [21], [27]. However, its transition towards online experimentation and remote/virtual labs presented learners with vast opportunities to control real or virtual equipment at a distance in order to perform scientific experiments [20]. It appears that different researchers refer to virtual laboratories differently; some view it as simulation-based learning, remote experimentation and remote inquiry-based learning [18], [31], [32].

Lack of reputable laboratory experiments contributed negatively to engineering education and resulted in students utilising outdated experiments [18]. The introduction of virtual laboratories gave students an opportunity to make use of good laboratory facilities operated by trained academics as they made use of remote experimentation [18]. It should be noted that virtual laboratories allow students to take ownership of the experiments and the quality control thereof [18]. Virtual laboratories become a multi-institution and yet multidiscipline initiative aimed at benefiting the students within the engineering fields [18]. In its nature, it addresses the issue of lack of good laboratories facilities, as well as trained teachers, by making remote experimentation possible [18]. It provides students with an opportunity to take ownership of teaching and learning space while academics provide minimal guidance. According to [18] and [31], Virtual Laboratory reached over 100 000 students, and are accessible after hours [18,31].

It has been proven that virtual experiments are proven to be more beneficial for teaching and learning [33].

Factually, learning is currently shifting from synchronous in-class settings to asynchronous online platforms so that education can be accessible to a wider audience [34]. It is believed that engineering education always requires laboratory work and hands-on practical components that are not possible to conduct online. However, it was only made possible by the introduction of simulations in fluid mechanics laboratories [34].

Virtual reality (VR) puts the user into a computer-created world that seems real [35]. In addition, VR creates an environment in which the user feels and seems to be moving inside a computer-created virtual world in the same way people move in and around a natural environment; while immersed in the virtual world, the user cannot perceive the real one which still surrounds him [36]. The findings on the study conducted by Luo depicted that students enjoyed taking measurements in the laboratory. However, other students on the other hand were sleeping around with nothing to show. As a result, virtual practicals are preferred as opposed to the traditional way of conducting practical work [26] for mechanical, electrical, civil engineering and other qualifications.

Augmented reality (AR) utilises smart devices and applications such as smart phones, tablets, wearable headsets and immersive technologies that enhance information in real-time to become intelligent [37]. It contemplates real-world events with computer-generated digital information by allowing the user to experience the interaction of the two worlds [38]. AR is capable of making required information available, presented and experienced by users 'augmenting the real world with digital information [39]. The practical application of augmented reality proved itself to be economically and technically viable for societal usage [39]. Hence, it is preferred within the mechanical, industrial, electrical, civil and other engineering disciplines.

2.3. Problem-based learning

Problem-based learning (PBL) enables the students to gain knowledge about and apply it to different problems they investigate [40]. As they seek solutions for the problems at hand, students have an opportunity to argue and debate based on their own understanding and the acquired facts [40]. PBL teaches students not to depend on their lecturers for information but to rather resort to other sources to make informed decisions. It is important to note that PBL is more suited for student-centred and community-based learning environments [41].

Additive manufacturing (AM) is the preferred technology of teaching used by academics to introduce topics that are industry and academic experts, laboratory and final engineering projects. AM is a PBL in nature due to its innovative features dealing with devices for people with disabilities. Other PBL-related projects include the development of an opener for medicine containers, a device for pouring liquids for people who have Parkinson's disease, and a 3D puzzle for people who have visual impairments.

[23]. It has the capability to encourage students' creativity, learning motivation and engagement [23].

The flipped classroom (FCM) has emerged as a learning model that increases students' active learning, collaboration and scaffolding during the learning process through better allocation of teaching time [43]. As such, the FCM method recognises that teaching activities within face-to-face environment should not be about academics teaching but should rather focus on creating an environment for students to effectively collaborate with each other while guided by the academics through scaffolding methods [44], [45]. It is believed that digital technologies have made it possible for teaching and learning to take place without students being confined to a classroom, and that learning can take place from anywhere as long as study material, videos, online quizzes etc. are available.

The FCM method is preferred by both practitioners and researchers in various subjects such as Mathematics, ICT, Social Studies, Humanities etc. [46]-[50], [89].

The results from various studies further confirmed the FCM method is an effective technology that raises the performance of students who face difficulties in face-to-face educational contexts, and equally assists academics and institutions to improve learning conditions [51]. The FCM's nature affords academics an opportunity to scaffolding and create an enabling environment for engagement with hands-on practices [51].

Traditionally, the Digital Engineering lectures were limited to the face-to-face mode of delivery where students only listened to the lecturer delivering the lecture [24]. It was indeed a tremendous challenge since there was no point of demonstrating the application of electronic circuits, digital electronics and high logic circuits for students to apply in industry [24]. The introduction of online teaching methods utilising the Proteus software to enhance skills of learners began to make a difference [24]. Both problem-based learning and competency-based learning suited the Digital Engineering module. This is due to the application of the software that demonstrates teaching and learning activities to students, which also requires both online and off-line learning with computer-based learning assistance [24].

Arguably, teaching practical science at a distance has been a challenge for a while in most institutions across the world. However, in July 2013, the Open University (OU) launched the Wolfson Open Science Laboratory (OSL), which offered a variety of practical science experiences to their distance learning students [13]. With this initiative, students are provided an opportunity to observe, investigate, gather and analyse data. In fact, numerous studies confirmed the effectiveness of PBL because it fostered active learning.

2.4. Case-based learning

Case-based learning (CBL) is an interactive group learning pedagogy preferable in various disciplines including business, law, health [14], [52] and in sciences [53]. Case-based learning is defined as an educational approach that allows students to work in small, collaborative groups to solve problems [54]. It is gaining popularity at a faster pace within the online fraternity [14]. (2018: 283-300). Furthermore, it stimulates students' thinking and discussions amongst students [14]. It is paramount to mention that the case presentation teaching method is frequently utilised in medicine [52].

Case-based is predominantly used where complex and interdisciplinary themes openly discussed [27]. Application oriented teaching platform became a reality in various university through online course construction [17].

According [55], civil engineering requires modern teaching methods such as hybrid, project, case and discussion into the classroom while utilising engineering examples to enhance teaching and learning.

2.5. Project-based learning

Project-based learning (PJBL) is a twenty-first century innovative learning approach that teaches various strategies to ensure students' success [56], particularly in higher learning environments. With this approach, students are afforded an opportunity to experience the reality of the real-world while boosting their communication skills and interdisciplinary learning [57]. However, PBL places more emphasis on learning through internal rather than external projects. Industrial collaboration projects build partnerships and provide opportunities to work on real problems, often seeking multidisciplinary approaches. This study explores and suggests the benefits that students derived from their participation in an external research and consultancy project at a University in Oman [57].

In 2016, Ralph conducted a research study looking into the post-secondary project-based learning in science, technology, engineering and mathematics [58]. The findings indicated that students benefitted from the project-based method regarding content knowledge, and interdisciplinary skills. Their learned skills would be instrumental in their future classes and career settings. However, most students did not enjoy the teamwork that PJBL brings but some learned from other group members [58]. All in all, the findings recommended that PJBL should be implemented institutionally rather than per discipline.

Typically, students using the project-based learning method work together as a team in order to solve the given problem by developing the product and development process [59]. PJBL is found to be an efficient teaching method in the twenty-first century as it promotes critical thinking, problem-solving skills, interpersonal communication, information and media literacy, collaboration, teamwork and leadership, creativity and innovation [60].

Other researchers conducted a study on Reform Practice of Engineering Drawing Courses in Chinese Colleges in 2019 [61]. The findings highlighted amongst others the emergence of the MOOC platform of outcomes-based education for engineering education models. All mechanical drawing courses implemented project-based teaching methods that were suited for the online environment. The MOOC platform catered from various online learning engagement models for a variety of questions, discussions sessions and projects [61].

2.6. Inquiry-based learning

The inquiry-based instructional approach is learner-centred and requires the learner to observe, generate questions, discover existing gaps and find appropriate resources to overcome these

gaps [62]. These students take charge of their own learning by planning, monitoring and evaluating their progress made [63]. In 2017, researchers conducted a research study about preparing teacher-students for the 21st century. An inquiry-based teaching method was utilised when 110 physics and chemistry teachers were involved. In their tasks, they managed to successfully explore and analyse the problems and formulated hypotheses and conducted their experiments. ICT applications were integrated by means of smart phones with cameras, laptop computers and software applications. As a result, the experiments were fruitful as they worked together with one common purpose [63].

It is a fact that in an inquiry-based phenomenon, students become their own teachers and take lead in the teaching and learning processes. The emergence of technology makes it easy for students to access their laboratory through the internet of things without them being in the physical laboratory [64-66]. It is imperative to note that virtual laboratory makes things bearable and cost-effective as compared to traditional laboratories as they are easy to maintain and safe to use. The science, engineering, and technology students find virtual laboratories user-friendly as they interact in inquiry-based classes where they can implement and analyse their own experiments and learn by using virtual objects and apparatus [67]. This teaching method also equips students with critical thinking, and innovative and team skills, which are valued for the twenty-first century job market [67]. A previous study initiated the STEM build website to coordinate online teaching and learning where instructors, markers, researchers, and Universal Design for Learning experts can easily interact [68].

Previously, engineering drawing was hand-based, now, online tutors are currently assisting these students. The utilisation of artificial intelligence makes it possible to provide real-time human-like feedback to students [69].

The emergence of the COVID-19 epidemic has seen the closure of campuses for schools and universities to ensure the safety of their students. As a result, online teaching and learning has become the reality in the 2020 academic year [70]. Online teaching and learning were found to be conducive for teaching students about circuits in electronic engineering [70] (Hu, 2020).

2.7. Competency-based learning

Over the past years, competency-based learning has rapidly gained momentum within the higher education space [71] and has been supported by ICT [74]. It is noticeable that over 896 institutions (universities and colleges) across the world prefer the competency-based learning method in their teaching and learning activities [71, 72]. Within competency-based learning, students depend more on their competency levels than on other people [21, 73]. Competency-based learning enables students to return to one or more competencies that are not yet mastered to avoid having to repeat courses as a result of failing [74]. Competency-based learning equips students with explicit and transferable learning objectives that provide them with personalised support and formative feedback so they can foster critical-thinking, problem-solving, communication and collaboration skills [74]. Competency-based models are dependent on measurable assessments linked to the related learning outcomes [75]. In most cases, competency-based learning is demonstrated through both formative and summative assessments [75]. However, at the Minnesota State University, both assessments methods (summative and formative) were successfully utilised for the Honors Program to assess the competency level of the students over time through electronic portfolios. As such, this assessment method replaces the standard research thesis required in numerous traditional Honors Programs [76].

In most instances, the competency-based method is used by professional bodies to measure the level of acquired competencies of graduates regarding their required training expected in industries [77]. Employers like to state to the institutions of higher learning what competencies they expect from the graduates. Hence, the importance of institutions of higher learning to always consult the companies based on what they expect their graduates to possess is imperative [78].

2.8. Communities of practice

The communities of practice approach combine experiential learning, social constructivism and connectivism, which allows people to work together and give each other moral support [79] and compassion [80] as they study. This method permits students to put their skills into practice and largely share their experiences [10, 81]. Communities of practice (COP) allow the combination of experiential learning, social constructivism and connectivism can while demonstrating the limitations of classifying learning theories. In addition, COP allows group of engineers to work together in order to solve general challenges [81-84]. It is argued that for teaching and learning to effectively take place, precedence should be on cognitive, social and teacher presence [85].

3. Research Methodology

As mentioned before, the current study aims to determine teaching approaches in SET, which are gaining popularity due to the technological evolution over the years. It also aims to ascertain the state of distance education evolution with reference to technology adoption within the Science, Engineering and Technology qualifications in the institutions of higher learning. As mentioned earlier, engineering students were traditionally afforded opportunities to experientially practice learned knowledge in a laboratory setup. It is therefore of vital importance to assess the literature if science, engineering and technology students in higher education are to continue receiving laboratory experiments while studying in an online environment. The study targeted literature from 2010 to 2020 to determine teaching approaches that are gaining popularity in SET within an online environment. The reviewed literature was obtained from various search engines such as Google Scholar, Scopus, Sabinet, ProQuest and EBSCO.

The following search strings were used to search the open-journal sites:

- Teaching approaches for science, engineering and technology
- Teaching approaches for distance education
- Teaching technologies for SET in distance education
- Integrative approach for distance education in the SET
- discipline
- Apprenticeship and experiential learning for distance

- education in SET
- Problem-based learning for distance education in SET
- Project-based learning for distance education in SET
- Competency-based learning for distance education in
- SET
- Inquiry-based learning for distance education in SET
- Case-based learning for distance education in SET
- Community of practice-based learning for distance
- education in SET
- AR in SET
- VR in SET
- Virtual laboratory in SET

The research study concentrated only on journals that focussed on institutions of higher learning within the SET discipline. All other journal papers focussing on primary schools and industry were excluded from the study. The results of the reviewed literature were presented in both table and graph formats for appropriate analysis. Pareto analysis was conducted to determine teaching methods and their respective technologies preferred in the SET qualifications in an online environment.

4. Research Findings and Discussion

Theories such as cognitive behaviour, social constructivism and connectivism formed the basis of the different approaches discussed in the current study. It is clearly indicated that teaching approaches for SET always pay attention to the practical implications on what students theoretically learn in a class setup. The evolution of distance education has seen traditional teaching approaches such as competency-based learning, project-based learning, problem-based learning, inquiry-based learning, case-based learning, and apprenticeship and experiential learning as becoming relevant in the twenty-first century through technologies such as VR, AR, virtual laboratories, AM, FCM, Proteus software, Wolfson Open Science Laboratory, Virtual Microscope etc. It should be noted that the findings of this study are entirely based on the reviewed literature sources. The results in Table 1, depict how different approaches transited from traditional face-to-face to online virtual practical so that students could continue putting their knowledge into practice.

The results from both Figure 1 and Table 1 show the popularity of these teaching approaches in an online environment emerging from 2010 until recently. The results also display that problem-based, apprenticeship and experiential learning, competency-based learning and inquiry-based learning are more popular in an online environment, particularly in SET. It suffices to mention that all highlighted teaching approaches are preferred in SET within online space but differ due to their popularity over the years.

Figure 1 elaborates on the popularity of teaching approaches suitable for online teaching and learning as per the author's contributions. The study results show that problem-based learning has gained popularity and exceeded all the other approaches, followed by apprenticeship and experiential approaches. The third important approach was found to be competency-based. However, case-based, project-based and community of practice-based approaches shared the fourth position for popularity. Last, but not the least, integrative approach proved to have less popularity. It is important to emphasise that all these approaches are applicable in an online environment within SET disciplines. It is evidence that from 2009,2010, 2011, 2012 these various teaching approaches were utilised mainly in face-to-face institutions. However, the emergence of the internet of things made it possible for utilising them for open and distance education, particularly 2017, 2019 and recently. In recent years as from 2015 until recent Virtual labs, virtual reality gave problem-based, apprenticeship and experiential learning as well as competency-based approaches impetus within the science and engineering space as seen in the table below. It is evidence from Figure 1 and Table 1 that other technologies such as LMS, FCM in education, 3D printers and AM in the manufacturing spaces continued to gain recognition as from 2016 and prominently in 2019. The study rightfully, depicted the gap in open and distance learning environment where similar teaching approaches in science and engineering where applicable.

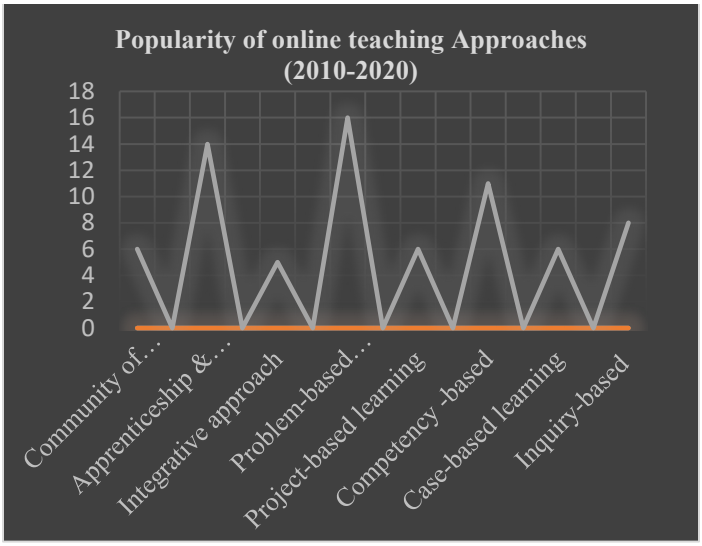


Figure 1: Popularity of teaching approaches

Table 1: Authors who endorsed teaching approaches

Authors endorsed these approaches	Learning Approaches & Preferred technologies	Used in SET	Online Learning delivery
[2011],[2015] [2017],[2018] [2019],[2019]	Community of Practice learning **Technologies:** Learning management system (LMS)	(6)	Yes
[2011][2012] [2013][2013] [2014][2015] [2017][2018] [2018][2018] [2018][2019] [2019][2020]	Apprenticeship & experiential **Technologies:** Virtual labs, AR, VR, LMS,	(14)	Yes

[2016][2017] [2018][2018] [2019]	Integrative approach **Technologies:** Virtual learning Systems; 3D printing	(5)	Yes
[2010][2017] [2017] [2017][2018] [2019][2020] [2020]	Inquiry-based learning **Technologies:** LMS	(8)	Yes
[2012][2014] [2014][2015] [2015][2017] [2015][2016] [2016][2017] [2017][2018] [2018][2018] [2019][2019]	Problem-based learning **Technologies:** AM; FCM; Proteus software; Wolfson Open Science Laboratory; Virtual Microscope	(16)	Yes
[2010][2016] [2016][2017] [2019][2020]	Project-based learning **Technologies:** FCM and LMS	(6)	Yes
[2015][2015] [2016][2016] [2016][2016] [2015][2016] [2017][2019] [2020]	Competency - based **Technologies:** FCM and LMS	(11)	Yes
[2011][2016] [2018][2018] [2019][2019]	Case-based learning **Technologies:** FCM and LMS	(6)	Yes

5. Conclusion

The emergence of the COVID-19 pandemic has provided an opportunity to many institutions to introduce various technologies to enhance teaching and learning activities for students and academics. Much more importantly, many institutions were not ready to offer non-venue-based examinations and remote laboratory practical. As a result, some institutions forged ahead and offered mid-year non-venue-based examinations while other institutions postponed all their examinations to October/November 2020. However, the issue of credible examination remained a concern to these institutions. These institutions want their students to write non-venue-based examinations, which are remotely invigilated by means of proctoring system to maintain their quality standards.

Equally important, institutions which offered campus-based practical experiments opted for alternative practical such as video-based and virtual practical to their students. Many of the institutions are only considering virtual practical now, as the COVID-19 situation does not show any sign of coming to an end soon. Hence, the importance of the current study so that institutions across the world can see various teaching approaches and technologies suitable for the online environment.

Traditionally, all practical experiments in science, engineering and technology institutions were conducted in physical laboratories. As such, most distance education institutions across the world collaborated with other campus-based institutions for their students to conduct laboratory work at those institutions, see Figure 1 and Table 1. However, such institutions always complained of the quality of practical work done in other institutions. In most cases, the Engineering Council of South Africa (ECSA) was concerned about the teaching and learning activities pertaining to engineering courses offered by distance education and those that were conducted online. ECSA, is a professional body that accredits all engineering qualifications in South Africa. They were more concerned about the practical work and the credibility of the assessments would ensue. Hence, the importance of the current study to publicise what other institutions across the world are doing regarding the virtual spaces utilising AR and VR technologies.

The findings of the current study indicate that virtual laboratories are the solution for most of the SET institutions operating in an online environment. The traditional teaching methods for civil engineering courses were no longer appropriate due to technological advancements [86]. Hence, the adoption of online courses for efficient learning. The introduction of micro-videos and final assessment methods made it possible for students in other areas to be involved [86]. Notably, most of the distance education institutions are already employing teaching approaches such as problem-based (16), apprenticeship and experiential learning(14), competency-based learning (11), inquiry-based learning (8), community of practice learning (6), case-based learning (6) and integrative-based learning (5).

Education technology has made it possible for students to learn outside of the classroom when they are at home. Both synchronous and asynchronous learning allows students to access teaching materials in their own comfortable space [87]. The study confirmed that many students in the engineering discipline developed an interest in teaching and learning where they were exposed to various assessment methods, linking theory for practice and case discussion [86]. This statement shows that the introduction of new technologies within the institutions of learning are enabling students to perform better, see Figure 1 and Table 1. The usage of different teaching methods plays an important role towards students' performances [88].

Both students and academics show that instant feedback provided by virtual laboratories yielded encouraging results [67]. Virtual laboratories began to gain popularity as early as 2012 and increasingly so in this digital era. This shows that many students and their academics prefer the introduction of the virtual laboratories in their learning environment. The institutions of higher learning are challenged by the current dispensation where

academics are expected to implement new pedagogical approaches and take advantage of ICT for teaching and learning, as can be seen in both Figure 1 and Table 1. As a result, these institutions need to equip the academics to be able to integrate technology within their teaching methods so that students continue to learn at ease. It is affirmed that the world-wide education system should incorporate the twenty-first century skills into curricula [63]. The current study acknowledges the challenges brought by the COVID-19 pandemic into the online institutions of learning, but the introduction of various teaching methods along with various technologies might bring relief if implemented correctly. Institutions are confronted with an opportunity to assess their current resources and equip their academics appropriately so that they can structure their assessments and practicals accordingly.

The current study proposes that a future study assess the implementation of various teaching approaches along their suitable technologies within the SET institutions in the southern Africa region.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was made possible by the funding received from the National Research Foundation of South Africa and the research unit of the University of South Africa.

References

[1] M. G. Kanakana-Katumba, R. Maladzhi, "Online Learning Approaches for Science, Engineering and Technology in Distance Education," In 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 930-934. IEEE, 2019, doi.org/10.1109/IEEM44572.2019.8978892.

[2] C. Hall, W. Swart, S. Duncan, "Balancing customer needs and standards in higher education," Quality Approaches in Higher Education, **3**(1), 2–7, 2012.

[3] E.I. Allen, J. Seaman, "Changing course: Ten years of tracking online education in the United States," Sloan Consortium. PO Box 1238, Newburyport, MA 01950, 2013.

[4] T. Kattoua, M. Al-Lozi, M., A.A. Alrowwad, "A review of literature on E-learning systems in higher education," International Journal of Business Management & Economic Research, **7**(5), 754-762, 2016.

[5] P.A. Cohen, B.J. Ebeling, J.A. Kulik, A "meta-analysis of outcome studies of visual-based instruction," Educational Technology Research and Development, **29**(1):26–36, 1981.

[6] J. Barbara, A. McNeil, "Meta-analysis of interactive video instruction: A 10 year review of achievement effects," PhD thesis, University of Idaho, 1989.

[7] S.W. Bonham, D.L. Deardorff, R.J. Beichner, "Comparison of student performance using web and paperbased homework in college-level physics," Journal of Research in Science Teaching, **40**(10):1050–1071, 2003. https://doi.org/10.1002/tea.10120

[8] H. Fynewever, "A comparison of the effectiveness of web-based and paper-based homework for general chemistry," The Chemical Educator, **13**(4):264–269, 2008.

[9] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, **46**(4), 197-221, 2011, doi.org/10.1080/00461520.2011.611369.

[10] T. Bates, "Second Wave of DE and History of the Open University," 2011.

[11] J.F. Heydenrych and P. Prinsloo, "Revisiting the five generations of distance education: Quo vadis ?," Progressio. **32**(1):5–26, 2010.

[12] L.S. Mbati, "Exploring media blends for constructivist learning in open and distance e-learning (ODeL) environments," Diss, 2013.

[13] T. W. Argles, M. Brodeur, N. St J. Braithwaite, "Delivering authentic online practical science teaching–geoscience perspectives from the OpenScience Laboratory," In International Geological Congress, Abstracts, 35, 2017.

[14] N. Wengrowicz, W. Swart, R. Paul, K. Macleod, D. Dori, Y.L. Dori, "Students' collaborative learning attitudes and their satisfaction with online collaborative case-based courses, " American Journal of Distance Education, **32**(4), 283-300, 2018, doi.org/10.1080/08923647.2018.1511509.

[15] S. Jiang, J. Xu, "Construction and Commissioning of Engineering Graphics Online Course," In Education Teaching Forum, **12**, 102-103, 2018.

[16] N. Brouwer, G. Fleerackers, N. H. Majcen, I. Maciejowska, C. McDonnell, and M. Mocerino, "Online course to improve university laboratory teaching practice," VIRT&L-COMM, 10, L-COMM, 2016.

[17] J. Wang, "Research on the classroom teaching reform of financial management" under the background of online course construction," Think Tank Era, **34**, 66-67, 2019.

[18] R. Bose, "Virtual Labs Project: A Paradigm Shift in Internet-Based Remote Experimentation," Access IEEE, **1**, 14, 2013, doi.org/10.1109/ACCESS.2013.2286202.

[19] E.G. Dada, A.H. Alkali, D. O. Oyewola, "An Investigation into the Effectiveness of Asynchronous and Synchronous E-learning Mode on Students' Academic Performance in National Open University (NOUN)," Maiduguri Centre. International Journal of Modern Education & Computer Science, **11**(5), 2019. doi: 10.1002/tea.10120

[20] A. Mikroyannidis, A. Gómez-Goiri, A. Smith and J. Domingue, "Online experimentation and interactive learning resources for teaching network engineering," IEEE. Global Engineering Education Conference (EDUCON), Athens, 181-188, 2017, doi.org/10.1109/EDUCON.2017.7942845.

[21] T. Bates, "Teaching in a Digital Age: Guidelines for Designing and Learning," Tony Bates Associates Limited, 2019.

[22] A. Y. Kolb, D.A. Kolb, "The learning way: Meta-cognitive aspects of experiential learning," Simulation & gaming, **40**(3), 297-327, 2009, doi.org/10.1177/1046878108325713.

[23] A. Stern, Y. Rosenthal, N. Dresler, Ashkenazi, D, "Additive manufacturing: An education strategy for engineering students," Additive Manufacturing, 27, 503-514, 2019, doi.org/10.1016/j.addma.2019.04.001.

[24] N. Kholis, M. S. Zuhrie, R. Rahmadian, "Innovation Online Teaching Module Plus Digital Engineering Kit with Proteus Software through Hybrid Learning Method to Improve Student Skills," Materials Science and Engineering, 336, 012036, doi.org/10.1088/1757-899X/336/1/012036.

[25] G. Guler, C. Sen, Z.S. Ay, A. Ciltas, "Engineering Skills that Emerge During Model-Eliciting Activities (MEAs) Based on 3d Modeling Done with Mathematics Pre-Service Teachers," International Journal of Education in Mathematics, Science and Technology, **7**(3), 251-270, 2019.

[26] L. Luo, "Analysis on Online and Offline Blended Teaching Design Taking Engineering Survey," Course of Non-surveying and Mapping Engineering Specialty as an Example. In 3rd International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2019), Atlantis Press, 2019.

[27] N. Mestrinho, B. Cavadas, "Innovation in Teacher Education: An Integrative Approach to Teaching and Learning Science and Mathematics," in Multidisciplinary Digital Publishing Institute Proceedings, **2**(21), 1343, 2018, doi.org/10.3390/proceedings2211343.

[28] A.O. Akande, J.P. Van Belle, "The use of Software as a Service to Facilitate Quality Learning Experience for Students in Higher Education Institutions," Simulation (ISSN 2518-0959 Online), **2**, 2017, doi.org/10.24178/ijsms.2017.2.2.08.

[29] P. Padayachee, S. Wagner-Welsh, H. Johannes, "Online assessment in Moodle: A framework for supporting our students," South African Journal of Higher Education **32**(5): 211-235, 2018, doi.org/10.1080/03043797.2015.1095162.

[30] R. Bhathal, "An appraisal of an online tutorial system for the teaching and learning of engineering physics in conjunction with contextual physics and mathematics, and relevant mathematics," European Journal of Engineering Education, **41**(5), 504-511, 2016.

[31] S. Frerich, D. Kruse, M. Petermann, A. Kilzer, "Virtual Labs and Remote Labs: Practical experience for everyone," Proc. IEEE Global Engineering Education Conference (ED UCON), 312-314, 2014, doi.org/10.1109/EDUCON.2014.6826109.

[32] J. M. Marquez-Barja, N. Kaminski, F. Paisana, C. Tranoris and L. A. DaSilva, "Virtualizing testbed resources to enable remote experimentation in online telecommunications education," Proc. IEEE Global Engineering Education Conference (EDUCON), 836-843, 2015, doi.org/10.1109/EDUCON.2015.7096069.

[33] L. Li, Y. Chen, Z. Li, D. Li, F. Li, H. Huang, "Online Virtual Experiment Teaching Platform for Database Technology and Application," 13th International Conference on Computer Science & Education (ICCSE), Colombo, 1-5, 2018. http://dx.doi.org/10.1109/ICCSE.2018.8468849

[34] G. Altuger-Genc, Y. Han, Y. Genc, "Towards Simulation Aided Online Teaching: Material Design for Applied Fluid Mechanics," International

Journal of Online and Biomedical Engineering (iJOE), **14**(12), 112-125. 2018, doi.org/10.3991/ijoe.v14i12.8630.
[35] M. Hincapié, A. Caponio, H. Rios, E.G. Mendívil, "An introduction to Augmented Reality with applications in aeronautical maintenance," In 2011 13th International Conference on Transparent Optical Networks, 1-4, IEEE, 2011, doi.org/10.1109/ICTON.2011.5970856.
[36] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.A. Vervier, I. Santos. "Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control," In Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, 340-351, 2019.
[37] H.K. Wu, S.W.Y. Lee, H.Y. Chang, J.C. Liang, "Current status, opportunities and challenges of augmented reality in education," Computers & education, **62**, 2013, 41-49, doi.org/10.1016/j.compedu.2012.10.024.
[38] R. Azuma, R. n.d, "A Survey of Augmented Reality. iitk.ac.in". Retrieved August 2020.
[39] T. Olsson, T. Kärkkäinen, E. Lagerstam, L.Ventä-Olkkonen, "User evaluated of mobile augmented reality scenarios" Journal of Ambient Intelligence and Smart Environments, **4**, 2012, 29-47.
[40] N.S. Ismail, J. Harun, M.A.Z.M. Zakaria, and S.M. Salleh, "The effect of Mobile problem-based learning application DicScience PBL on students' critical thinking," Thinking Skills and Creativity, **28**, 177-195. 2018, doi.org/10.1016/j.tsc.2018.04.002.
[41] S. Han, R. Capraro, M.M. Capraro, "How science, technology, engineering, and mathematics (STEM) project-based learning (PBL) affects high, middle, and low achievers differently: The impact of student factors on achievement," International Journal of Science and Mathematics Education, **13**(5), 1089-1113, 2015, doi.org/10.1007/s10763-014-9526-0.
[42] S. Ford, T. Minshall, "Invited review article: Where and how 3D printing is used in teaching and education," Additive Manufacturing, **25**, 131-150. 2019, doi.org/10.1016/j.addma.2018.10.028.
[43] J. Bergmann, A. Sams, "Flip your classroom: Reach every student in every class every day International Society for Technology in Education," Washington, DC, 2012.
[44] S.J. DeLozier, M.G. Rhodes, "Flipped classrooms: A review of key ideas and recommendations for practice," Educational Psychology Review (2016), 1-11, doi.org/10.1007/s10648-015-9356-9.
[45] Y. Chen, Y. Wang, Kinshuk, N.S. Chen, "Is FLIP enough? Or should we use the FLIPPED model instead?" Computers & Education, **79**, 16-27, 2014, doi.org/10.1016/j.compedu.2014.07.004.
[46] M. Katsa, S. Sergis, D.G, Sampson. "Investigating the potential of the flipped classroom model in K-12 mathematics teaching and learning Proceedings of the 13th international conference on cognition and exploratory learning in digital age", 2016.
[47] T. Muir, V. Geiger, "The affordances of using a flipped classroom approach in the teaching of mathematics: A case study of a grade 10 mathematics class," Mathematics Education Research Journal, **28**(1), 149-171, 2016 doi.org/10.1007/s13394-015-0165-8.
[48] M.N. Giannakos, J. Krogstie, N. Chrisochoides. "Reviewing the flipped classroom research: Reflections for computer science education," Proceedings of the computer science education research conference, 23-29, 2014, doi.org/10.1145/2691352.2691354.
[49] S. Reza, M.I. Baig, "A study of inverted classroom pedagogy in computer science teaching," International Journal of Research Studies in Educational Technology, **4**(2), 19-30, 2015, doi.org/10.5861/ijrset.2015.1091.
[50] V. Aidinopoulou, D.G. Sampson, "An action research study from implementing the flipped classroom model in primary school history teaching and learning," Journal of Educational Technology & Society, **20**(1), 237-247, 2017.
[51] S., Sergis, D.G. Sampson, L. Pelliccione, "Investigating the impact of Flipped Classroom on students' learning experiences: A Self-Determination Theory approach," Computers in Human Behavior, **78**, 368-378, 2018, doi.org/10.1016/j.chb.2017.08.011.
[52] B. Turk, S. Ertl, G. Wong, P.P. Wadowski, H. Löffler-Stastka, "Does case-based blended-learning expedite the transfer of declarative knowledge to procedural knowledge in practice?" BMC Medical Education, **19**(1), 1-10, 2019, doi.org/10.1186/s12909-019-1884-4.
[53] G. Marcus, R. Taylor, A. Ellis, "Implications for the design of online case based learning activities based on the student blended learning experience," in Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference, 557-586, 2004.
[54] P. Nicklen, J.L. Keating, S. Paynter, M. Storr, S. Maloney, "Remote-online case-based learning: a comparison of remote-online and face-to-face, case-based learning-a randomized controlled trial," Education for Health, **29**(3), 195, 2016.
[55] Newson, T.A, N.J. Delatte, "Case methods in civil engineering teaching," Canadian Journal of Civil Engineering, **38**(9), 1016-1030, 2011.
[56] S. Bell, "Project-based learning for the 21st century: Skills for the future," The clearing house, **83**(2), 39-43, 2010, doi.org/10.1080/00098650903505415.
[57] R. Belwal, S. Belwal, A.B. Sufian, A. Al Badi, "Project-based learning (PBL): outcomes of students' engagement in an external consultancy project in Oman," Education and Training, 2020.
[58] R. A. Ralph, "Post-secondary project-based learning in science, technology, engineering and mathematics," Journal of Technology and Science Education, **6**(1), 26-35, 2016, doi.org/10.3926/jotse.155.
[59] D. Kokotsaki, V. Menzies, A. Wiggins, "Project-based learning: A review of the literature Improving Schools," **19**(3) (2016), 267-277.
[60] S. K.W. Chu, R.B. Reynolds, N.J. Tavares, M. Notari, C.W.Y. Lee, "21st Century skills development through inquiry-based learning," Singapore: Springer Singapore. 2017, doi.org/10.1007/978-981-10-2481-8.
[61] Y. Dong, J. Song, S. Yi, X. Li, "Reform Practice of Engineering Drawing Courses in Chinese Colleges by the Blended Teaching Method with Different Teaching Modes and Resources Based on the XuetangX MOOC Platform," In 2019 2nd International Conference on Education, Economics and Social Science (ICEESS 2019). Atlantis Press, 2019, doi.org/10.2991/iceess-19.2019.70.
[62] C.E. Hmelo-Silver, C.A. Chinn, Chan, A.M. O'Donnell, "International handbook of collaborative learning," New York, NY: Routledge, 2013, doi.org/10.4324/9780203837290.
[63] P. Häkkinen, S. Järvelä, K. Mäkitalo-Siegl, A. Ahonen, P. Näykki, T. Valtonen, "Preparing teacher-students for twenty-first-century learning practices (PREP 21): A framework for enhancing collaborative problem-solving and strategic learning skills Teachers and Teaching," **23**(1), 25-41, 2017, doi.org/10.1080/13540602.2016.1203772.
[64] C. Soudien, "Transformation in higher education : A briefing paper," 2010.
[65] K. Kaen, "A Contextual O nline Game based on Inquiry Learning Approach for Improving Students' Learning Performance in a Chemistry Course," 601-608, 2017.
[66] V. Fruiht, "Project-Based Curriculum as a Guide for the Re-Sequencing of Discipline-Specific Statistics Courses," 2018.
[67] T. Lynch, I. Ghergulescu, "Review of virtual labs as the emerging technologies for teaching STEM subjects," In INTED2017 Proc. 11th Int. Technol. Educ. Dev. Conf. 6-8 March Valencia Spain, 6082-6091, 2017, doi.org/10.21125/inted.2017.1422.
[68] M. V. Ramirez, C.L. Gordy, "STEM build: an online community to decrease barriers to implementation of inclusive tactile teaching tools," Journal of Microbiology & Biology Education, **21**(1), 2020, doi.org/10.1128/jmbe.v21i1.1963.
[69] E. Hilton, B. Williford, W. Li, T. Hammond, J. Linsey, "Teaching Engineering Students Freehand Sketching with an Intelligent Tutoring. Inspiring Students with Digital Ink: Impact of Pen and Touch on Education", 135, 2019.
[70] H. Hu, S., Song, Y. Chai, Y. Gong, "Study on Online Teaching of High Frequency Circuit under COVID-19 Epidemic", Open Journal of Social Sciences, **8**(3), 493-496, 2020, doi.org/10.4236/jss.2020.83041.
[71] T.R. Nodine, "How did we get here? A brief history of competency-based higher education in the United States," The Journal of Competency-Based Education, **1**(1), 5-11, 2016, doi.org/10.1002/cbe2.1004.
[72] R, Kelchin, 2015, "The landscape of competency-based education: Enrollments, demographics, and affordability," Washington, DC: American Enterprise Institute.
[73] S. Directors, C. Technical, E. Consortium, "Building a strong competency-based pathways and career," 2015.
[74] P. Williams, "Does competency-based education with blockchain signal a new mission for universities?" Journal of higher education policy and management, **41**(1), 104-117, 2019, doi.org/10.1080/1360080X.2018.1520491.
[75] A.A. Kytmanov, M.V. Noskov, K.V. Safonov, M.V., Savelyeva, V.A. Shershneva, "Competency-based learning in higher mathematics education as a cluster of efficient approaches," Bolema: Boletim de Educação Matemática, **30**(56), 1113-1126, 2016, doi.org/10.1590/1980-4415v30n56a14.
[76] G. Walters, "Developing competency-based advising practices in response to paradigm shifts in higher education," NACADA Journal, **36**(1), 66-79, 2016, doi.org/10.12930/NACADA-15-033.
[77] S. Perera, S.O. Babatunde, J. Pearson, D, Ekundayo, "Professional competency-based analysis of continuing tensions between education and training in higher education," Higher Education, Skills and Work-Based Learning, 2017, doi.org/10.1108/HESWBL-04-2016-0022.
[78] J. Henrich, "Competency-based education: The employers' perspective of

higher education," The Journal of Competency-Based Education, **1**(3), 122-129, 2016, doi.org/10.1002/cbe2.1023.
[79] D, Brittany, J.E. Coryell, "Relationships, Participation, and Characteristics of a Community of Practice in S hort-Term Adult Study Abroad," Adult Learning, 2019.
[80] A. ettit, A. McVicar, P. Knight-Davidson, A. Shaw-Flach. "Releasing Latent Compassion through an Innovative Compassion Curriculum for Specialist Community Public Health Nurses," Journal of Advanced Nursing **75**(5), 53-62, 2019.
[81] B.W. Trayner, "Communities of practice a brief introduction," 1-8, 2015.
[82] A. Cox, "What are communities of practice? A critical review of four seminal works. Paper presented at the Fifth European Conference on Organizational Knowledge, Learning, and Capabilities," in Fifth European Conference on Organizational Knowledge, Learning, and Capabilities, 2004.
[83] B.W. King, "Communities of Practice," no. August, 2017.
[84] L. Leonard, C. Kies, N. Braaf, "Reflections: eLearning Colloquia enhance a Community of Practice," 2018, doi.org/10.4324/9780429283499-2.
[85] D. R. Garrison, M. C. Leveland-Innes, "Facilitating Cognitive Presence in Online Learning : Interaction Is Not Enough in Online Learning : Interaction Is Not Enough," Am. J. Distance Educ., **19**(3), 133-148, 2005.
[86] C.H.E.N. Tao, L.I. Ju-xuan, "Teaching Reform in the Course of Civil Engineering Introduction Under the Background of Online Courses," In International Conference on Modern Educational Technology and Innovation and Entrepreneurship (ICMETIE 2020), 76-80, Atlantis Press, 2020.
[87] M.M. Sahabadi, M. Uplane, "Synchronous and Asynchronous e-learning Styles and Academic Performance of e-students. Procedia - Social and Behavioral Sciences", **176**, 129-138, 2015.
[88] P. Ahmadi, K.I, Yousaf, "An Empirical Study of Teaching Methodologies and Learning Outcomes for Online and in-class Networking Course Sections," arXiv preprint arXiv:1711.11411, 2017.
[89] Y, Grossman, "Introduction to flavour physics," In LHC Phenomenology, 35-80. Springer, Cham, 2015, doi.org/10.1007/978-3-319-05362-2_2
[90] P. Fain, "Moving ahead with competency," Inside Higher Education, 2015, https://www.insidehighered.com/news/2015/02/17/two-year-colleges-washington-state-expand-competency-based-project-business-degree
[91] T. Mitchell, "Guidance to competency-based education experimental site released," Home Room, the official blog of the U.S. Department of Education, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1217-1224 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Development and Performance Analysis of HRPL Using 6LoWPAN CC2538 Module for IoT Ecosystem

Nin Hayati Mohd Yusoff*, Nurul Azma Zakaria

Center for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia, Durian Tunggal, Melaka, 76100, Malaysia

ARTICLE INFO

Article history:
Received: 09 September, 2020
Accepted: 01 December, 2020
Online: 16 December, 2020

Keywords:
6LoWPAN
Internet of Things
Smart Home
Performance Analysis

ABSTRACT

The Internet of Things (IoT) application has been experiencing increasingly progressive demand, especially for embedded devices (ED). However, the ED has limited capabilities, low power consumption resources, and low bandwidth in connecting to the Internet by using Wireless Sensor Networks (WSNs). Therefore, the WSNs form creates a necessity for new technologies and protocols for the IoT implementation. Thus, IPv6 Low Power Area Network (6LoWPAN) was designed by the Internet Engineering Task Force (IETF) to overcome the Internet Protocol (IP) based communication that allows direct communication between each ED. Nevertheless, the communication between ED using 6LoWPAN becomes challenging in designing routing protocols to achieve the efficient performance Quality of Service (QoS). Among the existing protocols for the 6LoWPAN network, RPL is considered effective for the 6LoWPAN system. However, the Power Consumption (PC) and routing overhead of RPL is high when it was implemented in a real scenario. Therefore, HRPL was proposed to enhance the RPL by introducing the rebroadcast technique in order to minimize the routing overhead at the same time reduced the PC usage. Thus, this paper is an extended version, in which 6LoWPAN Smart Home Testbed (6LoSH) was developed to investigate the impact of the number of nodes on PC usage for both protocols (HRPL and RPL) in a real scenario. The result shows for this instance, HRPL has succeeded in reducing the use of PC for both experiments (simulation and 6LoSH). On the other hand, the number of nodes had given an impact on PC usage. For further work, we plan to use multiple topologies and the different number of nodes to explore the HRPL for some performance metrics such as convergent time and latency.

1. Introduction

This paper is an extended version of work published in [1]. We are making this version available in order to have more results and discussions in comparison to its short version.

The 6LoWPAN was introduced by IETF to support a huge number of IP-based embedded devices that are connected to the Internet with a low power network capability. Moreover, the rapid development of the Internet of Things (IoT) ecosystem in multiple areas [2] such as healthcare, smart building automation [3], vehicle [4], etc. had made the embedded device to be in high demand [5]. The embedded devices are categorized as limited power usage, limited cost, limited bandwidth, and small memory usage [6]. Meanwhile, the development of IoT devices become a serious problem to enable interoperability and scalability among devices and systems [7,8]. Thus, the 6LoWPAN protocol was designed to complete and complement the technological constrain as an essential part of the IoT development application and ecosystem. 6LoWPAN that enables the ED which are mainly sensor nodes to connect to the Internet over IEEE 802.15.4 in Physical (PHY) layer and Medium Access Control (MAC) sublayer. Figure 1 presents the 6LoWPAN stack where the adaptation layer was added between the data link physical layer and network layer. 6LoWPAN is a wireless embedded internet [9]

*Corresponding Author: Nin Hayati Mohd Yusoff, nin6699@gmail.com

that integrates IPv6 with WSN [10] that allows the IP-based ED to connect and exchange information directly with other devices on the Internet [9]. Due to this matter, the energy used for connectivity will save, and for this condition, 6LoWPAN network becomes an essential part of the IoT deployment [11].

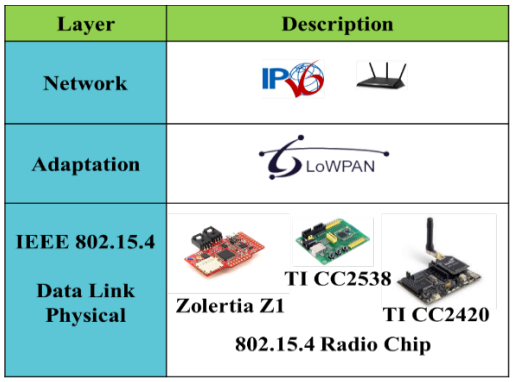


Figure 1: 6LoWPAN Stack

However, the infrastructure of IPv6/6LoWPAN and how IPv6 can integrate heterogeneous technology (system, tool, application, etc.) towards a new embedded application protocol (6LoWPAN) for IoT ecosystem become a challenging issue. This because the current standard of 6LoWPAN is limited to 250 kbps, and 127 bytes of frame length compared to IEEE 802.15.4 packet size is 1280 bytes. As a result, it is critically affected for the routing algorithm to optimize the path formation. Indirectly, it will degrade the interoperability of 6LoWPAN devices/applications. Besides, the implementation, integration and evaluation for large scale deployments of sensors networks in buildings, the potential heterogeneous combination of wireless and wired need to optimize the energy consumed by the system itself.

In addition, the 6LoWPAN protocol develops to optimize the routing scheme for converging cast traffic patterns for low power nodes [12]. One of the 6LoWPAN routing protocols is Protocol Over Low-Power and Lossy Networks (RPL) as defined in RFC6550 [13]. Based on previous research, the RPL protocol in 6LoWPAN is effective in terms of delay and Packet Delivery Rate (PDR). But, it can be further improved in terms of overhead as the original value is very high [14] [15]. According to Feschel, Huhn, & Schwetlick in 2012, the fragmentation of the frames should minimize the routing overhead and power consumption in order to save network lifetime that is a critical parameter to 6LoWPAN nodes. On the other hand, to optimize network performance, the routing overhead must be reduced. At the same time, the 6LoWPAN system offers the scalability and as well as interoperability between nodes. Therefore, this research project addresses algorithmic, networking, and application development using the 6LoWPAN CC2538 module in improving the 6LoWPAN protocol to achieve efficient performance QoS of the network in terms of PC. The proposed algorithm is implemented and tested using a simulation study and validated with the experimental approach by developing the 6LoWPAN Smart Home testbed (6LoSH). The initial phase of this study focused on a literature review that has been done in [17] to address the two (2) following challenges: i) Comparative analysis of existing systems and solutions and ii) Comparative analysis of existing standards and protocols. As a reminder of this paper, we present our work as follows. We describe a methodology used in Section II follows by Section III present the experiment setup. Then, Section IV describes the evaluation results. Finally, we emphasize the contribution of this paper and future works.

2. Methodology

The methodology is divided into three parts, as presented in Figure 2. The first part is the development process, followed by implementation, and the last part is testing and validity.

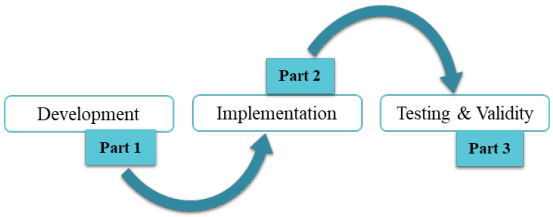


Figure 2: The Methodology Design

2.1. Development

The development of 6LoSH involves three (3) important aspects. First, the integration between the 6LoWPAN system requirement and building infrastructure with different technologies like sensors, actuators, and mobile devices. Figure 3 presents the proposed 6LoSH network design, which implements a simple node architecture [11], [9], [18]. 6LoWPAN nodes communicated with each other through border router, and share a backbone link to connect to the Internet. This allows all the 6LoWPAN nodes remain the same as the IPv6 prefix. Besides, 6LoWPAN connects to another IP network via a border router. At the same time, the border router is handling the 6LoWPAN compression and Neighbor Discovery (ND) for 6LoWPAN nodes. According to [11] the power consumption usage using simple architecture remain low when the intelligent objects are idle, and increases the data rate in both host and router mode. The end-users will be able to interact with the system by running several applications from their PCs, mobile phones, notebooks, and voice. These applications will make use of the services provided by the building system. They will allow end-user to obtain information about the state of the building and switched off/on the application over the Internet.

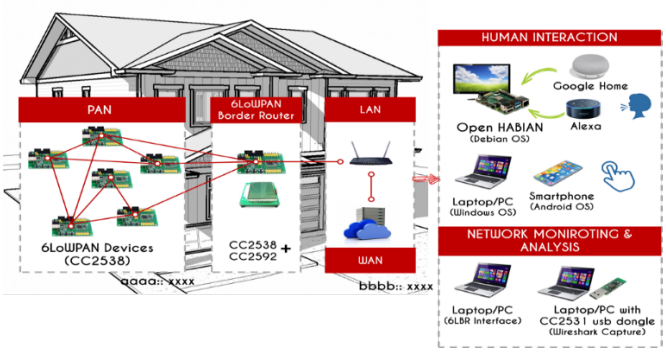


Figure 3: Simple 6LoWPAN Nodes (CC2538 Module) Architecture for 6LoSH

Second, the 6LoWPAN node, other hardware, and software requirements towards a new embedded application in developing 6LoSH. Table 1 shows the list of the equipment required for the implementation of the 6LoSH physical deployment. We used the TI CC2538 module [19] as a 6LoWPAN node and border router because it is a very low power wireless module for IoT applications [20], Figure 4. In contrast, Raspberry Pi 3 Model B is used for running the 6LoWPAN Border Router (6LBR). We observed 6LoWPAN packets by using a CC2531 traffic sniffer and the Wireshark packet capture tool with baud rate 115200. Besides, the Contiki OS [21] [22] by performed COOJA simulator is used to set up the CC2538 program and running simulation of the 6LoWPAN network.

Table 1: The equipment required for the implementation of the 6LoSH physical deployment

Hardware and Software Requirement		
Equipment	Count	Types, and Description
Node	10	• The CC2538 Texas Instrument
Border Router	1	
Sniffer	1	• CC2531 USB Dongle
Raspberry Pi 3	1	• Running 6LoWPAN Border Router (6LBR) • Model B (Debian OS)
Contiki OS (Cooja Simulator)		• Perform simulation testing • Writing a program for CC2538 module
Wireshark		• Packet capturing and analyzing the OTA packet

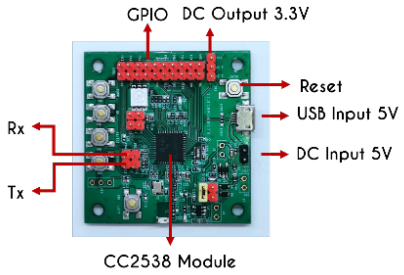


Figure 4: CC2538 Board

Finally, a novel algorithmic models and scalable solutions for the 6LoWPAN system. The on/off an application for embedded devices has been designed and flashed onto the CC2538 board, namely 6Node. The real-time arrangement of the 6Node setup is shown in Figure 5. Table 2 presents the proposed 6Node protocol parameter setup. The default trickle algorithm [15] is used to control the amount of DIO message in routing traffic by setting the parameter of DIO Interval Min as 12ms and DIO interval doublings as 16ms. To achieve the best Packet Delivery Ratio (PDR), the Minimum Rank with Hysteresis Objective Function (MRHOF, RFC6719) with Expected Transmission Count (ETX, RFC 6551) metrics are used as path selection mechanism [23]-[24].



Figure 5: Real-Time Arrangement of CC2538 setup

Table 2: System Parameter for CC2538 Protocol

Parameter	Algorithm/Coding

Trickle Algorithm	case RPL_TRICKLE_T_TIMEOUT:
	rpl_trickle_handle_t_timeout();
	break;
	case RPL_TRICKLE_I_TIMEOUT:
	rpl_trickle_handle_i_timeout();
	break;

Nodes rank	#define RPL_OCP_OF0 0 In #define RPL_OCP_MRHOF 1
RPL_Dio Interval Min	#ifdef RPL_CONF_DIOINTERVALMIN #define RPL_DIOINTERVALMIN RPL_CONF_DIOINTERVALMIN #else #define RPL_DIOINTERVALMIN 12 #endif
DIO interval doublings	#ifdef RPL_CONF_DIOINTERVALDOUBLINGS #define RPL_DIOINTERVALDOUBLINGS RPL_CONF_DIOINTERVALDOUBLINGS

2.2. Implementation

The implementation process involves the deployed hardware infrastructure, physical network, and integrated with software services for managing and controlling the hardware. Dealing with devices and technologies, it has explored existing platforms and solutions that could be relevant for 6LoSH to achieve effective communication (interoperability) between the nodes and systems [7], [8]. The home appliances are connected to nodes through a relay which in turn is connected to another digital pin of the board as illustrated in Figure 6. Figure 7 present the 6Node prototypes and Figure 8 show the development of 6Node prototype activities.

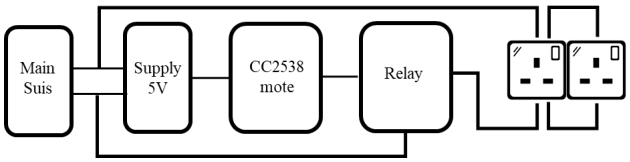


Figure 6: The Block Diagram of CC2538 Application (6Node)

2.3. Testing and validity

The performance of the 6LoSH platform has been evaluated and validated through simulation and calibrate the simulation result using an experiment to ensure the 6Node reach the efficiency of QoS. The HRPL protocol introduced in [1] and RPL

standard was used to assess the effectiveness of 6Node in reducing routing overhead toward new embedded applications for the 6LoWPAN environment.



Figure 7: The 6Node Prototypes

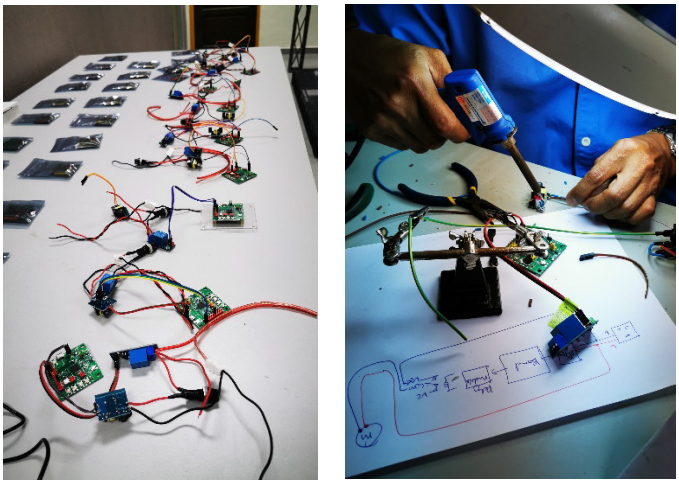


Figure 8: The development of 6Node Prototype Activities

3. Experiment Setup

Figure 6 shows the testing design and Table II describes a summary of the evaluation parameter setting for both experiments that was adapted from the existing model in [25] [26,27]. The comparative study in terms of power consumption has been done by using standard HRPL and RPL protocol to investigate the scalability and interoperability communication between devices (6LoWPAN appliances) and network performance. To be fair, the open-source implementations run on the same operating system (OS) which is Contiki OS. The HRPL and RPL protocol was simulated using the Cooja simulator [28,29] (Testing 1). Then, we compared the simulation results with the evaluation of HRPL and RPL through testbed experiments (testing 2). We created similar scenarios for both experiments. Lastly, we evaluated the performance of the HRPL protocol based on defined metrics. Figure 10 present the real-time arrangement of 6LoSH testing.

3.1. Simulation (Testing 1)

Tmote Sky CC2420 was used as a platform. The set of simulation scenarios (4, 6, 8 and 10 nodes) used in this study as shown in Figure 11. Firstly, the implementation of the RPL protocol from the standard examples of RPL collects file that contains source code of sink and sender nodes are simulated. We set a sink node in the middle of the network and deploy the sender node randomly. We set a transmission range (Tx) to 20m and interference range (INT) to 20m at Unit Disk Graph Model (UDGM) [30]. The radio messages were enabled to capture the traffic and saved as a pcap file. The saved pcap file would be used to analyze the network using Wireshark network analyzer. Then, the standard RPL collect file is modified by changing RPL library to HRPL library, in both source codes of sink and sender nodes. Next, the HPRL protocol is simulated in the same scenario. Lastly, the result between RPL and HRPL is compared with four types of power consumption: CPU Power, Transmit Power, Listen Power, and Low Power Mode (LPM). The simulation results will be evaluated for the different number of nodes (4, 6, 8, and 10 nodes).

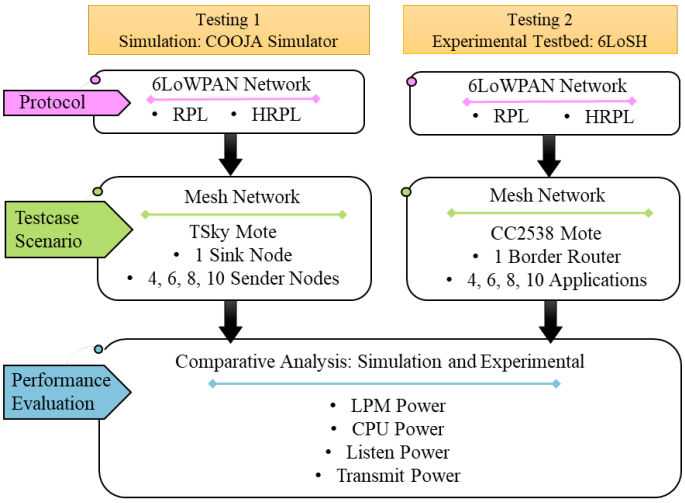


Figure 9: The 6LoSH Testing Design

Table 3: The Summary Parameter Seting for 6LoSH Testing

Method	Testbed (6LoSH)	Simulation
Protocol	HRPL	HRPL
Transport	UDP/IPv6	UDP/IPv6
Mote Type	CC2538 Module	Tmote Sky
Reception Ratio (Rx) and Transmission ratio (Tx)	-	100% & 100%
Packet Size	160bytes	160bytes
Queue	150 packets	250 packets
Transmission Power	-17dBm	0dBm
Transmission range	20m	20m
Network Scale	4, 6, 8, 10 nodes	4, 6, 8, 10 nodes
Length	60 minutes	60 minutes
Packet Capturing	Pcap (Wireshark)	Pcap (Wireshark)

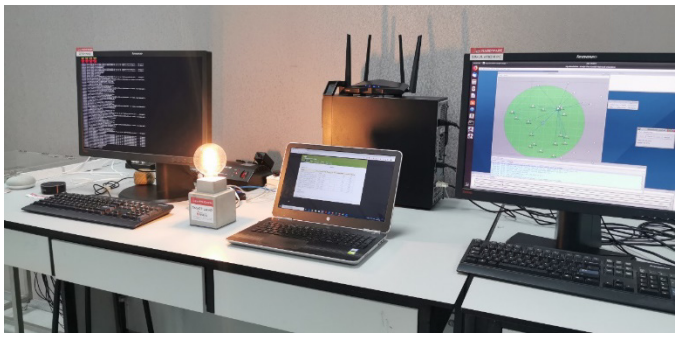


Figure 10: The Real-Time Arrangement of 6LoSH Testing

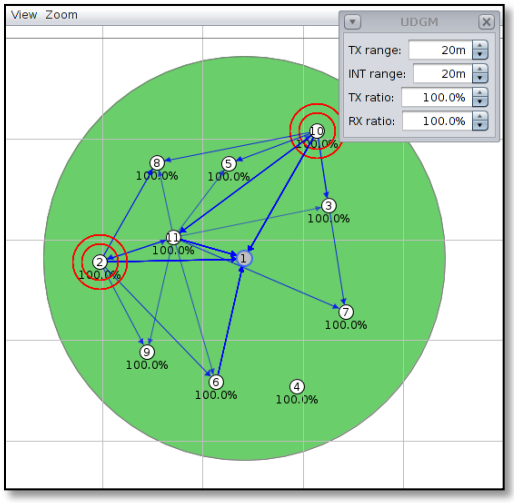


Figure 11: The Set of Simulation Scenario

3.2. Testbed (Testing 2)

The testbed results are used to calibrate the simulation results and also to validate the performance of HRPL in a real-scenario. In this scenario, one border router (6LBR) is located in the center and the different number of applications (4, 6, 8, and 10 nodes) were deployed randomly on the first floor in one of the houses in Malaysia, as presented in Figure 12.

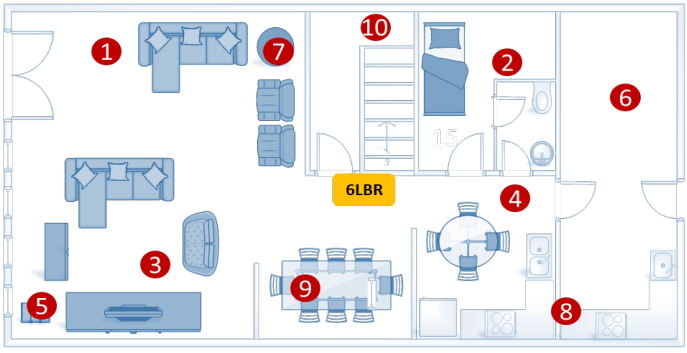


Figure 12: The 6Node Deployment

3.3. Performance Metric

The power consumptions are monitor in both experiments through the development of the Powertrace application in the Contiki operating system. There are four types of power are measured:

- CPU Power,
- Low Power Mode (LPM) Power
- Transmit Power and
- Listen Power.

CPU power refers to the power of computation required by the node in active mode while LPM power refers to the power used by the nodes in CPU sleep mode. Meanwhile, the energy used by each node to transmit the data package to the neighbors is called transmit power, and listening power refers to the energy required by each node in ready mode to receive the data package.

4. Results and Discussion

We assumed our experiment a random topology where the nodes were deployed randomly in an area of 20m x 20m (simulation) and 20m x 20m (testbed). We put the sink node in the center of the network. In order to enable interconnection between all the sender nodes, we set up each node with multi-hop transmission distances from the border router. For both experiments, we considered the DAG root at the border router. Then we set up the nodes to generate packets 1 packet/min randomly in a fixed interval time. We derived this analysis using a script to determine when all nodes had printed to the simulator output (simulator) and web interface (testbed) that they joined the DAG.

Table 4: Analysis of Total Power Consumption

Number of Nodes	Testing 1 (Simulation)			Testing 2 (6LoSH)		
	RPL (mW)	HRPL (mW)	%	RPL (mW)	HRPL (mW)	%
4	0.995	0.629	36.8	1.576	1.0405	34.0
6	1.015	0.641	36.9	1.607	1.0515	34.6
8	1.021	0.650	36.3	1.623	1.07225	33.9
10	7.239	3.175	56.1	7.906	4.1099	48.0

Figure 13 shows the results obtained from the analysis of HRPL and RPL power consumption usage in different nodes (4, 6, 8, and 10) for testing 1 and testing 2. By comparing the four different nodes, it is evident that the power consumption usage for Number of Nodes (NoN) = 4 is the lowest followed by NoN = 6 then NoN = 8 while NoN = 10 is the highest power consumption usage. The result shows that HRPL is better in all NoN and managed to reduce the overall power consumption by 36.8% (NoN=4), 36.9% (NoN=6), 36.3% (NoN=8) and 56.1% (NoN=10) lesser than RPL in testing 1. While in testing 2 HRPL has obtained a better result that can reduce power consumption by 34.0% (NoN=4), 34.6% (NoN=6), 33.9% (NoN=8) and 48.0% (NoN=10) lesser than RPL. In this instance, according to the observation, the HRPL had generated a smaller power consumption, and due to the number of nodes, however it had managed to obtain the increases of power consumption in real time scenario (6LoSH) compared in simulation. It is figured out which the power consumption value would have effect under different platform even used the same scenario and system setting. The cooja simulator enable the node which make a simple network and doesn't requirement gateway to connect to the internet. But

different condition in 6LoSH, raspberry pi (powered by battery) used as a 6LBR to deploy the node for physical network and this condition increase the radio transmission power received by each node. Besides, the retransmission activities detected leading to higher power consumption in real scenario.

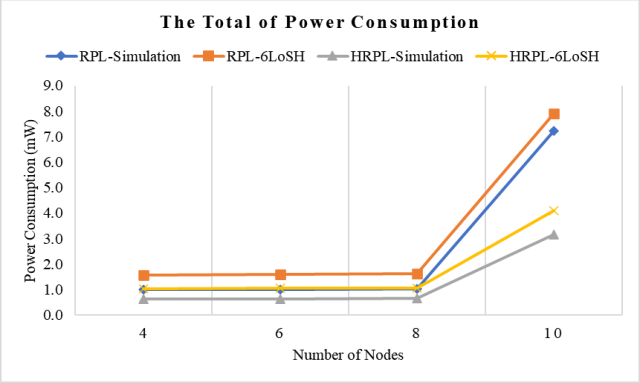


Figure 13: The Total Power Consumption for Both Experiment.

Table 5 presents the analysis of each type of power consumption for testing 1 and testing 2. The CPU Power, LPM Power, Listen Power, and transmit Power values obtained for each of the experiments were presented in Figure 14-17. It is observed that for NoN = 4 gave a minimum value of LPM Power, CPU Power, and Transmit Power. However, the value of Listen Power is the highest in comparison with others. Based on observation, the node turns on its radio during all its idle slot when it's set as a parent. This condition not only becomes the listening power increase but also wastes their energy for the whole network. Interestingly, the computation of each node is increased due to the way the node determines the best route from the source node to the destination node[31]. Meanwhile, the T-Sky CC2420 module that is used in simulation has less power compared to the CC2358 module that is used in 6LoSH.

Table 5: Analyses of Power Consumption for Testing 1 (Simulation) and Testing 2 (6LosH)

Types of Power	Number of Nodes	Testing 1 (Simulation)		Testing 2 (6LoSH)	
		RPL (mW)	HRPL (mW)	RPL (mW)	HRPL (mW)
CPU Power	4	0.332	0.209	0.453	0.285
	6	0.349	0.213	0.487	0.309
	8	0.357	0.218	0.514	0.328
	10	0.376	0.228	0.586	0.362
LPM Power	4	0.153	0.101	0.224	0.156
	6	0.153	0.104	0.218	0.143
	8	0.153	0.105	0.224	0.145
	10	1.519	0.988	0.551	0.315
Transmit Power	4	0.092	0.054	0.211	0.132
	6	0.086	0.056	0.214	0.134
	8	0.08	0.056	0.214	0.134
	10	0.88	0.504	2.101	1.211
Listen Power	4	0.418	0.265	0.688	0.468
	6	0.427	0.268	0.688	0.466
	8	0.431	0.271	0.671	0.465
	10	4.464	1.455	4.668	2.222

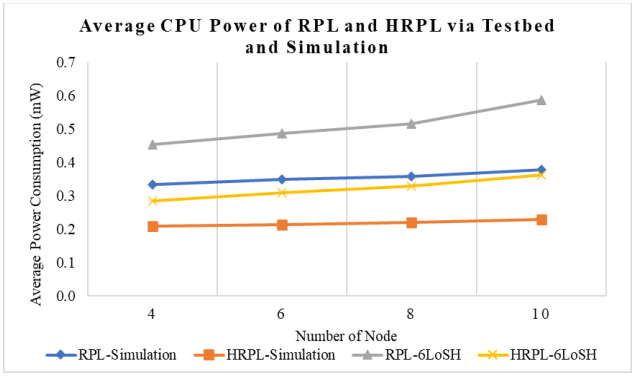


Figure 14: The CPU Power of HRPL and RPL

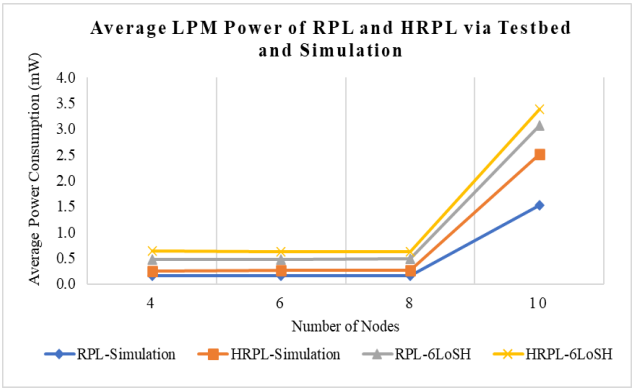


Figure 15: The LPM Power of HRPL and RPL

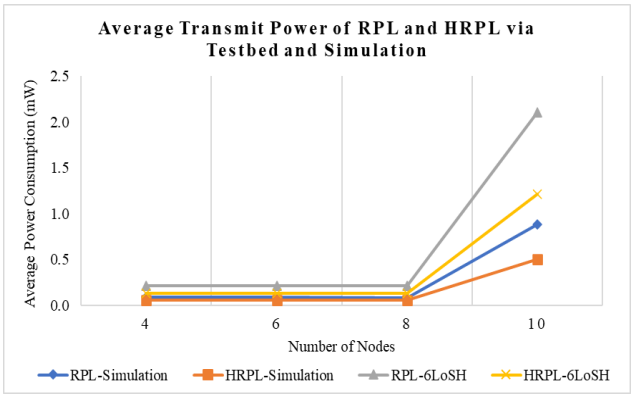


Figure 16: The Transmit Power Of HRPL and RPL

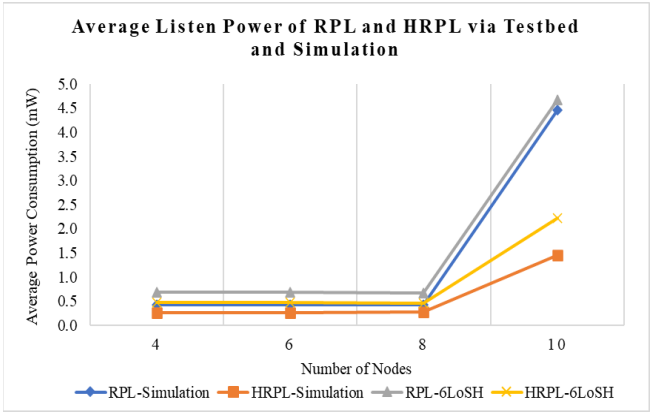


Figure 17: The Listen Power of HRPL and HRPL

5. Conclusion

6LoWPAN system enables the ED to communicate with each other using the IEEE 802.15.4 wireless network. Due to this matter, 6LoWPAN had become a priority network for IP-based communication for ED. However, the design of the 6LoWPAN system in the real-world scenario has become a challenging issue, such as interoperability, scalability, and integration among the nodes, and network. Thus this study aims to develop the 6LoWPAN smart home system (6LoSH) to evaluate the performance of HRPL and compared the result with RPL. The result of this study shows the PC of nodes reached its most minimum for NoN =4 scenarios, while low link quality is linked to NoN = 10 scenarios respectively for both experiments. These results confirm that the 6Node (6LoSH) reaches an efficiency of QoS and HRPL reliability to reduce the PC. The routing overhead for the number of nodes increases significantly with the increase of the PC. The main contribution of this study, the present result shows that HRPL outperforms the RPL standard of PC in both experiments. However, these results were derived from a limited number of nodes, the Transmission Range (Tx) is 5m, and it was tested for random topology. Further research should be conducted to investigate large-scale networks in multiple topologies for different distances and metrics.

Acknowledgment

The authors would like to thank the Ministry of Higher Education Malaysia and Universiti Teknikal Malaysia, in support the Fundamental Research Grant Scheme (FRGS) with reference number (FRGS/2018/FTMK-CACT/F00392).

References

[1] N.H. Mohd Yusoff, N.A. Zakaria, A. Sikora, J.S. Sebastian, "6LoWPAN Protocol in Fixed Environment: A Performance Assessment Analysis," Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019, **2**, 1142–1147, 2019, doi:10.1109/IDAACS.2019.8924283.

[2] H. Yadav, C. Agrawal, "A Review of Enhancing Energy Efficiency of Wireless Sensor Network," International Journal of Electrical, Electronics and Computer Engineering (IJEECE), **8**(1), 46–49, 2019.

[3] V.C. Thang, N. Van Tao, "A performance evaluation of improved IPv6 routing protocol for wireless sensor networks," International Journal of Intelligent Systems and Applications, **8**(12), 18–25, 2016, doi:10.5815/ijisa.2016.12.03.

[4] T. Abdelkader, K. Naik, A. Nayak, N. Goel, V. Srivastava, "A performance comparison of delay-tolerant network routing protocols," IEEE Network, **30**(2), 46–53, 2016, doi:10.1109/MNET.2016.7437024.

[5] J.A. Alzubi, R. Manikandan, O.A. Alzubi, N. Gayathri, R. Patan, "A Survey of Specific IoT Applications," International Journal on Emerging Technologies, **10**(1), 47–53, 2019.

[6] S.N.S. Vishalakshi R, Chaithra P, "6LoWPAN Routing Issues," International Journal Of Engineering And Computer Science, **3**(7), 2014.

[7] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, W. Zhao, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications," IEEE Internet of Things Journal, **4**(5), 1125–1142, 2017, doi:10.1109/JIOT.2017.2683200.

[8] S. Li, L. Da Xu, S. Zhao, "The Internet of Things: A Survey," Information Systems Frontiers, **17**(2), 243–259, 2015, doi:10.1007/s10796-014-9492-7.

[9] Z. Shelby, C. Bormann, 6LoWPAN : The Wireless Embedded Internet, 2009.

[10] P. Pongle, G. Chavan, "A survey: Attacks on RPL and 6LoWPAN in IoT," in 2015 International Conference on Pervasive Computing: Advance Communication Technology and Application for Society, ICPC 2015, 0–5, 2015, doi:10.1109/PERVASIVE.2015.7087034.

[11] S.N. Han, Q.H. Cao, B. Alinia, N. Crespi, "Design, implementation, and evaluation of 6LoWPAN for home and building automation in the Internet of Things," Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, **2016**–**July**, 2016, doi:10.1109/AICCSA.2015.7507264.

[12] A.J. Albarakati, J. Qayyum, K. a. Fakeeh, "A Survey on 6LowPAN & its Future Research Challenges," International Journal of Computer Science and Mobile Computing, **3**(10), 558–570, 2014.

[13] T. Winter, P. Thubert, A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, R. Alexander, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks Abstract Low-Power," IETF RFC6550, doi:10.1017/CBO9781107415324.004.

[14] H. Xie, G. Zhang, D. Su, P. Wang, F. Zeng, "Performance evaluation of RPL routing protocol in 6lowpan," in Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2014, doi:10.1109/ICSESS.2014.6933646.

[15] P. Levis, T.H. Clausen, J.W. Hui, O. Gnawali, J. Ko, "The Trickle Algorithm," IETF RFC6206, 2011, doi:http://dx.doi.org/10.17487/RFC6206.

[16] M. Feschel, A. Huhn, H. Schwetlick, "Routing protocols for 6LoWPAN," International Conference on IT Revolutions, **82 LNICST**, 71–83, 2012, doi:10.1007/978-3-642-32304-1_7.

[17] N.H. Mohd Yusof, N.A. Zakaria, N. Harum, "Design of Smart Building Automation System for 6LoWPAN Network : The Platform Structure," in Procedings National Symposium on Human Computer Interaction, 2019, 35–37, 2019.

[18] E. Toscano, L. Lo Bello, "Comparative assessments of IEEE 802.15.4/ZigBee and 6LoWPAN for low-power industrial WSNs in realistic scenarios," 9th IEEE International Workshop on, IEEE, 2012.

[19] CC2538 Powerful Wireless Microcontroller System-On-Chip for 2.4-GHz IEEE 802.15.4, 6LoWPAN, and ZigBee ® Applications 1 Device Overview.

[20] G.A. Naidu, J. Kumar, V. Garudachedu, P.R. Ramesh, "6LoWPAN Border Router Implementation for IoT Devices on RaspberryPi," SSRN Electronic Journal, 316–323, 2018, doi:10.2139/ssrn.3166729.

[21] The Official Contiki OS Blog: Contiki 3.0 Released, New Hardware from Texas Instruments, Zolertia.

[22] A. Velinov, A. Mileva, "Running and Testing Applications for Contiki OS Using Cooja Simulator," International Conference on Information Technology and Development of Education, 279–285, 2016.

[23] N. Pradeska, Widyawan, W. Najib, S.S. Kusumawardani, "Performance analysis of objective function MRHOF and OF0 in routing protocol RPL IPV6 over low power wireless personal area networks (6LoWPAN)," Proceedings of 2016 8th International Conference on Information Technology and Electrical Engineering: Empowering Technology for Better Future, ICITEE 2016, (7863270), 7863270, 2017, doi:10.1109/ICITEED.2016.7863270.

[24] J. Vasseur, M. Kim, K. Pister, N. Dejean, D. Barthel, "Routing Metrics Used for Path Calculation in Low-Power and Lossy Networks," IETF RFC6551, doi:10.1017/CBO9781107415324.004.

[25] Y. Tahir, S. Yang, J. McCann, "BRPL: Backpressure RPL for High-throughput and Mobile IoTs," IEEE Transactions on Mobile Computing, **2**(c), 1–1, 2017, doi:10.1109/TMC.2017.2705680.

[26] C.-A. La, M. Heusse, A. Duda, "Link reversal and reactive routing in Low Power and Lossy Networks," IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, 3386–3390, 2013, doi:10.1109/PIMRC.2013.6666733.

[27] E. Ancillotti, R. Bruno, M. Conti, E. Mingozzi, C. Vallati, "Trickle-L2: Lightweight link quality estimation through Trickle in RPL networks," Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014, WoWMoM 2014, 2014, doi:10.1109/WoWMoM.2014.6918951.

[28] A. Dunkels, B. Gronvall, T. Voigt, "Contiki - a Lightweight and Flexible Operating System for Tiny Networked Sensors," Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN'04), (January), 2010, doi:10.1109/LCN.2004.38.

[29] Y. Bin Zikria, M.K. Afzal, F. Ishmanov, S.W. Kim, H. Yu, "A survey on routing protocols supported by the Contiki Internet of things operating system," Future Generation Computer Systems, **82**, 200–219, 2018, doi:10.1016/j.future.2017.12.045.

[30] L. Sitanayah, C.J. Sreenan, S. Fedor, "A Cooja-based tool for coverage and lifetime evaluation in an in-building sensor network," Journal of Sensor and Actuator Networks, **5**(1), 1–22, 2016, doi:10.3390/jsan5010004.

[31] L. Lassouaoui, S. Rovedakis, F. Sailhan, A. Wei, "Evaluation of energy

aware routing metrics for RPL," 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 1–8, 2016, doi:10.1109/WiMOB.2016.7763212.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1225-1236 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Priority-based Scheduling Algorithm for NOMA-integrated V2X

Ala Din Trabelsi*, Hend Marouane, Faouzi Zarai

National School of Electronics and Telecommunications (ENET'COM), NTS'COM Laboratory, Sfax University, 3029, Tunisia

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 26 November, 2020
Online: 16 December, 2020

Keywords:
5G
V2X
Scheduling Algorithm
Resource allocation
PSA

ABSTRACT

The fifth-generation (5G) wireless system was implemented due to the massive connectivity, especially Internet of Things (IoT) including industrial automation, Vehicle to Everything (V2X) communications, which is shelter an extensive range of applications services and use cases and Intelligent Transportation Systems (ITS). A cellular Non-Orthogonal Multiple Access (NOMA) based on radio resource allocation schemes need to be carefully designed to serve and support the diversity of applications and services, through increased safety and traffic efficiency usage of connected vehicle and ranging from infotainment services. In this paper, we develop a novel scheduling and resource allocation algorithm for vehicular communications named Priority -based Scheduling Algorithm for NOMA integrated V2X (PSA). The novel scheme could make the limited radio Resource Blocks (RB) to be assigned in a more efficiently manner to the different traffic classes (safety and non-safety traffic) imposed by V2X exigency, in order to improve the performances of the system in terms of average throughput, Quality of Service (QoS), Average Blocking Rate (ABR) and to ensure fairness criteria between Vehicular User Equipments (VUEs). The performance of PSA is shown for urban macro-cell and rural macro-cell scenarios and we compared it with the conventional OMA and some recent works for NOMA based allocation.

1. Introduction

New 5G technologies supporting vehicular communications considered as a promising and vital role in the blueprint of Intelligent Transportation Systems (ITS), assisting to solve the main problems of transport engineering, improves road safety, traffic efficiency and comfort and infotainment services, try to prevent any fostering and dangerous situations investments in the automotive market [1-3].

V2X connectivity has been developed as a potential concept for 5G which strive to give communication models that can be supported by vehicles in several application contexts including vehicle-to-vehicle (V2V), vehicle-to-infrastructure/network (V2I/N) and vehicle-to-pedestrian (V2P) [4].

V2X will include two complementary sets of specifications for direct transmission which are 802.11p and cellular V2X (C-V2X) standards. The first one is proposed as the lower layers of Wireless Access in the Vehicular Environment (WAVE), it is standardized for V2V defined in USA and recognized as Dedicated Short-Range Communications (DSRC). 802.11p is also designed as the lower layers of the ITS-G5 standard in Europe. It is based on OFDM and which is known as the case of fourth generation (4G) technology LTE (Long Term Evolution). However, the adoption of DSRC has been delayed due to its Medium Access Control (MAC) and physical layer designs that have been planned for Wireless Local Area Networks (WLAN) and its poor scalability and the high-mobility environments that impose communication challenges, and lack of QoS guarantee. In addition, 802.11p indicates that it's restricted range, reliability and undefined performance especially in urban propagation scenarios when scaling up the number of transmitting VUEs [4].

Cellular-V2X is proposed to innovate and deliver an evolution path to 5G and it comes to handle the challenges of 802.11p. The main advantage of C-V2X that is based on a technology originally optimized for high-speed mobile services, which has been enhanced specifically for safety and automotive applications. C-V2X is outperformed 802.11p due to modulation and coding enhancement, to support a better communication range, greater capacity, better congestion control in denser environments with lower packet error rate, higher reliability and superior Non Line-of-Sight (NLoS) performance [4].

*Corresponding Author: Ala Din Trabelsi, trabelsi.ala93@gmail.com

Accordingly, 27 use cases for V2X communications are designed by third-Generation Partnership Project (3GPP) in Release 14 that including two categories of service requirements as safety related applications, like car platooning, autonomous driving, vulnerable road user warning, etc and non-safety related uses cases exemple, dynamic digital map update, mobile hot spot, mobile high data rate entertainment, office, home, etc [5].

More recently, 5G networks anticipate an explosive demand for massive connectivity driven by the ever-increasing number of connected vehicles, Internet of Things (IoT) and the explosive proliferation of mobile Internet, over limited radio resources [6].

The limitation of OMA-based scheme that it has not been efficiently and fully assigning the limited radio resource blocks causing to serious problems as lack of QoS guarantee for NLoS environment. This defiance can be exemplified with the following scenario, consider a single User Equipment (UE) with mediocre channel condition needs to be scheduled for fairness purposes, example this UE has superior priority information or has not been scheduled for a long time. This causes that one of the limited bandwidth resources is only allocated by this UE. In addition, the limitation of the conventional OMA is the number of users that can be served simultaneously. Towards these challenges, 3GPP is motivated to develop new technologies in order to maximizing the achievable throughput. Among various solutions, non-orthogonal multiple access (NOMA) is a very promising access technology. It has been envisioned and recognized as an innovation for 5G mobile networks to increase the number of links, enlarge the spectral efficiency and to ensure the high data rate necessity. It provides enhancements in terms of high reliability, low latency and communication's range for vehicular environments [7].

Using non-orthogonal radio resource allocation, make NOMA to multiplex more than a few users with different power levels into a same frequency resource and time to transfer information. Not only does NOMA support users with good channel conditions but also it serves both others with poor channel conditions and those operate in the same bandwidth resources [8][9].

To enable V2X communications, radio resource management is a vital specification of the MAC layer of the eNodeB, that aims to optimize the use of the limited availability of the spectrum while meeting the requirements of QoS and insure efficient resources scheduling, resource allocation and interference management. Even in 5G networks, a resource allocation technique requires to be carefully developed to satisfy different QoS requirements to achieve target throughput and to maximize spectrum efficiency for several V2X traffics and services.

The paper contributions are as following:

- Taking into account traffic classification, such as Safety Traffic (ST) and Non Safety one (NST) are considered
- A resource allocation algorithm called Priority -based Scheduling Algorithm for NOMA integrated V2X (PSA) is proposed.
- Adding a scheduling priority which is determined according to the channel condition for each VUE, users with higher priority will be selected for NOMA sharing.
- Due to the diversity and multitude of the propagation scenarios, various simulations are presented for performance evaluation for rural macro-cell and urban macro- cell, to demonstrate the efficiency and the applicability of the proposed PSA algorithm for any propagation scenarios.
- Validate the proposed PSA algorithm with several Matlab simulations showing the aggregate throughput, the fairness index and the ABR among different VUE traffic classes.
- Compare the proposed algorithm with the conventional OMA
- Evaluate the PSA algorithm with some recent works for NOMA based allocation.

This paper is presented as follows we study same proposed method as related works in section 2. We detail our approach in section 3. In the section 4 and 5 we portray its different steps. We evaluate and analyze the performances of the developed scheduler in section 6. Finally, in Section 7, we determine our work by conclusion and perspective as future work.

2. Related Work

The radio resources management task to enable V2X communications has been investigated in several research works such as:

In [10], the authors developed a novel resource allocation algorithm to maximize the number of scheduled V2N communications while satisfying the required latency for the several priority classes of the V2V links (two class of priority are defined, where the messages in the class 1 are more critical and aware to the latency comparing to messages in the second one). They perform the power allocation for potential resource reusing. Then, they selected the finest reusing pattern by employing the Hungarian scheme. However, this method does not take NOMA systems in consideration and the performance evaluations focused only to present the average latency.

In [11], the authors studied the resource allocation process when the periodic messages are not all at the same size in cellular V2X networks. They addressed an optimization problem for maximizing number of vehicles that can be allocated simultaneously without interfering each other. Then, they provided the same algorithm to solve optimization problem then evaluate its performance by increasing number of vehicles when a trade-off between range and density of vehicle is guaranteed. However, NOMA scenarios issue was usually ignored in design of the resource allocation algorithm.

In [12], the authors developed a novel resource allocation scheme for Power domain based a cooperative multi-cell network by using the cellular mode. This algorithm aims to improve the spectral efficiency of the system and achieves the desired QoS. They assume a communication scenario for two pairs of users sharing the same sub-carrier to transmit their information, where the primary users characterized by delay-sensitive communication that requires stringent QoS exigency as well as the minimum transmission rate, share the same radio RB with the secondary users which support non-delay sensitive or elastic traffic and exploit the physical resources of the network in a best-effort manner. Successive convex approximation (SCA) method is the

best solution to solve general optimization issues, they adopt the successive convex approximation (SCA). Furthermore, they investigate the computational complexity of the designed approach and they compare our work with existing transmission methods based on Orthogonal Frequency Multiple Access (OFDMA). However, this approach does not hit the V2X links requirements.

Some recent research has begun to explore efficient energy NOMA design were given in [13-16]. They presented a general overview of NOMA techniques and seek to optimize the power allocation problem in order to improve Energy Efficiency (EE). Then, they proposed a low complexity power allocation approach. However, those algorithms do not cover V2X links, traffic classifications and heterogeneous QoS requirements for each VUE link.

In [17], the authors focus on the problem of clustering users in groups of different size, and assigning RB to each cluster based on Power-Domain (PD-NOMA) scheme. Clustering and RB allocation are performed as stated by Proportional Fairness (PF) technique with the main objective that obtaining high cell throughput, low complexity and guaranteeing NOMA relevance. The performance evaluation of the clustering algorithm proposed Relevant Clustering (RC) compared with two other algorithms named Ordered Clustering (OC) and Disparity Clustering (DC). However, traffic classifications such safety information for the V2X links is not considered in this algorithm.

In [18], the authors improved proportional fair scheduling scheme for NOMA system, called a spectrum resource and power allocation algorithm with Adaptive Proportional Fair (APF) user pairing with low complexity. This algorithm aims to maximize system throughput and considers the fairness among UE by selecting the user pairing with higher priority expressed by the channel gain. For the optimizing the power allocation, they explore different power allocation schemes which focusing on the proportional rate constraint so as to ensure the best throughput performance. The limitation of this approach that does not considering V2X traffic classifications.

In [19], the authors have addressed and investigated the power allocation problem in downlink Multiple Carriers Non-Orthogonal Multiple Access (MC-NOMA) systems for maximizing Weighted Sum Rate (WSR) by considering imperfect successive interference cancellation, in order to present an iterative power allocation with low-complexity algorithm. Moreover, they have presented a joint user scheduling and power allocation scheme for the MC-NOMA systems. The proposed methods can achieve an optimal WSR performance and user fairness improvement compared to OMA and other existing schemes. However, the system model only considered each sub- channel can be multiplexed by two users no user priority and service classification are treated.

In [20], the authors investigated the benefits of 5G technologies based on NOMA to support massive connections services to group and assign the required sub-channel of the internet of things equipment. To show how the quality of experience requirements of IoT links affect the needed number of physical RBs in real time, they developed a novel method to improve spectrum distribution and to control IoT links mobility by using Non-Orthogonal Multiple Access technique for autonomous vehicles. The designed solution aims to maximize throughput, to contribute QoS improvement and exploiting k-means grouping to reply to the dynamics of the system and minimize the complexity of re-arranging NOMA clusters. The regrouping of IoT connections according to distance while considering the maximum supported devices by group. Therefore, they determined the requisite set of RB for each group of UEs based on the QoS requirement. However, IoT equipments are clustered without considering their type of traffic.

In [21], the authors is introduced an over view on optimization problems of scheduling and NOMA approaches. They planned a novel scheduling and resource allocation scheme based on NOMA-enhanced cellular network for device-to-device (D2D) communications called Tree-based Search power allocation Algorithm (TSA). This scheme is compared and evaluated with several statistic models for user scheduling algorithms, users selection problems for the D2D communications underlying cellular network based on NOMA technology and power allocation strategies such as FRPA (FRactional Power Allocation) and FIPA (FIxed Power Allocation algorithm) where their mean objectives are mitigating computational complexity of the algorithm, increase network capacity, guarantying the QoS and maintaining a better system throughput. The simulation results clearly demonstrate the superiority of TSA compared with traditional algorithms. However, the proposed algorithm does not consider V2X links and traffic classification (ST and NST).

In [22], the authors designed a new method entitled proportional-fairness based user pairing and power allocation algorithm for NOMA uplink transmission. This method is introduced to serve two scenarios: the first one is when user equipments are only distributed in the coverage area of the cell and the second one which is the more complex that is when the interference users are distributed outside of the cell. For scenario one, to optimize the problem they decoupled the User Pairing (UP) part and Power Allocation (PA) part. In the power allocation part, they classify the pairs of users by considering their channel condition and SINR constraint and they developed different power allocation techniques. While in the user pairing part, they designed a probability-based tabu search user-pairing algorithm (PTS). For scenario two, they carried out a prediction-based particle swarm optimization (PBPSO) power allocation algorithm with a low computational complexity to obtain a good proportional fairness performance under the constraints of outage probability. Then, they combine PBPSO and PTS to achieve high system performance in terms of average throughput and outage probability. However, this algorithm assigns the RB for the pairs of users independently of their type of traffic class.

In [23], the authors investigated the global throughput maximization for two UE with timeliness transmission over fading channels for partial and full Channel State Information at the Transmitter (CSIT). This approach supports certain level of fairness and the minimum average throughput constraint. In the full case of CSIT case, the dynamic resource allocation method was optimally determined by using lagrangian dual decay, while in the partial CSIT one, the power allocation technique was proposed focused on investigative results. However, the limitation of this algorithm that does not takes into account the demanded QoS requirements for users and the traffics classifications. The simulation results are investigated with limited parameters setting.

In addition, the required bandwidth to enable ST V2X services in NOMA has not been studied.

In [24], the authors are introduced an over view on the optimization scheduler problem joined with NOMA techniques. They developed and simulated some statistic models for user scheduling strategies and power allocation approaches which aim to achieve the different objectives as minimizing the computational complexity of the algorithm, satisfy the required QoS and obtaining a better system performance in terms of average sum rate and equity criteria. Also, they select Proportional Fairness (PF) and Round-Robin (RR) as a shared model of scheduling and NOMA techniques, to evaluate the optimization problem of power allocation for combined NOMA and round-robin scheduling scheme and proposing an optimization strategy for joint NOMA algorithms and Proportional Fairness scheduling which assure the required QoS. However, the proposed scheme doesn't reflect on the V2X link environments and traffic classifications.

The table 1 illustrates all the mentioned algorithms above. All related work mentioned above doesn't take charge the traffic classifications while seeking the throughput performance gain. Nevertheless, the benefits of V2X ST class at the road cannot be exploited successfully by these algorithms. Despite the existing literature, there is a lack of priority constraint-based scheduling scheme that marked the transmission ranking to each connection but it restricted just by the Channel Quality Indicator (CQI). In addition, diverse algorithms presented above didn't balance cell throughput, equity index and take the system-level QoS for V2X links, new solutions must be provided.

3. System model

The used input and variable parameters are given below.

3.1. System Model

In order to evaluate our proposal work, we consider NOMA scheme consists of a single base station and N connected VUEs denoted by I = {1,....j,…,N}. We suppose that the number of available RBs in the cell is M where the resource blocks set U = {1,…i,…,M}. To assure the required capacity of the N vehicular connections, we suppose that PD-NOMA (Power-Domain NOMA) technique linking by the base station. NOMA is a scheme used to appreciate multiple accesses by utilizing power domain in a spectrally resourceful way and can schedule multiple vehicular user equipments in the same radio resource. This can be maintained by assigning dissimilar power levels to VUEs. Moreover, NOMA scheme is expected to allocate more power level to the VUEs with poor channel qualities. Denote the power of VUE_j on subcarrier k by $P_{j,k}$, satisfying $\sum_{j \in N} P_{j,k} \leq P_j^{max}$, where $P_j^{max} = P$ is the maximum power constraint allocated for VUE_j. Therefore, we assume that $N \geq 2M$ for dense environment and each vehicular user assign to one RB in the same Transmission Time Interval (TTI) while the entire other not scheduled vehicular users are not allocated. In addition, we assume that the distribution and the position of links are supposed random in the coverage area.

To maintain the improvements of V2X link, radio resource allocation and Transmitter -Receiver selection are managed by the eNodeB in each TTI.

Table 1: Comparison between different algorithms studied above

Ref	NOMA systems	Throughput optimization		Qos satisfaction	Services classifications ST / NST	Fairness
		Average	Instantaneous			
[10]	√	√				√
[11]	√	√				
[12]	√	√		√		
[13]	√	√				
[14]	√	√				
[15]	√	√				
[16]	√	√				√
[17]	√	√				
[18]		√				
[19]	√	√		√		
[20]	√	√				
[21]	√	√	√	√		
[22]	√		√			√
[23]	√	√		√		√
[24]			√		**Only ST**	

Table 2: Notations

Symbol	Description
N	Connected VUEs.
I	Set of vehicular uses equipments.
M	Available RBs.
U	Set of resources blocks

P_{ST}, P_{MT}, P_{IT}	Percentage of safety, management and infotainment class of traffic.
N_{ST}, N_{MT}, N_{IT}	Number of active VUEs for each class of traffic.
φ, ψ	Set of safety and non safety VUEs.
$N_{NST}= N_{MT}+N_{IT}$	Number of NST VUEs.
Ć, Ŕ	Set of ST and NST VUEs candidates that can be served to different resource.
P_{factor}	The power level for VUE.
H_ST	Channel coefficient of safety VUE.
H_NST	Channel coefficient of non-safety VUE.
thr_min_safety	Minimum throughput for safety traffic requirement.
H_safety	Matrix contains the channel coefficient of safety VUEs.
H_nonsafety	Matrix contains the channel coefficient of non- safety VUEs.

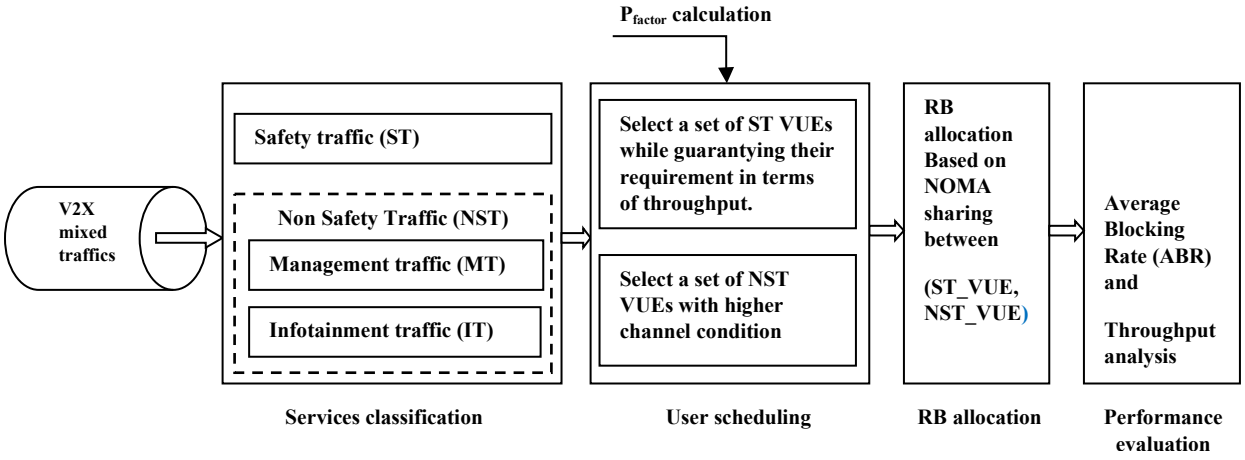


Figure 1: PSA system model

Using NOMA system, the eNodeB superimposes the message waveforms with different assigned power level for its selected users. At the Rx, successive interference cancellation (SIC) schemes are employed to cancel multi-user interferences and each user detects its own signal. The receivers with higher channel gain apply SIC for signal decoding. Therefore, those the modulated symbols are subtracted and removed from the united data until the associated UE obtains the desired message [6].E.g., consider two Tx: VUE_{j1} and VUE_{j2} occupying the RB_i in the current TTI transmitting to receiver VUE_m satisfying $P_{j1,k}^{(i)} \left|H_{j1,m,k)}^{(i)}\right|^2 < P_{j2,k}^{(i)} \left|H_{j2,m,k}^{(i)}\right|^2$, where $P_{jn,k}^{(i)}$ representing the transmit power of the transmitter VUE_{jn}, n $\in\{1, 2\}$. $H_{j2,m,k}^{(i)}$ presentes the channel coefficient of the subchannel k which allocated for the vehicular user VUE_{jn} in the RB_i. The receiver VUE_m deals the signal of the great channel condition VUE_{j2} as a noise, first, detects and remodels the signal of the inferior channel gain of VUE_{j1}, then, exploits the SIC technique to cancel or withdraw the signal of VUE_{j1} which has the mediocre channel condition to get signal of VUE_{j2}.

Generally, the signal x_k, transmitted by the eNodeB at a certain sub-band k is a simple power multiplexing of the symbols generated by the scheduled users.

$$x_k = \sum_{j\in N_m} x_{j,m,k} = \sum_{j\in N_m} \sqrt{P_{j,k}^{(i)}}\, s_j^{(i)} \quad (1)$$

Denotes $N_m=\{1 \le j \le N \mid \max (SINR_{j,m}^{(i)})\}$: presents the set of VUEs inside receiver vehicular UE m's communication range of interest. $P_{j,k}^{(i)}$ is the power allocated to user j at sub-band k and $S_j^{(i)}$ indicates the transmitted signal of Tx VUE_j.

The received signal vector corresponding to the time slot i can be defined by:

$$Y_{m,k}^{(i)} = H_{j,m,k}^{(i)} x_{j,m,k} + n_m^{(i)}$$
$$= \sum_{j\in N_m} \gamma_j^{(i)} \cdot \mu_{j,k}^{(i)} \sqrt{P_{j,k}^{(i)}} H_{j,m,k}^{(i)} S_j^{(i)} + n_m^{(i)} \quad (2)$$

where:

$\gamma_j^{(i)}$: Binary variable denotes that RB_i is assigned by the BS to VUE_j.

$\mu_{j,k}^{(i)}$: Binary variable indicates that subchannel k is assigned by the eNodeB to VUE_j.

$H_{j,m,k}^{(i)}$: presents the channel coefficient of subchannel k between Tx_j and Rx_m in RB_i.

$n_m^{(i)}$: CN~$(0,\sigma^2)$ is the additive white Gaussian noise (AWGN) for reciever VUE_m, and σ^2 is the noise variance. Furthermore, the channel gain can be defined as:

$$H_{j,m}^{(i)} = G \cdot h_{j,m,k}^{(i)} \cdot \beta \cdot \left(d_{j,m}^{(i)}\right)^{-\alpha} \quad (3)$$

Denotes:

G: Constant power gain factor implied by the amplifier and antenna.

$h_{j,m,k}^{(i)}$: Complex Gaussian variable CN (0,1) indicates the rayleigh fading.

β: Shadowing fading.

$d_{j,m}^{(i)}$: Distance between VUE_j and VUE_m.

α: Path loss exponent.

3.2. Propagation Scenario

Many covered propagation scenarios are detailed in [25] and [26]. It is indispensable to extend such modeling and measurement efforts to various environments that are also very important for V2X communications, mainly when safety applications are supported. Moreover, modeling a radio channel has been one of the most challenges in the mobile radio system design, and it is based on measurements specifically made for an intended communication system. Propagation models are the base for channel modeling, as they try to describe the way a radio signal changes during its travel distances from the Tx to the Rx and to mitigate interference. Also, radio propagation has a significant impact on the performance of cellular mobile radio systems due to increased data rate, bandwidth, mobility and QoS.

Then more, the trajectory between the Tx and the R can differ from a simple Line Of Sight (LOS) to a line that is obstructed by mountains, buildings, or foliage, named Non line of sight (NLOS).

The main characteristics of the radio channel - such as path loss, fading and time-delay spread and the parameterization vary from each scenario to an other one similar environments.

The macro-cell propagation can be divided into two classes as urban macro-cell propagation and rural macro-cell one. As the surrounding building heights, the density of obstacles and the number of users is very high in urban area, the cellular network is more robust and complicated where one macrocell antenna serves many microcells and picocells to ensure a better service compared to rural area because of the presence of less buildings and obstacles in the area [26].

For our work, we focus on rural macro-cell and urban macro-cell senarios with vehicles travelling on multiple lanes and on random position in the cell, including other pedestrians. Since we are considering the network-controlled allocation under independent Rayleigh fading channel and the line of sight (LOS) propagation properties of the vehicular networks, without loss of generality, we only incorporate the effect of the distance dependent path loss between the eNodeB and each VUE into analysis. Therfore, Path-loss models play an important role in the design of cellular systems to specify key system parameters such as antenna heights, frequency, transmission power and so on.

According to [26], The path-loss models are parameterized and controlled using results from a wide-ranging set of measurement campaigns, and they are extensively used and accepted.

Path loss (PL) is a measure of the average radio frequency attenuation suffered by a transmitted signal when it reaches the receiver, after traversing a path of several wavelengths [25]. It is defined by:

$$PL=\alpha*\log10(d[m])+B+C*\log10(fc[Ghz]/5) \quad (4)$$

Were, the fitting parameter α includes the path-loss exponent, B is the intercept, C describes the path loss frequency dependence and fc is the center frequency.

However, the path loss frequency dependencies model used in our scenarios are defined below:

$$PL_{Rural}[dB] = 40*\log10(d[m])+10.5-18.5*\log10(h_{BS}[m])-18.5*\log10(h_{MS}[m])+1.5*\log10(fc*1e-9/5) \quad (5)$$

$$PL_{Urban}[dB] = 40*\log10(d[m])+13.47-14*\log10(h_{BS}[m])-14*\log10(h_{MS}[m])+6*\log10(fc*1e-9/5) \quad (6)$$

where, h_{BS} and h_{MS} are the antenna heights at the base station and the mobile station respectively. d denotes the distance between the Tx and the Rx.

4. Proposed Algorithm

This proposed technique aims achieve a higher cell throughput, consider the heterogeneity of VUEs channel gains and the equity criteria. This algorithm trys to increase the number of served safety V2X links in the cell while a minimum requirement must be guaranteed on terms of throughput and to maximize the average data rate of the NST VUEs.

Our proposed algorithm named Priority -based Scheduling Algorithm for NOMA integrated V2X (PSA), can be considered as an assignment solution that aims to obtain an efficient and optimal spectrum allocation for VUE. Firstly, we begin by services classification, also, the incoming links of the VUE are classifyed according to their types of services as: safety traffic and non safety traffic . The NST is divited into two type of services as Management Traffic (MT) and Infotainment Traffic (IT).

Then, we assign the physical resources blocks to the pairs VUEs which can occupy the same spectrum by taking into account the channel condition of the transmitters on each RB. The designed approach enhances the equity index in the processus of resource allocation among diverse VUEs class of services. In fact, if the selected VUE has already assigned to an RB during the current TTI, this one will have't another radio resource block in the same TTI while all the selected users are not served. This, at the same time, raises the spectral efficiency and the capacity of the system inside the coverage area, and ensures a better compromise between the average throughput and the fairness index.

Once, the RB allocation is performed, a power control process is used to treat the interference and minimize the binary error rate by calculating per each resource block the P_{factor} that allows the safety user to obtain the needed throughput as well as maximizing throughput of the non-safety user, also, each VUE can attain on the allocated RB without minimizing its achieved throughput. The non-scheduled safety VUEs in the present TTI will be selected with the superior priority to be served for the subsequently TTI. The scheduling algorithm of our system is shown in Figure 1.

The mean objective of our proposed algorithm PSA is maximizing: the cell throughput, the number of served VUEs, the number of served safety VUEs, the sum rate of non-safety communications. We try to minimize the interference inter VUEs using the same RB by imposing a power constraint for different vehicular UEs.

5. Description of the Proposed Algorithm

In this part, we present more details for all steps of our proposed strategy.

5.1. Step 1: Service Classification

NOMA resource allocation schemes need to be designed to serve the ever-increasing number of vehicles that are connected for different services, such as safety features and passengers multimedia amenities.

V2X services are a real time traffics and arrive with strict exigency on throughput, reliability and latency. Also, we consider two types of service classes as Safety Traffic (ST) that imposes timeliness requirements, network scalability, and stringent reliability and refers to the case when the transmit signals need high QoS and requires to be selected almost momentarily in order to minimize road accidents, such as traffic condition warning, curve speed warning, vulnerable road user warning and emergency vehicle warning [27].

The NST (Non Safety traffic) included two categories of services: MT and IT. The Management traffic required to be scheduled with most importance thanks to their significance which aims to enhance traffic flow by minimizing traffic congestion and travel time, examples traffic information and recommended, intersection management, regulatory speed limit and automated parking system [28].

Infotainment Traffics are characterized by the delay tolerance and provides convenience applications and comfort for VUEs to improve their driving experience, exemples media downloading, vehicle software update and data provisioning [28].

The idea of the first step is to sort the arriving vehicular links according to their type of service that are: ST, MT and IT.

5.2. Step 2: User Scheduling

Scheduler functionality is to allocate the RBs to the UEs according to their requested requirements and their radio conditions at every TTI. So, the available resources blocs are very limited and it has to be allocated with an optimal manner that highest cell throughput and equity index are ensured among all VUEs, applications or service classes.

The system comprises two categories of V2X communications. Let $\varphi=\{1\ldots j,\ldots N_{ST}\}$ the set of Vehicular users with Safety Traffic (VST) and $\psi=\{1\ldots j',\ldots N_{NST}\}$ the set of VUEs with Non Safety Traffics (VNST), where $\{\varphi, \psi\} \in \{I\}$.

Moreover, the purpose of this step is: firstly, to select a set of VST by $\acute{C} \in \{\varphi\}$ while guarantying their requirement in terms of throughput. The idea is to search the combination (VST_j, RB_i) where $1 \leq j \leq N_{ST}$ and $1 \leq i \leq M$, with the highest metric on the RB_i. This metric expresses the channel gain between the VST_j, $j \in N_{ST}$ and the receiver m on subcarrier k, denotes by $\left|H_{j,m,k)}^{(i)}\right|^2$, which includes both small-scale Rayleigh fading and large-scale path loss.

Secondly, to select a set of VNST by $\acute{R} \in \{\psi\}$ that will be shared the same $RB_i$with VST. The selection of VNST is based on the higher channel condition expressed by the estimated SINR which corresponds to the CQI (Channel Quality Indicator) index sent by the VNST to the eNodeB.

Also, priority users scheduling considered as an important tasks to enhance the network performances in terms of aggregate sum rate and equity index before the resource allocation.

5.3. Step 3: RBs Allocation

The third step is designed based on NOMA to maximize the aggregate throughput of the system by assigned the limited physical RB to pair of VUEs (VSTj, VNSTj') selected in the above step. Exploiting the knowledge on vehicle positions and propagation characteristics at the eNodeB, this latter allocate the resources to guarantee the required quality of service and maximize the network capacity. Moreover, the proposed RB allocation method is divided on two processes. Starting with the RB allocation for ST communications which aims to allocate the candidate $\acute{C} \in \{\varphi\}$ safety users to different resource blocs and set a minimum requirement must be guaranteed for those. Then, the RB allocation for VNST candidate $\acute{R} \in \{\psi\}$

The proposed scheme offered superior priority to the safety links than others type of traffics because they can't tolerate delay and must be scheduled and served immediately without forgetting to taking into considiration the equity index. Moreover, in this step the scheduler checks if each VST has the minimum requirement in terms of throughput. In fact, if the VUE is't satisfied, for the scheduler, this vehicular user is reselected to transmit for the next TTI. This allows us to maximize the served VUEs number by trying to give each one his requested requirement. The algorithm for RB allocation is summarized as following.

Algorithm 1: RB allocation

```
{Initialization};
// select the set B safety VUE candidates can be served to
different resource blocs;
Check each line of the matrix H_safety and calculate the
matrix rb_acceptable_safety : if a single allocation is
acceptable that is used (where an allocation can be used is '1'
or can't be used is '0');
at_least_one_allocated = true;

While i=1 to N at_least_one_allocated do
    at_least_one_allocated = false;

    For i=1 to all the vehicle of safety traffic do
        if sum(rb_acceptable_safety(i,:))==1 then
            Find the best VUE candidate to be served at the
            rb_to_allocate;
            Select the best VUE candidate to be served at the
            rb_to_allocate;
```

```
        Store the number of selected VUE corresponds to
        the index the rb_to_allocate which granting the
        minimum requirement in terms of throughput;
        at_least_one_allocated = true;
      end
    end
  End
  // Allocate the different RBs for the selected Ć safety
  VUEs.
  While sum(sum(rb_acceptable_safety(:,:))) > 0 do
    Calculate H_safety matrix of the channel coefficient per
    each selected VUE at the rb_to_allocate;
    Find (ue_to_allocate,rb_to_allocate) = max h_matrix
    (ue_to_allocate,rb_to_allocate), with
    (ue_to_allocate,rb_to_allocate) ∈ Ć X U;
    Selecting the best VUE with highest channel gain;
    Remove the ue_to_allocate row from h_matrix;
    Assign rb_to_allocate to ue_to_allocate ;
    Update U: U = U \ { rb_to_allocate };
    Update Ć: Ć = Ć \ { ue_to_allocate };
  End
```

We define an indicator defined by a variable s_j^i , (i∈M, j∈N) which indicates if the ith RB is assigned to j^{th} VUE or not.

$$S_j^i = \begin{cases} 1 & \text{if } RB_i \text{ is used by } VUE_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

This step is designed to maximize non safety vehicular users throughput (for both management and infotainment VUEs) by prioritizing the users with good channel condition and not to starve others with bad channels condition.

Specifically, the SINR at the receiver Rx with the highest channel gain which can cancel the interference is computed as:

$$SINR_{j,k} = \frac{s_j^{(i)} P_{factor}^{(i)} P \rho_{j,m,k}^{(i)}}{1 + \sum_{l \in L_n, l \neq j} (1 - P_{Factor}^{(i)}) P \rho_{l,m,k}^{(i)}} \quad (8)$$

where Ln represents the subset of the interfered VUEs when the receiver m exploiting the resource block for the transmission.

5.4. Step 4: Performance Evaluation

To collect the benefits of the output of our proposed scheme, certain performance evaluations are need to be addressed such as cell throughput and average blocking rate.

Algorithm 2 briefly describes this process.

Due to the unavailability of radio resources, ABR is a typical constraint for VUE with safety traffic (VST) in PSA which aims to maximize served VST thanks to their high priority, great QoS requirement and due to better road utilization and surroundings. The average blocking rate is calculated as following:

$$ABR = \frac{\text{Number of blocked transmitters}}{\text{Total number of VST}} \quad (9)$$

Algorithm 2: Throughput calculation

```
For i=1 to N do
  RBi is used by safety VUE then;
  Determinate the channel coefficient H_ST of the selected
  safety VUE;
  Determinate the channel coefficient H_NST of the selected
  non safety VUE;

  If H_ST > H_NST then
    % safety is the primary VUE that cancels the
    interference
    Calculete P_factor;
    P_factor= (N0/(P*H_ST))*(2^(thr_min_safety/Bw_rb)-
    1);
    % N0: Additive White Gaussian Noise (AWGN);
    Calculate the throughput of safety VUEi;
    thr_user_safety=(Bw_rb)*log2(1+(P*(P_factor)*H_ST/(P
    *(1-P_factor)* H_ST + N0)));
    Calculate the throughput of non safetyVUEi;
    thr_user_nonsafety= (Bw_rb)*log2(1+(P*(1-P_factor) *
    H_NST / (P * P_factor * H_NST+N0)));
  else
    % safety is the secondary VUE that cannot cancel the
    interference;
    P_factor= (2^(thr_min_safety/Bw_rb)-1)*(P*H_ST+N0) /
    ( 2^( thr_min_safety / Bw_rb)*(P*H_ST));
    Calculate the throughput of safety VUEi:
    thr_user_safety=(Bw_rb)*log2(1+(P*(P_factor)*H_ST/(
    P*(1-P_factor)*H_ST+N0)));
    Calculate the throughput of non- safety VUEi;
    thr_user_nonsafety = (Bw_rb)*log2(1+(P*(1-P_factor)*
    H_NST / N0 ));

  end
end
```

In this step we calculate per each resource block the Pfactor that allows the safety VUE to obtain the needed throughput and to maximize the sum rate of non-safety VUE at the same time. Indeed, the objective is to take into account power split between the two vehicular users which share the same RB to minimize the interference.

The achievable data rate obtained from VUE_j over sub-channel k at the RB_i can be mathematically formulated as:

$$R_{j,m,k}^{(i)} = B\_W \log_2 \left(1 + \frac{s_j^{(i)} P_{Factor}^{(i)} P \rho_{j,m,k}^{(i)}}{1 + \sum_{l \in L_n, l \neq j} (1 - P_{Factor}^{(i)}) P \rho_{l,m,k}^{(i)}} \right) \quad (10)$$

6. Simulation Results

In this section, simulation results based on Matlab are presented and performed to demonstrate and validate the efficiency of the developed PSA schemes. We assume that the position and the distribution of vehicular users are randomly in the

coverage area for the current TTI. Simulation results provide a better insight of various issues i.e. average throughput, BER and fairness analysis. In addition, we vary the percentage of ST and NST services, we take as an example three scenarios that are mentioned in table 3where scenario 1 is oriented Non Safety Traffic (NST), scenario 2 is marked by equity of percentage of traffic and the last one is oriented Safety Traffic (ST), in order to evaluate the system performance and show that our algorithm is efficient and applicable for any scenarios. So the objective of these scenarios is to highlight the effect of the priority on the average blocking rate, the cell throughput and the equity index. Moreover, the performance results of the developed algorithm are compared with the conventional OMA based allocation which is detailed in [29] and some recent works for NOMA based allocation to demonstrate the benefits of our proposed approach.

Table 3: Parameter setting

Parameter	Value		
Bandwidth	10 MHz		
Available resource blocks	50		
connected VUEs in the cell	[100, 150]		
eNodeB in the coverage area	1		
Transmitted power	15 dBm		
Number of VUEs allocated the same spectrum	2		
TTI	1 ms		
	Scenario 1	Scenario 2	Scenario 3
Percentage of ST services	30%	50%	70%
Percentage of MT services	40%	30%	20%
Percentage of IT services	30%	20%	10%

The used parameters for simulation results are announced below in Table 3.

6.1. ABR Analysis

ABR is an important Key Performance Indicators (KPIs) [30]. It is presenting the ratio between the number of blocked transmitter VUEs due to the limited spectrum and the total number of vehicular users defined for each scenario. Moreover, the average blocking rate achieved by the PSA algorithm was calculated by their average values for all scenarios and traced as illustrated in Figure 2.

While increasing the number of connected vehicular UEs in the coverage area, the average blocking rate raises a little or will be perpetually starved in scenario 1 due to the limitation of radio resource blocks. We mark that the raise of average blocking rate is less important for safety traffic in scenario 1 than the others one, those explain that the RBs are assigned based on the type of traffic connections. Scenario 1 is oriented NST traffics, so the number of VST is less than the number of VNST. In addition, the main objective of the proposed approach is to maximize the number of vehiclar users served by assigning a physical RB to two VUEs and by giving more priority to the safety links which must be served immediately due to their timeliness requirement. In scenario 1 and 2, the user with good channel condition can guarantee the minimum required QoS, has a superior priority to be scheduled while VUE with a mediocre channel gain waiting their access, which conducts to a equity problem and high latency. PSA aims to allocate multiple VUEs with diverse channel qualities simultaneously. Moreover, it can supply an enhancement by ensuring massive connectivity as indicated in Figure 2.

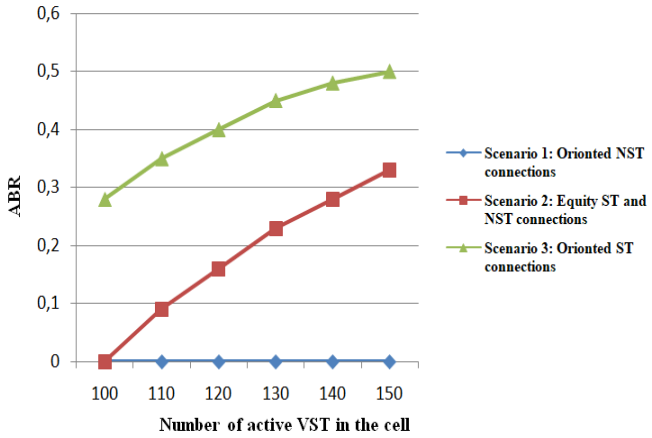


Figure 2: ABR analysis for ST communications

6.2. Cell Throughput

Maximizing average throughput is the principal goal of our algorithm. Figure 3 illustrates the performance of PSA scheme in terms of overal throughput for the three already mentioned scenarios versus the number of connected links in the cell.

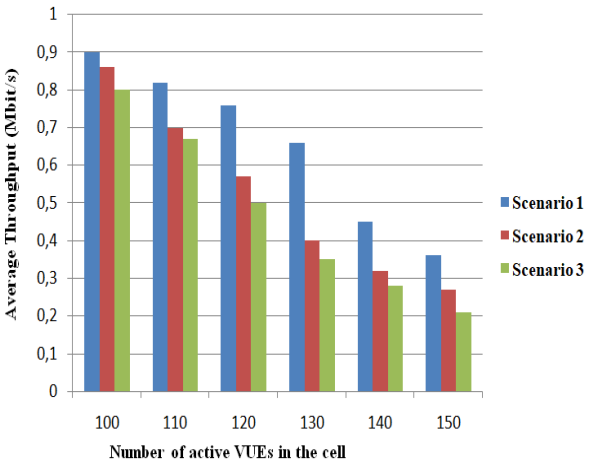


Figure 3: Average throughput analysis for NOMA (Rural Macro-cell)

We show that the raise in cell throughput is more better for scenario 1 than for scenario 2 and scenario 3 which is less efficient. From the Figure 3, it is observed that the total throughput of the proposed algorithm gradually decreases when the number of VUEs increases. This is due to the large demand on resources blocs. Since the proposed algorithm cancels the waste of the RBs by updating the metric before each new allocation and by maximizing the use of available RBs. Indeed, the proposed algorithm achieve 0,9 Mbps (scenario 1) when the targeted active VUEs in the cell is

equal to 100. This scenario oriented to be NST, while this type of traffics needs a high throughput to provide desired QoS. On the other hand, scenario 3 is oriented safety traffic. Taking into account the type of communications, this increase affects the throughput because in real world scenarios not all V2X connected links are safety communication and not all VUEs are close to the base station. ST connections often requires stringent exigency as will as low latency and high reliability. Indeed, safety traffics representing small data packets compared to non safety traffics. For ST connection like control loss warning, pre-crash sensing warning, forward collision warning and queue warning, the short transmit packets have a size of 50-300 Bytes [31] [32]. Figure 3 shows that the acheived throughput for ST can support several packets sizes. So, PSA allows exchanging packets size along to 120 Bytes when the average sum rate is equal to 0.9 Mbps. Moreover, the results satisfy the recommendation which is imposed by 3GPP organization when the lowest message size is 50 Bytes.

Indeed, Figure 4 shows how the cell throughput varies according to the number of active connections in the coverage area for the conventional OMA.

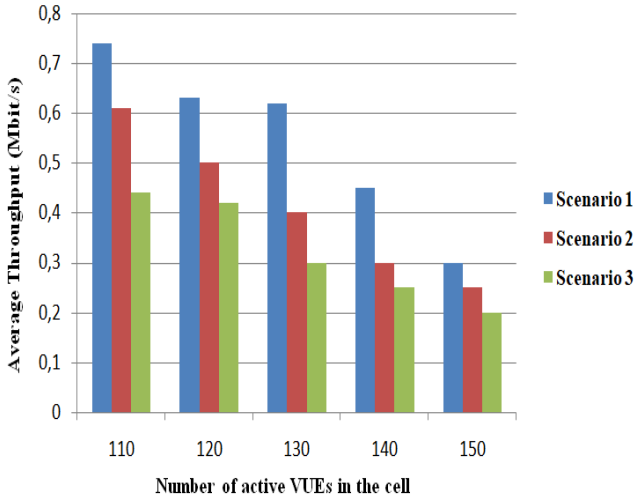


Figure 4: Average throughput analysis for OMA (Rural Macro-cell)

It's obvious that overall throughput decreases when VUEs increases. From that, it is clear that NOMA outperforms OMA even in all scenarios, which verifies that PSA based on NOMA can attain a higher cell throughput than OMA which has a gap to the greatest attainable cell throughput due to it uselessness on active V2X links selection. So, our proposed algorithm out performs OMA-based scheme due to a better resources allocation.

This is because PSA effectively preserves user diversity by its RB allocation metric. Unlike OMA scheduler, NOMA can simultaneously ensure optimal use a huge number of VUEs as well as performing better QoS requirements.

System model of our proposed algorithm does not consider only rural propagation scenario but also urban environment. Figure 5 illustrates the average throughput varies with the number of connected vehicular users in the cell for urban macro-cell propagation scenario evaluated for three different scenarios.

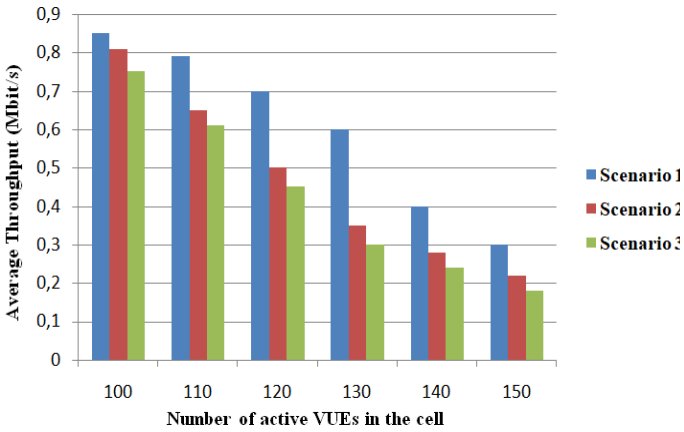


Figure 5: Average throughput analysis (Urban Mcro-cell)

Urban macro-cell propagation will face much path loss due to various properties, such as diffraction, reflection, refraction, absorption, free space loss, coupling loss and aperture-medium compared to the rural macro-cell propagation. This is confirming that the average sum rate for rural macro-cell outperforms the urban macro-cell average throughput as shown in figure 5. Moreover, all propagation scenarios based on NOMA algorithms are much better compared to OMA one, which confirming the convergence and stability of our proposed algorithm and verifying that NOMA can obtain a higher system throughput than OMA.

In addition, Figure 6 displays the system throughput of the developed resource allocation approach PSA compared with the attainable system throughput in [21] versus of the number of vehicular UEs in the system. Our algorithm is compared to [21], already detailed in related work section, because we find that this reference present the most high throughput compared to other studied related works.

Tree-based Search power Allocation algorithm (TSA) can both solve the link selection problems, user scheduling and power allocation with a low computational complexity for Device to Device underlaying cellular network based on NOMA techniques to increase the average throughput [21].

For FIxed (channel-Independent) Power Allocation algorithm (FIPA), the set of users is sorted in the decreasing normalized channel gain order, the system reaches to assign more power level to users experiencing high channel qualities based on the fixed weighted sum of the instantaneous throughput [33].

For FRactional Power Allocation (FRPA), the transmission power control depends on channel condition variations among users. Respective users must be informed by allocated power through explicit control signaling, which differentiates this allocation from the fixed power allocation [33].

For our proposed PSA algorithm and TSA defined in [21], the cell throughput increases with a varying speed that increasingly slows down due to increasing the number of connected VEUs in the system.

From Figure 6, the increment of the average throughput, PSA is better than TSA and therefore outperforms the other three algorithms FRPA and FIPA.

Figure 6 also clearly demonstrates the superiority of PSA that outperforms the compared scheme in terms of sum rate since the proposed algorithm takes into account the service classification and the channel gains of the respective non-orthogonally multiplexed VUEs.

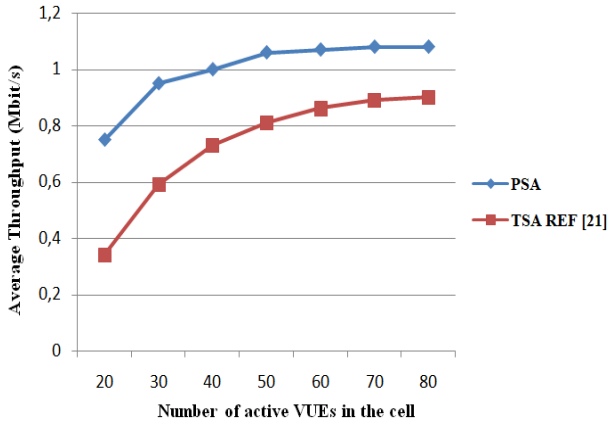


Figure 6: Average throughput analysis

6.3. Fair Index

The developed solution focuses on priority management to serve users fairly and increase the number of served VUEs. Fairness is a key factor for evaluating the simulation results of scheduling and resource allocation schemes, its calculates the equity index among VUEs. As PSA aims to maximize the minimum achievable instantaneous data rate, this can ensure fairness among users. To measure the equity of the system, we exploite the well-known techniques, it is Jain's fairness index [34] given by equation "(11)".

$$f(D_1, D_2, D_3, \dots D_K) = \frac{\left[\sum_{K=1}^{N_s} D_K\right]^2}{N \sum_{K=1}^{N_s} (D_K)^2} \quad (11)$$

where N_s is the number of served VST users and D_K is the throughput of the Kth VST.

We can observe on Figure 7 that the designed scheme has a good level of equity index for VST because it maximizes the use of the available RB between different V2X communications. Moreover, PSA tend to minimize the rate gap among scheduled users, thus maximizing fairness among them. Comparing with OMA algorithm, the proposed algorithm achieves not only higher system throughput but also better fairness between subscribers.

It was shown that the fairness index of the proposed resource allocation scheme with NOMA system is higher than others paper [18]. The equity of all algorithms decreases when the number of connected links raises. The authors simulate this factor of performance with limited active connections in the system. This demonstrated that the proposed algorithm can ensure the best fairness performance by the different number of connected vehicular users comparing with resultas shown on [18].

Simulation results have illustrate that the proposed algorithm PSA has various proprieties that maintain remarkable enhancement in terms of average throughput, Average Blocking Rate (ABR) and to guarantee fairness index between different class of VUEs comparing to the conventional OMA assisted relay scheme and some recent work for NOMA based allocation. According to [31] and [32], the simulation results guarantee the 3GPP recommendation when the minimum small packet size is 50 bytes. Hence, our proposed algorithm is capable to support variable packets sizes meeting the requirements imposed by 3GPP standard.

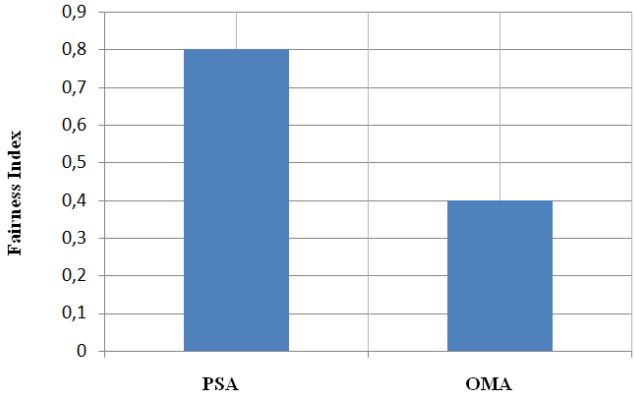


Figure 7: Fairness index analysis for ST connections

7. Conclusion

This paper, presents a novel scheduling and resource allocation approach named PSA for NOMA systems. PSA aims to maintain the highest average throughput by considering the heterogeneity of channels quality of each V2X link. The proposed work increases the number of served safety V2X links in the cell while a minimum requirement must be guaranteed on terms of throughput and maximize the average throughput of NST VUEs. The simulations results verify the optimality of our proposed algorithm in comparison with conventional scheme OMA that lack of caring fairness between vehicular users. The proposed scheme performs amelioration in term of throughput compared to other studied publication papers. A better compromise between cell throughput and fairness index has been achieved. As perspective for this research work, we plan to study the possibilities offered by Quanser platforms for reel environment implementation. In fact Quanser is an open-architecture system integrating plants with realistic dynamics sufficient for physically-relevant testing, real-time computation, and modeling for rapid testing of algorithms and concepts in MATLAB/Simulink and LabVIEW environments. We can also adopt such improvement to optimize the proposed scheme in terms of complexity.

References

[1] A. Al-Dulaimi, X. Wang, C. Lin , "Machine-Type Communication in the 5G Era: Massive and Ultrareliable Connectivity Forces of Evolution, Revolution, and Complementarity," Wiley-IEEE Press, 2018, doi:10.1002/9781119333142.ch14.

[2] Y. Han et al., "Periodic Radio Resource Allocation to Meet Latency and Reliability Requirements in 5G Networks," IEEE Vehicular Technology Conference (VTC Spring), 2018, doi:10.1109/vtcspring.2018.8417636.

[3] B. Di et al., "Non-Orthogonal Multiple Access for High-Reliable and Low-Latency V2X Communications in 5G Systems," IEEE Journal on Selected Areas in Communications, 2017, doi:10.1109/jsac.2017.2726018 .

[4] Z. MacHardy et al., "V2X Access Technologies: Regulation, Research, and Remaining Challenges," IEEE Communications Surveys & Tutorials, 2018, doi:10.1109/comst.2018.2808444.

[5] A. Masmoudi et al., "A Survey on Radio Resource Allocation for V2X Communication," Wireless Communications and Mobile Computing, 2019, doi:10.1155/2019/2430656.

[6] F.A. Giovanni, "An Overview of Vehicular Communications," Future Internet, 2019, dio:10.3390/fi11020027.
[7] B. Di et al., "V2X Meets NOMA: Non-Orthogonal Multiple Access for 5G-Enabled Vehicular Networks," IEEE Wireless Communications, 2017, doi:10.1109/mwc.2017.1600414.
[8] S. Rai et al., "Result Analysis of Efficient MIMO Framework for NOMA Downlink and Uplink Transmission based on Signal Alignment," International Journal of Computer Applications 2018, doi: 10.5120/ijca2018916907
[9] A. Agarwal et al., "Performance analysis for non-orthogonal multiple access (NOMA)-based two-way relay communication," the Institution of Engineering and Technology (IET Communication), 2019, doi:10.1049/iet-com.2018.5641.
[10] X. He et al., "Spectrum and Power Allocation for Vehicular Networks with Diverse Latency Requirements," IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 2019, doi:10.1109/smartgridcomm.2019.8909694.
[11] A. Bazzi et al., "Optimizing the Resource Allocation of Periodic Messages With Different Sizes in LTE-V2V," IEEE Access, 2019, doi:10.1109/access.2019.2908248.
[12] A. Nazari et al., "Resource allocation in power domain NOMA based cooperative multicell networks," the Institution of Engineering and Technology (IET Communication), 2020, doi: 10.1049/iet-com.2019.0624.
[13] N. Glei et al., "Power Allocation for Energy-Efficient Downlink NOMA Systems," international conference on Sciences and Techniques of Automatic control & computer engineering (STA), 2019, doi: 10.1109/STA.2019.8717240.
[14] K. Huang et al., "Energy efficient resource allocation algorithm in multi-carrier NOMA systems," International Conference on High Performance Switching and Routing (HPSR), 2019, doi:10.1109/HPSR.2019.8807992.
[15] Z. Yuan "Energy Efficient Power Allocation for Multi-carrier Non-orthogonal Multiple Access (NOMA) Systems with Proportional Rate Constraints," IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019, doi:10.1109/ITNEC.2019.8729390.
[16] M. Zeng et al., "Spectral- and Energy-Efficient Resource Allocation for Multi-Carrier Uplink NOMA Systems," IEEE Transactions on Vehicular Technology, 2019, doi:10.1109/TVT.2019.2926701.
[17] M. Pischella et al., "NOMA-Relevant Clustering and Resource Allocation for Proportional Fair Uplink Communications," IEEE Wireless Communications Letters, 2019, doi:10.1109/lwc.2019.2897796.
[18] K. Long et al., "Spectrum Resource and Power Allocation With Adaptive Proportional Fair User Pairing for NOMA Systems," IEEE Access, 2019, doi:10.1109/access.2019.2908673.
[19] X. Wang et al., "Low-Complexity Power Allocation in NOMA Systems With Imperfect SIC for Maximizing Weighted Sum-Rate," IEEE Access, 2019, doi:10.1109/access.2019.2926757.
[20] R. Abozariba et al., "NOMA based Resource Allocation and Mobility Enhancement Framework for IoT in Next Generation Cellular Networks," IEEE Access, 2019, doi:10.1109/access.2019.2896008.
[21] J. Wang et al., "A Novel Resource Allocation Scheme in NOMA-Based Cellular Network with D2D Communications," Future Internet, 2020, doi:10.3390/fi12010008.
[22] L. Chen et al., "Proportional Fairness-Based User Pairing and Power Allocation Algorithm for Non-Orthogonal Multiple Access System," IEEE Access, 2019, doi:10.1109/access.2019.289618.
[23] H. Xing et al., "Sum-rate Maximization Guaranteeing User Fairness for NOMA in Fading Channels," IEEE Wireless Communications and Networking Conference (WCNC), 2018, doi:10.1109/wcnc.2018.8376949.
[24] J. He et al., "Design and Optimization of Scheduling and Non-orthogonal Multiple Access Algorithms with Imperfect Channel State Information," IEEE Transactions Vehicular Technology, 2018, doi:10.1109/tvt.2018.2869305.
[25] METIS D1.4, METIS Channel Models, ICT-317669-METIS/D1.4, METIS project, 2013.
[26] WINNER II D1.1.2, Channel models, V1.2, 2008.
[27] G. Dimitrakopoulos et al., "Vehicular Communications Standards, Current Technologies in Vehicular Communication. Springer, Cham, 2017.
[28] C.M. Silva et al., "A Survey on Infrastructure-Based Vehicular Networks," Hindawi Mobile Information Systems, 2017, doi:10.1155/2017/6123868.
[29] A.D. Trabelsi et al., "Dynamic Scheduling Algorithm based on Priority Assignment for LTE-V2X vehicular Networks," International Wireless Communications & Mobile Computing Conference (IWCMC), 2019, doi:10.1109/iwcmc.2019.8766632.
[30] G. Cecchini et al., "Localization-Based Resource Selection Schemes for Network-Controlled LTE-V2V," International Symposium on Wireless Communication Systems (ISWCS), 2017, doi: 10.1109/ISWCS.2017.8108147.
[31] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects: Study on LTE support for Vehicle to Everything (V2X) services(Release 14) , 3GPP TR 22.885 V14.0.0, 2015.
[32] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects; Study on enhancement of 3GPP Support for 5G V2X Services (Release 15) , 3GPP TR 22.886 V15.1.0, 2017.
[33] N. Otao et al., "Performance of Non-orthogonal Access with SIC in Cellular Downlink Using Proportional Fair-Based Resource Allocation," International Symposium on Wireless Communication Systems (ISWCS), 2012, doi:10.1109/iswcs.2012.6328413.
[34] R. Jain, The Art of Computer Systems Performance Analysis, Wiley, 1991.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1266-1272 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Agricultural Data Fusion for SmartAgro Telemetry System

Ioana Marcu[1,*], Ana-Maria Drăgulinescu[1], Carmen Florea[1], Cristina Bălăceanu[2], Marius Alexandru Dobrea[2], George Suciu[2]

[1]*Telecommunication Department, University Politehnica of Bucharest, 010082, Romania*

[2]*R&D Department, Beia Consult International, Bucharest, 010082, Romania*

ARTICLE INFO

Article history:
Received: 16 July, 2020
Accepted: 15 October, 2020
Online: 24 October, 2020

Keywords:
Precision Agriculture
Data Fusion
Reliability
Sensors

ABSTRACT

Smart agriculture concept uses innovative solutions including IoT and Cloud storage features, dedicated sensors for monitoring basic agricultural parameters, new communications protocols, etc. SmartAgro architecture comprises a telemetry system for Key Performance Indicators (KPIs) such as air & soil temperature, air & soil relative humidity, leaf wetness, etc. The current paper outlines the reliability of the implemented system by comparing and analyzing data collected in spring 2019 and spring 2020. The relevance of this season consists in great air variations due to the transition from winter to summer. Being monitored in a vine area near Bucharest, these data may be useful for different statistics related to grapes culture in this season and can be used by interested parties for future predictions related to vine crops. Moreover, in this paper, data fusion will allow advanced data management and coherence achievement among collected data.

1. Introduction

This paper is an extension of work originally presented in SIITME'19 conference [1]. In [1] the authors presented the telemetry system with its main advantages consisting in solar panel supply and data reliability. The current work extends the demonstration of the system's reliability by processing data from Spring 2019 and Spring 2020. In addition to the previous work, data fusion will be used to fill the gaps between the recorded data and to ensure a proper system's management.

Considering all climate changes, the evolution of the agriculture plays an important role in the lives and well-being of people, since it represents a source of food for population and, respectively for domestic animals. Climate change influences agriculture in different ways. Changes in temperature and precipitation are already affecting crop yields [2]. Consequently, people must adapt to and implicitly adjust the solutions used to ensure food or water quality, for irrigation and daily use [3]. The health of soils and crop, also, is very important, as it affects the quality and quantity of agricultural crops [4, 5].

When required to evaluate successful analysed crops, the key performance indicators (KPIs) must be considered since they are quantitative, practical, directional, and actionable. Depending on the analysed corps, KPIs differ. For example, based on KPIs crop usage can be tracked to evaluate the production and to monitor the overall costs. The most significant impacts of KPIs on agriculture are increased productivity, profit and time save [6].

An important role in the precision agriculture is played by IoT platforms. Using them, the quality of the crops can be enhanced by real time data acquisition, processing and decision making. These data are converted, thus, in useful information for farmers, and, also, in a facile understandable manner [7]. Capturing, transmitting, storing and processing the volume of information collected by sensors connected on the IoT platform, show a number of challenges, in particular, regarding integration technologies, communications, databases and computing. A middleware platform which alleviates these issues is FIWARE. FIWARE is a technology supported by the European Commission to make possible the IoT in the context of the Future Internet [8].

SmartFarmNet is an IoT platform that automatically collects data from soil, as fertilization and irrigation. The data is then automatically correlated, and the invalid data is filtered-out from the perspective of assessing crop performance. Also, with the help of the platform the crop forecasts can be computed, and the farmers receive personalized crop recommendations for any farm [9].

There are also platforms which are specialised on only one aspect such as SWAMP. Within SWAMP project there was developed and assessed an IoT-based smart water management platform for precision irrigation in agriculture. The platform was

*Corresponding Author: Ioana Marcu, Email: ioana.marcu@upb.ro

built in such manner that it can be configured and deployed in different ways. Thus, the platform can deal with the requirements and limitations of different countries, climate, soils, and crops, which require flexibility to adapt to a range of deployment configurations involving mixed technologies [10].

Data fusion techniques applied on data collected by different sensors used in agricultural area allow a better understanding of parameters' evolution and advanced data management, especially in cases where the data volume is huge [11], [12]. Yet, the applicability domain is not limited only to agriculture, but it can comprise different applications that are sensor-based and that imply multiple data sources.

The current paper aims to emphasize the role of the implemented SmartAgro telemetry system in ensuring reliable data for further use in statistics and specialized predictions. Data fusion methods will allow additional processing that will offer a global perspective of the monitored parameters. The paper is organized as follows: Section 2 presents the related work on data fusion solution for key parameters monitored in precision agriculture; Section 3 contains the description of SmartAgro telemetry system and in Section 4 system's setup and relevant monitoring results are presented. Data fusion implications are outlined in Section 5 and Section 6 comprises Conclusions.

2. Data Fusion for Agricultural Area

According to [13], when acquirred data present high and diverse information, data fusion considers the juxtaposition of large set of data to ensure reliable, homogeneously and fair overview of the collected information. The advantage of data fusion of data received from multiple different sensors relies in "an improved estimate of physical phenomenon via redundant observations" [14]. The efficiency of data fusion was previously demonstrated in precision agriculture domain [15-18]. In [15], authors present the benefits on crop monitoring of 2D and 3D data fusion for a vineyard monitoring and use the results in order to classify vines in serveral classes by processing data from multiple sources (different sensors, Unmanned Aerial Vehicle (UAVs), etc). In 2012, in [16] different data fusion methods were used (e.g. multiple linear regression (SMLR), partial least squares regression (PLSR) and principal components analysis combined with stepwise multiple linear regression (PCA+SMLR) techniques) to predict multiple soil properties. Authors' conclusions indicate that data fusion techniques are more relevant in clayey field and worse in sandy field and, in addition, these methods can improve the quality of soil sensing in precision agriculture if appropriate sensors are selected. Later, in 2017, sensor data fusion for soil health assessment was applied in [17] and, as a result, faster determination of soil health was achieved by merging data gathered from all sensors. In a more advanced manner, in [18] sensing data fusion methods are involved in crop detection. Authors use an efficient method of fusing multi-source remote sensing images with a convolution neural networks (CNN) for semantic segmentation to identify crops (93% succesful rate) in detecting and identifying crops.

In this paper we use data fusion technique on agricultural KPIs to fill the gaps and to create a complete picture of their variation even in the absence of their recording by the telemetry system.

3. ADCON-based Architecture of Telemetry System

Monitoring of KPIs (such as air and soil temperature, crop state, air and soil relative humidity) for a vine located in a residential area close to Bucharest was performed using an ADCON-based telemetry system (called SmartAgro). The selected season was spring since it is a season in which high variations may be observed because of the transition from winter to summer (two seasons with extreme temperatures). Figure 1 illustrates the new concept of SmartAgro telemetry system in which different dedicated agricultural sensors are interconnected for main parameter's monitoring.

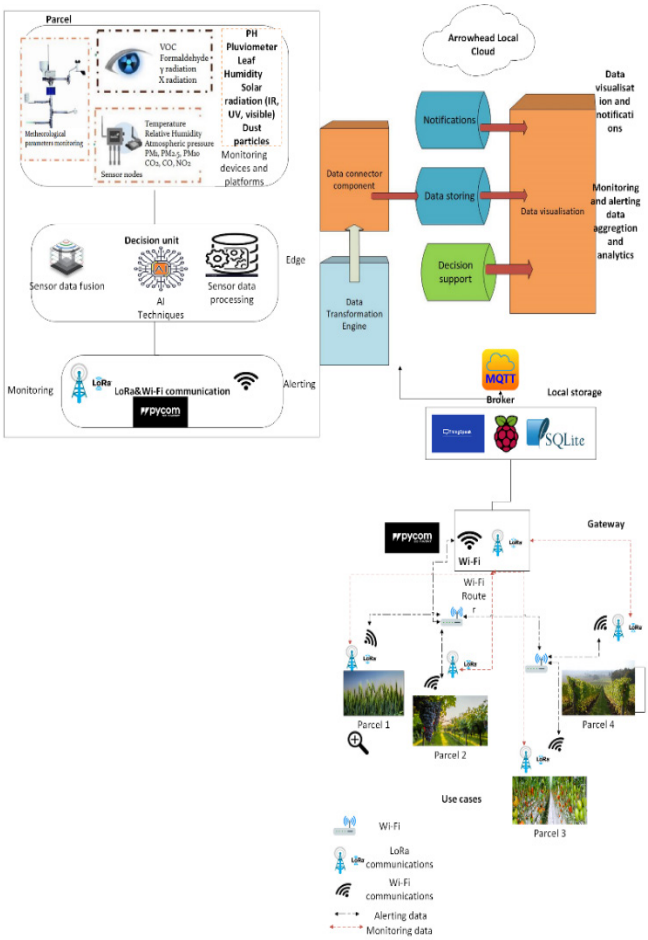


Figure 1: SmartAgro innovative architecture [19]

The data acquired from the agricultural sensors were centralized into a database and were used to highlight the impact of measured parameters on crops [19]. The architecture differs from the traditional ones by introducing two new levels consisting in: *The Edge level*: at this level telemetry data are passed through a decision-making system based on artificial intelligence techniques for data analysis and detection of abnormal values. Also, at this level, the data are classified as belonging to alert scenarios or simple monitoring. This determines the optimization of communication, in terms of traffic and energy consumption. LoRa technology is proposed for monitoring data, a technology known for extremely low energy consumption and for the field very extensive coverage. For scenarios involving alerts and critical change of parameters, Wi-Fi (short range) or 4G technology can

still be used for a large coverage area. Further, *the Local Storage level* has a role in storing relevant, processed, analysed, and labelled data to reduce latencies in alert scenarios and for applications. on-field, off-line. Additional functionalities of the proposed telemetry system are given in [19].

4. Extended monitoring results

To demonstrate the reliability of the system, extended monitoring results are presented. The measurements were performed in 2019 and 2020 and the data were acquired each 6 hours per day, starting with 8 a.m.

4.1. Air temperature monitoring

Figure 2 illustrates the variation of air temperature during Spring of 2019 and 2020.

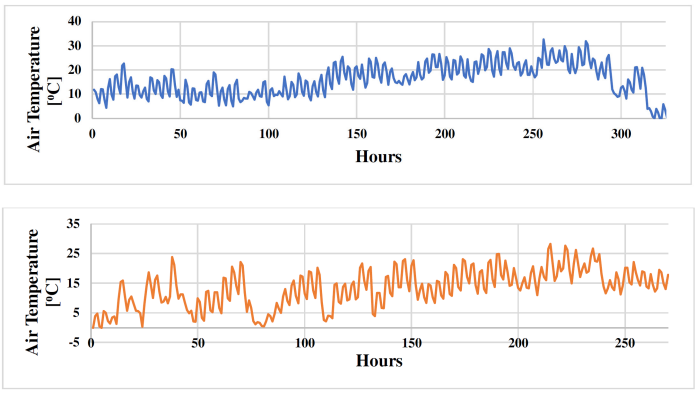


Figure 2: Air temperature variation in Spring 2019 (up), Spring 2020 (down)

Based on results in Figure 2, it can be seen the patterns of the day-night air temperature variation, also called Day/Night Differential (DIF). Higher temperature peaks can be observed in Figure 3. DIF value has multiple significances: firstly, DIF values were related to plant growth. Moreover, values of DIF around 8°C were proved to provide the best plant growth, whereas DIF values between 12°C and 22°C showed a low correlation to the predicted results in [20] for Chrysanthemum.

From the air temperature data provided by SmartAgro platform in the two seasons (Spring 2019 and Spring 2020), we can state that in Spring 2019 the plant growing should have been more pronounced, as the DIF was lower than 12°C.

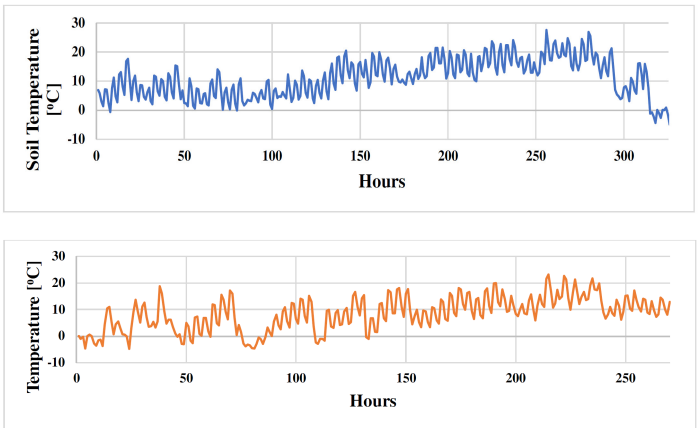


Figure 3: Soil temperature variation in Spring 2019 (up), Spring 2020 (down)

4.2. Soil temperature monitoring

From Figure 3, which outline the variation of soil temperature for Spring 2019 and Spring 2020, there can be observed that the patterns are identical with their correspondent air temperature data in Spring 2019 and, respectively, in Spring 2020, with the exception of a temperature offset of -5°C for both Spring 2019 and for Spring 2020.

4.3. Relative air humidity monitoring

From Figure 4, which depicts the variation of relative air humidity for Spring 2019 and Spring 2020, it can be observed that the supersaturation phenomenon was similarly frequent in both years. This is related to the prediction of the rainfall and appears when the air humidity reaches 100%.

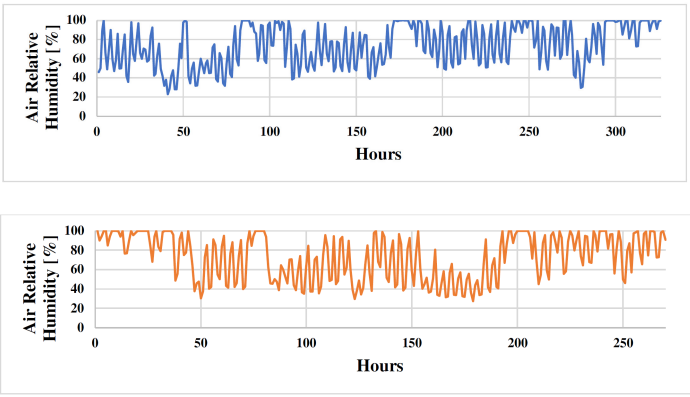


Figure 4: Relative air humidity variation in Spring 2019 (up), Spring 2020 (down)

4.4. Relative soil humidity monitoring

Figure 5 emphasizes the variation of relative soil humidity for Spring 2019 and 2020.

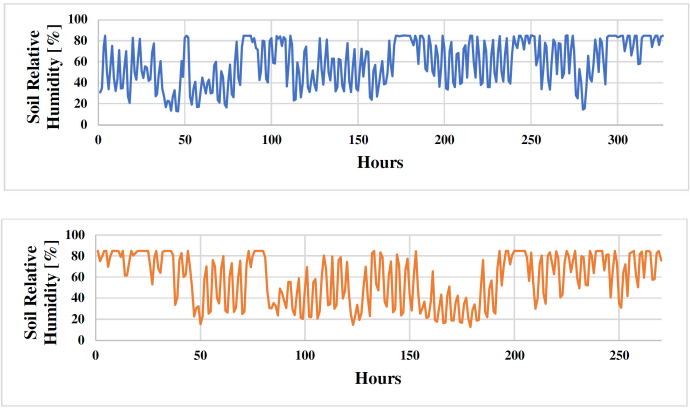


Figure 5: Relative soil humidity variation in Spring 2019 (up), Spring 2020 (down)

From Figure 5, it can be noticed that, similar to the soil and ambient temperature, the soil and ambient humidity follows the same pattern, with an offset of -15% both in Spring 2019 and in Spring 2020.

4.5. Leaf wetness monitoring

Figure 6 presents the leaf wetness variation for Spring 2019 and 2020. It can be observed that several peaks occur especially

after 21:00 p.m. From the analysed data, there is no evidence of correlation between the air humidity and the leaf wetness, nor between air temperature and leaf wetness.

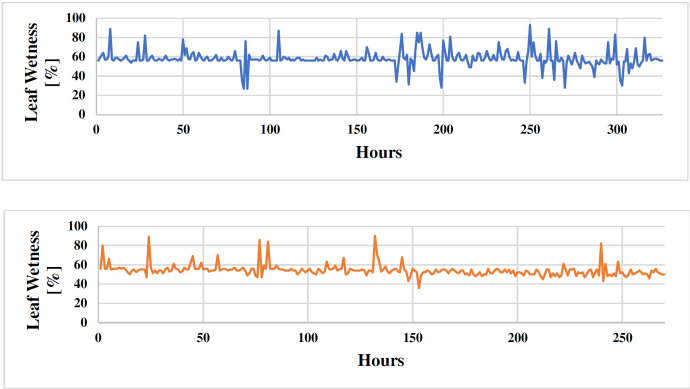


Figure 6: Leaf wetness variation in Spring 2019 (up), Spring 2020 (down)

5. Agricultural Data Analysis

The previous graphical representations did not reveal the absence of some data or the heterogeneity of the data. For example, from the previous graphs there cannot be noticed the absence of the samples for certain moments of the day or even the fact that the samples were not acquired at the same moment every day. Moreover, the data were not acquired each day.

The time intervals for data collection are March 4th, 2019 and May 31st, 2019 (Spring 2019) and March 1st, 2020 and May 28th, 2020 (Spring 2020), respectively. Analysing data of a real acquisition, it can be remarked that the samples were collected at the time moments given in Table 1.

Table 1: Classification of the samples based on the moment of acquisition

Hours (0-24)	No. of Days (ND)	ND (2019)	ND (2020)	Total Samples
3, 9, 15, 21	98	57	41	392
2, 8, 14, 20	48	23	25	192
3, 9	1	0	1	2
9, 15, 21	3	2	1	9
2	1	0	1	1
	Total			**596**

Unfortunately, missing data determine different issues concerning the predictions and forecasts or the decisions taken by the support decision systems integrated in the overall architecture. Nevertheless, sensor data fusion techniques can bring many benefits such that the faults triggered by the data gaps can be mitigated. For exemplification, two variables are considered: air temperature and soil temperature. They were chosen after the analysis of the samples acquired because the vectors storing their values comprise NaN values, that is, there are missing samples in air temperature and soil temperature data because of system failures.

In Figure 7, it is illustrated the graphical representation of the air temperature variation at 3 a.m. and 2 a.m. for each of the days in which the data were collected, more precisely, in the intervals March 4th, 2019 and May 31st, 2019 (88 days) and March 1st, 2020 and May 28th, 2020 (88 days). Figure 7 highlights small, but important, data gaps.

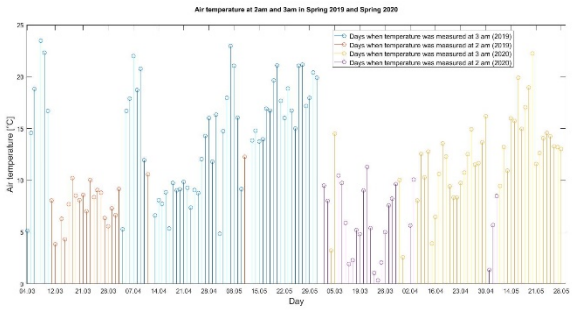


Figure 7: Air temperature measured at 2am and 3am in Spring 2019 and Spring 2020

In Figure 8, it is highlighted the soil temperature variation during the same seasons, and it can be noticed that also soil temperature data is missing. In addition, as previously mentioned, the soil temperature follows the same variation pattern as in the case of air temperature and, by computing the difference between air and soil temperature, a constant value of 5°C is determined. Therefore, finding a method to determine an approximate value for air temperature will also succeed in determining the approximate value for soil temperature, too, and vice versa.

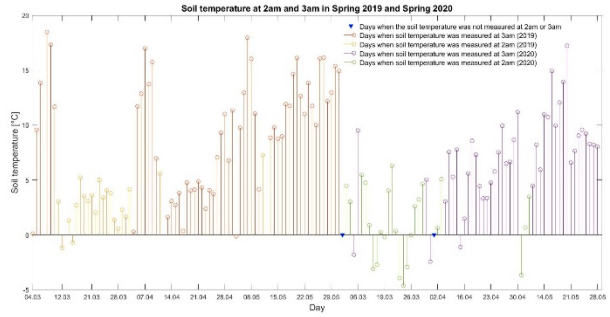


Figure 8: Soil temperature measured at 2am and 3am in Spring 2019 and Spring 2020

Further, the choice of soil temperature and air temperature variables will be justified with respect to the variation of the other parameters. Analyzing the graphical representation of air relative humidity (Figure 9), soil humidity (Figure 10) and leaf wetness (Figure 11) variations, missing data can be also observed (samples missing before and after March 13, 2019), but the data gaps are affecting all variables. The only variables that experience isolated data gaps when all the other parameters are represented are soil temperature and air temperature.

In Figure 9, air relative humidity variation for the same moment of day for Spring 2019 and Spring 2020 is represented. It can be observed that the minimum air relative humidity recorded in Spring 2019, at 2 a.m., is 37.78 % and it was recorded on March 24th, 2019, while the maximum value of 100% was reached 7 times (March 11-12, 2019; March 15, 2019; April 11th-12,2019; May 11th-12,2019).

Regarding the air relative humidity measured at 3 a.m., in Spring 2019, the minimum value was 55.74% on March 9th, 2019,

while the maximum value (100%) was reached 10 times (10% in March, 40% in April, 50% in May).

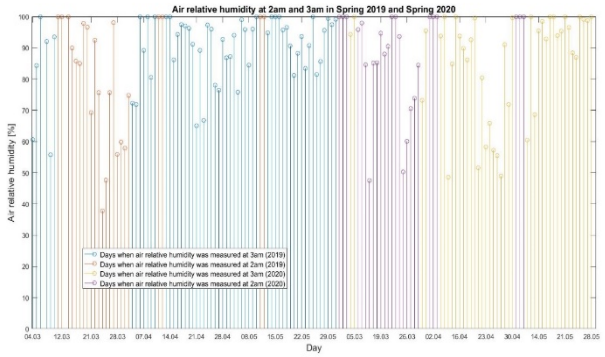


Figure 9: Air relative humidity measured at 2am and 3am in Spring 2019 and Spring 2020

In Spring 2020, the minimum value of the air relative humidity measured at 2 am was 47.45% on March 16, 2020, whereas the maximum value (100%) recorded at 2 a.m. was reached 10 times (40% in March, 30% in April and 30% in May).

Concerning the samples acquired at 3 a.m. in Spring 2020, the minimum value of air relative humidity was recorded on 13.04.2020 (48.57%), whereas the maximum value of 100% was reached 9 times (22.2 % in March, 11.1% in April and 66.7% in May).

Computing the difference between soil humidity and air humidity, a constant difference of 15% is achieved and it can be seen that the variations patterns are identical (Figure 10). This can be proven also by computing the standard deviation (SD) for each season for the two possible time moments of acquisition (2 a.m. and 3 a.m.) for both variables. For both soil and air humidity variables, the values of the standard deviation coincide and are given in Table 2.

Table 2: Standard deviation of soil and air humidity for Spring 2019 and Spring 2020 (at 2am and 3 am)

Season	SD (2am)	SD (3am)
Spring 2019	19.49	10.92
Spring 2020	15.76	16.85

Therefore, by finding the approximate value of the air relative humidity, the approximate value of the soil humidity can also be determined.

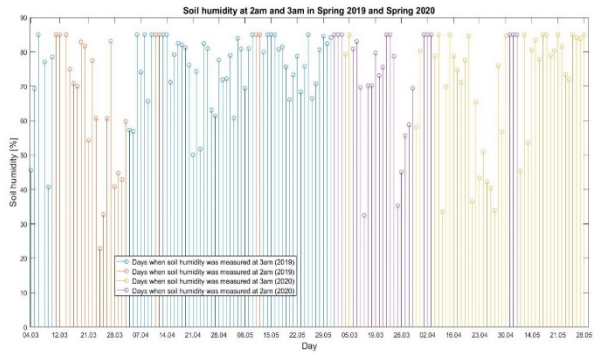


Figure 10: Soil humidity measured at 2am and 3am in Spring 2019 and Spring 2020

Finally, in Figure 11, the variation of the leaf wetness is depicted. The standard deviation (STD) of the leaf wetness parameter was computed.

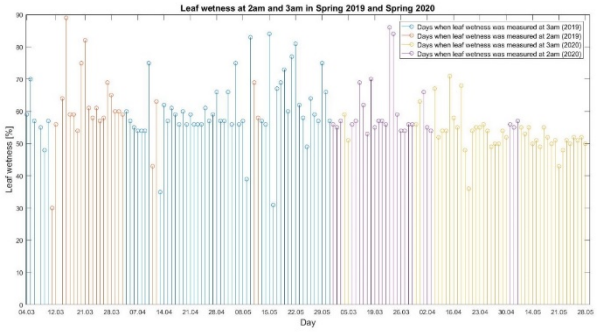


Figure 11: Leaf wetness measured at 2am and 3am in Spring 2019 and Spring 2020

The lowest values of SD were obtained for Spring 2020 3 a.m. (a small value with respect to Spring 2019 was also obtained for 2 a.m.), while the highest one is obtained for Spring 2019 (2 a.m.). The small variation of leaf wetness values in Spring 2020 is highlighted in the graphical representation, too. All values of the standard deviation are given in Table .

Table 3: Standard deviation of leaf wetness for Spring 2019 and Spring 2020 (at 2am and 3 am)

Season	SD (2am)	SD (3am)
Spring 2019	11.66	10.00
Spring 2020	8.64	6.04

6. Proposed data fusion algorithm

Next, based on the data analysis performed in Section 5, we proposed a data fusion algorithm based on a hybrid decision tree. Here, the hybrid attribute is given due to the fact that, with respect to the traditional binary tree approach, in the proposed algorithm three cases may arise: the value of the parameter is in range (1), the value of the parameter is out of range (2), the value of the parameter is not available (3). When a sample of a parameter is not available at the querying moment (case 3), the algorithm commands the estimation block that computes an estimated value of the parameter based on current samples of the other parameters and based on previous samples of the required parameter. In Figure 12, the proposed algorithm was depicted. In Table 4, the parameters acronyms and their meanings are given.

Table 4: Parameters acronyms and significance

Acronym	Significance
t_air	Normal air temperature range
t_soil	Normal soil temperature range
rh_air	Normal air relative humidity range
rh_soil	Normal soil relative humidity range
leaf_wet	Normal leaf wetness range
T_air	Current air temperature sample
T_soil	Current soil temperature sample
RH_air	Current air relative humidity sample
RH_soil	Current soil relative humidity sample
Leaf_w	Current leaf wetness sample

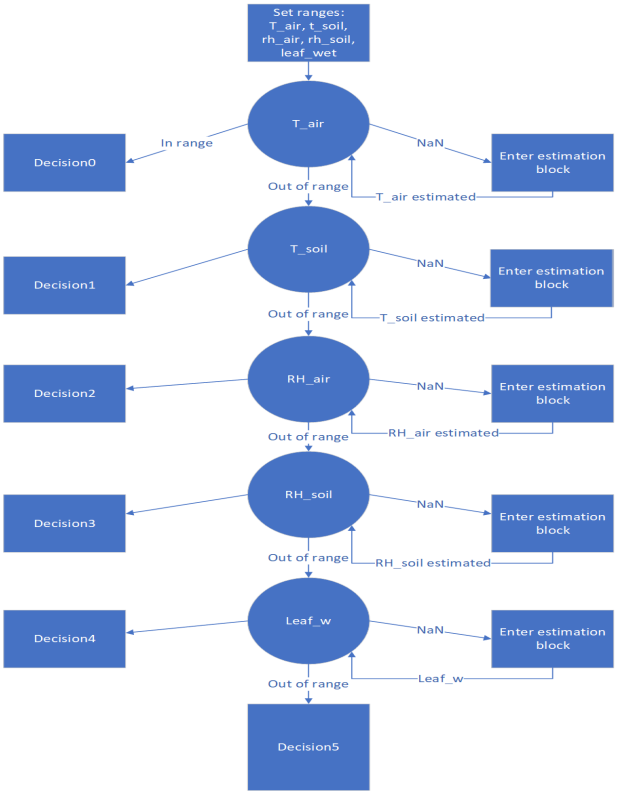


Figure 12: Hybrid decision tree for data fusion-based decision making

The correlation between the parameters is depicted in Figure 13-18.

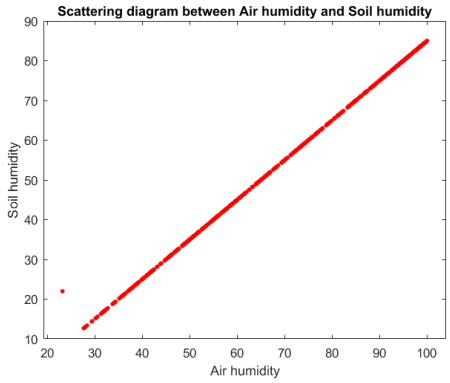


Figure 13: Scattering diagram of air humidity and soil humidity

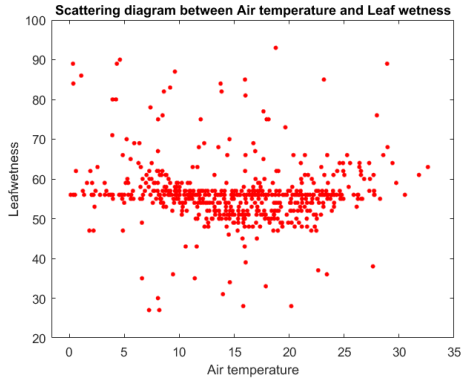


Figure 14: Scattering diagram of air temperature and leaf wetness

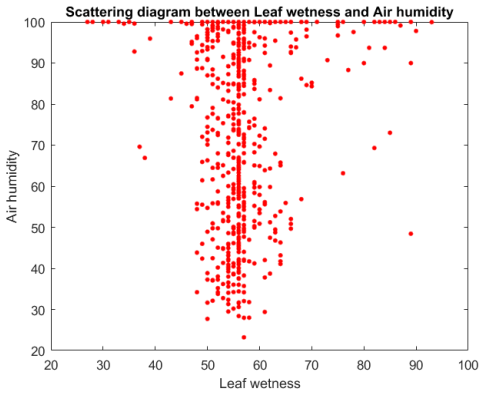


Figure 15: Scattering diagram of leaf wetness and air humidity

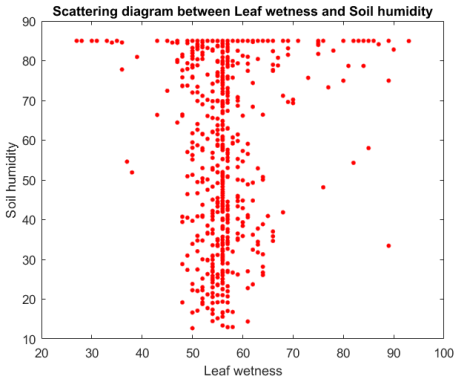


Figure 16: Scattering diagram of leaf wetness and soil humidity

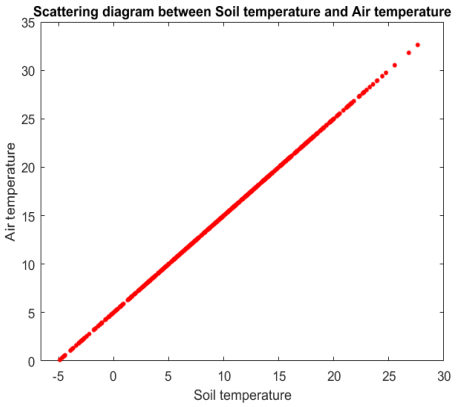


Figure 17: Scattering diagram of soil temperature and air temperature

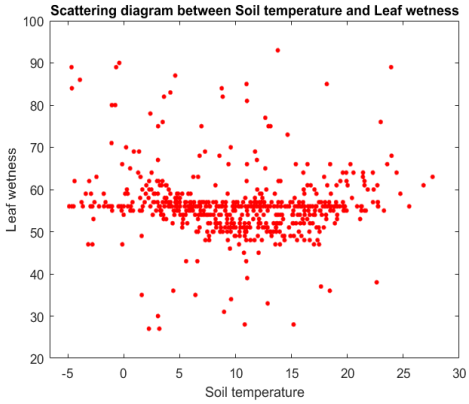


Figure 18: Scattering diagram of soil temperature and leaf wetness

7. Conclusions

Agriculture is an on-going evolving domain since worldwide survival depends in on it a great manner. Telemetry systems developed for crop and field monitoring (such as SmartAgro) play important roles in providing crucial KPIs related to air & soil temperature and/or air & soil humidity, as well as crop diseases detection. The architecture of the implemented SmartAgro system is highlighted by two relevant levels: The Edge level and the Local Storage level. They enable data classification and use parameters improvement in various cases such as alert scenarios and for applications. on-field, off-line.

Being equipped with multiple different, SmartAgro provides massive quantity of data for the monitored parameters. In this paper, the reliability of the recorded data provided by the telemetry system is outlined by performing measurements in 2 consecutive years (2019 and 2020) with a frequency of 6h/ day.

Based on the collected data related to the air temperature in Spring 2019 and Spring 2020, it can be noticed that Spring 2019 was a more favourable season for plant growing since the DIF was lower than 12°C. The variation of soil temperature led to the observation that the patterns are identical with their correspondent air temperature data in both seasons (Spring 2019 and Spring 2020). Further, by recording data on relative soil humidity and leaf wetness there was observed that there is no evidence of correlation between the monitored parameters. Yet, the similarities in variations during Spring 2019 and Spring 2020 demonstrates the reliability of the data recorded using SmartAgro telemetry system.

Since it was noticed that the monitored parameters have not been collected at the same moment in time with regularity, data fusion technique was used to fill the gaps and to provide a global overview on the behaviour of the system. Two KPIs have been considered: air temperature and soil temperature. Using data fusion, it was proven that:

- By finding a method to determine an approximate value for air temperature will also succeed in determining the approximate value for soil temperature, too, and vice versa.
- By finding the approximate value of the air relative humidity, the approximate value of the soil humidity can also be determined.

As conclusion, the goals of data fusion referring to advanced data management and coherence achievement among collected data were achieved within this research.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The work presented in this paper was funded by SmartAgro project subsidiary contract no. 8592/08.05.2018, from the NETIO project ID: P_40_270, MySmis Code: 105976 and DISAVIT project, code EUROSTARS-2019-E!113655-DISAVIT/2020 and Operational Programme Human Capital of the Ministry of European Funds through Financial Agreement 51675/09.07.2019, SMIS Code 125125.

References

[1] I. Marcu, G. Suciu, C. Balaceanu, A.-M. Dragulinescu, M. A. Dobrea, "IoT Solution for Plant Monitoring in Smart Agriculture", 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME), Cluj-Napoca, Romania, Romania, Oct. 2019, DOI: 10.1109/ECAI46879.2019.9041952

[2] European Environment Agency, Climate change adaptation in the agriculture sector in Europe, EEA Report No 04/2019 (ISBN 978-92-9480-072-5)

[3] M. Uzunoglu, M. S. Alam, "Dynamic modeling, design, and simulation of a combined PEM fuel cell and ultracapacitor system for stand-alone residential applications" in IEEE Trans. Ener. Conv., **21**(3), 767–775, 2006. https://doi.org/10.1109/TEC.2006.875468

[4] I. Hadade, T. M. Jones, F. Wang, L. di Mare, "Software Prefetching for Unstructured Mesh Applications" in 2018 IEEE/ACM 8th Workshop on Irregular Applications: Architectures and Algorithms (IA3), Dallas, TX, USA, 2018. https://doi.org/10.1109/IA3.2018.00009

[5] H. Lihua, "Analysis of Fuel Cell Generation System Application," Ph.D Thesis, Chongqing University, 2005.

[6] https://blog.agrivi.com/post/measuring-kpis-for-increasing-productivity

[7] M. H. Nehrir, C. Wang, Modeling and Control of Fuel Cells: Distributed Generation Applications, Wiley-IEEE Press, 2009.

[8] R. Martínez, JA Pastor, B. Álvarez, A. Iborra, "A Testbed to Evaluate the FIWARE-Based IoT Platform in the Domain of Precision Agriculture" Sensors (Basel). 2016; **16**(11):1979, 2016. doi:10.3390/s16111979

[9] F.J. Ferrández-Pastor, J.M. García-Chamizo, M. Nieto-Hidalgo, J. Mora-Martínez, "Precision Agriculture Design Method Using a Distributed Computing Architecture on Internet of Things Context" Sensors (Basel). **18**(6):1731, 2018. doi:10.3390/s18061731

[10] C. Kamienski, J.P. Soininen, M. Taumberger, et al. "Smart Water Management Platform: IoT-Based Precision Irrigation for Agriculture" Sensors (Basel). 2019; **19**(2), 276. Published 2019. doi:10.3390/s19020276

[11] E-agriculture in Action: Big Data for Agriculture. Bangkok, FAO and ITU. 2019. Available online: http://www.fao.org/3/ca5427en/ca5427en.pdf

[12] X. Shi, X. An, Q. Zhao, H. Liu, L. Xia, X. Sun, Y. Guo, "State-of-the-Art Internet of Things in Protected Agriculture", Sensors (Basel). 2019 Apr; **19**(8): 1833, 2019

[13] A. de Juan, R. Tauler, Data Fusion Methodology and Applications, Elsevier, 2019

[14] P. Rashinkar, V. S. Krushnasamy, "An overview of data fusion techniques" in 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2017, DOI: 10.1109/ICIMIA.2017.7975553

[15] L. Comba, A. Biglia, D. Ricauda Aimonino, P. Barge, C. Tortia, P. Gay, "2D and 3D data fusion for crop monitoring in precision agriculture" in 2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Portici, Italy, 2019, DOI: 10.1109/MetroAgriFor.2019.8909219

[16] H. S. Mahmood, W. B. Hoogmoed, E. J. van Henten, "Sensor data fusion to predict multiple soil properties", Precision Agric **13**, 628–645, 2012. DOI 10.1007/s11119-012-9280-7

[17] K. S. Veuma, K. A. Suddutha, R. J. Kremerb, N. R. Kitchena "Sensor data fusion for soil health assessment" Geoderma **305,** 53–61, 2017.

[18] J. Pena, Y. Tan, W. Boonpook, "Semantic segmentation based remote sensing data fusion on crops detection", Journal of Computer and Communications, **7**(7), 53-64., 2019 DOI: 10.4236/jcc.2019.77006

[19] I. Marcu, G. Suciu, C. Bălăceanu, A. Vulpe, and A.-M. Drăgulinescu, "Arrowhead Technology for Digitalization and Automation Solution: Smart Cities and Smart Agriculture" Sensors, **20**(5), 1464, Mar. 2020.

[20] SM Carvalho, E Heuvelink, R Cascais, O. van Kooten. Effect of day and night temperature on internode and stem length in chrysanthemum: is everything explained by DIF?. Ann Bot. **90**(1):111-118, 2002. doi:10.1093/aob/mcf154

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1273-1281 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Laser Deprocessing Technique and its Application to Physical Failure Analysis

Yanlin Pan*, Jia Rui Thong, Pik Kee Tan, Siong Luong Ting, Chang Qing Chen

GLOBALFOUNDRIES Singapore Pte. Ltd., QRA-EFA, Singapore, 738406, Singapore

ARTICLE INFO

Article history:
Received: 26 August, 2020
Accepted: 18 October, 2020
Online: 24 October, 2020

Keywords:
Laser deprocessing
PFA
Bit-counting
TEM

ABSTRACT

This paper is an extension of work originally presented in IPFA 2019. In the original work, a new memory bit-counting method in physical failure analysis (PFA) using laser deprocessing technique (LDT) is introduced. In the present paper, LDT will be further exploited and the methodology applied to PFA will be fully discussed. Compared to the conventional methods that involve high-cost equipment such as focused ion beam (FIB) and reactive ion etcher (RIE), the novel LDT method using a laser system instead lowers the cost by more than 5 times and shortens the failure analysis (FA) cycle time by up to 45%. The new improved methodology can significantly increase PFA throughput especially in semiconductor foundries, and facilitate more applications in other types of FA labs.

1. Introduction

In modern electronic industry, the development of the integrated circuit (IC) design and technology node of semiconductor devices poses increasing challenges to PFA techniques due to the higher density of transistors and metal layers in a chip [1]. As one of the failure analysis (FA) branches for defect identification and yield improvement, PFA acts an important role in analyzing memory failure devices, where bit-counting and marking on the failed bits are necessary for the final physical imaging of the defect in transmission electron microscopy (TEM).



Figure 1: Illustration of typical PFA workflow for memory failure devices.

A typical PFA workflow for memory failure devices is illustrated in Figure 1. The workflow starts from RIE removing the inter-metal dielectric to expose the metal of each layer. Then it is followed by mechanical polishing using polisher with polishing slurry on the rotating cloth platen. During the polishing, optical microscope (OM) and scanning electron microscope (SEM) inspection are engaged to monitor and inspect the sample surface. When the sample reaches certain layers, bit-counting in the memory blocks of the device will be attempted to locate and mark the failed bits for TEM analysis.

In static random access memory (SRAM) devices, bit-counting in the memory blocks can be done as the sample reaches Metal2 (M2) or lower metal layer in which both bit lines (BLs) and word lines (WLs) are identifiable and countable. However, in the cases that request for critical dimension (CD) measurements on Via2 (V2) layer or cases where V2 deformation/void defect is highly suspected, traditional method of mechanically delayering away M3 that will damage part of V2 is no longer suitable. In non-volatile memory (NVM) devices, the memory bit-counting is even more challenging, since the metal layer (here it is M5) blocking the WLs directly lands on the memory cells (Figure 2a, b). There have been numerous FA studies on SRAM/NVM devices, from hard short/open failure [2-4] to subtle defect induced marginal failure [5-7]. After the electrical fault isolation (EFI) or bit-map analysis, memory bit-counting is the key step for TEM analysis or transistor-level probing to locate the defect. Prior to the laser deprocessing technique (LDT) assisted method was developed, there were two conventional methods of bit-counting used in SRAM/NVM device.

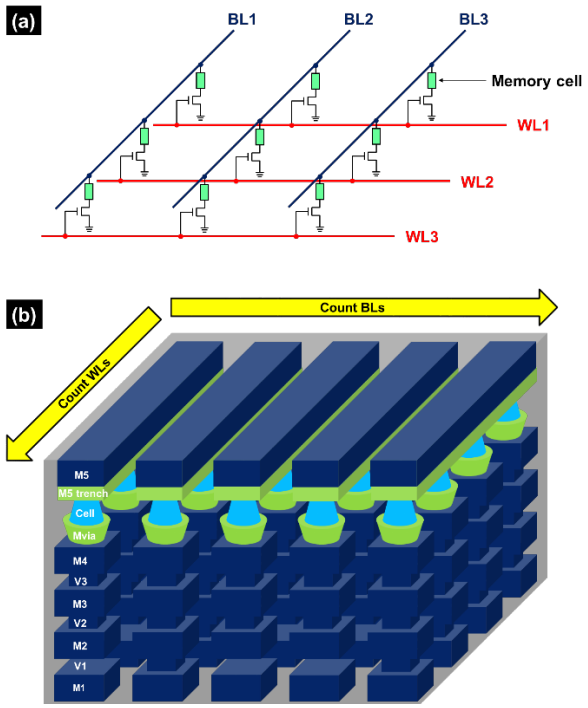


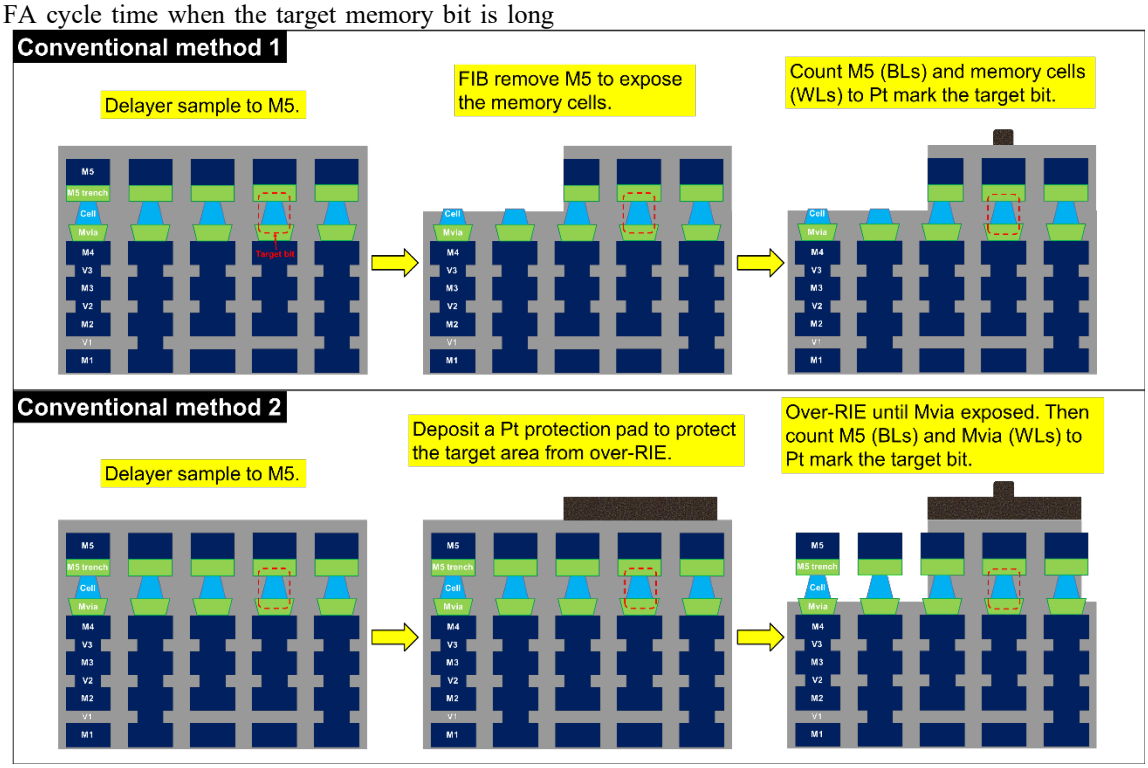
Figure 2: Typical circuit diagram (a) and bit-counting (b) of memory arrays in NVM devices.

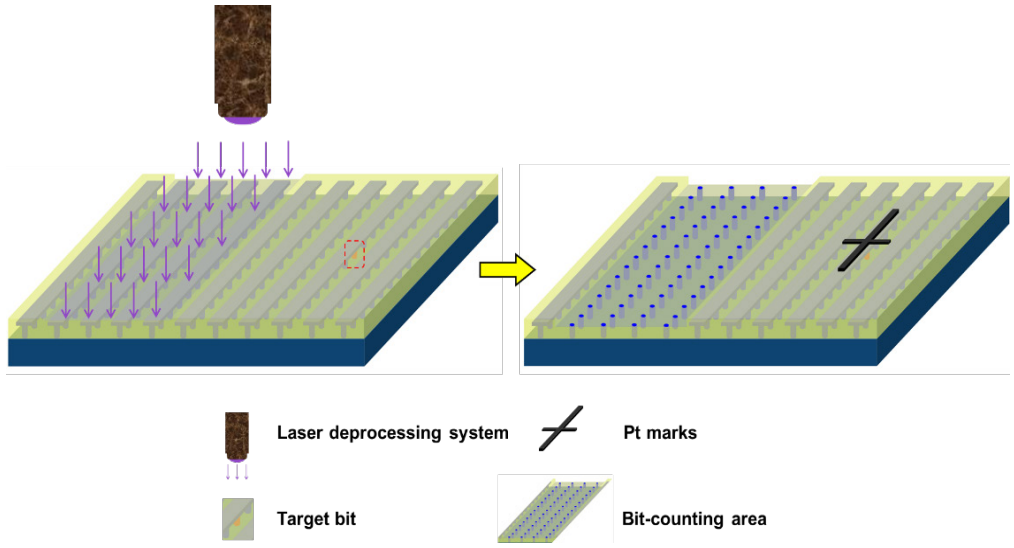
One method is using focused ion beam (FIB) to remove M5 near the target BL to count the exposed cells as counting WLs. The other method is over-RIE the sample surface to expose and count the Memory Via (Mvia) as WLs (Figure 3). The first method has the advantages of real-time monitoring and localized milling, however it requires a FIB-SEM dual-beam system of which the cost is much higher than that of a SEM single-beam system, and much longer FA cycle time when the target memory bit is long distance from the memory block edge. The second method needs only a SEM single-beam system, but it comes with possible sample damage from the Pt protective pad missing the target memory bit. Besides this, for advanced technology node such as 28nm or below, the Mvia size deceases smaller than that of M5, which makes the method by over-RIE not workable. Therefore, new methods with fast cycle time and low cost for the bit-counting in memory devices are highly desired.

In this paper, we will introduce an innovative LDT assisted memory bit-counting method that is applicable to both SRAM and NVM devices. LDT is fast, simple, and has been used to secure an uneven sample surface and remove unwanted layers [1, 8]. LDT employs pulse laser with high peak power to irradiate the sample and physically remove material from the surface, which is economical, efficient, and user-friendly. Moreover, through a systematical parametric study, LDT is able to create uniform deprocessed surface in a large area on various types of materials.

2. Experiments

The experiments were conducted on a 40 nm SRAM device and a 28 nm NVM device. A mechanical polisher (ALLIED HighTech TwinPrep 5) with polishing cloth (Spec-Cloth), and polishing slurry (0.05 µm or 3 µm diamond suspension) was used for mechanical polishing. A SEM single-beam system (FEI Magellan^{TM} 400L) equipped with a gas injection system was used for sample inspection and Pt marking. An optical microscope (ZEISS Axiotron) was used for optical inspection on the sample surface. For LDT, a laser system (New Wave Research Ezlaze3) was used for the sample deprocessing. The laser source emits a pulse laser (wavelength: 532 nm, energy range: 0 to 0.6 mJ, pulse width: ~5 ns) which was integrated into a probe station (Cascade Microtech PM5).





Parameters including laser output power, number of laser shots, and magnification of optical lens were studied on the dummy area of the NVM sample at M5 level. For different devices and materials, the same workflow can be followed to obtain the optimal parametric condition for laser deprocessing in each case. The window size of the laser is adjustable without power shift. Maximum size for 20× lens magnification was 150 µm × 150 µm. TEM sample preparation was performed using a FIB-SEM dual-beam system (FEI Helios NanoLab 450S). TEM analysis and energy dispersive X-ray (EDX) mapping were performed using a 200 kV Field Emission TEM (JEOL JEM-2100F) equipped with a SDD-EDX detector (Oxford X-MaxN 100TLE).

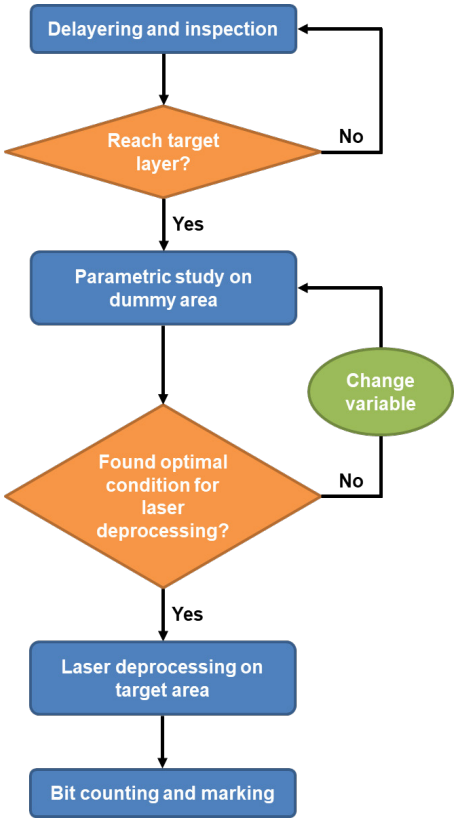


Figure 5: LDT assisted PFA workflow for memory failure devices.

A schematic illustration of LDT in NVM devices is shown in Figure 4. Before the laser deprocessing on the target location, a series of parametric experiments were carried out on the dummy area. Figure 5 shows the LDT assisted PFA workflow for memory failure devices. Here the parametric variables include the work mode of laser energy, attenuation level, lens magnification, and the number of laser shots. For the selection of each parameter, we used parametric arrays to determine the optimal condition in terms of the surface uniformity of the deprocessed sample. A uniform sample surface is critical to the bit-counting accuracy. Any false counting could result in missing the defect.

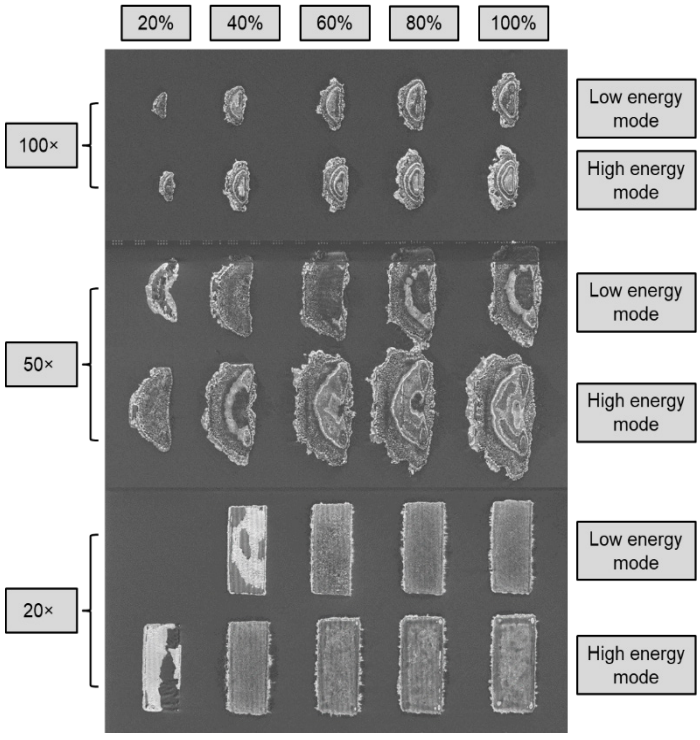


Figure 6: SEM images of the laser deprocessed surfaces of the NVM devices under different energy mode, attenuation level (%), and lens magnification.

2.2. Selection of laser power and lens magnification

Figure 6 shows the SEM images of the laser deprocessed surfaces of the NVM devices under different laser power and lens magnification. From the results, we can see that 100× and 50× lens created a very uneven and deformed surface with multiple layers exposed, regardless of energy mode and attenuation level. So we narrowed down the selection to 20× lens arrays. In low energy mode, 80% attenuation level delivered the most uniform deprocessed surface at the NVM cell level. In high energy mode, the best results were delivered by 40% attenuation level.

The surface uniformity of low energy mode at 80% attenuation level and high energy mode at 40% attenuation mode are comparable for the bit-counting, as shown in Figure 7. However, in low energy mode, the 100% attenuation level makes use of 0.2 mJ of energy while in high energy mode, at 100% attenuation level, the machine makes use of 0.6 mJ of energy. Therefore, output energy of 80% attenuation level in low energy mode is 0.8 × 0.2 mJ = 0.16 mJ, which is lower than 0.4 × 0.6 mJ = 0.24 mJ for 40% attenuation level in high energy mode. Considering the fact that lower laser energy is safer for the sample, low energy mode and 80% attenuation level were selected for the laser deprocessing for memory bit-counting.

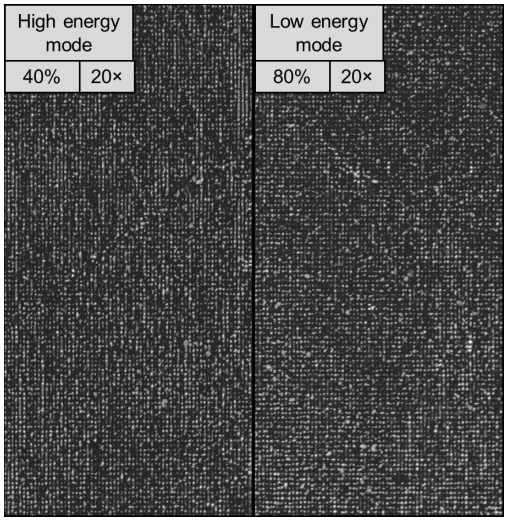


Figure 7: High magnification SEM images of the laser deprocessed surfaces of the NVM devices under 40% attenuation level in high energy mode and 80% attenuation level in low energy mode.

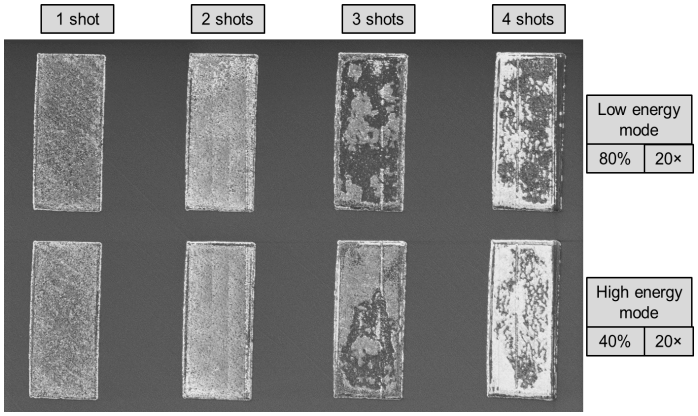


Figure 8: SEM images of the laser deprocessed surfaces of the NVM devices under different number of laser shots.

The experiments above were all from a single laser shot. In this section, using the lens and the attenuation level selected for each energy mode, a boundary test was performed on the maximum number of laser shots on the same surface before the surface becomes uneven again. In some cases, the layer on top of the bit-counting layer might be too thick and need more than one shot of laser to remove. Figure 8 shows the SEM images of the laser deprocessed surfaces of the NVM sample under different numbers of shots. It can be seen that the surfaces for each shot started to become uneven after the second shot was made on the same surface. This could be due to that the third shot was shone on the critical interface between the non low-k dielectric layer (memory cells and Mvia) and the ultra low-k dielectric layer (M4 and below). The unevenness was greatly increased during the third shot because of the big difference in the mechanical removal rate (up to 10 times) between the non low-k and the ultra low-k materials [9]. The experimental results would differ from case to case. For this NVM device, one shot of laser which created uniform bit-counting area (Figure 7) was selected.

2.4. Boundary limitations of laser shots

To understand the damage induced by laser shots to the memory bits, TEM analysis was done on the boundary of the laser deprocessed area. The results (Figure 9) show that four bits (Bit 1, 2, 3, 4) adjacent to the deprocessing area were damaged by the laser shot due to the thermal effect. Similar to other parameters, boundary limitations are dependent on the sample materials and surface conditions, and should be studied on the dummy area first to avoid damaging to the target bits.

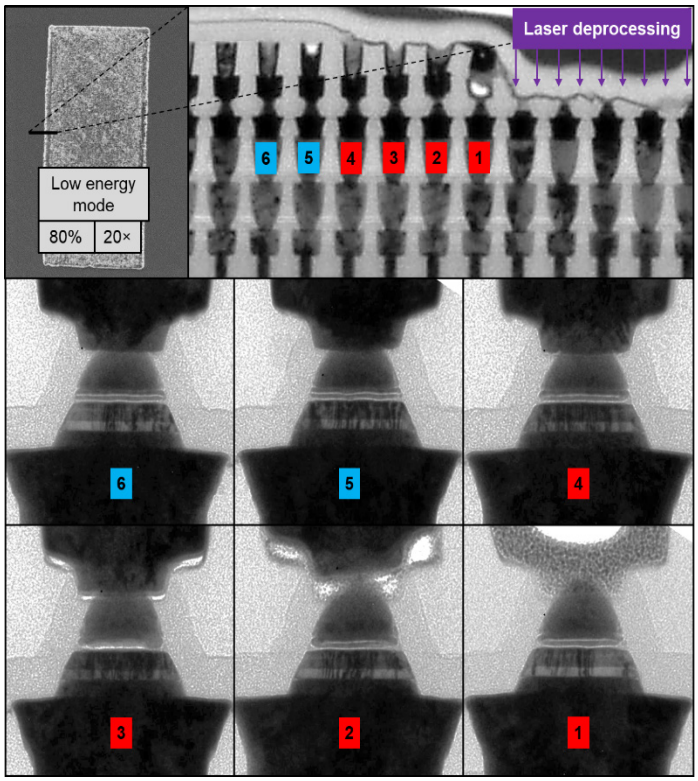


Figure 9: SEM images of the laser deprocessed surfaces of the NVM devices under different energy mode, attenuation level (%), and lens magnification.

2.5. Operation procedure of LDT assisted memory bit-counting

The procedure steps below describe how to utilize LDT to perform the memory bit-counting in SRAM/NVM devices.

1) Delayer the sample and stop at one layer before the bit-counting layer (M5 for NVM, M3 for SRAM) by mechanical polishing and inspection.
2) Make a SEM Pt mark at the target BL as a "sight" for the laser deprocessing process in OM (this step can be skipped for SRAM devices with small memory blocks).
3) Remove M5 (NVM) or M3 (SRAM) layer at certain distance (referring to the boundary limitations of laser shots) from the target BL to expose the memory cells by laser deprocessing (Figure 4). For SRAM devices, an extra deprocessed area for counting WLs is needed.
4) Water-polish the sample to clean the laser burst debris off the deprocessed area.
5) In the deprocessed area, count the exposed memory cells as WLs and M5 lines as BLs to locate the target bit from the NVM memory block (Figure 4). For SRAM devices, both WLs and BLs will also be identifiable and countable from M2 or below.
6) Make SEM Pt marks on the located target bit accordingly for TEM analysis (Figure 4).

Table 1: Operation time usage (in hours) of the conventional methods and the new LDT method for NVM devices.

	Pt mark target BL	Deposit Pt protection layer	Over RIE	FIB milling	Laser deprocessing	Counts WLs	Total
Conventional method 1	N/A	~0.6	~0.2	N/A	N/A	~0.7	~1.5
Conventional method 2	~0.3	N/A	N/A	~0.8	N/A	~0.5	~1.5
New method	~0.3	N/A	N/A	N/A	~0.1	~0.5	~0.9

In Table 1, the operation time usage of the conventional methods and the new LDT method for NVM devices are compared, which is evaluated under the same experimental conditions. The time usage of each operation step may vary with different devices, requirements of the cases and experiences of the operators. It can been seen that the LDT method can shorten the time usage by up to 45%.

Table 2: Summary of the advantages and disadvantages of the conventional methods and the new LDT method.

	Advantages	Disadvantages
Conventional method 1	Localized milling, real-time monitoring	High cost, Long FA cycle time
Conventional method 2	Low cost	Long FA cycle time, High risk of sample damaging, Not suitable for advanced technology node
New method	Low cost, Short FA cycle time	Sample dependent setting of laser parameters

Table 2 shows the advantages and disadvantages of the conventional methods and the new LDT method. Although the new method needs parametric study on the sample, the one-time effort eventually significantly saves FA time and FA cost.

3. Results and Discussion

3.1. Case study on NVM device

In this case, a 28 nm NVM device was performed with PFA using LDT. Figure 10 is the M5 graphic data system (GDS) layout of the device. After wafer sort, the bitmap information was provided in the data log. According to the e-test results, the NVM chip suffered from single-bit short failure (Figure 10).

The sample was polished to M5 layer, and then continued with LDT for bit-counting (Figure 11). First, a 45° angled e-beam Pt line mark was deposited in SEM as the reference for the laser deprocessing below the target BL (located by M5 GDS layout), and was deposited long and thick (SEM: 10 kV at 0.8 nA; Rate of deposition: 10 μm [length] × 2 μm [thickness] in 5 mins) to be clearly visible in OM. Second, at around 20 BLs to the right of the target bit, LDT was used to create a laser exposed area at the level of memory cells, in which the WLs were counted and marked according to the WL address. Finally, Pt marks were deposited on the located target bit for the following TEM analysis.

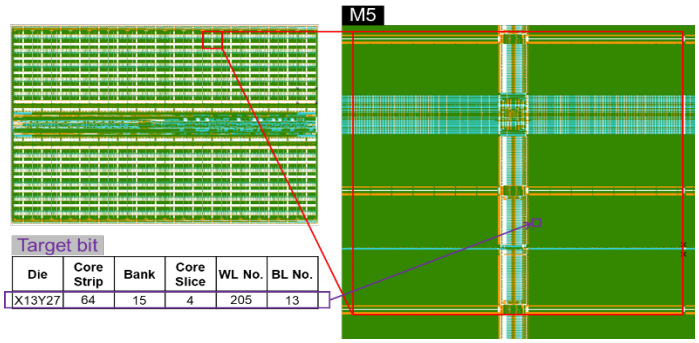


Figure 10: M5 GDS layout of 28 nm NVM device.

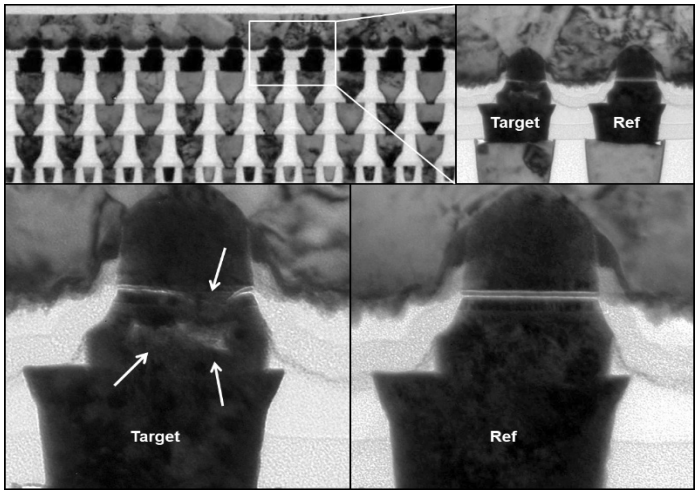


Figure 12: TEM images (Y direction along M5 as shown in Figure 11) of the failed bit and the reference bit.

TEM results of the sample are shown in Figure 12. From the TEM images, metal void and collapsed insulator layer were found at the failed bit. The metal void in the lower cell electrode should form in the first place, and then induce the uneven deposition of the insulation layer and the breakdown of the memory cell.

Figure 13 shows the scanning transmission electron microscope (STEM) images of another case where the TEM sample was cut across M5 to compare between different BLs from

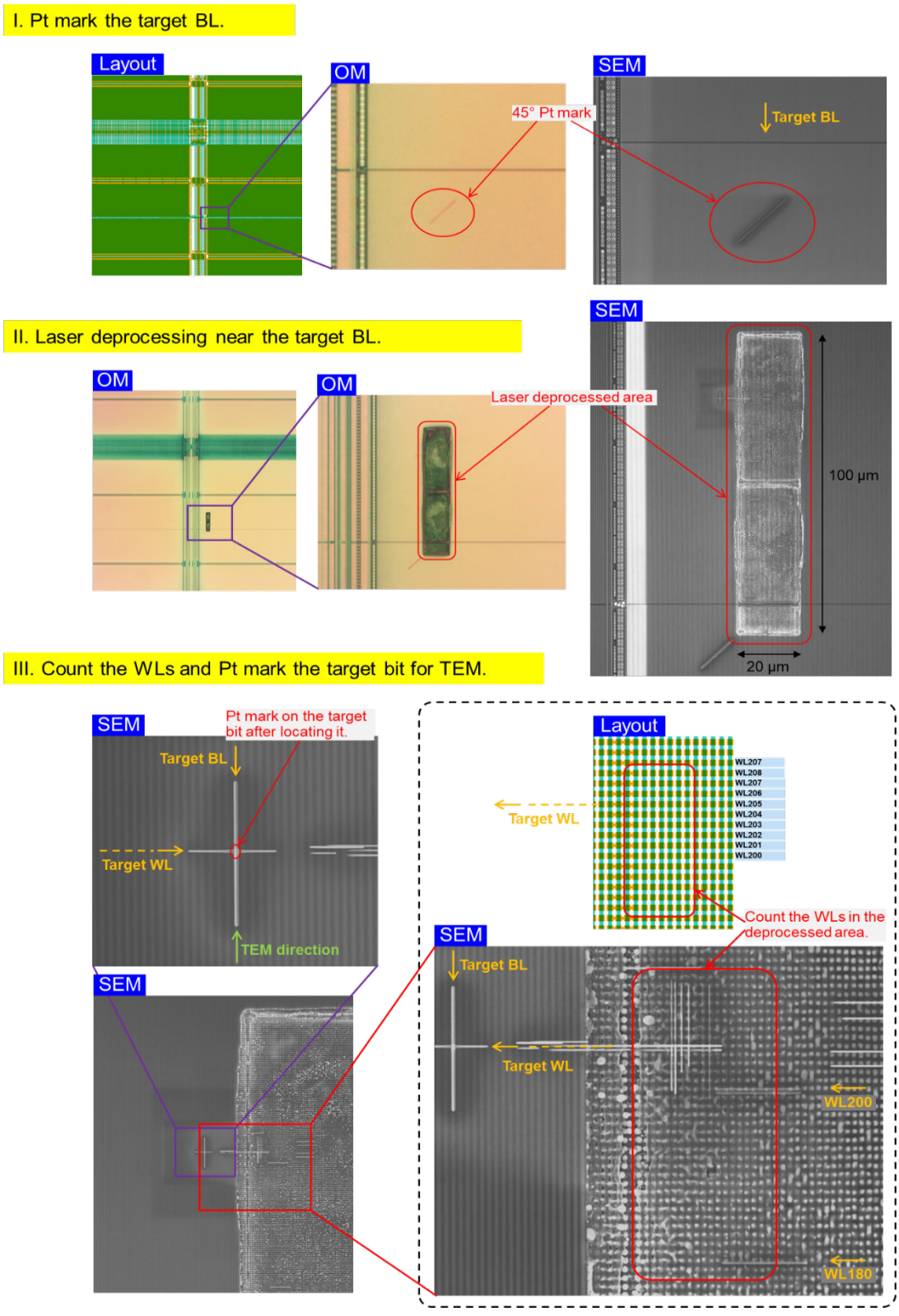


Figure 11: Schematic illustration of LDT assisted bit-counting at M5 in NVM devices.

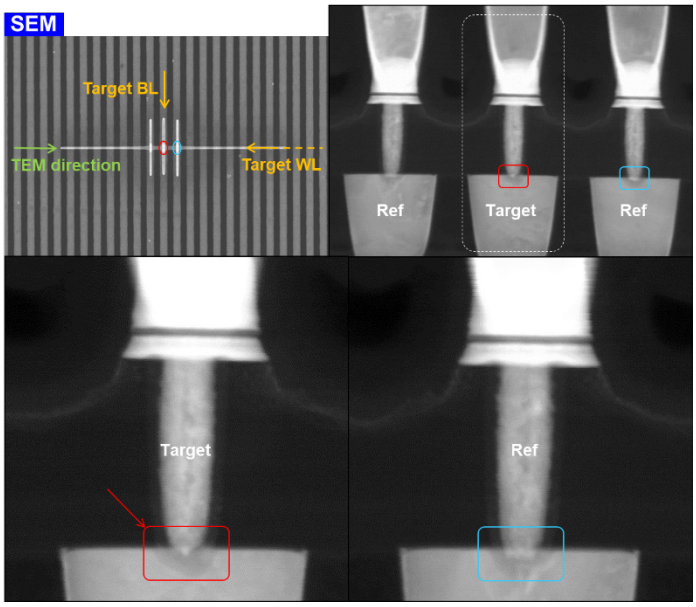


Figure 13: STEM images (X direction across M5) of the failed bit and the reference bit.

the same WL. From the results, abnormal profile at the bottom of the Mvia was found at the failed bit, which caused high landing resistance of the bit. It is worth to mention that, cutting along Y direction, even when the conventional methods aim at wrong WL by a few bits, the real target bit is usually still covered in the TEM sample so it can be corrected in the TEM analysis. However, in the X direction cut, wrong counting due to the insufficient exposing of the memory bits using conventional methods will result in missing the defect. Therefore, the LDT method offers the advantage of efficiently creating a bit-counting area and locating the target WLs in a high success rate.

3.2. Case study on SRAM device

In the SRAM case, a 40nm device with SRAM blocks was found to suffer from a 3-bits cluster failure. Through a series of analysis, Cu void defect in V1/V2 is suspected to result in the random wafer pattern, so the delayering must stop at M3 level to measure the V2 CD. However, without a reference, bit-counting at M3 is not practical, as seen from the GDS layout of SRAM devices (Figure 14). At M3, BLs and WLs are all not able to be properly identified. If we all the way polish to V2, the top portion of V2 will be damaged and the CD measurements will become impossible.

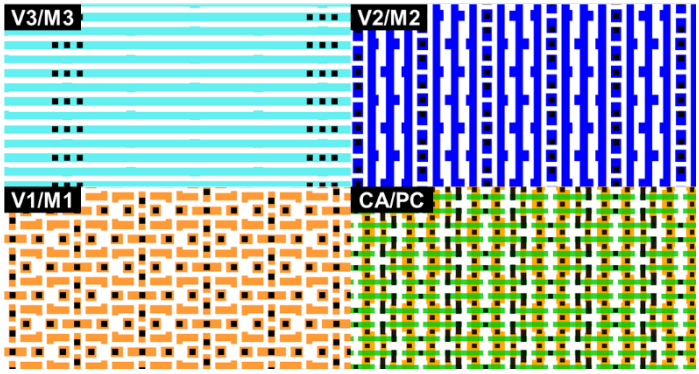


Figure 14: Typical GDS layout of SRAM devices.

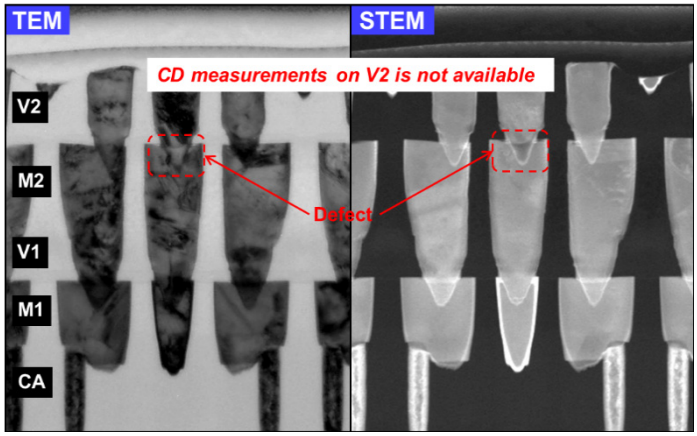


Figure 15: TEM/STEM images of a failed SRAM sample preparation. Top portion of V2 has been damaged.

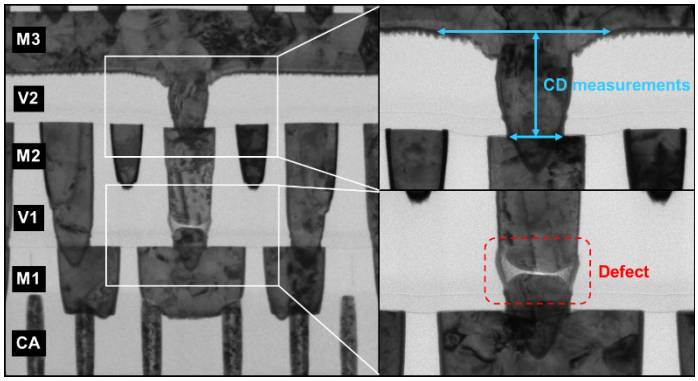


Figure 16: High/low magnification TEM imaging on the failed bits.

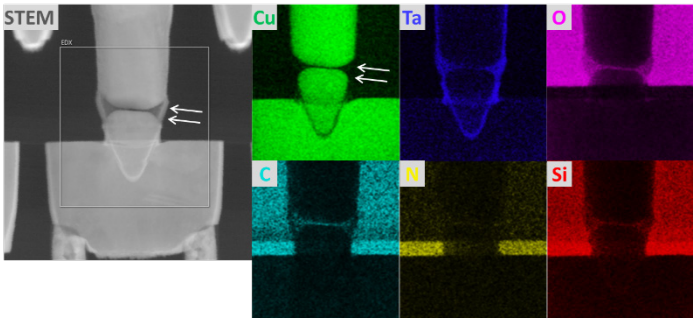


Figure 17: EDX mapping analysis on the failed bit. Defect of Cu void is observed, as arrowed in the Cu quant map.

Figure 15 shows a failed SRAM sample preparation. To solve the problem, we used LDT to create two reference areas (one for counting BLs and the other one for counting WLs) near the failed bit (Figure 18). In the reference areas, bits at V1/M1 level are clearly visible for the counting of WLs and BLs, while the target area is still at M3.

By counting and marking the WLs/BLs in the reference area, the failed bit was located and marked. TEM sample preparation was then performed on the marked bit. The TEM results are shown in Figure 16. From the high/low magnification TEM images, metal void was found at the bottom of V1. CD measurement on V2 was also taken. EDX analysis was then performed on the void area (Figure 17), in order to study the mechanism of the void formation. The EDX quant maps showed that Cu diffusion/migration occurred at the arrowed void area. Therefore, oxidation/corrosion

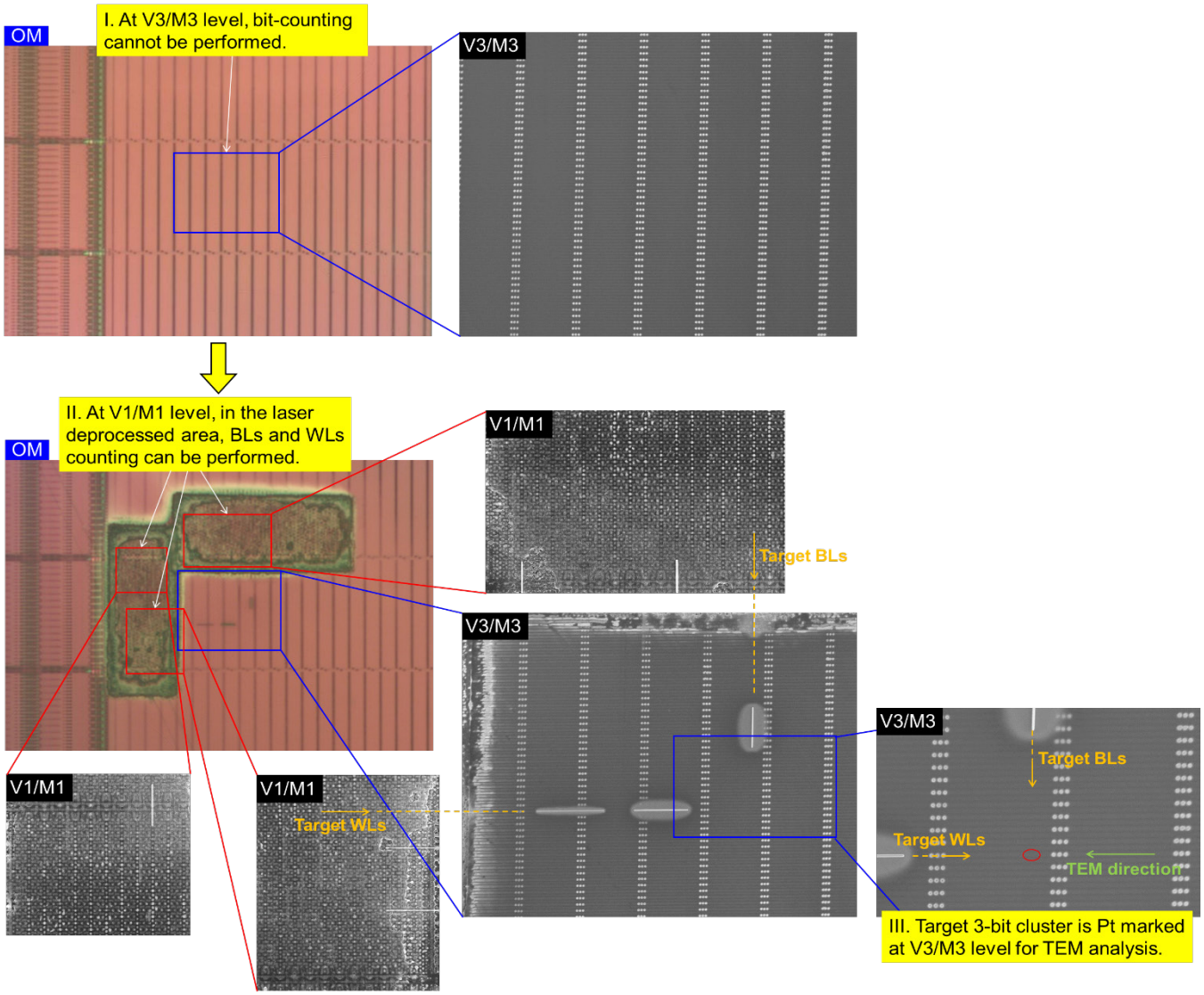


Figure 18: Schematic illustration of LDT assisted bit-counting at M3 in SRAM devices.

of the seed Cu layers could be the root cause of the failure. The high aspect ratio of the metal trench in the dual damascene process could introduce thinner seed Cu layer at the trench bottom. In the subsequent bulk Cu plating, the H+ ions in the electrolyte may attack the Cu seed layer as an acid etchant, since the electrons in the current flowing through the thinning portion were not enough to neutralize the excess H+ ions. This would finally induce the formation of the void at the bottom of either V1 or V2.

4. Conclusion

In this paper, an innovative LDT assisted PFA technique is fully exploited and discussed. Compared to the conventional methods that involve high-cost equipment, this new method employing a pulse laser system instead significantly lowers the FA cost (by more than 5 times) and shortens the FA cycle time (by up to 45%), and facilitates potential applications in other FA areas such as pre-EFI sample preparation, large size cross-sectional analysis, etc. Moreover, only basic FA skills such as using OM and SEM are required to manipulate this new method which is suitable for FA engineers from all different levels.

Conflict of Interest

The authors declare no conflict of interest.

Abbreviations

BL	Bit-Line
CD	Critical Dimension
EDX	Energy Dispersive X-ray
EFI	Electrical Fault Isolation
FA	Failure Analysis
FIB	Focused Ion Beam
GDS	Graphic Data System
IC	Integrated Circuit
LDT	Laser Deprocessing Technique
Mvia	Memory Via
NVM	Non-Volatile Memory
OM	Optical Microscope
PFA	Physical Failure Analysis

RIE Reactive Ion Etching
SEM Scanning Electron Microscope
SRAM Static Random Access Memory
STEM Scanning Transmission Electron Microscope
TEM Transmission Electron Microscope
V*x* Via*x*, *x* is layer number

References

[1] H. H. Yap, P. K. Tan, L. Zhu, H. Feng, Y. Z. Zhao, R. He, H. Tan, B. H. Liu, Y. M. Huang, D. D. Wang, J. Lam, Z. H. Mai, "Application of Laser Deprocessing Technique in PFA on Chemical Over-etched on Bond-pad Issue," Microelectronics Reliability, **64**, 357-361, 2016, doi:10.1016/j.microrel.2016.07.057.

[2] S. Ikeda, Y. Yoshida, K. Ishibashi, Y. mitsui, "Failure Analysis of 6T SRAM on Low-voltage and High-frequency Operation," in 2003 IEEE Transactions on Electron Devices, 1270-1276, 2019, doi:10.1109/TED.2003.813474.

[3] Z. G. Song, "SRAM Failure Analysis Evolution Driven by Technology Scaling," in 2014 IEEE International Symposium on Physical and Failure Analysis of Integrated Circuits (IPFA), 2014, doi:10.1109/IPFA.2014.6898207.

[4] P. Egger, C. Burmer, "SRAM Failure Analysis Strategy," in 2003 International Symposium for Testing and Failure Analysis (ISTFA), 177-183, 2003.

[5] C. Q. Chen, P. T. Ng, S. P. Neo, P. K. Tan, A. C. T. Quah, J. Zhu, "Application of Nanoprobing on Subtle Defects in the Embedded Non-volatile Memory Device," in 2019 International Symposium for Testing and Failure Analysis (ISTFA), 2019.

[6] W. T. Chang, T. E. Hsieh, "Advance Static Random Access Memory Soft Fail Analysis Using Nanoprobing and Junction Delineation Transmission Electron Microscopy," Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena, **25**, 202, 2007.

[7] T. Mizutani, T. Saraya, K. Takeuchi, M. Kobayashi, T. Hiramoto, "Transistor-level Characterization of Static Random Access Memory Bit Failures Induced by Random Telegraph Noise," Japanese Journal of Applied Physics, **55**, 04ED05, 2016, doi:10.7567/JJAP.55.04ED05.

[8] Y. Z. Zhao, Q. J. Wang, P. K. Tan, H. H. Yap, B. H. Liu, H. Feng, H. Tan, R. He, Y. M. Huang, D. D. Wang, L. Zhu, C. Q. Chen, F. Rivai, J. Lam, Z. H. Mai, "Application of Fast Laser Deprocessing Techniques on Large Cross-section View Area Sample with FIB-SEM dual beam system," Microelectronics Reliability, **64**, 362-366, 2016.

[9] Y. L. Pan, P. K. Tan, S. L. Ting, C. Q. Chen, "Advanced Sample Preparation Techniques for Rescuing Damaged Samples with Cracks, Scratches, or Unevenness in Delayering," in 2020 IEEE International Symposium on Physical and Failure Analysis of Integrated Circuits (IPFA), 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1282-1287 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

An Overview on CryptDb and Word2vec Approaches

Hana Yousuf[1], Asma Qassem Al-Hamad[2], Said Salloum[1,3,*]

[1]*Faculty of Engineering & IT, The British University in Dubai, 345015, UAE*

[2]*Librarian and information department, Imam Abdulrahman Bin Faisal University, 1982, KSA*

[3]*Research Institute of Sciences & Engineering, University of Sharjah, 27272, UAE*

ARTICLE INFO

Article history:
Received: 24 June, 2020
Accepted: 07 October, 2020
Online: 24 October, 2020

Keywords:
Big data
CryptDB
SQL databases
Word2Vec
Secure vectorization

ABSTRACT

Big data is a vast data set that was used in many areas. Online applications are subject to theft of confidential information because opponents can exploit software errors to access private data, and because curious or malicious officials can capture and lose data. CryptDB is a functional system that provides security and confidentiality through a set of operations. The obvious confidentiality of these attacks is for applications supported by SQL databases. It works by executing SQL queries on encrypted data using robust coding systems that support SQL. Word2Vec outputs word vectors that can be displayed as a large piece of text, or even we first train data. Word2Vec forms and word similarity assessment. Without a doubt, this article calls for proper research that sheds light on the security features using CryptDB to prevent data theft and privacy breaches in the server. The motivation of this research is to have an overview of CryptDB and Word2Vec implementation on the existing research approaches.

1. Introduction

The theft of private data is a big issue [1–3], especially for online applications [4–11]. The SQL databases may be attacked and vulnerable to sensitive information or theft since they can exploit the bugs to gain access to private information. Also, the attackers can capture and leak the data to those who require them. Hence, the dataset must be safeguarded. CryptDB provides confidentiality against the attackers through SQL databases [10]. The processes are encrypted in SQL by collecting effective SQL aware encryption schemes. The process can also be used to encrypt the credential keys so that an item can be decrypted with simply a password. Since the processes and data are encrypted, the database administrator will not view the data. Even if there is an attack on the server, the attacker will not access the decrypted data. CryptDB can handle multiple queries simultaneously and has less overhead.

Word2vec is one type of word embedding technique used to represents a string in a set of real numbers. This is a technique used in Natural Language Processing (NLP) through deep learning to extract data from the requested document. Word2Vec represents the input text as a statistical or vector form using a two-layer neural net that process text using. Word2vec's applications extend on the far side than analyzing sentences; it can be used in an application that has a well-defined pattern

The motivation to carry out this research is to have an overview of Word2Vec and CryptDB approaches and identify the challenges in the current process in both techniques.

1.1. CryptDb

One of the most widely known machines is employed in various shifts in classification tasks. Two threats are tackled by CryptDB [12]. The first of these is a curious database administrator (DBA) that attempts to learn private data (such as financial statements, health records, personal information, etc.) by sneaking into the DBMS server; however, CryptDB does not allow the DBA to access the private data. The other threat is an adversary that acquires full control of the application and DBMS services. Here, no guarantees can be offered by CryptDB for users logged in the application when an attack occurs; however, it can make sure that the data of the users who are not logged in is protected from threats. A simple solution is to create a different database encryption key

*Corresponding Author: Said Salloum, University of Sharjah, UAE. Tel: +971507679647 Email: ssalloum@sharjah.ac.ae

for every user not applicable for their data in applications, including shared data, such as conference review sites and bulletin boards [13].

The CryptDB's function is to perform queries on encrypted information. The main reason for this practice is that SQL uses a specific set of factors; each can effectively support encrypted data [13]. CryptDB combats with these challenges using three core ideas:

- The first one is to run SQL questions on encrypted information. CryptDB applies this concept by employing an encryption methodology that supports SQL and takes advantage of the truth that all SQL inquiries comprise a well-defined set of primitive factors, like Equality Confirmation, framework comparisons, total (amounts), and links. By adjusting known encryption systems (for correspondence, increments, and work confirmation) and employing a new encryption strategy to ensure the protection of joins, CryptDB encrypts each data component so that DBMS can execute the converted information [14]. CryptDB is practical security since it uses essentially symmetric-key encryption in order to prevent completely identical encryption, and runs an unmodified DBMS program (utilizing custom capacities).
- The second strategy is flexible encryption based on the query. A few encryption systems pass more data to the DBMS server than others but should undergo some queries. To avoid leaking all information encryption to the database management system, CryptDB carefully alters the SQL encryption framework for any specific data element, based on queries. To execute these modifications effectively, CryptDB utilizes 'onions of encryption'. Onions are a new approach to store numerous encryption texts within the database and maintain a distance from exorbitant re-encryption [13].
- The third concept is to link encryption keys to users' passwords so that every information within the database can be decrypted by employing a keychain established within the client's password who can approach that data. If the client does not log on to the app and the rival does not know the password, the adversary will not decode the user's information even if the DBMS and the application server are completely under threat [13]. To form a critical chain that captures the app's protection and sharing arrangement, CryptDB permits the developers to give policy comments through the SQL app chart and identifying clients who have access to each part of the information.

1.2. Word2Vec

Google developed Word2vec in 2013 [12], which is a neural network through which text data is processed. Word2vec is not a single algorithm, it is made up of two learning models, which are the Common Bag of Word bag (CBOW) and Skip grams. The word is predicted by CBOW based on its context, while in Skip-Gramm, context determines a word. Word2Vec ultimately developed in a model word vector by feeding the text. Word2Vec first generates a vocabulary from group training text and learning vector representations for every word [12]. The architecture of the word2vec algorithm is given in Figure 1.

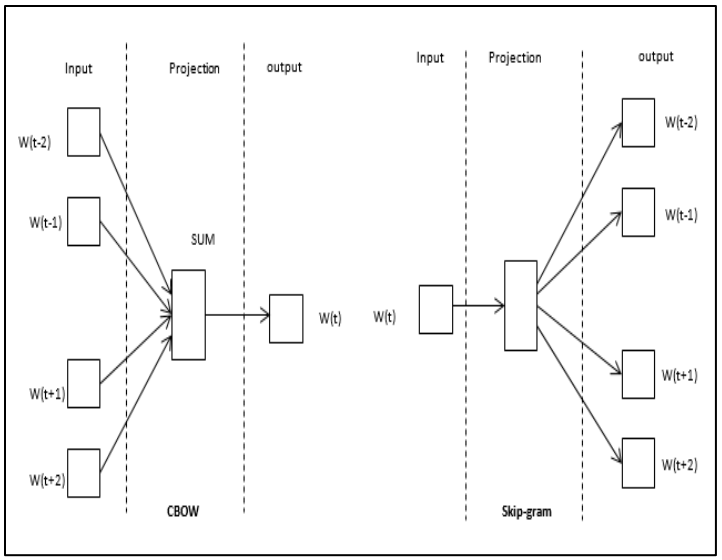


Figure 1: Word2Vec Architecture

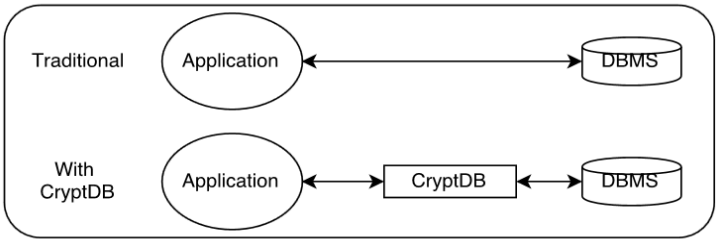


Figure 2: CryptDB along with traditional DB [12].

The purpose of using Word2vec is to cluster the vectors of similar words along in vector-space, then detect the similarity by measuring cosine similarity; it creates vectors that are distributed numerical representations of the word. Word2vec will build extremely accurate guesses of a word's definition based on previous presences. The guesses are used to construct a word that associate with the other word in the sentence.

2. Literature Review

A study by [15] has used effective Galois Field mathematical environments for encrypting algorithms effectively. The increase in network sensor systems and network databases has led to more interest in using cryptology in sensors algorithms and databases. Vectorized Advanced Encrypted Standard (AES) has been implemented for the database systems. Due to vectorization, the implementation is very small and requires 100 times smaller code than ordinary AES implementation. It is also fast and has an effective design. The implementation achieves higher speed which is comparable to real-time prototyping with OpenSSL with good database analytics and processing.

Researchers of [16] have introduced a ciphertext policy with constant sizes known as Cipher-text Policy Hidden Vector Encryption (CP-HVE). An HVE is a unique type of anonymous IBE, which uses identity as the main parameters where the attributes linked with the ciphertext or secret keys contain wild-cards. Two different schemes have been used, one with a composite ordered bi-linear group and a prime ordered bilinear group. Both of these schemes have high security by differentiating

the ciphertext from plaintext. This vectorized encryption has the ability to have a regular size than the other HVE methods.

HVE was also implemented by [17], a type of predictive encryption technique that supports conjunctive equality and ranges on the encrypted data. A novel HVE scheme has been built that is fully secure under standard assumptions. The proposed scheme is based on bilinear maps that are more advantageous for the private keys and pairing computations for decryptions. Tag-based dual system encryption has been developed to hide the vector components and compress the tag's values into one. It is challenging to extend the HVE along with IBE for supporting a higher hierarchical mechanism. Hence, it is still challenging to construct HVE for providing encryption-based security. The number of pairings in the computations is high, which must be reduced in the future to increase the speed of the processes.

In [18], author has described the types of CryptDB. Random (RND) is a type of CryptDB that relatively provides the most security that is selected under a probabilistic adaptive plain text attack. Here, two equal halves of data are mapped into different ciphertexts. RND ensures that computations are performed on ciphertexts directly. RND uses encryptions like AES and Blowfish simultaneously with a random initialization vector. AES is used more than Blowfish for 64-bit block sizes since AES has 128-bit blocks that cause the ciphertexts to be very long. For high security, it is assumed that the server does not change results.

3. CryptDB is a solution for facing the threats

CryptDB works by intercepting all SQL queries in that database proxy and rewrites queries to perform for encrypted data. All inquiries pass through the agent. The agent encrypts and decrypts all data, and some parameters change while maintaining the query's semantics. The DBMS server never receives decryption. A clear text key so that confidential data is not displayed to ensure that strange DBA cannot access private information (Threat 1).

To protect themselves from compromises in applications, proxy servers, and DBMS servers (Threat 2), developers comment on the specific SQL schema different principles; its keys to decode different parts database. They are also making a slight change to their apps, providing encryption keys to the proxy. The agent specifies which parts of the database are to encrypt under any key. The result is that CryptDB guarantees the confidentiality of data owned by users who are not logged in within one. The compromise that just logs in the admin recognizes and corrects the settlement. CryptDB protects data confidentiality, but it does not guarantee its safety, freshness, or completeness of the results returned to implementation. The opponent who threatens the application, the agent, or a malicious DBMS or DBA server that can delete one or all data stored in the database. Attacks on users' computers, Cross-site scripting outside of CryptDB. The security guarantees are provided for in these threat models [19].

3.1. Threat 1: DBMS Server compromise

CryptDB provides security against a curious DBA or another external attacker in this threat while offering complete access to the DBMS server's data. It focuses on confidentiality (data privacy) rather than availability or integrity. It is presumed that the attacker is passive, wanting to get access to confidential data without modifying queries presented by the application, query results, or the data within DBMS. This threat comprises DBMS software being compromised, gaining root access to DBMS, and access to the RAM of physical machines. There is now a greater degree of database consolidation in enterprise data centers, databases are being outsourced to public cloud computing infrastructures, and third party-DBAs are being used, which is why it is becoming very important to deal with this threat [19–21].

The goal of CryptDB is to ensure that the data remains secure from this threat by carrying out SQL queries over the encrypted data DBMS server. The agent uses secret keys for encrypting the entire data; it includes or incorporates them in the outgoing DBMS queries [22]. This tries to make it possible to use the DBMS server to process queries for encrypted data, similar to an unencrypted database. This is done by activating the particular functions for data elements required for encrypted data. The DBMS should have the ability to identify the factors it consists of. There is the same column; however, the actual content items are different. Hence, the DBMS server should be activated by the proxy to identify the relationships among the data required for query processing. SQL-enabled encryption should be used as it is capable of adapting vigorously. CryptDB, when issuing requests, manages the relationships it reveals between lines to the server. The order of the items in the column is not known, which is not even required to learn more about the rest of the columns. When DBMS is needed, CryptDB will identify it by carrying out an ORDER command or determine MAX or MIN items within this column, and not by any other method [22].

3.2. Threat 2: Arbitrary threats

It is possible that the proxy and DBMS server infrastructure experience arbitrary security breaches. Due to the opponent, the method used in Threat 1 is not adequate. It is now possible to access the keys used to encrypt the complete database. To deal with this, different data elements (for example, data for various users) should be encoded using different keys. The application database schema should be suspended to present further privacy guidelines to choose the key developers who would be using each data item. It is still impossible for strange DBA to obtain private data on a DBMS server (Threat 1). The application server or proxy can become decrypted—data from only the presently registered (stored in the proxy). Data from the users who are not active will not be encrypted with the keys available to the attacker and will stay confidential. CryptDB provides substantial guarantees in this configuration in dealing with the arbitrary server-side breaches, including the ones that obtain root access to the proxy or application. CryptDB leaks the majority of the data users that have been active for some time. Concerning an SQL attack, the compromise duration comprises of the SQL queries of the attacker. The system is considered vulnerable until the attacker's email address stays within the database [22], [23].

4. CryptDB Implementation

SQL queries are carried out by CryptDB using encrypted data. There is a lack of trust in the DBMS machines and administrators; however, the application and the proxy can be trusted. CryptDB allows the DBMS server to perform SQL queries on encrypted data, just like how it would be carrying out the same queries on

plaintext data. No modifications in the existing applications are required. Usually, the DBMS query plan for an encrypted query is similar to the actual query, except that the operators forming the query, like projections, selections, aggregates, joins, and orderings, are done on cipher tests. In a few scenarios, modified operators are used. A secret master key MK is stored by CryptDB's proxy, in addition to the database scheme and the existing encryption layers of each column [24]. An anonymized schema is observed by the DBMS server (where the names of tables and columns are substituted with opaque identifiers), encrypted user data, and certain auxiliary tables utilized CryptDB. CryptDB-specific user-defined functions (UDFs) are also provided by CryptDB to the server, using which the server can use ciphertexts for specific functions [24]. Four steps are followed when processing a query in CryptDB.

- A query is presented by an application, which is intercepted and rewritten by the proxy: each table and column name is anonymized. With the master key MK's help, each constant is encrypted in the query using an encryption scheme that is most appropriate for the required operation.
- It is determined by the proxy whether the DBMS server should be provided keys to modify the encryption layers before issuing the query. If this is the case, then an UPDATE query is issued at the DBMS server that brings about a UDF to modify the relevant columns' encryption layer.
- The encrypted query is forwarded by the proxy to the DBMS server, which performs it using standard SQL (often bringing about UDFs for performing aggregation or keyword search).
- The encrypted query result s returned by the DBMS server that is decrypted by the proxy and sent back to the application.

5. Word2Vec Implementation

We prepare our data to be trained using the Word2Vec model. We are taking the below sentence as an example:

We must love life

Step-1: Consider the word 'we' as input and 'must listen' is the output word as follows:

Input	Output
We	must
We	love

Step-2: Now consider 'must' as input word; the output will be the close words as follows

Input	Output
We	must
We	love
Must	we
Must	love
Must	life

Step-3: The same will be done for the rest of the sentence as follows:

Input	Output
We	must
We	love
Must	we
Must	love
Must	life
Love	we
Love	must
Love	life
Love	must
Love	live

At the end of creating a training sample for the sentence, we gotten samples. After that, we will obtain Word2Vec embedding by having a data set consists of 4000 unique words and word vectors size 80 each.

- Vocabulary Size = 4000
- Word vector size =80

The input will be our input vector, and the output will be the probability of nearby vectors. The learned weight matrix can be extracted after the model finishes the training process, and the word vector can be extracted, as shown in Figure 2 and Figure 3.

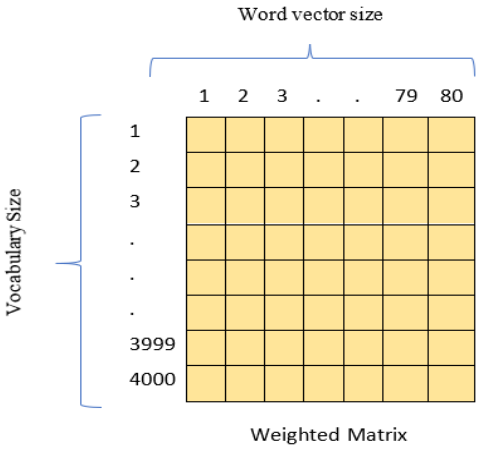


Figure 2: The weight matrix has a size of 4000 x 80.

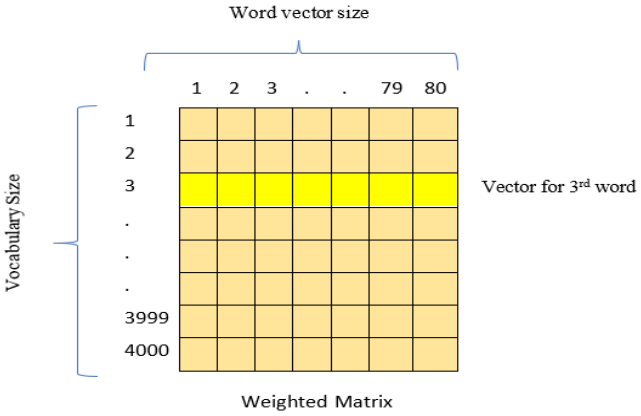


Figure 3: Vector representation for the 3rd word

The above shows how to have a word embedding using Word2Vec. Similar words in the vocabulary set will have similar vectors pointing toward the same direction for example, life and living will have the same direction as shown in Figure 4.

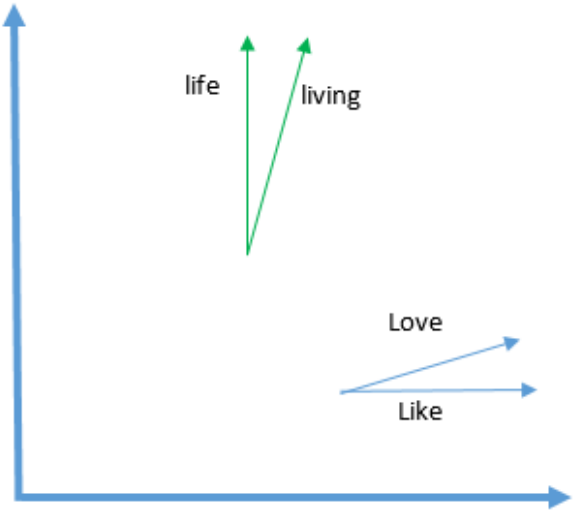


Figure 4: The vector direction for similar words

6. Conclusions and Future Work

CryptDB is a system that offers a logical and powerful level of confidentiality against two considerable threats that attack database-backed applications, i.e., curious DBAs and arbitrary compromises of the application server and the DBMS. Three steps are followed by CryptDB to achieve its objectives: efficiently performing queries over encrypted data through a novel SQL-aware encryption strategy, dynamically modifying the encryption level using onions of encryption to decrease the information given to the untrusted DBMS server, and linking encryption keys to under passwords such that only authorized users can get access to encrypted data. Word2Vec method is an effective data processing technique. It takes into account the significant data dimension issue when handling large-scale training data to offer a means of grouping similar data. This method may be employed to decrease the data dimension.

The CryptDB and Word2Vec approaches will be used together to provide a secure word embedding.

Acknowledgment

This is a part of project done in British University in Dubai.

References

[1] M. Alshurideh, B. Al Kurdi, S.A. Salloum, I. Arpaci, M. Al-Emran, "Predicting the actual use of m-learning systems: a comparative approach using PLS-SEM and machine learning algorithms," Interactive Learning Environments, 1–15, 2020.
[2] M. AlShamsi, S.A. Salloum, M. Alshurideh, S. Abdallah, Artificial Intelligence and Blockchain for Transparency in Governance, Springer: 219–230.
[3] J. Almaazmi, M. Alshurideh, B. Al Kurdi, S.A. Salloum, "The Effect of Digital Transformation on Product Innovation: A Critical Review," in International Conference on Advanced Intelligent Systems and Informatics, Springer: 731–741, 2020.
[4] A.Y. Zainal, H. Yousuf, S.A. Salloum, "Dimensions of Agility Capabilities Organizational Competitiveness in Sustaining," in Joint European-US Workshop on Applications of Invariance in Computer Vision, Springer: 762–772, 2020.
[5] H. Yousuf, S. Salloum, "Survey Analysis: Enhancing the Security of Vectorization by Using word2vec and CryptDB."
[6] S.K. Yousuf H., Lahzi M., Salloum S.A., "Systematic Review on Fully Homomorphic Encryption Scheme and Its Application.," In: Al-Emran M., Shaalan K., Hassanien A. (Eds) Recent Advances in Intelligent Systems and Smart Applications. Studies in Systems, Decision and Control, Vol 295. Springer, Cham, 2021.
[7] S.A. Salloum, R. Khan, K. Shaalan, "A Survey of Semantic Analysis Approaches," in Joint European-US Workshop on Applications of Invariance in Computer Vision, Springer: 61–70, 2020.
[8] S.A. Salloum, M. Alshurideh, A. Elnagar, K. Shaalan, "Machine Learning and Deep Learning Techniques for Cybersecurity: A Review," in Joint European-US Workshop on Applications of Invariance in Computer Vision, Springer: 50–57, 2020.
[9] S.A. Salloum, M. Alshurideh, A. Elnagar, K. Shaalan, "Mining in Educational Data: Review and Future Directions," in Joint European-US Workshop on Applications of Invariance in Computer Vision, Springer: 92–102, 2020.
[10] S.A. Salloum, A.Q. AlHamad, M. Al-Emran, K. Shaalan, A survey of Arabic text mining, 2018, doi:10.1007/978-3-319-67056-0_20.
[11] S.A. Salloum, M. Al-Emran, A.A. Monem, K. Shaalan, Using text mining techniques for extracting information from research articles, 2018, doi:10.1007/978-3-319-67056-0_18.
[12] Y.-C. Chang, M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in International Conference on Applied Cryptography and Network Security, Springer: 442–455, 2005.
[13] K.C. M. Viktor, "Big data: A revolution that will transform how we live, work, and think," Houghton Mifflin Harcourt, 2013.
[14] N. Aburawi, CryptDB mechanism on graph databases, 2020.
[15] J. Kepner, V. Gadepally, B. Hancock, P. Michaleas, E. Michel, M. Varia, "Parallel vectorized algebraic AES in Matlab for rapid prototyping of encrypted sensor processing algorithms and database analytics," in 2015 IEEE High Performance Extreme Computing Conference (HPEC), IEEE: 1–8, 2015.
[16] T.V.X. Phuong, G. Yang, W. Susilo, "Efficient hidden vector encryption with constant-size ciphertext," in European Symposium on Research in Computer Security, Springer: 472–487, 2014.
[17] J.H. Park, K. Lee, W. Susilo, D.H. Lee, "Fully secure hidden vector encryption under standard assumptions," Information Sciences, **232**, 188–207, 2013.
[18] R.A. Popa, C.M.S. Redfield, N. Zeldovich, H. Balakrishnan, "CryptDB: protecting confidentiality with encrypted query processing," in Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, 85–100, 2011.
[19] K. Lang, "Newsgroups Data Set," Available at: Qwone. Com/~ Jason/20Newsgroups/.[Accessed 30-Sep-2015], 20AD.

[20] H.H.O. Nasereddin, A.J. Darwesh, "An Object Oriented Programming on Encrypted Database System (CryptDB)," Journal of Talent Development and Excellence, **12**(1), 5140–5146, 2020.
[21] X. Jiang, X. Kong, Z. Xu, "Research on order-preserving encryption scheme based on CryptDB," in Journal of Physics: Conference Series, IOP Publishing: 32106, 2020.
[22] S.S.M. Chow, J.-H. Lee, L. Subramanian, "Two-Party Computation Model for Privacy-Preserving Queries over Distributed Databases.," in NDSS, Citeseer, 2009.
[23] H. Yousuf, A.Y. Zainal, M. Alshurideh, S.A. Salloum, Artificial Intelligence Models in Power System Analysis, Springer: 231–242.
[24] S. Rizvi, A. Mendelzon, S. Sudarshan, P. Roy, "Extending query rewriting techniques for fine-grained access control," in Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 551–562, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 1288-1298 (2020)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

A Circular Invariant Convolution Model-Based Mapping for Multimodal Change Detection

Redha Touati[*,1,2], Max Mignotte[1], Mohamed Dahmane[2]

[1]*Vision lab, Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal, QC, H3T 1J4, Canada*

[2]*R & D vision Département, Centre de Recherche Informatique de Montréal (CRIM), Montréal, QC, H3T 1J4, Canada*

ARTICLE INFO

Article history:
Received: 02 September, 2020
Accepted: 18 October, 2020
Online: 26 October, 2020

Keywords:
Multisource
Multisensor
Mutilooking
Unbalanced satellite images
Circular invariant convolution model
Line minimization process
Change detection
Mapping or projection model

ABSTRACT

The large and ever-increasing variety of remote sensing sensors used in satellite imagery today explains why detecting changes between identical locations in images that are captured, at two separate times, from heterogeneous capturing systems is a major and challenging recent research problem in the field of satellite imaging for fast and accurate determination of temporal changes. This work presents an original concentric circular invariant convolution model that aims at projecting the first satellite image into the imaging modality of the second image. This allows the two images to have identical statistics so that one can then effectively use a classical monomodal change detection method. The invariant circular convolution kernel parameters are estimated in the least squares sense using a conjugate gradient routine whose optimal direction is determined by a quadratic line search algorithm. After the projection of the before image into the imaging modality domain associated with the after image is achieved, a basic pixel-by-pixel difference permits the estimation of a relevant soft difference map which is segmented into two classes by a multilevel Markovian technique. A series of experiments conducted on several pair of satellite images acquired under different imaging modalities, resolution scales, noise characteristics, change types and events, validates the effectiveness of this strategy. The experimentation shows that the proposed model can process different image pairs with less re-striction about the source images and natural event, coming from different sensors or from the same sensor, for detecting natural changes.

1 Introduction

The ever-increasing number of Earth observation satellites, which use today new high-tech sensors, are often technologically quite different from those that have provided the satellite images so far archived. This heterogeneity of satellite image data has thus lately contributed to the advent and development of a novel and rapidly growing research interest in remote sensing and geoscience imaging commonly known as multi-modal (or heterogeneous) change detection (MCD). The purpose of the multimodal CD (MCD) [1, 2] is to detect and precisely locate any change in land cover between at least two satellite images taken in the same place, at different times and under different acquisition conditions. It usually concerns CD models that process heterogeneous satellite images, *i.e.* provided by different satellite sensor types which may combine active and passive sensors such as SAR/optical or by one sensor type, with two heterogeneous SAR, optical or others or finally by different satellites using the same satellite sensor but under different specifications, looks, wavelengths, or calibrations [3].

Recently, MCD has aroused a lot of attention and interest in satellite imagery thanks to its consistency and coherence with our environment favouring the emergence of increasingly heterogeneous images. Nevertheless, it is a difficult procedure because MCD has to be flexible enough to treat a multi-modal image pair, for solving the problems usually considered by the traditional single-modality CD methods [4, 5] including damage, land, environmental monitoring or city planning.

In the literature, the proposed MCD approaches can be classified under four classes. The first class includes the simplest methods which use local descriptors [6] or similarity measures [6]–[9], using invariance properties relative to the imaging modality processed. The second category gathers the methods that do not have assumptions about the data distributions and are thus non-parametric. This includes algorithms using machine learning techniques that learn from the training samples [10]–[19] or from unsupervised parameter,

[*]Corresponding Author: Redha Touati, Email: touatire@iro.umontreal.ca

without requiring any training phase, for example the least squares (LSQ) Multidimensional scaling (MDS) and energy-based model, integrating data-driven pixel pairwise constraints, introduced in [20]. The third group relies on parametric methods trying to model common statistical features or relationship between the different imaging modalities or between the different multisource data *via* a set of multivariate distributions [21]–[27] or through a pixel pairwise modeling integrated in a Markov Random Field (MRF) model [28]. Finally, the last category regroups algorithmic procedures essentially relying on a projective transformation of the heterogeneous images to a common intermediate feature space, where the projected satellite images would ultimately share the same statistical characteristics and on which any conventional homogeneous CD models could be subsequently used [29]–[38]. This category can encompass also the procedures projecting the first image in the imaging modality of the second image or inversely [39]. The method presented in this work fits fully into this sub-category.

More precisely, this work presents a concentric circular invariant convolution mapping which aims at projecting the *before* image onto the imaging modality of the *after* image. In this way, we ensure that the statistics of the pre and after-event satellite images are nearly identical and that a classical monomodal CD procedure can then be efficiently used, as easily as if the two images came from the same imaging modality or sensor with the same settings (specifications/calibrations/wavelength).

To this end, we will show that, once the mapping is done by the proposed convolutive model, the pixel-wise difference image estimated from the image pair shows good likelihood distributions properties corresponding to the "change" and "no change" label classes; that is to say, presenting a mixture of distributions not too far from normal distributions and also not too flat (*i.e.* large or non-informative) so that a binary segmentation with an unsupervised Markovian model to be efficient, despite the big difference between the imaging modalities that could be confronted in multi-modal satellite imagery and which will be evaluated in this paper.
As an amelioration of our convolution model-based mapping [1], in the present research we propose a circular invariant model. Furthermore, a three dimensional (3D) convolution based mapping strategy was adopted in the specific case of MCD using two color images (or multispectral). A battery of tests was conducted to quantify the benefits of such improvements.

It is worth mentioning that convolution model estimation has also been widely experienced and analyzed in different image processing problems as image restoration [40] or 2D or 3D deconvolution issues [41]–[44].

The reminder of this paper is structured as follows: Section 2 describes the proposed projection model based on a concentric circular invariant convolution representation and the unsupervised Markovian approach. Section 3 presents the results and an experimental comparison with state of the art multimodal CD methods. Section 4 concludes the paper.

2 Proposed MCD Model

Let I_b and I_a, be the bi-temporal remote sensing image pair of N pixels, captured in the same area, at two consecutive times (*i.e*, before and after a given event), and coming from different imaging systems or sensors. Let $I_b^\sharp$ and $I_a^\sharp$ also be the informational part of the image or the semantic information of the scene concretely representing the set of real objects and materials imaged in I_b and I_a. The acquisition system, related to respectively the pre-event and post-event image can be modeled by the following linear models: $I_b = I_b^\sharp * u_b$ and $I_a = I_a^\sharp * u_a$ where "$*$" is the convolution operator and u_b and u_a mathematically model the underlying structure of the Point Spread Function (PSF) of the pre-event and post-event data acquisition system [45].

This modeling framework defines a reliable and efficient way for projecting the before-change image to the domain or imaging system of the after-change image image. It consists to apply the following operation to the pre-event image: $(I_b * u_b^{*-1}) * u_a = I_b^\sharp * u_a$ where "u_b^{*-1}" represents the inverse convolution operator of u_b (giving $u_b^{*-1} * u_b = \delta$, such as δ is the Dirac impulse function) or by commutativity and associativity of the convolution product, applying the convolution operator $u_{b\mapsto a} = (u_b^{*-1} * u_a) = (u_a * u_b^{*-1})$ to I_b. This operation allows us to convert the original MCD issue into a conventional monomodal CD method which is used in the case of a pair of images having the same acquisition mode and therefore having the same statistics.

In order to reliably find $u_{b\mapsto a}$, the simplest strategy is to search for the convolution filter parameters that will minimize in the LSQ context, the energy function given by equation 1:

$$\hat{u}_{b\mapsto a} = \arg\min_{u} \overbrace{\sum_{(x,y)\in S_{nc}} \Big([I_b(x,y) * u_{b\mapsto a}(x,y)] - I_a(x,y)\Big)^2}^{E(u_{b\mapsto a})} \tag{1}$$

where S_{nc} is the set of pixels that do not belong to the class label "change" and which must be evaluated in the final change detection map. As first approximation for $\hat{u}_{b\mapsto a}$, we can take for S_{nc}, the set of all pixels in the image. This can be justified because the area of change is usually quite small in proportion to the image which, also, can be considered (in a very first approximation) as noisy observations or outliers in the image data with which it is shown that the LSQ estimator performs well (especially thanks to the fact that the function $E(u_{b\mapsto a})$ to be minimized is convex with regard to the parameter vector of the convolution filter $u_{b\mapsto a}$). As a result, different efficient minimization procedures can be adopted [47].

In this work, we use a combination of conjugate gradient and a line minimization search routine guaranteeing a considerably increased convergence speed compared to other minimization methods. Another advantage of such a minimization technique consists on its ability to integrate some hard constraints on $u_{b\mapsto a}$. In our case, since the two imaging modalities are circular invariant, we constraint $u_{b\mapsto a}$ to be also circular invariant. To this end, at each iteration k of the conjugate gradient descent (CGD), we simply average each of the $(sz \times sz)$ parameter values of $u_{b\mapsto a}^{[k]}$ that are equidistant (in the L_1 distance sense) from the center of the convolution filter

Table 1: Heterogeneous datasets description

No	Dates	Area	Image size	Event	Satellite
1	Sept. 1995 - Jul. 1996	Sardinia, It	412×300	Lake overflow (30 m.)	Landsat-5 (NIR band) / Optical
2	Nov. 1999 - Oct. 2000	Gloucester, Uk	330×590	Flooding (≈ 40 m.)	SAR-ERS / Spot
3	July 2006 - July 2007	Gloucester, Uk	2325×4135	Flooding (0.65 m.)	TerraSAR-X / QuickBird 2
4	Feb. 2009 - July 2013	Toulouse, Fr	4404×2604	Construction (2 m.)	TerraSAR-X / Pleiades
5	May 2012 - July 2013	Toulouse, Fr	2000×2000	Construction (0.52 m.)	Pleiades / WorldView 2
6	2001 - Jan. 2002	Congo, Af.	400×800	Nyiragongo volcano eruption (10 m.)	Radarsat (3 / 5-looks)
7	June 2008 - June 2013	Island town, Chn.	415×403	Urbanization (8 m.)	Radarsat-2 / Google Earth
8	June 2008 - Sept. 2012	Shuguang village, Chn	419×342	Building construction (8 m.)	Radarsat-2 / Google Earth
9	June 2008 - June 2009	Yellow river, Chn.	409×367	River's drying up (8 m.)	Radarsat-2 (1 / 4-looks)
10	Aug. 2013 - Aug. 2015	Wehe river, Xi'an city, Chn.	392×303	flooding (2 m.)	GF-1 (≠ bands) [46]
11	1999 - 2000	Gloucester, Uk	990×554	flooding (≈ 25 m.)	SPOT/NDVI [17]
12	May 2002 - July 2009	Hubei Campus, Chn	240×240	Building construction (2.44 m.)	QuickBird / IKONOS
13	April 2005 - July 2009	Wuhan University, Chn.	400×400	Building construction (3.28 m.)	QuickBird / IKONOS

(see Algorithm 1). This allows us to integrate this constraint with a linear complexity with respect to the $(sz \times sz)$ parameters of the convolution filter. This overall minimization routine is presented in our basic model [1].

Algorithm 1 Concentric Circular Constraint

$u(x, y)$ Convolution filter of dim. $(sz \times sz)$ **[input]**
$u^{\circ}(x, y)$ Circular invariant filter **[output]**
TAB$[x][y]$ 2D table (length,width)= $(2sz,2)$ of floats

- Fill the table TAB$[i][j]$ with zeros

for *each cell of* $u(i, j)$ *with coordinates* (i, j) **do**

- $d \leftarrow L_1$distance from (i, j) to the center of u
- TAB$[d][0] \leftarrow$ TAB$[d][0] + u(i, j)$
- TAB$[d][1] \leftarrow$ TAB$[d][1] + 1$

for *each cell of* $u^{\circ}(i, j)$ *with coordinates* (i, j) **do**

- $d \leftarrow L_1$distance from (i, j) to the center of u°
- $u^{\circ}(i, j) \leftarrow$ TAB$[d][0]$/TAB$[d][1]$

In order to improve this approximation of $\hat{u}_{b\mapsto a}$, we can formulate the LSQ estimation problem of the convolution filter parameters as a fixed point problem involving a contraction mapping. Technically speaking, the first estimation of $\hat{u}_{b\mapsto a}$ allows us to obtain $I_{b\mapsto a}(x, y) = I_b(x, y) * u_{b\mapsto a}(x, y)$, corresponding to the image pre-event image projected on the modality of the post-event image. A simple pixel-by-pixel difference between $I_{b\mapsto a}(x, y)$ and $I_a(x, y)$ enables us then to produce an output image of absolute difference that is subsequently binarized to the "change" and the "non-change" class label. To this end, a Gaussian kernel is used to model the distribution of each class and the Expectation Maximization (EM) [48] algorithm is used both to estimate the parameters of this weighted distribution mixture but also to give an approximate estimate of S_{nc} (*i.e.*, the pixels that does not belong to the class label "change") with a binarization scheme in the Maximum Likelihood (ML) sense. This process allows us to finally better estimate $\hat{u}_{b\mapsto a}$ (see Eq. 1). This process is repeated until the stability of the algorithmic steady state fixed point thus defined is reached. This fixed point-based strategy yields to the following iterative procedure:

$$\hat{u}_{b\mapsto a}^{[k+1]} = \arg\min_{u} \sum_{(x,y)\in S_{nc}^{[k]}} \Big([I_b(x, y) * u_{b\mapsto a}(x, y)] - I_a(x, y) \Big)^2 \quad (2)$$

$$\Phi_D^{[k+1]} = \mathrm{EM}\Big\{ \big| \overbrace{[I_b(x, y) * u_{b\mapsto a}^{[k+1]}(x, y)]}^{I_{b\mapsto a}^{[k+1]}(x,y)} - I_a(x, y) \big| \Big\}_{D^{[k+1]}(x,y)} \quad (3)$$

$$S_{nc}^{[k+1]} = \mathrm{ML}_{\text{binarization}}\{\Phi_D^{[k+1]}, D^{[k+1]}(x, y)\} \quad (4)$$

where $S_{nc}^{[0]}$ stands for the whole image pixels and $\Phi_D^{[k+1]}$ is the parameter vector of the two weighted normal kernels evaluated by the EM procedure (at iteration $k + 1$) on the difference image $D(x, y)$. $\mathrm{ML}_{\text{binarization}}\{\Phi_D^{[k+1]}, .\}$ represents the binarization algorithm, based on $\Phi_D^{[k+1]}$, in the Maximum Likelihood sense.

A second enhancement can be incorporated in this iterative procedure by simply inverting the direction of temporality of the two (before and after) satellite images in the algorithm introduced in our basic model [1]. Concretely, the estimation of $\hat{u}_{a\mapsto b}$ can be done by $I_{a\mapsto b}(x, y) = I_a(x, y) * u_{a\mapsto b}(x, y)$. This allows us to obtain another difference map between $I_{a\mapsto b}(x, y)$ and $I_b(x, y)$. The reverse estimate can be used along with the original one to obtain an average estimate ensuring an improved (less noisy) difference image $D(x, y)$ (see Algorithm 2).

In the case where the image before or after is a single-band image and the second image (possibly a multi-band) is then converted into gray levels and this pair of greyscale images are used as explained in Algorithm 3. A third improvement is proposed in the particular case where we have at our disposal two images before and after in colors (or with $b > 1$ color bands). In this case, the mapping and a difference map $D(b, x, y)$ is estimated for each of the b images individually and sequentially. A difference map is defined finally as

the L_∞ distance, for each pixel (x, y), between the b-components of each pixel ($D(x, y) \leftarrow \max_b D(b, x, y)$) as illustrated in Algorithm 2.

After the fixed point becomes stable (concretely when $S_{nc}^{[k]} \approx S_{nc}^{[k+1]}$, or after $L_{\max}^{P}$, a maximal number of iterations is reached), the parameter vector of the Gaussian kernel mixture Φ_D (given by the EM procedure) is used to classify the difference map into two classes, once again, in the Sequential Maximum *A Posteriori* (SMAP) sense by using the multiscale, fine-to-coarse and coarse-to-fine segmentation procedure [49].

This overall minimization routine is summarized in Algorithm 2 and in Algorithm 3 for a N-band image pair.

Algorithm 2 Convolution-Based Multimodal Change Detection

$S(x, y)$	Binary change detection map [**output**]
$I_b(x, y)$	The before image [**input**]
$I_a(x, y)$	The after image [**input**]
$D(x, y)$	Difference map
Φ_D	Vector parameters
$L_{\max}^{EM}$	Maximal number of iterations for EM
$L_{\max}^{P}$	Maximal number of fixed-point iterations

1. Initialization

• Set $S(x, y)$ with labels "no change" for each pixel

2. Fixed Point Estimation Step

Repeat

• Estimate $u_{b\mapsto a}$ and $u_{a\mapsto b}$ (see [1])

• Circular constraint on $u_{b\mapsto a}$ and $u_{a\mapsto b}$

• $D(x, y) = |[I_b(x, y) * u_{b\mapsto a}(x, y)] - I_a(x, y)|$
. $+ |[I_a(x, y) * u_{a\mapsto b}(x, y)] - I_b(x, y)|$

• $\Phi_D \leftarrow$ EM estimation ($L_{\max}^{EM}$ iter.) on $D(x, y)$

• $S^{[l]} \leftarrow$ ML binarization of $D(x, y)$ based on Φ_D

• $k++$

Until $k < L_{max}^{P}$ (**or** $S(x, y)^{[l]} \neq S(x, y)^{[l-1]}$);

3. Segmentation Step

• $S(x, y) \leftarrow$ SMAP on $D(x, y)$ based on Φ_D

3 Experimental Results

3.1 Heterogeneous Dataset Description

Our strategy is validated by conducting a set of evaluation scenarios on thirteen dissimilar heterogeneous datasets exhibiting different multi modality imaging data such as cross-sensor or multisensor (#1,#5,#12,#13), multisource (#2,#3,#4,#7,#8), multi-looking (#6,#9,#11) or multispectral images (#10)) with different resolution levels and image sizes and imaging a wide variety of changed events which are degraded by a wide variety of both noise type and levels as depicted in Table 1.

In all the experimental results, we recall that if one of the two images is in grayscale, we must convert the other image (colors or with b (> 1) bands) into grayscale to apply algorithm 2. Otherwise, if the two images contain the same number of bands (b), we rely on algorithm 3.

In addition, for computational reasons, we have reduced the image size so that the widest edge of it is at most equal to 500 pixels. For the estimation stage, the size sz of the convolutional (square) filter was set to 9, with a maximal iteration number $L_{\max}^{G}$ set to 200 and a fixed-point iteration number $L_{\max}^{P}$ equal to 2. The iteration number $L_{\max}^{EM}$ used in the EM algorithm is 12 iterations (see our basic model [1]). For the segmentation stage, the depth of the SMAP was set to $d = 9$ [49]. The regularization parameter θ was fixed to 0.9 (a value commonly used for this SMAP fine-to-coarse, coarse-to-fine segmentation algorithm [49, 54]).

Algorithm 3 CBMCD For N-Band Image pair

$S(x, y)$	Binary change detection map [**output**]
$I_b(b, x, y)$	The before image with $b < N$ bands [**input**]
$I_a(b, x, y)$	The after image with $b < N$ bands [**input**]
$D(x, y)$	Difference map
Φ_D	Vector parameters
$D(b, x, y)$	Temporary map with $b < N$ bands
$L_{\max}^{EM}$	Maximal number of iterations for EM
$L_{\max}^{P}$	Maximal number of fixed-point iterations

1. Initialization

• Set $S(x, y)$ with labels "no change" for each pixel and fill $D(k, x, y)$ with zeros

2. Fixed Point Estimation Step

Repeat

for *each band b of* $[I_b(b, x, y), I_a(b, x, y)]$ **do**

• Estimate $u_{b\mapsto a}$ and $u_{a\mapsto b}$

• Circular constraint on $u_{b\mapsto a}$ and $u_{a\mapsto b}$

• $D(b, x, y) = |[I_b(b, x, y) * u_{b\mapsto a}] - I_a(b, x, y)|$
. $+ |[I_a(b, x, y) * u_{a\mapsto b}] - I_b(b, x, y)|$

• $\Phi_D \leftarrow$ EM ($L_{\max}^{EM}$ iterations) on $D(b, x, y)$

• $D(x, y) \leftarrow \max_b D(b, x, y)$

• $S \leftarrow$ ML binarization of $D(x, y)$ based on Φ_D

• $k \leftarrow k + 1$

Until $k < L_{max}^{P}$ (**or** $S(x, y)^{[l]} \neq S(x, y)^{[l-1]}$);

• $\Phi_D \leftarrow$ EM ($L_{\max}^{EM}$ iterations) on $D(x, y)$

3. Segmentation Step

• $S(x, y) \leftarrow$ SMAP on $D(x, y)$ based on Φ_D

3.2 Convolution Filter Estimation Step Result

For all the test scenarios, the maximum iteration number used in the conjugate gradient descent combined with the minimization line

process, is sufficient to ensure the convergence of the minimization procedure. As example, we show, in Figure 1, how evolves the error rate along the minimization process. In addition, for all the scenarios, we estimate for $u_{b\mapsto a}$ or $u_{a\mapsto b}$ a convolution filter (that can be likened to a point spread function (PSF)) which is quite different from one dataset to another since the frequency response (or modulation transfer function) of the imaging modality is also quite different for each pair of imaging multi-modalities. Besides, the convolution filter size sz is also justified, enough because small values of parameters are estimated at the edges of the rectangular filter in all the tested cases (see Figure 2) for an example of convolution filter estimate).

3.3 Results & Evaluation

We evaluate and compare the obtained results using the classification rate accuracy that measures the correct changed and unchanged pixels percentages: ACC and the F-measure Fm which can be defined in the equations 5 and 6.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Fm = \frac{2TP}{2TP + TN + FN + FP} \tag{6}$$

where TP and TN are the true positives and negatives. FN and FP designate the false negatives and positives.

Table 2 shows a comparison between our segmentation results and the different supervised and unsupervised state-of-the-art methods [9, 12, 16, 17, 20, 23]–[27, 32, 38, 51]–[53].

Figures 3, 4 and 5 compare qualitatively the obtained results. The average accuracy rate of our proposed model is higher than the supervised and unsupervised state-of-the-art methods, obtained by our proposed MCD model on the thirteen heterogeneous dataset is 87.78% with a F-measure equals to 0.444. Compared to these results, our preliminary model (non-circular invariant) [1] gave essentially the same result in terms of rate accuracy but with a much lower F-measure which did not exceed 0.389.

Globally, we notice that our approach works efficiently when a) the considered MCD problem tends towards the simple monomodal CD problem; in other words, in the case of a simple imaging modality which is not too different between the pre-event and post-event image (*e.g.*, when there are similar sensors involving slightly different noise laws between the before and after images), or b) in the case of a couple of images with high resolution, or c) in the case of

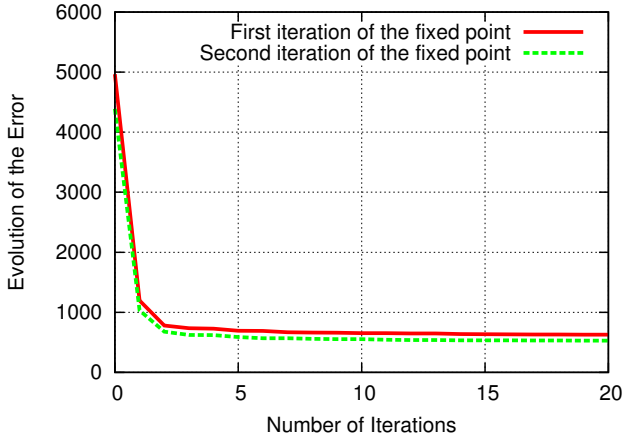


Figure 1: Cost function evolution (see Eq. (1) for the estimation of $u_{b\mapsto a}$ associated to dataset #1

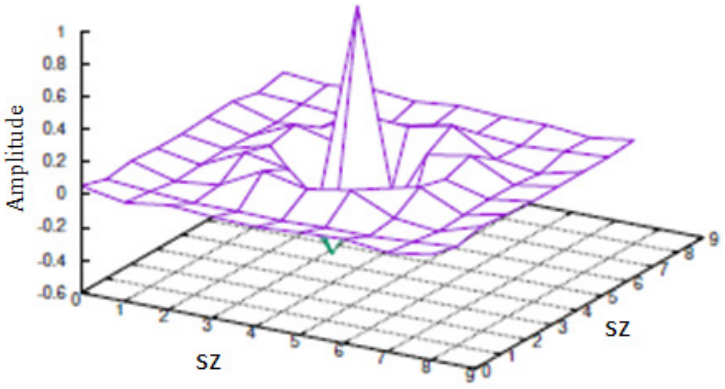


Figure 2: 9×9 Convolution filter estimate $u_{b\mapsto a}$ associated to dataset #1

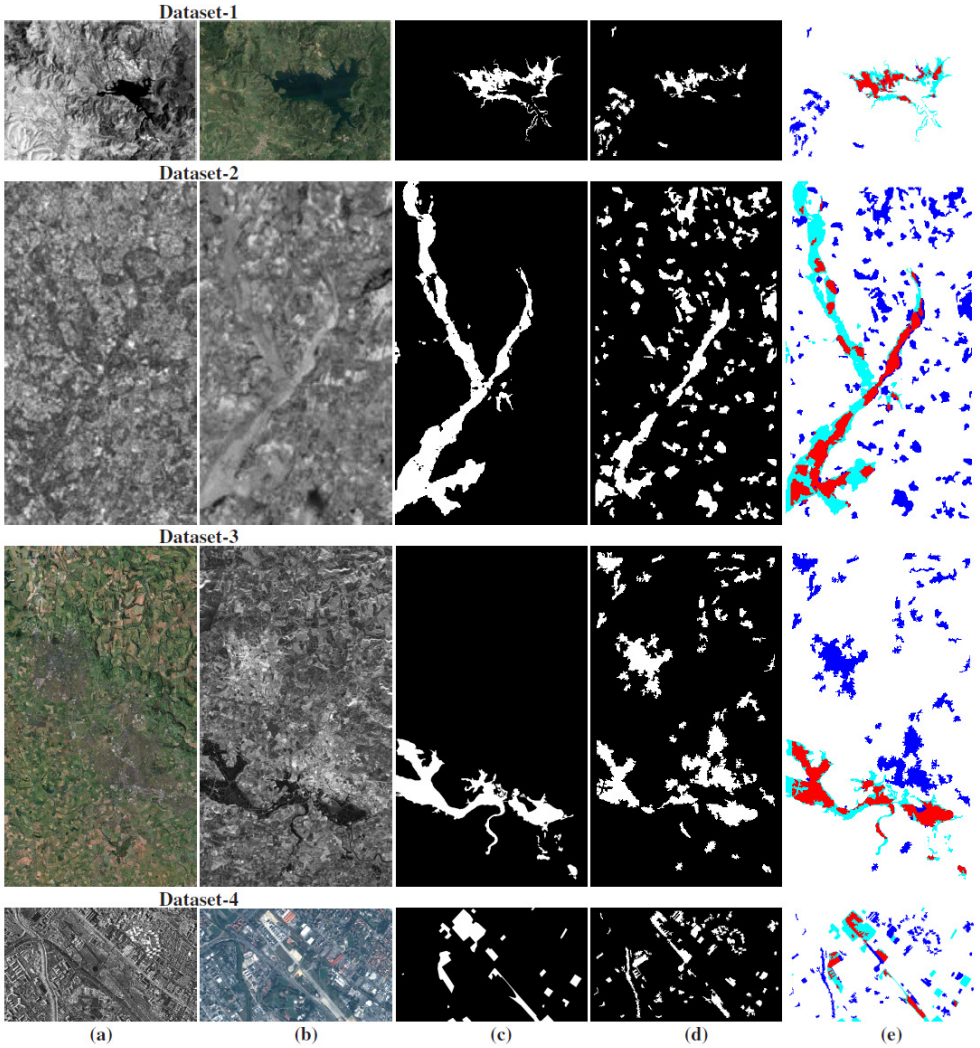


Figure 3: Multimodal datasets. (a) image t_1, (b) image t_2, (c) the ground truth; (d) binary map segmentation, (e) confusion map estimated by our model with white (TN), red (TP), blue (FP), cyan (FN) colors

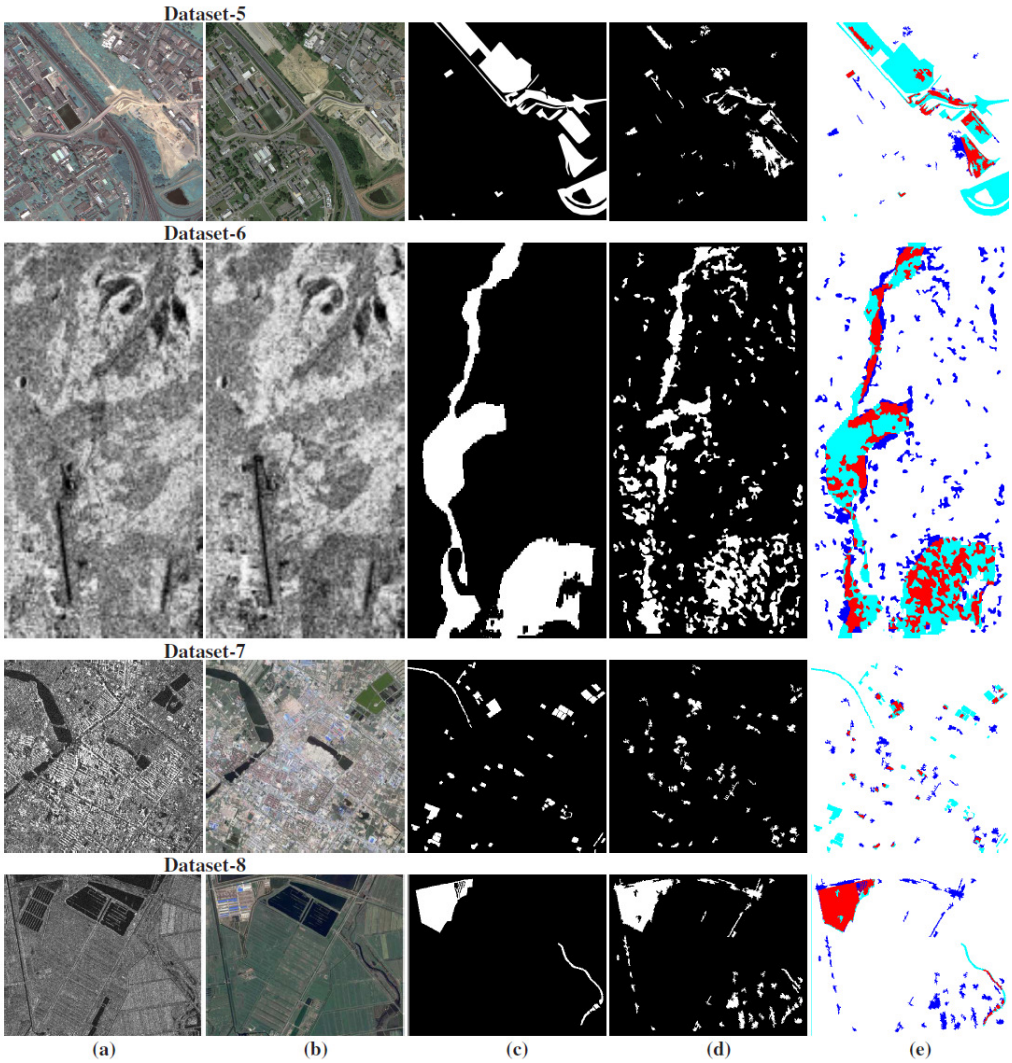


Figure 4: Multimodal datasets. (a) image t_1, (b) image t_2, (c) the ground truth; (d) binary map segmentation, (e) confusion map estimated by our model with white (TN), red (TP), blue (FP), cyan (FN) colors

simple changed events images with a simple uniform texture; as example, a city is flooding by a river (inducing an uniform texture for a new zone or also in building construction or structure with a uniform texture very different from the rest of the image). Conversely, our approach is a little less suitable when it comes to processing a bi-temporal images in which the variability between the before and after images is more important. Otherwise said, in the case of an imaging modality resulting from the combination of active and passive sensors, thus inducing a combination of multiplicative and additive noises in the before and after images with a dissimilar noise law. Also for images pairs with a low spatial resolution and showing complex changed events; as example, urban site construction (replacing a complex texture zone by the new complex texture area).

It requires about 54 seconds to the MCD to process an image. The processing time depends on the estimation problem complexity, the size of image, and the mono or multichanel based estimation (or 660 seconds for the set of 13 image pairs considered in this validation study) using a code with non-optimized C++ version running on $i7-930$ Intel CPU with 2.8 GHz Linux machine.

4 Conclusion

We have presented in this work an efficient mapping model for the change detection problem in multimodal imagery relying on the parametric estimation of an invariant circular convolution CD model. Optimization of the model was performed on a convex quadratic error function using a conjugate gradient routine, whose optimal direction is determined by a quadratic line minimization algorithm, and formulated as a fixed point problem involving a contraction mapping. This allows us to project the before image in the imaging modality space of the second image and inversely so that a difference change map is then computed pixel by pixel and efficiently binarized with a multiscale Markov model. Our proposed strategy seems well suited for the multimodal change detection issue and robust enough against images provided under different resolution levels, type or level of noise and exhibiting a variety of natural event.

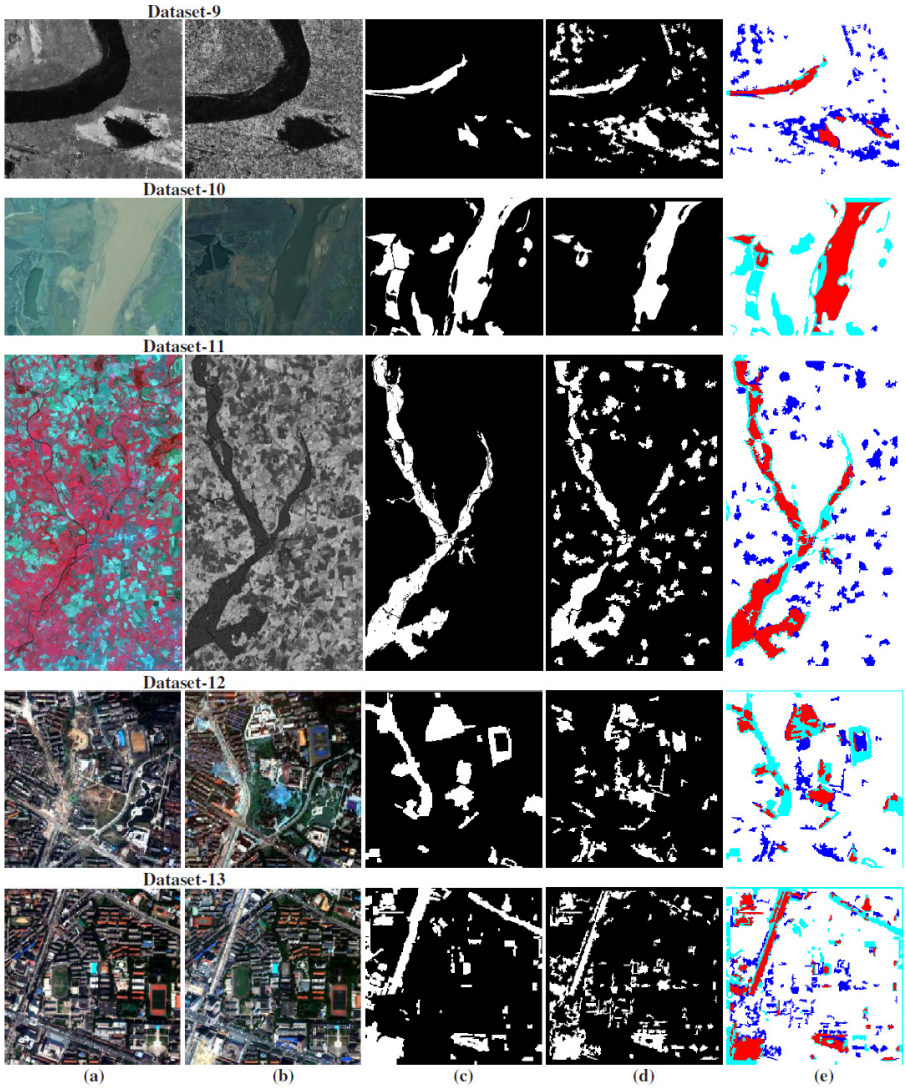


Figure 5: Multimodal datasets. (a) image t_1, (b) image t_2, (c) the ground truth; (d) binary map segmentation, (e) confusion map estimated by our model with white (TN), red (TP), blue (FP), cyan (FN) colors

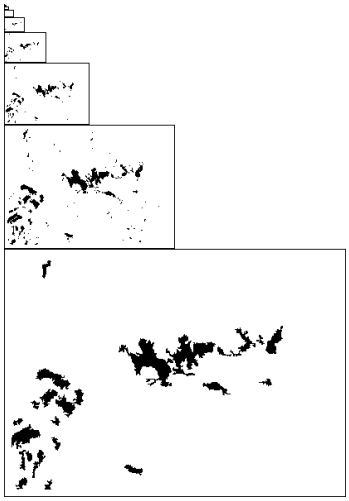


Figure 6: SMAP-based segmentation obtained for the dataset #1

Table 2: Results comparison.

Dataset [#1]	Accuracy rate
Proposed model	**0.942**
Touati et al. [38]	0.942
Zhang et al. [12]	0.975
PCC [12]	0.882

Dataset [#2]	Accuracy rate
Proposed model	**0.835**
Liu et al. [32]	0.818
Liu et al. [32]	0.655

Dataset [#3]	Accuracy rate
Proposed model	**0.896**
Touati et al. [20]	0.949
Touati et al. [9]	0.932
Touati et al. [38]	0.943
Prendes et al. [26]	0.844
Correlation [26]	0.670
Mutual Information [26]	0.580

Dataset [#4]	Accuracy rate
Proposed model	**0.882**
Touati et al. [20]	0.867
Touati et al. [38]	0.878
Touati et al. [38]	0.881
Prendes et al. [50, 27]	0.918
Prendes et al. [25]	0.854
Copulas [23, 25]	0.760
Correlation [23, 25]	0.688
Mutual Information [23, 25]	0.768
Difference [51, 25]	0.782
Ratio [51, 25]	0.813

Dataset [#5]	Accuracy rate
Proposed model	**0.866**
Touati et al. [20]	0.853
Touati et al. [9]	0.870
Touati et al. [38]	0.877
Prendes et al. [26, 27]	0.844
Correlation [26, 27]	0.679
Mutual Information [26, 27]	0.759
Difference. [51, 27]	0.708
Ratio [51, 27]	0.661

Dataset [#6]	Accuracy rate
Proposed model	**0.818**
Chatelain et al. [24]	0.749
Correlation [24]	0.713
Ratio edge [24]	0.737

Dataset [#7]	Accuracy rate
Proposed method	**0.946**
zhan et al. [52]	0.979

Dataset [#8]	Accuracy rate
Proposed model	**0.949**
Touati et al. [38]	0.967
Liu et al. [16]	0.976
PCC [16]	0.821

Dataset [#9]	Accuracy rate
Proposed model	**0.882**
zhan et al. [52]	0.991

Dataset [#10]	Accuracy rate
Proposed model	**0.862**
zhan et al. [52]	0.937

Dataset [#11]	Accuracy rate
Proposed model	**0.882**
HPT [17]	0.826

Dataset [#12]	Accuracy rate
Proposed model	**0.835**
Tang et al. [53]	0.986
Multiscale [53]	0.991

Optical/Optical [#13]	Accuracy rate
Proposed model	**0.815**
Tang et al. [53]	0.959
Multiscale [53]	0.966

Acknowledgement

The authors thank the researchers whose shared the *multi-modal* dataset.

References

[1] R. Touati, M. Mignotte, M. Dahmane, "Multimodal change detection using a convolution model-based mapping," in eighth International Conference on Image Processing Theory, Tools and Applications, IPTA 2019, Istanbul, Turkey, November 2019, 2019, doi:10.1109/IPTA.2019.8936127.

[2] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, Q. Du, "Multi-Modal Change Detection, Application to the Detection of Flooded Areas: Outcome of the 2009-2010 Data Fusion Contest," IEEE J. Sel. Topics Appl. Earth Observ., **5**(1), 331–342, 2012, doi:10.1109/JSTARS.2011.2179638.

[3] L. Su, M. Gong, P. Zhang, M. Zhang, J. Liu, H. Yang, "Deep Learning and Mapping Based Ternary Change Detection for Information Unbalanced Images," Pattern Recognition, **66**(C), 213–228, 2017, doi:10.1016/j.patcog.2017.01.002.

[4] R. Hedjam, M. Kalacska, M. Mignotte, H. Z. Nafchi, M. Cheriet, "Iterative Classifiers Combination Model for Change Detection in Remote Sensing Imagery," IEEE Transactions on Geoscience and Remote Sensing, **54**(12), 6997–7008, 2016, doi:10.1109/TGRS.2016.2593982.

[5] R. Touati, M. Mignotte, "A multidimensional scaling optimization and fusion approach for the unsupervised change detection problem in remote sensing images," in sixth IEEE International Conference on Image Processing Theory, Tools and Applications, IPTA'16, Oulu, Finland, December 2016, 1–6, 2016, doi:10.1109/IPTA.2016.7821021.

[6] G. Liu, J. Delon, Y. Gousseau, F. Tupin, "Unsupervised change detection between multi-sensor high resolution satellite images," in 24th European Signal Processing Conference, EUSIPCO 2016, Budapest, Hungary, August 29 - September 2, 2016, 2435–2439, 2016, doi:10.1109/EUSIPCO.2016.7760686.

[7] V. Alberga, "Similarity Measures of Remotely Sensed Multi-Sensor Images for Change Detection Applications," Remote Sensing, **1**(3), 122–143, 2009, doi:10.3390/rs1030122.

[8] D. Brunner, G. Lemoine, L. Bruzzone, "Earthquake Damage Assessment of Buildings Using VHR Optical and SAR Imagery," IEEE Transactions on Geoscience and Remote Sensing, **48**(5), 2403–2420, 2010, doi:10.1109/TGRS.2009.2038274.

[9] R. Touati, M. Mignotte, M. Dahmane, "A new change detector in heterogeneous remote sensing imagery," in 7th IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA 2017), Montreal, Canada, Qc, 2017, doi:10.1109/IPTA.2017.8310138.

[10] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. L. Rojo-Álvarez, M. Martínez-Ramón, "Kernel-Based Framework for Multitemporal and Multisource Remote Sensing Data Classification and Change Detection," IEEE Trans. Geoscience and Remote Sensing, **46**(6), 1822–1835, 2008, doi:10.1109/TGRS.2008.916201.

[11] P. Du, S. Liu, J. Xia, Y. Zhao, "Information Fusion Techniques for Change Detection from Multi-temporal Remote Sensing Images," Inf. Fusion, **14**(1), 19–27, 2013, doi:10.1016/j.inffus.2012.05.003.

[12] P. Zhang, M. Gong, L. Su, J. Liu, Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, **116**, 24–41, 2016, doi:10.1016/j.isprsjprs.2016.02.013.

[13] M. Gong, P. Zhang, L. Su, J. Liu, "Coupled Dictionary Learning for Change Detection From Multisource Data," IEEE Trans. Geoscience and Remote Sensing, **54**(12), 7077–7091, 2016, doi:10.1109/TGRS.2016.2594952.

[14] W. Zhao, Z. Wang, M. Gong, J. Liu, "Discriminative Feature Learning for Unsupervised Change Detection in Heterogeneous Images Based on a Coupled Neural Network," IEEE Trans. Geoscience and Remote Sensing, **55**(12), 7066–7080, 2017, doi:10.1109/TGRS.2017.2739800.

[15] N. Merkle, P. F. S. Auer, R. Muller, "On the Possibility of Conditional Adversarial Networks for Multi-Sensor Image Matching," in Proceedings of IGARSS 2017, IGARSS 2017, 1–4, Fort Worth, Texas, USA, 2017, doi:10.1109/IGARSS.2017.8127535.

[16] J. Liu, M. Gong, K. Qin, P. Zhang, "A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images." IEEE Transactions Neural Netw. Learn. Syst., 2018, doi:10.1109/TNNLS.2016.2636227.

[17] Z. Liu, G. Li, G. Mercier, Y. He, Q. Pan, "Change Detection in Heterogenous Remote Sensing Images via Homogeneous Pixel Transformation," IEEE Transactions on Image Processing, **27**(4), 1822–1834, 2018, doi:10.1109/TIP.2017.2784560.

[18] R. Touati, M. Mignotte, M. Dahmane, "Anomaly feature learning for unsupervised change detection in heterogeneous images: A deep sparse residual model," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **13**(1), 588–600, 2020, doi:10.1109/JSTARS.2020.2964409.

[19] R. Touati, M. Mignotte, M. Dahmane, "Partly uncoupled siamese model for change detection from heterogeneous remote sensing imagery," Journal of Remote sensing and GIS, **9**(1), 2020, doi:10.35248/2469-4134.20.9.272.

[20] R. Touati, M. Mignotte, "An energy-based model encoding non-local pairwise pixel interactions for multi-sensor change detection," IEEE Transactions on Geoscience and Remote Sensing, **56**(1), 2018, doi:10.1109/TGRS.2017.2758359.

[21] B. Storvik, G. Storvik, R. Fjørtoft, "On the Combination of Multisensor Data Using Meta-Gaussian Distributions," IEEE Trans. Geoscience and Remote Sensing, **47**(7-2), 2372–2379, 2009, doi:10.1109/TGRS.2009.2012699.

[22] G. Mercier, G. Moser, S. Serpico, "Conditional copula for change detection on heterogeneous SAR data," in 2007 IEEE International Geoscience and Remote Sensing Symposium, 2394–2397, 2007, doi:10.1109/IGARSS.2007.4423324.

[23] G. Mercier, G. Moser, S. Serpico, "Conditional copulas for change detection in heterogeneous remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, **46**(5), 1428–1441, 2008, doi:10.1109/TGRS.2008.916476.

[24] F. Chatelain, J. Y. Tourneret, J. Inglada, "Change Detection in Multisensor SAR Images Using Bivariate Gamma Distributions," IEEE Transactions on Image Processing, **17**(3), 249–258, 2008, doi:10.1109/TIP.2008.916047.

[25] J. Prendes, M. Chabert, F. Pascal, A. Giros, J. Tourneret, "A New Multivariate Statistical Model for Change Detection in Images Acquired by Homogeneous and Heterogeneous Sensors," IEEE Transactions on Image Processing, **24**(3), 799–812, 2015, doi:10.1109/TIP.2014.2387013.

[26] J. Prendes, M. Chabert, F. Pascal, A. Giros, J. Tourneret, "Performance assessment of a recent change detection method for homogeneous and heterogeneous images," Revue Française de Photogrammétrie et de Télédétection, **209**, 23–29, 2015.

[27] J. Prendes, New statistical modeling of multi-sensor images with application to change detection, Ph.D. thesis, Toulouse, 2015.

[28] R. Touati, M. Mignotte, M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov Random Field model," IEEE Trans. on Image Processing, **29**(1), 757–767, 2020, doi:10.1109/TIP.2019.2933747.

[29] M. Xu, C. Cao, H. Zhang, Y. Xue, Y. Li, J. Guo, C. Chang, Q. He, M. Gao, X. Li, "Change detection of the Tangjiashan barrier lake based on multi-source remote sensing data," in 2009 IEEE International Geoscience and Remote Sensing Symposium, volume 4, IV–303–IV–306, 2009, doi:10.1109/IGARSS.2009.5417373.

[30] M. Volpi, F. de Morsier, G. Camps-Valls, M. Kanevski, D. Tuia, "Multi-sensor change detection based on nonlinear canonical correlations," in 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS, 1944–1947, 2013, doi:10.1109/IGARSS.2013.6723187.

[31] C. Wu, B. Du, L. Zhang, "Slow Feature Analysis for Change Detection in Multispectral Imagery," IEEE Transactions on Geoscience and Remote Sensing, **52**(5), 2858–2874, 2014, doi:10.1109/TGRS.2013.2266673.

[32] Z. G. Liu, G. Mercier, J. Dezert, Q. Pan, "Change Detection in Heterogeneous Remote Sensing Images Based on Multidimensional Evidential Reasoning," IEEE Geoscience and Remote Sensing Letters, **11**(1), 168–172, 2014, doi: 10.1109/LGRS.2013.2250908.

[33] M. Volpi, G. Camps-Valls, D. Tuia, "Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis," ISPRS Journal of Photogrammetry and Remote Sensing, **107**, 50–63, 2015, doi:10.1016/j.isprsjprs.2015.02.005.

[34] X. Chen, J. Li, Y. Zhang, L. Tao, "Change Detection with Multi-Source Defective Remote Sensing Images Based on Evidential Fusion," ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 125–132, 2016, doi:10.5194/isprsannals-III-7-125-2016.

[35] D. Tuia, D. Marcos, G. Camps-Valls, "Multi-temporal and Multi-source Remote Sensing Image Classification by Nonlinear Relative Normalization," ISPRS Journal of Photogrammetry and Remote Sensing, 2016, doi: 10.1016/j.isprsjprs.2016.07.004.

[36] Z. Liu, L. Zhang, G. Li, Y. He, "Change detection in heterogeneous remote sensing images based on the fusion of pixel transformation," in 20th International Conference on Information Fusion, FUSION 2017, Xi'an, China, July 10-13, 2017, 1–6, 2017, doi:10.23919/ICIF.2017.8009656.

[37] L. T. Luppino, S. N. Anfinsen, G. Moser, R. Jenssen, F. M. Bianchi, S. B. Serpico, G. Mercier, "A Clustering Approach to Heterogeneous Change Detection," in Image Analysis - 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12-14, 2017, Proceedings, Part II, 181–192, 2017, doi:10.1007/978-3-319-59129-2_16.

[38] R. Touati, M. Mignotte, M. Dahmane, "Change detection in heterogeneous remote sensing images based on an imaging modality-invariant MDS representation," in 25th IEEE International Conference on Image Processing (ICIP'18), Athens, Greece, 2018, doi:10.1109/ICIP.2018.8451184.

[39] L. T. Luppino, F. M. Bianchi, G. Moser, S. N. Anfinsen, "Remote sensing image regression for heterogeneous change detection," CoRR, **abs/1807.11766**, 2018, doi:10.1109/MLSP.2018.8517033.

[40] S. Benameur, M. Mignotte, J.-P. Soucy, J. Meunier, "Image restoration using functional and anatomical information fusion with application to SPECT-MRI images," International Journal of Biomedical Imaging, **2009**, 12 pages, 2009, doi:10.1155/2009/843160.

[41] D. Kundur, D. Hatzinakos, "Blind image restoration via recursive filtering using deterministic constraints," in Proc. International Conference on Acoustics, Speech, and Signal Processing, volume 4, 547–549, 1996, doi: 10.1109/ICASSP.1996.547737.

[42] M. Mignotte, J. Meunier, "Unsupervised restoration of brain SPECT volumes," in Vision Interfaces, VI'00, 55–60, Montréal, Québec, Canada, 2000.

[43] M. Mignotte, J. Meunier, "Blind deconvolution of SPECT images using a noise model estimation," in SPIE Conference on Medical Imaging, volume 3979, 1370–1377, San Diego, CA, USA, 2000, doi:10.1117/12.387647.

[44] M. Mignotte, J. Meunier, "Three-dimensional blind deconvolution of SPECT images," IEEE trans. on Biomedical Engineering, **47**(2), 274–281, 2000, doi: 10.1109/10.821781.

[45] N. B. Smith, A. Webb, Introduction to Medical Imaging: Physics, Engineering and Clinical Applications, Cambridge Texts in Biomedical Engineering, Cambridge University Press, 2010, doi:10.1017/CBO9780511760976.

[46] T. Zhan, M. Gong, J. Liu, P. Zhang, "Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, **146**, 38 – 51, 2018, doi: 10.1016/j.isprsjprs.2018.09.002.

[47] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, Numerical Recipes in C, Cambridge University Press, Cambridge, USA, 2nd edition, 1992.

[48] A. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Royal Statistical Society, 1–38, 1976, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[49] C. Bouman, M. Shapiro, "A Multiscale Random Field Model for Bayesian Image Segmentation," IEEE Trans. on Image Processing, **3**(2), 162–177, 1994, doi:10.1109/83.277898.

[50] J. Prendes, M. Chabert, F. Pascal, A. Giros, J.-Y. Tourneret, "Change detection for optical and radar images using a Bayesian nonparametric model coupled with a Markov random field," in Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP'15, Brisbane, Australia, 2015, doi:10.1109/ICASSP.2015.7178223.

[51] O. D. Team, "The ORFEO Toolbox Software Guide," 2014, available at http://orfeo-toolbox.org/.

[52] T. Zhan, M. Gong, J. Liu, P. Zhang, "Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, **146**, 38–51, 2018, doi: 10.1016/j.isprsjprs.2018.09.002.

[53] Y. Tang, L. Zhang, "Urban Change Analysis with Multi-Sensor Multispectral Imagery," Remote Sensing, **9**(3), 2017, doi:10.3390/rs9030252.

[54] M. Mignotte, C. Collet, P. Pérez, P. Bouthemy, "Sonar image segmentation using an unsupervised hierarchical MRF model," IEEE Trans. on Image Processing, **9**(7), 1216–1231, 2000, doi:10.1109/83.847834.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 1293-1305 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Empirical Probability Distributions with Unknown Number of Components

Marcin Kuropatwiński*, Leonard Sikorski

Talking2Rabbit Ltd., R & D Department, 80-266, Poland

ARTICLE INFO

Article history:
Received: 28 August, 2020
Accepted: 09 December, 2020
Online: 16 December, 2020

Keywords:
probability mass function
probability density function
multinomial distribution
Laplace estimator
Gaussian mixture model

ABSTRACT

We consider the estimation of empirical probability distributions, both discrete and continuous. We focus on deriving formulas to estimate number of categories for the discrete distribution, when the number of categories is hidden, and the means and methods to estimate the number of components in the Gaussian mixture model representing a probability density function given implicitly in terms of its realizations. To reach the stated goals, we solve certain combinatorial problems for discrete distribution and develop methods to compute the expected Kullback-Leibler divergence for Gaussians. The last mentioned result is needed to develop the theory of continuous distributions. Sample applications and an extensive numerical study are given.

1 Introduction

This paper is an extension of work originally presented in 2019 Signal Processing Symposium [1]. The extension includes:

- measure theoretic account,
- derivation of the results first presented in the paper [1],
- new section devoted entirely to estimation of the number of components in continuous distributions,
- a numerical study of algorithms dealing with estimation of the continuous distributions,
- a specialized algorithm for counting number of monomials in a trace of covariance matrix raised to a power.

Empirical probability distributions are a crucial element of many applications such as machine learning [2], source coding [3], data compression [4], speech recognition [5], speaker recognition [6], image recognition [7], noise reduction [8], bandwidth extension [9], and many others. It is also a scientific discipline in itself, which is studied independently [10], [11]. In this paper, we deal with the means and methods to estimate probability distributions, both discrete and continuous. That being said, we also focus on a particular problem encountered during probability distributions estimation, which is the problem of an unknown number of components. In case of discrete distribution, the number of components is the number of categories, and in case of continuous probability distributions, it is the number of Gaussian components in the Gaussian mixture model (GMM) [2]. We provide estimators and methods to determine the number of components. Toward this, in case of GMMs, we formulate an information-theoretic criterion that allows the selection of an optimal, in some sense, number of components.

To strengthen the ties between discrete and continuous probability distribution, we provide a short account of measure theory and probability spaces. In particular, we show that the difference lies in the type of the underlying measurable space, which is a *standard space* with finite countable alphabets in case of discrete distributions and a *Polish Borel space* (where the *Polish space*, i.e., the complete separable metric space, is the sample space) in the case of continuous distributions. The name *Polish space* originates from the pioneering work by Polish mathematicians on such spaces.

1.1 Measure Theoretic Account

This account is based on the presentation provided by Robert M. Gray in an excellent book [12].

*Corresponding Author: Marcin Kuropatwiński, marcin@talking2rabbit.com

1.1.1 Measurable Space

A measurable space is a pair, $(\Omega, \mathcal{B})$ consisting of a sample space Ω with a σ-field $\mathcal{B}$ of subsets of Ω (also called the event space). A σ-field or σ-algebra $\mathcal{B}$ is a collection of subsets of Ω with the following properties:

$$\Omega \in \mathcal{B} \tag{1}$$

$$\text{If } F \in \mathcal{B}, \text{ then } F^c = \{\omega : \omega \notin F\} \in \mathcal{B} \tag{2}$$

$$\text{If } F_i \in \mathcal{B};\ i = 1, 2, \ldots, \text{ then } \cup F_i \in \mathcal{B} \tag{3}$$

The type of the measurable space is what differentiates discrete from continuous distributions. More precisely, it is the topology of the sample space. The discrete topology gives rise to discrete probability distributions, and the euclidean topology gives rise to continuous probability distributions.

Let $\mathcal{F} = \{F_i, i = 0, 1, 2, \ldots, n-1\}$ be a finite field of sample space Ω, that is, $\mathcal{F}$ is a finite collection of sets in Ω that are closed under finite set theoretic operations. To make it more concrete, let us give an example of such a field, which is the power-set of a finite, countable set of atoms (also called categories in the sequel). A set F in $\mathcal{F}$ will be called an *atom* only if its subset is also a field member of itself and the empty set, that is, it cannot be broken up into smaller pieces that are also present in the field. Let $\mathcal{A}$ denote a collection of atoms of $\mathcal{F}$. Then, one can show that $\mathcal{A}$ consists of exactly all nonempty sets of the form

$$\bigcap_{i=0}^{n-1} F_i^* \tag{4}$$

where F_i^* is either F_i or F_i^c. Let us call such sets *intersection sets*; we observe that any two intersection sets must be disjoint since for at least one i, one intersection set must lie inside F_i and the other within F_i^c. In summary, given any finite field $\mathcal{F}$ of the sample space Ω, we can find a unique collection of atoms $\mathcal{A}$ of the field such that the sets in $\mathcal{A}$ are disjoint, nonempty, and have the space Ω as their union. Thus, $\mathcal{A}$ is a *partition* of Ω. Such a sample space has a *discrete topology* with the basis given by the set of all atoms $\mathcal{A}$.

Now, we turn our attention to continuous distributions. As already mentioned, the underlying space is the *Polish Borel space*. An example of such a space, which will be used in this paper, is the *Euclidean space* endowed with the Euclidean topology. The basis for the Euclidean topology is the set of all open balls in that space. The *Polish space* is a complete, separable, metric space. We will explain what the listed properties of the *Polish spaces* mean.

A space is called *metric* if it is a set A, with elements called *points*, such that for every pair of points in A, there is an associated non-negative number $d(a, b)$ with the following properties:

$$d(a, b) = 0 \text{ if and only if } a = b \tag{5}$$

$$d(a, b) = d(b, a) \quad \textit{symmetry} \tag{6}$$

$$d(a, b) \leq d(a, c) + d(c, b) \text{ all } c \in A \quad \textit{triangle inequality} \tag{7}$$

A set F is said to be *dense* in A if every point in A is a point in F or a limit point of F.

A metric space A is called *separable* if it has a countable dense subset, that is, if there is a discrete set, say B, such that all points in A can be well approximated by points in B. This means that all the points in A are points in B or limits of the points in B. For example, n-tuples of rational numbers are dense in $\mathbb{R}^n$.

A sequence $\{a_n; n = 0, 1, 2, \ldots\}$ in A is called a *Cauchy sequence* if for every $\epsilon > 0$, there is an integer N such that $d(a_n, a_m) < \epsilon$ if $n \geq N$ and $m \geq N$. A metric space is *complete* if every Cauchy sequence converges, that is, if a_n is a Cauchy sequence, then there is $a \in A$ for which $a = \lim_{n\to\infty} a_n$.

1.1.2 Probability Spaces

A *probability space* $(\Omega, \mathcal{B}, P)$ is a triple consisting of a sample space Ω, a σ-field $\mathcal{B}$ of subsets of Ω, and a probability measure P defined on the σ-field; $P(F)$ assigns a real number to every member F of $\mathcal{B}$ so that the following conditions are satisfied:

Nonnegativity:

$$P(F) \geq 0, \text{ all } F \in \mathcal{B}, \tag{8}$$

Normalization:

$$P(\Omega) = 1. \tag{9}$$

Countable Additivity:

$$\text{If } F_i \in \mathcal{B},\ i = 0, 1, 2, \ldots \text{ are disjoint, then}$$
$$P(\bigcup_{i=0}^{\infty} F_i) = \sum_{i=0}^{\infty} P(F_i). \tag{10}$$

1.1.3 Densities

The probability density function (PDF) is a measure theoretic term defined through the *Radon-Nikodym theorem*.

A measure m is said to be *absolutely continuous* with respect to another measure P on the same measurable space, formally $m \ll P$, if $P(F) = 0$ implies $m(F) = 0$.

Theorem (*Radon-Nikodym theorem*)

Given the two measures m and P on a measurable space $(\Omega, \mathcal{F})$ such that $m \ll P$, there exists a measurable function $h : \Omega \to \mathbb{R}$ with the property that $h \geq 0$ such that

$$m(F) = \int_F h dP, \text{ all } F \in \mathcal{B}. \tag{11}$$

The function h is called the Radon-Nikodym *derivative* or *density* of m w.r.t. P and is denoted by $\frac{dm}{dP}$. If $\int f dP = 1$, then f is called a *probability density function*.

2 Discrete Distributions

For the statement of all results, as well as notations conventions, we refer the reader to the paper [1]. The results in the article [1] include discussion of the sun rise problem by Pierre-Simon Laplace from the essay [13] and some applications of the theory. We reference the paper [1] in its entirety. Here, only extensions of the material from the paper [1] are included, which have not found place there due to the form of a conference paper. However, for the readers convenience we provide a short summary of the notations used in the derivations:

- K - number of categories, possibly a hidden variable
- M - number of random experiments

- S - the partition, set of all categories
- $\{p_i\}_{i\in S}$ - multinomial proportions
- $X = \{x_i\}_{(i\in[1,M],x_i\in S)}$ - observations from the random experiments
- $|A|$ - cardinality of the set A
- UNIQUE(X) - unique elements in the multiset X
- $Z = |\text{UNIQUE}(X)|$ - the diversity index
- $P_i = |\{x_j : x_j = i\}|$ - counts of the observations

2.0.1 Derivations of the results for the uniform distribution

Assumption of uniform distribution is restrictive. However, it gives initial insight into the problem and, thus, is briefly presented here. As already introduced by X, we denote the sequence of observations. We can show that conditional probability of this sequence given the hypothetical K is equal to:

$$p(X|K) = \binom{K}{Z}\binom{M}{P_1 P_2 \cdots P_3}\frac{1}{K^M}. \tag{12}$$

It can be seen that the maximum likelihood estimate for the hypothetical number of categories, does not depend on the middle term, which includes the multinomial coefficient. Thus, the estimate can be obtained by,

$$K_{\text{ML}} = \underset{K}{\text{argmax}}\left[\binom{K}{Z}K^{-M}\right]. \tag{13}$$

It is convenient to introduce another quantity, which plays an important role in our theory. This quantity is the *generalization coefficient* N, which equals by definition:

$$N \equiv \frac{M}{Z}. \tag{14}$$

The ML estimate of K can then be obtained by solving the following equation:

$$\frac{1}{v}\ln\left(\frac{1}{1-v}\right) = N, \tag{15}$$

where:

$$v = \frac{Z}{K_{\text{ML}}}, \tag{16}$$

is the fraction of the number of observed categories to the number of all categories. The condition for the likelihood to be monotonically decreasing is as follows:

$$M > \log_{\left(\frac{Z+1}{Z}\right)}(Z+1). \tag{17}$$

If this condition holds, the ML estimate for K is equal to Z. The derivations of the above conditions are in place next. First, we note a property that establishes the link with the known in the statistical literature problem of coupon collector [14]:

$$\lim_{K\to\infty}\left(\frac{K\times H(K)}{\log_{\left(\frac{K+1}{K}\right)}(K)}\right) = 1, \tag{18}$$

where one can easily recognize that $K \times H(K)$ is the expected number of trials before the coupon collector collects the whole collection. In the above expression, $H(K)$ is the harmonic number, equal by definition:

$$H(K) \equiv \sum_{i=1}^{K}\frac{1}{i}. \tag{19}$$

The main vehicle of the derivation is the following expression that is valid for the harmonic numbers [15]:

$$\sum_{i=1}^{K}\frac{1}{i} = C + \ln K + \frac{1}{2K} - \sum_{i=2}^{\infty}\frac{A_i}{K(K+1)\cdots(K+i-1)}, \tag{20}$$

where C is the Euler-Mascheroni constant. It can be seen that asymptotically, as K approaches infinity, the terms after the logarithmic term vanish to zero. This leads to the following property:

$$\lim_{K\to\infty}\left(\sum_{i=1}^{K}\frac{1}{i} - \ln K\right) = C. \tag{21}$$

The derivation for the condition (15) begins with taking the logarithm of the considered expression (13):

$$\ln[p(X|K)] = \sum_{i=1}^{K}\ln i - \sum_{i=1}^{Z}\ln i - \sum_{i=1}^{K-Z}\ln i - M\ln K. \tag{22}$$

Suppose, i is a continuous variable, which setting follows from allowing that variable to take on non-integer values. It can be seen that the middle sum does not depend on K; therefore, derivative w.r.t that variable reads:

$$\frac{d}{dK}\ln[p(X|K)] = \sum_{i=1}^{K}\frac{1}{i} - \sum_{i=1}^{K-Z}\frac{1}{i} - \frac{M}{K} = \ln K - \ln(K-Z) - \frac{M}{K}. \tag{23}$$

The last expression allows us to immediately state the (15) and (16). Equations (15) and (16) allow us to conclude that the sample length needed to learn a given percent of the categories S is a multiple of K. Setting $Z = M$, we see that, indeed, this ($Z = M$) is the sufficient and necessary condition for optimal K approaching infinity. This is due to the following identity:

$$\lim_{K\to\infty}\frac{K}{M}\ln\left(\frac{1}{1-\frac{M}{K}}\right) = 1. \tag{24}$$

It remains to prove the (17). In this case, the maximum of the likelihood should be attained at $K = Z$. Thus, we have the following inequality:

$$\ln[p(X|Z)] > \ln[p(X|Z+1)], \tag{25}$$

which implies:

$$\ln(Z+1) - M\ln(Z+1) < -M\ln Z, \tag{26}$$

and, after some algebra, we attain at the (17). From the above reasoning all the results pertaining to the discrete, uniform distribution from the paper [1] can be deduced. In Figure 1, we illustrate the dependence of the data amount M needed to learn a given percent of all categories.

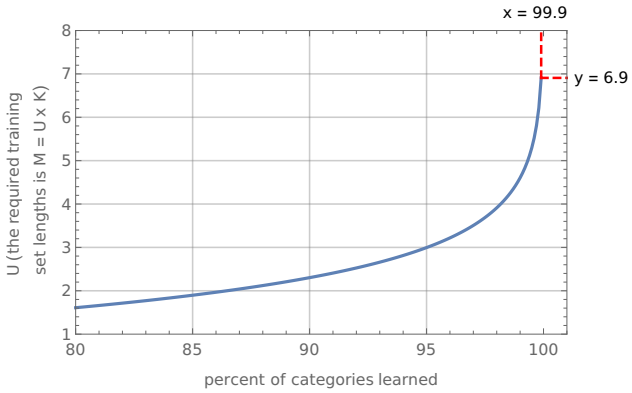


Figure 1: Data amounts requirement to learn a given percent of categories in case of uniform distribution

Next, we discuss the case of the non-uniform distribution, which is more relevant for real-world applications.

2.0.2 *Derivations of the results for the non-uniform distribution*

The next step will be the derivation of the conditions analogous to that introduced in the previous section, which are distribution-free (we relax the assumption of categories of equal probabilities). The desired effect will be achieved by using the earlier introduced multinomial proportions $\{p_i\}_{i\in S}$. Since we do not impose a constraint on the multinomial proportions, the considered probability function is now in the form,

$$p(X|K, p_1, \cdots, p_K) = \binom{M}{P_1 P_2 \cdots P_Z} \sum_{\{\mathbf{k}\,:\,\text{combinations of } Z \text{ objects out of } K \text{ objects}\}} \prod_{i=1}^{Z} p_{k_i}^{P_i}, \tag{27}$$

where $\mathbf{k} = [k_1, \cdots, k_Z]$. We integrate the above function over a unit simplex D:

$$D = \left\{ \mathbf{p} : \sum_{i=1}^{K} p_i = 1, p \in \mathbb{R}_+^K \right\}. \tag{28}$$

This corresponds to the assumption that all PMFs are equally likely, meaning we assume that we do not know the true PMF and attach to each possible $\mathbf{p} = [p_1, \cdots, p_K]$ an equal weight (we assume they are equally probable):

$$p(X|K) = \frac{1}{\text{vol}(D)} \int_D p(X|K, \mathbf{p}) d\mathbf{p} = \binom{K}{Z}\binom{M}{P_1 P_2 \cdots P_Z} \frac{1}{\text{vol}(D)} \int_D p_1^{P_1} \times \cdots \times p_Z^{P_Z} d\mathbf{p}, \tag{29}$$

where the expression (29) follows the fact, that the value of the integral does not depend on the choice and order of the probabilities in the monomial integrand. To proceed with the derivation, we compute the integral in (29) using Brion's formulae; please see [16]:

$$\frac{1}{\text{vol}(D)} \int_D p_1^{P_1} \times \cdots \times p_Z^{P_Z} d\mathbf{p} = (K-1)! \frac{\prod_{i=1}^{Z} P_i!}{(M+K-1)!}. \tag{30}$$

In light of the above expression, we get,

$$p(X|K, M) = \frac{K! M! \Gamma(K)}{Z!(K-Z)!\Gamma(K+M)}, \tag{31}$$

and next,

$$p(Z|K, M) = \frac{M\Gamma(K)\Gamma(K+1)\Gamma(M+Z)}{\Gamma(Z+1)^2\Gamma(K+M)\Gamma(K-Z+1)}. \tag{32}$$

To get the probability of the hypothetical number of categories, given M and Z, we apply the following derivation:

$$p(K|Z, M) = \frac{p(Z|K, M)p(K)}{p(Z|M)} = \frac{p(Z|K, M) \times C}{\sum_{K=Z}^{\infty} p(Z|K, M) \times C} = \frac{p(Z|K, M)}{\sum_{K=Z}^{\infty} p(Z|K, M)}, \tag{33}$$

where C is some constant (not to be confused with the Euler-Mascheroni constant used earlier in this document). Evaluation of (33) yields the following:

$$p(K|Z, M) = \frac{\Gamma(K)\Gamma(K+1)\Gamma(M-1)\Gamma(M)}{\Gamma(Z)\Gamma(Z+1)\Gamma(K+M)\Gamma(K-Z+1)\Gamma(M-Z-1)}. \tag{34}$$

From the reasoning presented above we can state the following:

$$s = \frac{u}{u+1}, \quad s = \frac{Z}{K}, \quad u = \frac{M}{K}. \tag{35}$$

Also, all results pertaining to non-uniform distribution, stated in the paper [1], are a consequence of the above reasoning. Based on (35) we can plot the dependence of the amount of data needed to learn given percent of all categories. This dependence is shown in Figure 2.

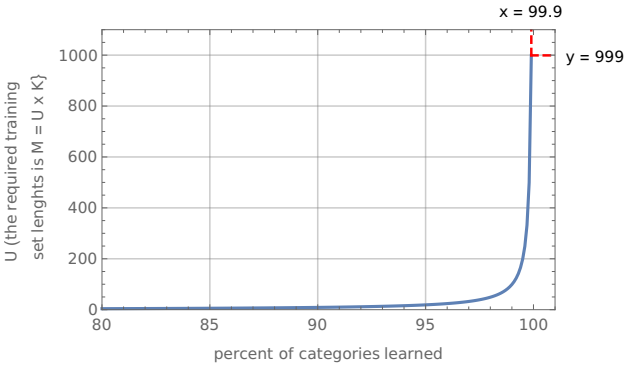


Figure 2: Data amounts requirement to learn a given percent of categories in case of non-uniform distribution

3 Continuous Distributions

The prominent method of density estimation for continuous distribution are Gaussian mixture models (GMM). The GMMs can approximate a PDF given by some training vectors drawn from that PDF. A wide-spread fitting procedure of the GMMs to the given training vectors is based on the Expectation-Maximization (EM) algorithm. Both, the mathematical definition of GMMs and the fitting procedure based on the EM is explained in detail in the book [2]. The neat property of the GMMs is that their expressive power allows modeling any PDF to any accuracy, given enough training

vectors are available. However, the traditional fitting procedure is susceptible to overtraining if too many components are chosen. Thus, in this paper, we study a fitting method that does not overfit, and, secondly, allows for the estimation of the number of components given the newly introduced information-theoretic criterion. To reach the goal of formulating the criterion, we first develop a mathematical approach to compute the expected Kullback-Leibler (KL) divergence for the single multivariate Gaussian, whereas the expectation is taken over given number M of samples used to estimate the multivariate Gaussian. In this section, we first develop the theory of expected KL divergence; next, we provide the modified training procedure and end the section with a numerical study.

3.1 Computation of expected Kullback-Leibler divergence

To the best of the authors' knowledge, the computation of the expected Kullback-Leibler divergence for multivariate Gaussian variables has been never previously explored in the literature. Thus, this paper deals with this problem, which can be formally expressed as,

$$\mathrm{E}_{\mathrm{KL}}(M) = \mathbb{E}_{x_i\sim p(x)}\left[D_{\mathrm{KL}}(p(x)\|\hat{q}(x,x_1,\cdots,x_M))\right], \tag{36}$$

where $p(x)$ and $q(x)$ are multivariate Gaussians.

The solution to the problem indicated above can be used in the context of the Minimum Description Length (MDL) cf. [17] or Minimum Discrimination Information (MDI), cf. [18]. MDI and MDL principles have been used often for solving computational learning problems (see [19]–[23]).

The well-known result from the statistics is the following formula for the KL divergence between two multivariate Gaussians:

$$D_{\mathrm{KL}}(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x, \tag{37}$$

for $p(x)$ and $q(x)$ given as:

$$p(x) = \mathcal{N}(x|\mu_1,\Sigma_1) \quad \text{and} \quad q(x) = \mathcal{N}(x|\mu_2,\Sigma_2), \tag{38}$$

where:

$$\mathcal{N}(x|\mu,\Sigma) = \det(2\pi\Sigma)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \tag{39}$$

and finally (see [24]):

$$D_{\mathrm{KL}}(p\|q) = \frac{1}{2}\Big[\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) - d + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\Big], \tag{40}$$

where d is the dimension of PDFs p and q.

Next, let us assume that the positively defined and symmetric Σ_1 is given, μ_1 is a zero vector, and Σ_2 and μ_2 are estimated from samples x_i, where $x_i \sim \mathcal{N}(x|\mu_1,\Sigma_1)$:

$$\mu_2 = \frac{1}{M}\sum_{i=1}^{M} x_i \quad \Sigma_2 = \frac{1}{M}\sum_{i=1}^{M}(x_i-\mu_1)(x_i-\mu_1)^T = \frac{1}{M}\sum_{i=1}^{M} x_i x_i^T, \tag{41}$$

where M is the number of samples (length of the observation). We also assume that Σ_2 is positively definite and thus well-conditioned (otherwise KL divergence would go to infinity). This assumption leads to the requirement that M has to be greater than the dimension of Σ_1 when Σ_2 is non-diagonal.

As already indicated in (36), we are looking for the expectation of the Kullback-Leibler divergence over finite sample:

$$\mathbb{E}_{x_i\sim p(x)}\left[D_{\mathrm{KL}}(p(x)\|\hat{q}(x,x_1,\cdots,x_M))\right],$$

where $p(x)$ and $q(x)$ are given in (38).

We consider two cases: the first case is when the covariance matrix is diagonal, and the second case, which is a more complicated one, is when the covariance matrix is full.

3.1.1 Diagonal Case

At first, it is worth noticing that the expectation of the Kullback-Leibler divergence does not depend on a specific form of the covariance matrix Σ_1. Let us introduce the following substitution:

$$x_i = \Sigma_1^{\frac{1}{2}} y_i \quad \text{where} \quad y_i \sim \mathcal{N}(y|0,I), \tag{42}$$

which, in turn, under the assumption given by the (41), gives:

$$\mathbb{E}\left[\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right)\right] = \mathbb{E}[\mathrm{tr}(\log\Sigma)], \tag{43}$$

where:

$$\Sigma = \frac{1}{M}\sum_{i=1}^{M}\mathrm{diag}(y_i\odot y_i), \tag{44}$$

where $\odot$ denotes the element-wise product, the Hadamard product. It is worth mentioning that we have used the following identity (see [25]):

$$\log(\det(\Sigma)) = \mathrm{tr}(\log\Sigma). \tag{45}$$

Since Σ is diagonal and y_i has a zero mean vector, we conclude,

$$\begin{aligned}\mathbb{E}[\mathrm{tr}(\log\Sigma)] = \\ \mathbb{E}\left[\sum_{k=1}^{d}\log\left(\frac{1}{M}\sum_{i=1}^{M} y_i(k)^2\right)\right] = \\ \mathbb{E}\left[\sum_{k=1}^{d}\log(\sigma_M(k)^2)\right] = \\ = d\cdot\mathbb{E}[\log(\sigma_M(1)^2)] = \\ d\left(\psi\left(\frac{M-1}{2}\right) + \log 2 - \log(M-1)\right),\end{aligned} \tag{46}$$

where $\psi(n)$ is the digamma function.

The next expression, which appears in (40), is:

$$\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) = \mathrm{tr}(\Sigma^{-1}), \tag{47}$$

where Σ is defined as in (44). After a number of transformations, we get

$$\mathbb{E}[\mathrm{tr}\Sigma^{-1}] = \mathbb{E}\left[\sum_{k=1}^{d}\frac{1}{\sigma_M^2(k)}\right] = d\cdot\mathbb{E}\left[\frac{1}{\sigma_M^2(1)}\right] = d\frac{1}{M-2}. \tag{48}$$

The last expression, which needs to be calculated, is

$$\mathbb{E}[\mu_2^T \Sigma_2^{-1} \mu_2] = \mathbb{E}\left[\sum_{k=1}^{d} \frac{\bar{x}(k)^2}{\sigma_M(k)^2}\right] = d \cdot \mathbb{E}\left[\frac{\bar{x}(1)^2}{\sigma_M(1)^2}\right] = d\frac{M-1}{M\cdot(M-2)}. \quad (49)$$

And finally, our derived formula reads,

$$\mathrm{E_{KL}}(M) = \frac{d}{2}\left[\psi\left(\frac{M-1}{2}\right) + \log 2 - \log(M-1) - 1 + \frac{M}{M-2} + \frac{M-1}{M\cdot(M-2)}\right]. \quad (50)$$

3.1.2 Non-diagonal Case

In the non-diagonal covariance matrix case, we first introduce the same substitution as in the diagonal case, which is shown in (42):

$$x_i = \Sigma_1^{\frac{1}{2}} y_i \quad \text{where} \quad y_i \sim \mathcal{N}(y|0, I). \quad (51)$$

We start by calculating the logarithmic expression, which appears in (40). By using identity in (45), and transforming the expression using (41), as previously, we get,

$$\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) = \mathrm{tr}(\log \Sigma). \quad (52)$$

To calculate the expectation of this expression, we use the expansion of the matrix logarithm into the power series, see [26]:

$$\log A = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(A-I)^k}{k}, \quad (53)$$

and, after a number of transformations, we get,

$$\mathbb{E}[\mathrm{tr}(\log \Sigma)] = \mathbb{E}[\mathrm{tr}(\log m\Sigma - \log mI)] = \mathbb{E}\left[\mathrm{tr}\left(\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}\sum_{j=0}^{k}\binom{k}{j}(-1)^j m^{k-j}\Sigma^{k-j} - \log mI\right)\right], \quad (54)$$

where m is chosen so as to $\|m\Sigma - I\| < 1$ is satisfied.

We further develop our approximation by introducing the following auxiliary function:

$$z(k, k_{max}) = \begin{cases} \sum_{i=1}^{k_{max}} (-1)^{k+1} \cdot \binom{i}{k} \cdot \frac{1}{i} & \text{when } k = 0 \\ \sum_{i=k}^{k_{max}} (-1)^{k+1} \cdot \binom{i}{k} \cdot \frac{1}{i} & \text{when } k > 0 \end{cases}, \quad (55)$$

and, consequently:

$$\mathbb{E}[\mathrm{tr}(\log \Sigma)] = \sum_{k=0}^{k_{max}} z(k, k_{max}) \cdot m^k \cdot \mathbb{E}\left[\mathrm{tr}\left(\Sigma^k\right)\right] - d \log m, \quad (56)$$

where d is a dimension of the covariance matrix. To proceed further, we calculate the expectation of $\mathrm{tr}(\Sigma^k)$. It is worth noticing that the integral

$$\int_{-\infty}^{+\infty} t^n \mathcal{N}(t|0,1)\mathrm{d}t = \frac{1}{\sqrt{2\pi}} 2^{\frac{n-1}{2}}((-1)^n + 1)\Gamma\left(\frac{n+1}{2}\right), \quad (57)$$

vanishes for odd n. At this point, we need to calculate the number of monomials with only even exponents that appear in $\mathrm{tr}(\Sigma^k)$.

Now we can describe the algorithm for calculating the number of such monomials. First, we introduce the following auxiliary matrix:

$$A = \frac{1}{\sqrt{M}}\left[y_1 \middle| \ldots \middle| y_M\right]. \quad (58)$$

One can easily notice that:

$$\Sigma = \frac{1}{M}\sum_{i=1}^{M} y_i y_i^T = A \cdot A^T, \qquad \Sigma^n = \underbrace{(AA^T)\ldots(AA^T)}_{n}, \quad (59)$$

which leads to formula for the specific element, ${\Sigma^n}_{ij}$, of the covariance matrix raised to the n-th power:

$${\Sigma^n}_{ij} = A_{i*} \cdot A^T \ldots A \cdot A^T_{*j} = \sum_{s_1=1}^{d}\sum_{k_1=1}^{M} a_{ik_1} \cdot a^{(T)}_{k_1 s_1} \ldots \sum_{s_{n-1}=1}^{d}\sum_{k_n=1}^{M} a_{s_{n-1}k_n} \cdot a^{(T)}_{k_n j} = \sum_{s_1,\ldots s_{n-1}} \sum_{k_1,\ldots,k_n} a_{ik_1} a_{s_1 k_1} a_{s_1 k_2} \ldots a_{s_{n-1}k_n} a_{jk_n}, \quad (60)$$

The number of the above configurations with only even exponents is equal to the number of even partitions of $2n$. For such a partition, we have to calculate the number of monomials separately. The detailed description of the algorithm for counting monomials is presented in section A. Implementation of the algorithm for counting monomials and computing expected KL divergence is available (see [27]). Below, we have shown examples of the formulas derived by the above-mentioned algorithm, for the power of $n = 3$, where d is a dimension of the covariance matrix and M is the number of samples used to estimate the matrix:

$$\begin{gathered} \text{Partition: 2 2 2} \\ dM(M-1)(M-2) + dM(d-1)(d-2) + 3dM(M-1)(d-1) \\ \\ \text{Partition: 4 2} \\ 3dM(M-1) + 3dM(d-1) \\ \\ \text{Partition: 6} \\ dM \end{gathered} \quad (61)$$

Once we have calculated the number of monomials for each power n, it is relatively easy to calculate the expectation in (56).

The second expression, which has to be calculated, is

$$\mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right). \quad (62)$$

We start in the same way as in the diagonal case:

$$\mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) = \mathrm{tr}\Sigma^{-1}. \quad (63)$$

The next step is to expand inversion of the covariance matrix into the Neumann series (see [28]):

$$\Sigma^{-1} = m \sum_{k=0}^{\infty} (I - m\Sigma)^k, \tag{64}$$

where m is selected so that $\|I - m\Sigma\| < 1$ is satisfied. After several transformations, we get

$$\Sigma^{-1} \approx \sum_{k=0}^{k_{max}} (-1)^k m^{k+1} \binom{k_{max}+1}{k+1} \Sigma^k, \tag{65}$$

which finally gives,

$$\mathbb{E}[\mathrm{tr}\Sigma^{-1}] \approx \sum_{k=0}^{k_{max}} (-1)^k m^{k+1} \binom{k_{max}+1}{k+1} \mathbb{E}[\mathrm{tr}\Sigma^k]. \tag{66}$$

We note that the resulting expression has a similar form to (56), with the difference of the k-dependent expression. The last expression for which the expectation needs be calculated is

$$(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1). \tag{67}$$

By using substitution (41), and subsequently (42), we get

$$\mu_2^T \Sigma_2^{-1} \mu_2 = \left(\frac{1}{M} \sum_{i=1}^{M} \Sigma_1^{\frac{1}{2}} y_i \right)^T \Sigma_2^{-1} \left(\frac{1}{M} \sum_{i=1}^{M} \Sigma_1^{\frac{1}{2}} y_i \right) = \left(\frac{1}{M} \sum_{i=1}^{M} y_i \right)^T \Sigma^{-1} \left(\frac{1}{M} \sum_{i=1}^{M} y_i \right), \tag{68}$$

then, given (58), and by expanding the inverse of the covariance matrix as in (64) into the Neumann series, we get:

$$\mu_2^T \Sigma_2^{-1} \mu_2 = \sum_{i,j} \left[\frac{1}{M^2} A^T \Sigma^{-1} A \right]_{ij} \approx \frac{1}{M^2} \sum_{i,j} \sum_{k=0}^{k_{max}} (-1)^k m^{k+1} \binom{k_{max}+1}{k+1} A_{i*}^T \Sigma^k A_{*j}. \tag{69}$$

Now, we need to calculate the following expression:

$$A_{i*}^T \Sigma^n A_{*j}. \tag{70}$$

By multiplying out we get:

$$A_{i*}^T \Sigma^n A_{*j} = \sum_{s_0=1}^{d} \sum_{s_n=1}^{d} a_{is_0}^{(T)} \left[\Sigma^n\right]_{s_0 s_n} a_{s_n j} = \sum_{s_0} \sum_{s_n} a_{s_0 i} \sum_{s_1,\ldots,s_{n-1}} \sum_{k_1,\ldots,k_n} a_{s_0 k_1} a_{s_1 k_1} a_{s_1 k_2} \ldots a_{s_{n-1} k_n} a_{s_n k_n} a_{s_n j} = = \sum_{s_1,\ldots,s_{n+1}} \sum_{k_1,\ldots,k_n} a_{s_1 i} a_{s_1 k_1} \ldots a_{s_{n+1} k_n} a_{s_{n+1} j}. \tag{71}$$

An important observation is in place, namely, the monomials cannot have only even exponents if $i \neq j$. This allows us to bring down the calculation of the expectation of the above expression to the expectation of the trace of Σ^{n+1}. Thus, finally, we get

$$\mathbb{E}[\mu_2^T \Sigma_2^{-1} \mu_2] \approx \frac{1}{M} \sum_{k=0}^{k_{max}} (-1)^k m^{k+1} \binom{k_{max}+1}{k+1} \mathbb{E}[\mathrm{tr}\Sigma^{k+1}]. \tag{72}$$

Obviously, the algorithm for calculating the above expression is the same as in the case of the two previously examined terms.

3.2 Procedure

Here, we show how to estimate the GMM using the modified training procedure. The introduced procedure solves the problem of overtraining. First, we focus on training the GMM with fixed K - the number of components in the mixture. Based on this development, we formulate a method to select the optimal number of components for a given training set without resorting to the development sets. The classical Expectation-Maximization algorithm, cf. [2], produces in each iteration more and more accurate estimates of the mean vectors, covariance matrices, and weights. We denote the trajectory of these parameters in subsequent iterations as,

$$\theta = \left[\theta^{(1)}, \cdots, \theta^{(J)}\right], \tag{73}$$

where J is the maximum acceptable number of iterations (or the number of iterations until convergence), and

$$\theta^{(j)} = \left[\Sigma_1^{(j)}, \cdots, \Sigma_K^{(j)}, \mu_1^{(j)}, \cdots, \mu_K^{(j)}\right], \tag{74}$$

with j the iteration number, Σ_i denoting covariance matrix for the $i-$th component, μ_i denoting the mean vectors for the $i-$th component, and K the number of components in the Gaussian mixture. Moreover, we denote $\theta_k^{(j)} = [\Sigma_k^{(j)}, \mu_k^{(j)}]$. To proceed further, we introduce and define some mathematical entities:

- the *entropy* in nats of the categorical distribution $\{r_i\}_i$ is defined as $H(\{r_i\}) = -\sum_i r_i \log r_i$
- the *kernel width* is the effective number of samples used to estimate a given component. Given the samples $\{x_i\}_{i=1}^M$, we compute the *kernel width*, $w(\theta, \{x_i\})$, also abbreviated by $w(\theta)$, using the following set of formulas:

$$r_i = \frac{\mathcal{N}(x_i|\theta)}{\sum_l \mathcal{N}(x_l|\theta)}, \tag{75}$$

$$w(\theta) = \exp\left(H(\{r_i\})\right), \tag{76}$$

- Gaussian *component differential entropy* in nats, $H(\theta)$, will be defined as:

$$H(\theta) = \frac{1}{2} \log\left(\det(2\pi e \Sigma)\right), \tag{77}$$

- the *score* or *cross entropy* for the component is expressed by the following formula:

$$s(\theta) = H(\theta) + \mathrm{E}_{\mathrm{KL}}\left(w(\theta)\right). \tag{78}$$

We see that the expression for *cross entropy* depends only on the observed quantities.

Equipped with the above definitions, we can formulate the algorithm,

Algorithm 1: train GMM

Require: starting parameters values $\theta^{(0)}$, maximal number of iterations J, initial scores $s\left(\theta_k^{(0)}\right) = \infty$, set all weights to $\rho_k = \frac{1}{K}$.
Ensure: $\theta^{(J)}$ and weights $\{\rho_k\}$.
1: **for** $j \in [1, \ldots, J]$ **do**
2: compute $\theta^{(j)}$ update according to EM (keeping weights unchanged)
3: **for** $k \in [1, \ldots, K]$ **do**
4: **if** $s\left(\theta_k^{(j)}\right) > s\left(\theta_k^{(j-1)}\right)$ **then**
5: $\theta_k^{(j)} = \theta_k^{(j-1)}$ (backtracking)
6: **end if**
7: **end for**
8: **end for**
9: run few EM updates of weights alone
10: **return** $\theta^{(J)}$ and weights $\{\rho_k\}$

Algorithm 2: train GMM

Require: starting parameters values $\theta^{(0)}$, maximal number of iterations J, initial scores $s\left(\theta_k^{(0)}\right) = \infty$, set all weights to $\rho_k = \frac{1}{K}$.
Ensure: $\theta^{(J)}$ and weights $\{\rho_k\}$.
1: **for** $j \in [1, \ldots, J]$ **do**
2: compute $\theta^{(j)}$ update according to EM (keeping weights unchanged)
3: **for** $k \in [1, \ldots, K]$ **do**
4: **if** $s\left(\theta_k^{(j)}\right) > s\left(\theta_k^{(j-1)}\right)$ **then**
5: $\theta_k^{(j)} = \theta_k^{(j-1)}$ (backtracking)
6: **end if**
7: **end for**
8: **end for**
9: ~~run few EM updates of weights alone~~
10: **return** $\theta^{(J)}$ and weights $\{\rho_k\}$

The mechanism behind the algorithm 1 is as follows:

- if the number of components is large in comparison to the number of training vectors, the precision of each component goes up with each iteration until it eventually reaches infinity; this means at the same time, $H(\theta)$ goes toward $-\infty$ (in practice, it will stop at a low value due to the constraint we put on the minimal eigenvalue of the covariance matrix)
- the second term in the expression for *score*, see 78, to the contrary, grows toward $+\infty$ as *kernel width* approaches dimension of the training vectors d. Thus, even if *componenent differential entropy* goes toward $-\infty$, the *cross entropy* has a minimal extreme point and at the very last grows toward $+\infty$.
- thus, the statements from 3-7 in the algorithm 1 prevent the collapse of the components to something resembling a dirac delta
- in effect, the algorithm does not overtrain

We also examine the algorithm 2, which does not run a few EM updates of weights (statement 9 in algorithm 1).

3.2.1 Number of components selection criterion

The *cross entropy* in 78 is the expression on number of nats needed to code from the component. Based on the *cross entropy*, we can get the number of bits needed to code with given fidelity using the Shannon-Lower-Bound [29].

Given the component *weights* $\{\rho_k\}_{k=1}^K$, the average number of nats needed to code from the GMM modeled source PDF is equal to

$$C_1\left(K, \theta^{(J)}, \{\rho_k\}\right) = C_1(K) = \underbrace{H\left(\{\rho_k\}_{k=1}^K\right)}_{\text{components indices entropy}} + \underbrace{\sum_{k=1}^{K} \rho_k s\left(\theta_k^{(J)}\right)}_{\text{mean per component cross entropy}} . \quad (79)$$

We also consider a second criterion, which seems to give meaningful results:

$$C_2\left(K, \theta^{(J)}, \{\rho_k\}\right) = C_2(K) = \underbrace{\sum_{k=1}^{K} \rho_k s\left(\theta_k^{(J)}\right)}_{\text{mean per component cross entropy}} . \quad (80)$$

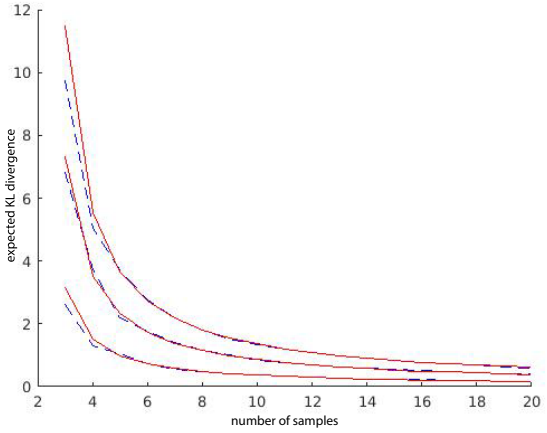


Figure 3: Comparison of the empirically and analytically obtained expectation of the KL divergence for Gaussians with the diagonal covariance matrices. Dashed lines represent the Monte Carlo result. Dimensions of covariance matrices from the beginning of the coordinate axes: 3, 7, and 11, respectively.

To find the optimal number of components K (which minimizes the average number of nats needed to code from the source), we proceed as follows:

- we swap the K from 1 to M
- for each K we run the algorithm 1 or 2.

- we smooth with smoothing splines, see [30], the resulting curve $C_*(K)$, obtaining a smoothed curve $\tilde{C}_*(K)$
- we select $\hat{K}$ for which $\tilde{C}_*(\hat{K})$ attains minimum and returns the optimal number of components.

Above, the asterisk means either 1 or 2.

3.3 Numerical experiments

3.3.1 Expected Kullback-Leibler Divergence

In this section, the comparison our analytical results with the Monte-Carlo obtained curves of expected Kullback-Leibler divergence will be presented. First, the results for the diagonal covariance matrix will be shown in Figure 3.

As can be seen, the analytical expectation shows high accuracy for the diagonal covariance matrices.

The results are a bit worse for the non-diagonal matrix. The following issues affected the accuracy of the result:

- the calculation of the higher degree series expansions was too hard as the computation of the number of even exponent monomials grew exponentially with n. We were able to compute the number of monomials for n as high as eight (the formulas for $n = 8$ when written on A4 page taking 44000 rows).
- The Monte-Carlo curves on the left are sensitive to the threshold for the detection of the semi-definiteness of the matrices. It sometimes happens that one of the eigenvalues of the covariance matrix is very small. Then value of term $\mathrm{tr}\Sigma^{-1}$ grows enormously. It is clear that the result depends on the choice of the threshold.

Figure 4 depicts the results for non-diagonal covariance matrices. To obtain the plot, we rejected matrices with the smallest eigenvalue less than 0.01 (which is around $\frac{1}{100}$ of the largest eigenvalue).

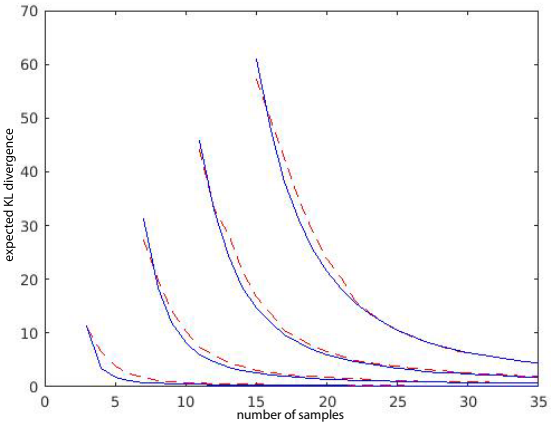


Figure 4: Comparison of empirically and analytically obtained expectation of the KL divergence for Gaussians with the non-diagonal covariance matrices. Dashed lines represent the Monte Carlo estimation. Dimensions of covariance matrices from the beginning of the coordinate axes: 3, 7, 11, and 15, respectively.

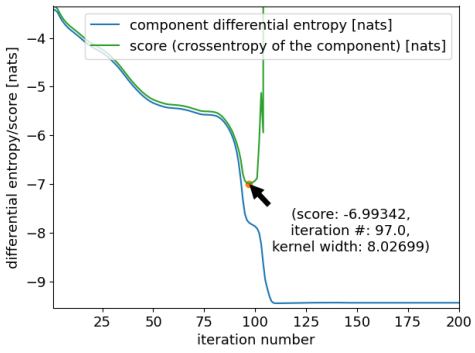


Figure 5: Illustration of the overtraining prevention mechanism. Plot shows that score prevents the overtraining, differential entropy of the Gausssian component goes down and the score starts increasing after kernel width exceeds 8 training points

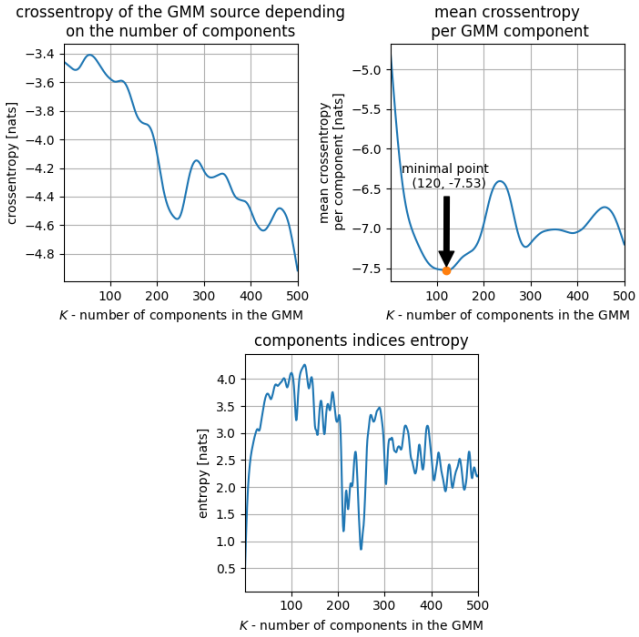


Figure 6: Results of the algorithm 1 for $d = 2$

3.3.2 Overtraining Prevention Mechanism

Here, we present a drawing (Figure 5) showing the trajectory of the *score* and *component differential entropy*. The plot has been obtained by by recording the parameters for subsequent iterations for one chosen component of the GMM. We see that the *score* attains minimum but *component differential entropy* goes to $-\infty$. This allows to use the backtracking procedure present in algorithms 1 and 2.

3.3.3 Numerical Analysis of the Number of Components Selection Criteria

All numerical experiments of this section has been carried out using the Line Spectral Frequencies (LSF) as the data to which the GMMs are fitted. The LSFs were computed from the LibriSpeech database. In all subsequent experiments the training set length M equals 1000.

First, the workings of the algorithm 1 will be presented. The Figure 6 shows the criteria C_1 and C_2 and the *components indices entropy* on pictures from left to right and from top to bottom, respectively. We see erratic behaviour of the *components indices entropy* that falls down, meaning it silences some components by driving some weights to zero. The criterion C_1 suggests that the optimal number of components is > 500, what is unlikely and probably caused by silencing of some components by zero weights. The C_2 criterion, instead, gives a sensible result. We see that for algorithm 2, selected by the C_2 criterion, number of components is similar.

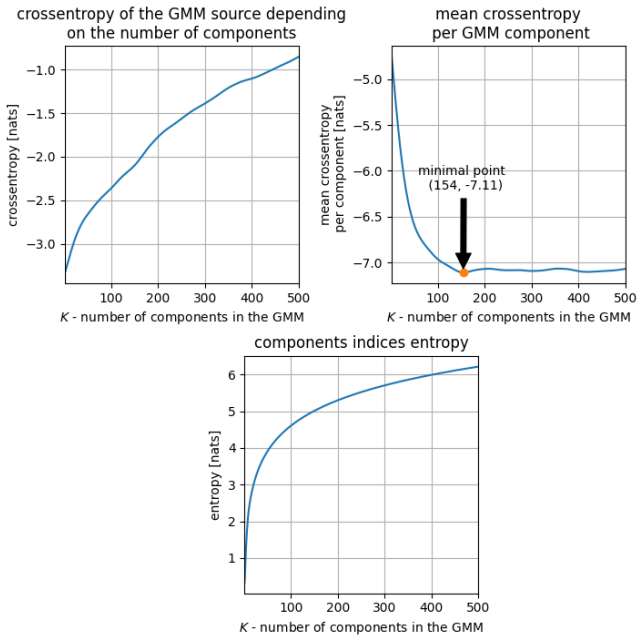


Figure 7: Results of the algorithm 2 for $d = 2$

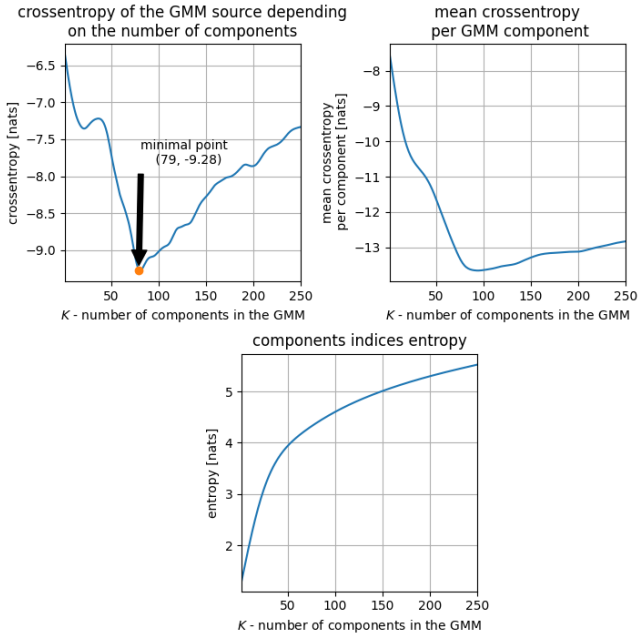


Figure 8: Results of the algorithm 2 for $d = 4$

As the second example, we show the workings of the algorithm 2. The plots in Figure 7 show the same curves as in the case of the Figure 6 (see previous paragraph). The C_1 criterion points as the optimal number of components exactly one component, that is, a single Gaussian. Above $K = 1$, the *components indices entropy* penalizes the *score* to the extent that the score curve is monotonically increasing. Another optimal number of components is suggested by C_2. We computed the GMMs for the number of components selected using the C_2 criterion, for which the contour plots are shown on Figure 10. We see that the GMM is tightly fitted to the data for the selection criterion C_2. However, the composite criterion C_1 suggests a single component, which seems to by trivial, but still optimal, as indicated by the methodology proposed.

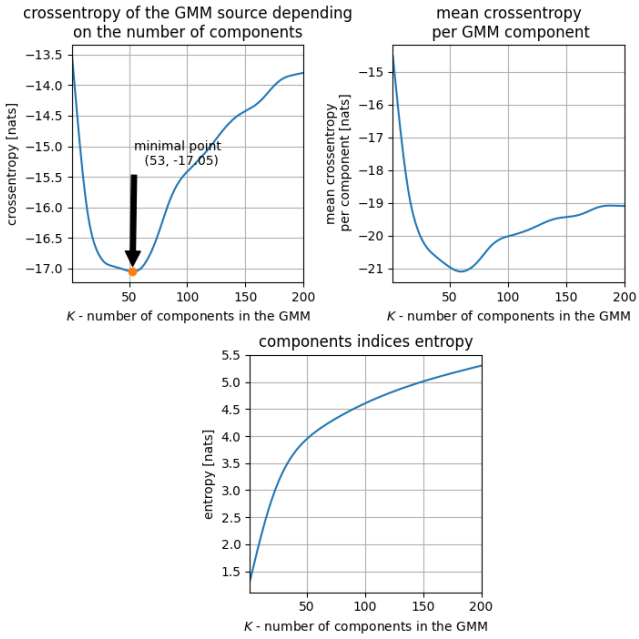


Figure 9: Results of the algorithm 2 for $d = 8$

As the third example, we show the working of the algorithm 2 for the problem dimension $d = 4$ (first four line spectral frequencies). In this case, the criterion C_1 returns trustworthy results, the minimum on the C_1 curve is clearly noticeable. The returned result is sensible and indicates that for $d = 4$, the training vectors distribution needs more components to be modeled accurately; the distribution is probably far from Gaussian. The curves are presented in Figure 8. This experiment is an indication that the proposed methodology is sound. To make the evidence even stronger we present more experiments for dimensions $d = 8$, Figure 9, and $d = 16$, Figure 12. As expected the optimal number of components decreases with the dimension of the problem. This is to maintain proper generalization.

The last result shows the behavior of the classical number of components selection criterion, that is the maximization of the likelihood on the development set. For this experiment, from 1000 samples of the training set, we excluded 100 samples as the development set. In Figure 13, we show the plot of the log-likelihood on the development set as a function of the number of components. It is evident that the plot is quite erratic of what decreases trust in this method of selection of the number of components. The results can

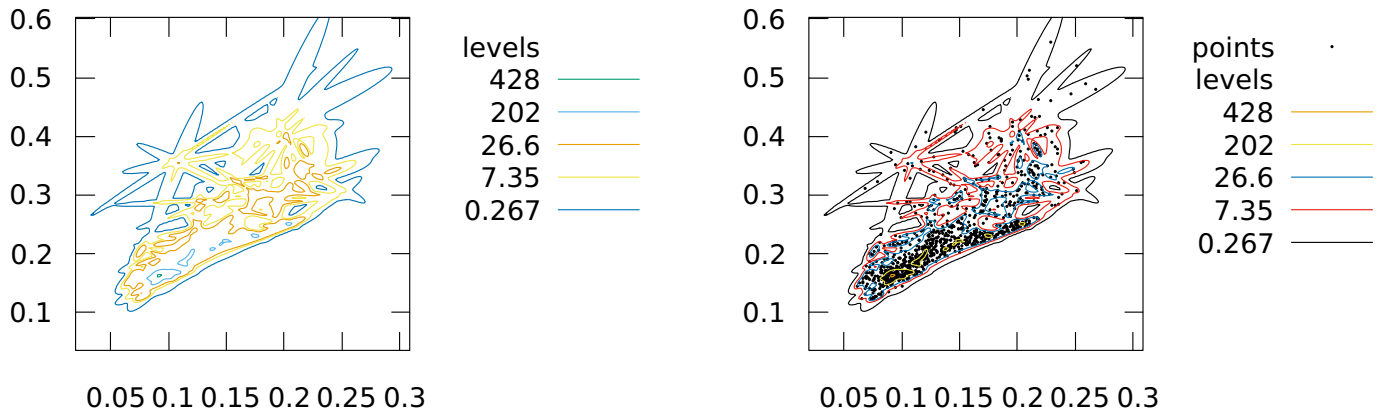


Figure 10: GMM fitted using algorithm 2 with number of components selected using the criterion C_2

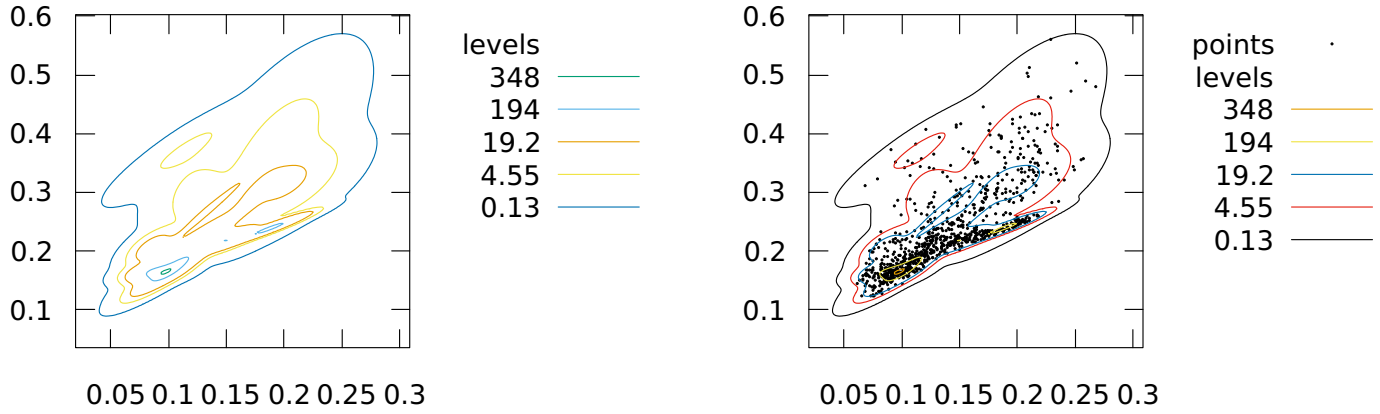


Figure 11: GMM fitted using traditional EM algorithm with number of components selected using tuning on the development set

be made more stable if we allow the development set to be larger. Such a measure will come at the cost of reducing the number of training points. Here, our proposed method shows the advantage of not using development sets. The contour plots for the GMM fitted using traditional method with the number of components tuned on the development set are shown in Figure 11.

4 Conclusions

In this paper, the theory of both discrete and continuous distribution with unknown number of components has been developed. The measure theoretic reason for the inherent similarity between PMFs and PDFs, with the difference in the structure of the underlying sample space, has been given. The difference causes the two distribution categories to be treated with much different mathematical tools. The main result of the paper are the means and methods of computing the number of components, that is, the number of categories in the PMF case and the number of Gaussian components in the GMM modeling in a PDF case, which is given implicitly in terms of its realizations.

Development of the theory required considering the expected Kullback-Leibler divergence - a difficult problem on its own rights. Especially, the algorithm for counting monomials with only even exponents in the expression for a trace of a matrix raised to a power has been developed. This algorithm allows to compute certain integrals analytically without resorting to the Monte-Carlo experiments. This algorithm can also be of interest on its own rights.

We believe that the theory presented in this paper will find many practical applications in diverse fields of science, technology, and engineering as a convenient tool of data analysis.

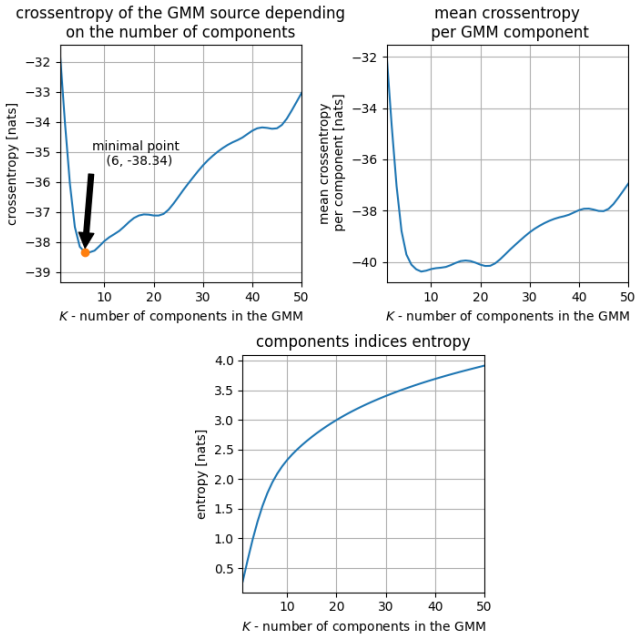


Figure 12: Results of the algorithm 2 for $d = 16$

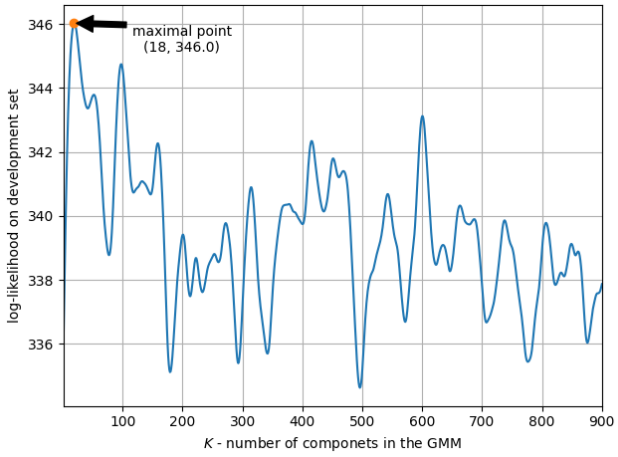


Figure 13: Results of the traditional EM algorithm $d = 2$

References

[1] M. Kuropatwiński, "Estimation of Quantities Related to the Multinomial Distribution with Unknown Number of Categories," in 2019 Signal Processing Symposium (SPSympo), 277–281, 2019, doi:10.1109/sps.2019.8881992.

[2] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[3] R. M. Gray, Source coding theory, Springer, 2012.

[4] R. M. Gray, "Gauss Mixture Vector Quantization," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 1769–1772, 2001, doi:10.1109/icassp.2001.941283.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey et al., The HTK book, Cambridge University Engineering Department, 2002.

[6] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, **10**(1), 19–41, 2000, doi:10.1006/dspr.1999.0361.

[7] H. Permuter, J. Francos, I. Jermyn, "A study of Gaussian mixture models of color and texture features for image classification and segmentation," Pattern Recognition, **39**(4), 695–706, 2006, doi:10.1016/j.patcog.2005.10.028.

[8] A. Kundu, S. Chatterjee, A. S. Murthy, T. Sreenivas, "GMM Based Bayesian Approach To Speech Enhancement in Signal/Transform Domain," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 4893–4896, 2008, doi:10.1109/icassp.2008.4518754.

[9] M. Nilsson, H. Gustaftson, S. V. Andersen, W. B. Kleijn, "Gaussian Mixture Model Based Mutual Information Estimation Between Frequency Bands in Speech," in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 525–528, 2002, doi:10.1109/icassp.2002.1005792.

[10] B. W. Silverman, Density estimation for statistics and data analysis, CRC Press, 1986.

[11] L. Devroye, G. Lugosi, Combinatorial methods in density estimation, Springer, 2012.

[12] R. M. Gray, Probability, random processes, and ergodic properties, Springer, 1988.

[13] P. S. Laplace, Philosophical essay on probabilities, Springer, 1998.

[14] P. Flajolet, D. Gardy, L. Thimonier, "Birthday paradox, coupon collectors, caching algorithms and self-organizing search," Discrete Applied Mathematics, **39**, 207–229, 1992, doi:10.1016/0166-218x(92)90177-c.

[15] S. Gradshteyn, I. Ryzhik, Table of integrals, series, and products, Academic Press, 2014.

[16] V. Baldoni, N. Berline, J. De Loera, M. Köppe, M. Vergne, "How to integrate a polynomial over a simplex," Mathematics of Computation, **80**(273), 297–325, 2011, doi:10.1090/s0025-5718-2010-02378-6.

[17] J. Rissanen, "Modeling by shortest data description," Automatica, **14**(5), 465–471, 1978, doi:10.1016/0005-1098(78)90005-5.

[18] S. Kullback, Information theory and statistics, Courier Corporation, 1997.

[19] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, C. J. Taylor, "A minimum description length approach to statistical shape modeling," IEEE Transactions on Medical Imaging, **21**(5), 525–537, 2002, doi:10.1109/tmi.2002.1009388.

[20] R. M. Gray, J. C. Young, A. K. Aiyer, "Minimum Discrimination Information Clustering: Modeling and Quantization with Gauss Mixtures," in 2001 International Conference on Image Processing, 14–17, 2001, doi:10.1109/icip.2001.958039.

[21] M. H. Hansen, B. Yu, "Model selection and the principle of minimum description length," Journal of the American Statistical Association, **96**(454), 746–774, 2001, doi:10.1198/016214501753168398.

[22] Y. Ephraim, A. Dembo, L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," IEEE Transactions on Information Theory, **35**(5), 1001–1013, 1989, doi:10.1109/18.42209.

[23] R. M. Gray, A. Gray, G. Rebolledo, J. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," IEEE Transactions on Information Theory, **27**(6), 708–721, 1981, doi:10.1109/tit.1981.1056410.

[24] J. Duchi, "Derivations for Linear Algebra and Optimization," Technical report, University of California at Berkeley, 2007.

[25] C. S. Withers, S. Nadarajah, "log det A = tr log A," International Journal of Mathematical Education in Science and Technology, **41**(8), 1121–1124, 2010, doi:10.1080/0020739x.2010.500700.

[26] B. Hall, Lie groups, Lie algebras, and representations: An elementary introduction, Springer, 2015.

[27] L. Sikorski, M. Kuropatwiński, "Counting Monomials Algorithm," `https://www.codeocean.com/`, 2019, doi:10.24433/CO.3748956.v1.

[28] D. Zhu, B. Li, P. Liang, "On the Matrix Inversion Approximation Based on Neumann Series in Massive MIMO Systems," in IEEE International Conference on Communications (ICC), 1763–1769, 2015, doi:10.1109/icc.2015.7248580.

[29] T. Berger, Rate distortion theory: A mathematical basis for data compression, Prentice-Hall, 1971.

[30] C. De Boor, A practical guide to splines, Springer, 1978.

A The algorithm for counting monomials

The solution of the problem from section 3.1.2 is divided into six short algorithms

Algorithm 3: Counting monomials with only even exponents for given even partition of number n

Require: Even partition $\mathcal{P} = \{p_1, ..., p_k\}$ of the number n, problem dimension d,number of random samples m
Ensure: Number of monomials with only even exponents
1: $S := 0$
2: **for** every multipermutation of $\mathcal{P}$ **do**
3: Create $\mathcal{T} = \{\underbrace{1, ..., 1}_{p_1}, ..., \underbrace{k, ..., k}_{p_k}\}$
4: **for** every multipermutation of $\mathcal{T}$ **do**
5: Check if $\mathcal{T}$ is correct with Algorithm (7).
If Algorithm (7) returned false, go to the next multipermutation of $\mathcal{P}$.
6: Create matrices D and M for the table $\mathcal{T}$ with Algorithm (4).
7: If there are unfilled places in matrices D and M, create pairs of matrices D' and M' by filling unfilled places in D and M in every possible way.
8: From the set of every generated in previous step pairs of matrices, discard these which are incorrect. Correctness of the pair of matrices is checked with Algorithm (5).
9: For every pair of matrices calculate it's numerical value with Algorithm (6) and add it to S.
10: **end for**
11: **end for**
12: **return** S

Algorithm 4: Creating matrices D and M

Require: $\mathcal{T} = \{t_1, ..., t_n\}$, where $t_i \in \{1, ..., k\}$
Ensure: Filled matrices D and M
1: Create matrices D and M of a dimension k
2: **for** $i = 1$ to n **do**
3: **if** i is odd **then**
4: To $M_{t_i t_{i+1}}$ and $M_{t_{i+1} t_i}$ write "=".
5: To $D_{t_i t_{i+1}}$ and $D_{t_{i+1} t_i}$ write "$\neq$".
6: **end if**
7: **if** i is even **then**
8: To $D_{t_i t_{i+1}}$ and $D_{t_{i+1} t_i}$ write "=".
9: To $M_{t_i t_{i+1}}$ and $M_{t_{i+1} t_i}$ write "$\neq$".
10: **end if**
11: **end for**
12: **return** matrices D and M
Note: If there is an occurrence of writing "$\neq$" or "=" in previously filled place, stop the algorithm and return nothing.

Algorithm 5: Checking if given pair of matrices D and M is correct

Require: Matrices D and M
Ensure: *True*, if given pair is correct; *False* in other cases
1: **for** every pair of indexes i, j of the matrix **do**
2: **if** $D_{ij} = M_{ij} =$ "=" **then**
3: **return** *False*
4: **end if**
5: **if** the negation of any from implications listed below is true:
$[(X_{ij} =$ "=") $\wedge$ $(X_{ik} =$ "=")$] \Rightarrow X_{kj} =$ "=" or $[(X_{ij} =$ "=") $\wedge$ $(X_{ik} =$ "$\neq$")$] \Rightarrow X_{kj} =$ "$\neq$"
where X is a matrix D or M **then**
6: **return** *False*
7: **end if**
8: **end for**
9: **return** *True*

Algorithm 6: Calculating a component of sum from formula for number of monomials

Require: Pair of matrices D and M, problem dimension d, number of random samples m
Ensure: A component of sum from formula for number of monomials
1: $q = 1$
2: **for** $i = 1$ to dim D **do**
3: **if** $i = 1$ **then**
4: $q = q \cdot d \cdot m$
5: **else**
6: Calculate c_d oraz c_m with Algorithm (8) with input data: (D, i) and (M, i)
7: **if** there exists $i < j$, such that $D_{ij} =$ "$\neq$" **then**
8: $q = q \cdot (d - c_d)$
9: **end if**
10: **if** there exists $i < j$, such that $M_{ij} =$ "$\neq$" **then**
11: $q = q \cdot (m - c_m)$
12: **end if**
13: **end if**
14: **end for**
15: **return** q

Algorithm 7: Checking if table $\mathcal{T}$ is correct

Require: $\mathcal{T} = \{t_1, ..., t_n\}$, where $t_i \in \{1, ..., k\}$
Ensure: *True*, if $\mathcal{T}$ is correct; *False* in other cases
1: $A = \emptyset$
2: **for** $i = 1$ to n **do**
3: **if** there exists $s < i$, such that $t_s > t_i$ and $t_i \notin A$ **then**
4: **return** *False*
5: **end if**
6: insert t_i into A
7: **end for**
8: **return** *True*

Algorithm 8: Calculating auxiliary minuend

Require: Matrix X, number i
Ensure: Auxiliary minuend
1: **if** $i = 1$ **then**
2: **return** 0
3: **end if**
4: **for** each $j < i$ **do**
5: **if** If $X_{ij} =$ "$\neq$" **then**
6: Calculate c_j with Algorithm (6) with input data (X, j)
7: **end if**
8: **end for**
9: $c = \min_j c_j$
10: **return** $c + 1$

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1299-1307 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

C-Band FMCW Radar Design and Implementation for Breathing Rate Estimation

Mohammad Mohammad Abdul-Atty[1], Ahmed Sayed Ismail Amar[1,*], Mohamed Mabrouk[2]

[1]*Department of Electronics and Communications Engineering, Ain Shams University, Cairo, 18655, Egypt*

[2]*Electrical Engineering Department, Alexandria University, Alexandria, 22301, Egypt*

ARTICLE INFO

Article history:
Received: 31 July, 2020
Accepted: 28 September, 2020
Online: 26 October, 2020

Keywords:
FMCW
Radar
Human
Breathing
Detection

ABSTRACT

In this paper, a portable Frequency Modulated Continuous Wave (FMCW) radar system was designed and implemented for human movements and breathing detection. The radar operates with a frequency band ranges from 4.7 to 4.9GHz. The radar sub-systems were designed and simulated using up to date computer-aided-design tools before implementation. The Voltage Controlled Oscillators (VCO), high gain antenna, low loss power divider/combiner, and a high selectivity bandpass filter were implemented, and their parameters were measured using a microwave analyzer. The simulated results and the measured results show a significant correlation. The new RF front end module components enhanced the radar signal-noise-ratio (SNR), and the breathing rate detection. The Digital Signal Processing (DSP) is implemented on the Field Programmable Gate Array (FPGA) board for movement and breathing signals detection processing in real-time. Highly sensitive detection, configurability, low cost, low power consumption, and portability were considered in the designed system. We believe that the enhanced reconfigurable radar system will be helpful in several biomedical monitoring applications.

1. Introduction

Human breathing estimation is very important in security, medical, and homecare applications [1]. In security applications, special military troops use the human breathing detection devices to detect human behind walls in case of building enforcement operations. In medical applications, vital signs monitoring was proposed in US patent in 1990 [2]. After that, human breathing and heartbeat estimation instruments for medical application were introduced since 1993 [3]. Hospitals use human breathing radars to estimate the breathing especially for burned skin or COVID-19 patients where it is preferred not to contact the medical crew. In homecare applications, the population of Americans edged more than 65 years will grow to 90 million people in 2050 [4]. These old peoples who lives alone need to be monitored using breathing and heartbeat detection radars to detect sudden medical problems and report them to an emergency medical center.

Radar systems may be designed to be pulsed or continuous wave transmitters depending on the required application. Frequency modulated continuous wave (FMCW) radars outperforms the pulsed radars in transmitting higher average power with its capability of measuring the target range, especially for small cross section area targets. Because of that, FMCW radars were used in human breathing and heartbeat estimation and human movements detection applications [5]. In [6], a 24.15 GHz FMCW radar with a scanning bandwidth of 72 MHz was used to extract the human heartbeat rate signal. In [7], a 24 GHz FMCW radar was used for heartbeat rate estimation. The clutter noise was canceled but the effect of human limbs motion decreased the human breathing estimation rate accuracy. According to the selection of the operating frequency band, the human chest cross section area has its highest value in the frequency band 3-4 GHz and the received signal-to-noise ratio is high in the frequency band 4-6 GHz [8], [9] .However, many FMCW radars for human breathing rate estimation were designed and implemented in other frequency bands such as 2.4 GHz, 9.6GHz, 24GHz, and 77GHz [7], [10]-[13]. The harmonics of the human breathing frequency may interfere with the heartbeat rate that result in heartrate estimation failure [10]. One of the most effective methods is to use comb filters to remove these harmonics [14]. For C-band biomedical radars, the state-of-the-art presented the radar systems without considering configurability, cost, and design reliability which is covered in this paper.

*Corresponding Author: Ahmed Sayed Ismail Amar, 434 king Faisal street, Cell No. +0201065664849 & ahmed.s.i.amar@ieee.org

Most uses of radar can be classified as detection, tracking, or imaging. In this research, the scope is on radar detection. The most fundamental problem in radar is detection of an object or physical phenomenon. This requires determining whether the receiver frontend output at a given time represents the echo from a reflecting object or only noise. Detection process starts with using the techniques of signal acquisition and interference reduction necessary to perform these tasks. These techniques were discussed briefly in [15] and the signal is then processed for a decision. Detection decisions are usually made by comparing the amplitude of the signal to a threshold which may be set a priori in the radar design, which is not preferred because noise changes over time, or may be computed adaptively from the radar data.

Digital signal processing on the radar returns can be implemented using main task functions such as: Moving Target Indicators (MTI), Constant False Alarm Rate (CFAR), digital integrators, and target information packet extractors [16]. These main functions can be set to be configurable. For example, CFAR techniques have been developed in the last few decades such as order statistics CFAR (OS-CFAR) [17], cell average CFAR (CA-CFAR) [18], greatest of CFAR (GO-CFAR) [19], and smallest of CFAR (SO-CFAR) [20]. Each one of these techniques is suitable for specific environmental conditions. The DSP can be design to select the desirable technique by the user.

In this paper, a configurable C-band FMCW radar is designed and implement for human breathing detection. Section 2 presents the principal of operation. Section 3 present the Radio Frequency (RF) components design and implementation. Section 4 is the measurements results, section 5 presents the conclusion, and section 6 discusses the future work .

2. Principles of Operation

The FMCW radar range resolution depends on the transmitted signal bandwidth while the maximum range can be calculated from the chirp time interval. (Figure 1.a) shows the linear variation of the transmitted frequency from f_start to f_stop where the bandwidth is equal to $f_{stop} - f_{start}$. The chirp time is T_chirp, that the maximum range is equal to $C \times T_{chirp}/2$, where C is the light propagation velocity. The transmitted signal $x_t(t)$ can be presented by:

$$x_t(t) = A_T \cos(2\pi f_{start}.t + \pi \frac{BW}{T_{chirp}} t^2 + \varphi(t)) \quad (1)$$

where, A_T is the amplitude of the transmitted signal, BW is the bandwidth, and $\varphi(t)$ is the transmitter phase noise. When a transmitted signal hit a target, it reflects with attenuated amplitude and time shift related to the target range. The received signal $x_t(t)$ can be presented by:

$$x_r = A_R \cos(2\pi f_{start}(t - \tau_d) + \pi \frac{BW}{T_{chirp}}(t - \tau_d)^2 + \varphi(t - \tau_d)) \quad (2)$$

where, A_R is the attenuated received signal amplitude, and τ_d is the round-trip delay from the radar antenna to the target and back again to the radar, and is given by:

$$\tau_d = \frac{2R}{c} \quad (3)$$

where, R is the target range. When the received signal is mixed with the transmitted one, the output can be presented by:

$$y(t) = \frac{1}{2} * A_T * A_R \cos[(2\pi f_{start} t + \pi \frac{BW}{T_{chirp}} t^2 + \varphi(t) - 2\pi f_{start}(t - \tau_d) - \pi \frac{BW}{T_{chirp}}(t - \tau_d)^2 - \varphi(t - \tau_d))] \quad (4)$$

$$y(t) = A \cos(2\pi f_{start}\tau_d + 2\pi \left(\frac{BW}{T_{chirp}}\tau_d\right) t - \pi \frac{BW}{T_c}\tau_d^2 + \Delta\varphi(t)) \quad (5)$$

After simplifying $y(t)$, it can be expressed as,

$$y(t) = A \cos[(2\pi f_b t + \varphi_b(t) + \Delta\varphi(t))] \quad (6)$$

where, A it the amplitude of the received signal (t) , f_b is the beat frequency and it is equal to $\frac{BW}{T_{chirp}}\tau_d$, $\varphi_b(t)$ is the received signal phase and it equals to $2\pi f_c \tau_d - \pi \frac{BW}{T_c}\tau_d^2$, $\Delta\varphi(t)$ is the phase noise and it equals to $\varphi(t) - \varphi(t - \tau_d)$. When the target is moving with a specific speed, Doppler frequency adds to the beat frequency and becomes $f_b + f_D$, where f_D is the Doppler frequency. (Figure 1.b), shows the phase difference between the two mixed signals ($f_b + f_D$).

The phase noise is not considered in this research. The term $\pi\,(BW/T_{chirp})\,\tau_d^2$ is very small compared to f_b and can be neglected. Then, the beat signal which is equal to $2\pi f_{start}\tau_d + \frac{BW}{T_{chirp}}\tau_d$, is processed to extract the target parameters.

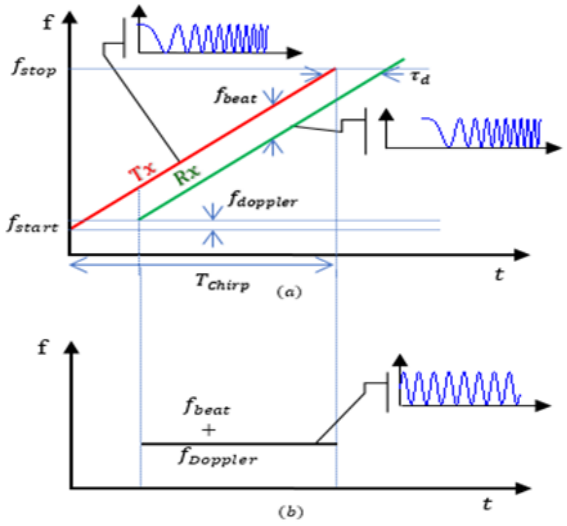


Figure 1: Sawtooth based FMCW radar basic concept: (a) Transmitted and received signal in the frequency-time domain, and (b) single target beat signal

3. System Design

In this section, the design of a C-band FMCW radar for human breathing rate estimation is presented. The system is implemented using RF components and configurable digital signal processing subsystem.

3.1. General Overview of the Proposed Radar

(Figure 2) shows the block diagram of the proposed radar. A single channel architecture is utilized in this design, in order to reduce hardware and memory resources, computation complexity, hence; a cost effective design can be achieved

A function generator board is used to control the Voltage Controlled Oscillator (VCO) to generate a chirp signal with a frequency ranged from 4.7-4.9 GHz. The VCO output is divided into two different ports using 3dB equally power divider. The output of the first port of the divider is connected to a bandpass filter (BPF) from 4 GHz to 5 GHz. The filter output then is sent to the power amplifier with a 20dB gain, and transmitted through a horn antenna with 16.5dB gain. The receive antenna which has the same gain of the transmitter antenna, receives the signal, then a 24dB Low Noise Amplifier (LNA) is used to amplify this signal, then the signal is sent to a BPF operates from 4 GHz to 5 GHz, then this signal is mixed with a replica of the transmitted signal coming from the second port of the power divider, to perform the frequency down conversion. A Low Pass Filter (LPF) then is used to filter the mixer output signal from the unwanted signals. An 8-bit Analog-to-Digital Converter (ADC) is used to digitize the analog signal, then the digital data is passed to the FPGA for signal processing. A PC is connected to the FPGA output for further processing, and a screen is connected directly to the FPGA to show the targets and the target history.

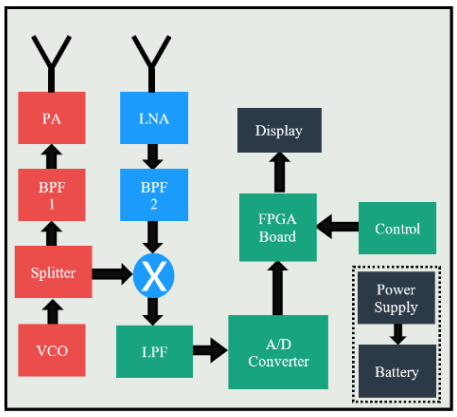


Figure 2: Proposed Radar general block diagram.



Figure 3: Fabricated radar system.

3.2. Radar System Parameters

A 75 cm range resolution is achieved by using 200 MHz sweep bandwidth from the VCO.

The following equation is used to determine the range resolution.

$$R_{res} = \frac{c}{2*BW} \quad (7)$$

where C is the light speed, and BW is the proposed bandwidth. (Figure 3) shows the fabricated radar system, which considers the radar components fixation.

In this design, a chirp duration T_c of 1 ms is chosen. Table 1 shows a listed ssystem parameters. A sampling rate of 2.048 MHz is used for sampling the signal with a band of frequencies that includes the targeted f_b. The beat frequeny can be calculated using equation (6), by measuring the delay time between the transmitted signal and the received signal, the mixer can do this job by mixing these two signals, then the range can be measured. (Figure 4) shows the collected radar data: the x-axis is the fast time direction, each column represents one range cell, while y-axis is the slow time direction, each row represents one complete received chirp or one Coherent Processing Interval (CPI). To get the target's range, then Fast Fourier Transform (FFT) can be performed in the fast time direction, while performing FFT in the slow time direction gets the target's doppler information. (Figure 4) shows the target's movement in many range cells, while another stationary target exists in the same range cell across the slow time direction.

Table 1: Radar system parameters

Parameter	Value
Output Power	16 dBm
Center Frequency (f_c)	4.8 GHz
Sweep Bandwidth (BW)	200 MHz
Sampling Rate (f_s)	2.048 MSPS
FFT Points	2048 point
Chirp Duration T_c	1 ms
Frequency resolution	1 KHz
Range resolution	0.75 m
Maximum Number of Available Range Cells	1024 Range Cell
Range of Interest	15 m
Battery Used	11.1v/5A
Power Consumption	19.98 W
Radar operation time for a single charge	2.5 Hours
Device Cost	800 $

3.3. Radar RF Front End Module

In our previous work, the (VCO), bandpass filter, power divide/combiner, and the antenna were designed and fabricated for a human breathing radar system [21]. In this work, a step forward in the design and fabrication was accomplished by introducing

more analysis and design tuning. The radar Radio Frequency-Front End Module (RF-FEM) circuits are enhanced to improve the (SNR) of the received signal, and therefore improving the breathing rate detection. The new (RF-FEM) circuits are improved to be low cost, compact, high selectivity, low insertion loss in the passband, high isolation in the stopband, excellent return loss, equal phase, and equal amplitude. The (RF-FEM) architecture consists of a single transmitter and a single receiver as presented in (Figure 2). In the transmitter subsystem, a monolithic (VCO) chip TGV2562-SM is utilized to generate frequencies from 4.7 to 4.9 GHz In order to generate a chirp waveform with a preselected bandwidth, a sawtooth signal is used. The VCO printed circuit board is designed, fabricated on Rogers RO4003 substrate material with dielectric material $\varepsilon_r = 3.38$, thickness (h) of 0.203 mm, and loss tangent ($tan\delta = 0.0022$) as shown in (Figure 5). The VCO circuit is measured using N9918A spectrum analyzer, and the VCO output sweeps over the range from 4.7 to 4.9GHz.

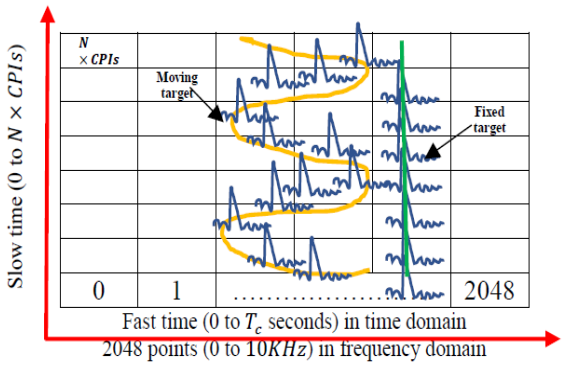


Figure 4: Radar data collection.

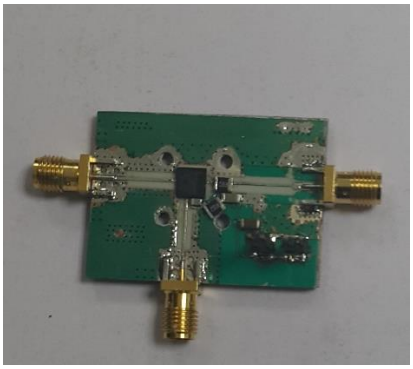


Figure 5: Fabricated printed circuit board of the VCO

equations (Figure 2) shows that the VCO output is connected to a 2-way power splitter. the output signal divides into two paths, a BPF in the transmitter subsystem, and a mixer in the receiver subsystem. (Figure 6) shows the design of a compact microstrip two-way equal Wilkinson RF power divider resonates at 4.8 GHz using CST software. the simulation results within the passband have good matching, isolation, and less than 0.15dB losses as shown in (Figure 6). The proposed power splitter is fabricated on Rogers RT6006 substrate with dielectric constant $\varepsilon_r = 6.15$, thickness (h) of 0.64 mm, and loss tangent (tanδ= 0.0023) and tested using N9918A Vector Network Analyzer (VNA). The measured results show minimum insertion loss 0.2dB, Return Loss, isolation below -23dB in the operating frequency band as shown in (Figure 7).

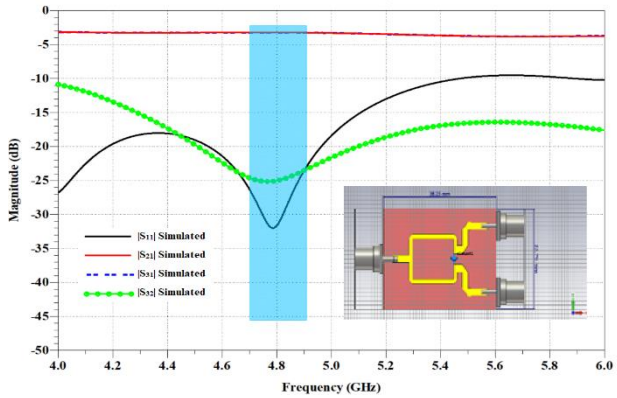


Figure 6: Power divider simulation results and the design structure

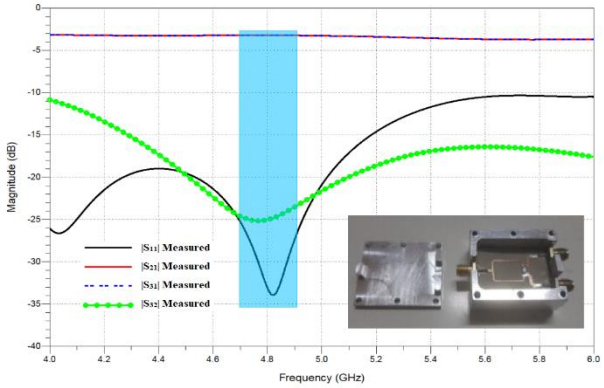
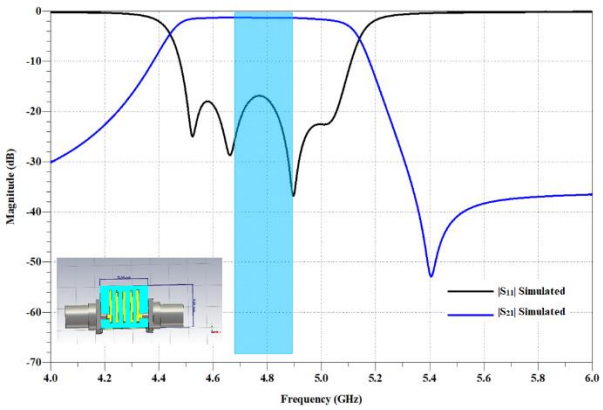


Figure 7: Measured S-parameters results of the fabricated power divider.

A high selectivity BPF is used to select the proposed RF signal on the transmitter subsystem and the receiver subsystem. The design of the presented BPF is based on a microstrip interdigital filter using CST software as shown in (Figure 8). The order of the interdigital filter is five. the simulation results of the proposed wideband BPF is illustrated in (Figure 8). The filter is fabricated on the epoxy Rogers 6006 substrate. Both measured and simulated results of the proposed wideband BPF have low insertion loss -1.1 dB, and good return loss. A good match is found between them the simulated and measured results. The simulated S-parameters are centered at 4.8 GHz while the measured results are shifted slightly and centered at 4.75 GHz as shown in (Figure 9). Any small difference between simulated and measured results from maybe a result of the standard manufacture defects and the copper surface roughness [22].

The BPF output on the transmitter subsystem is amplified via the use of a ZRON-8G+ Mini-Circuits power amplifier. The output of the amplifier is attached to a high gain transmitter antenna. The transmitter antenna is a pyramid horn antenna operating at C-band. (Figure 10) shows the design model and the simulated return loss of the horn antenna using CST software. The simulated results of 3-D radiation patterns of the antenna are given in (Figure 11). The antenna gain is more than 16.5 dB over the operating bandwidth. The horn antenna is fabricated from an aluminum sheet with 3mm thickness as shown in (Figure 12). The simulated results using the CST simulator are compared with the measured one and it achieves

good agreements. The antenna return loss was measured using N9928A Field Fox handheld microwave VNA. As illustrated in (Figure 12), the simulated and measured return loss results of the antenna are less than -10 dB in the operating bandwidth.



Figurer 8: CST structure of the interdigital filter and the simulation results.

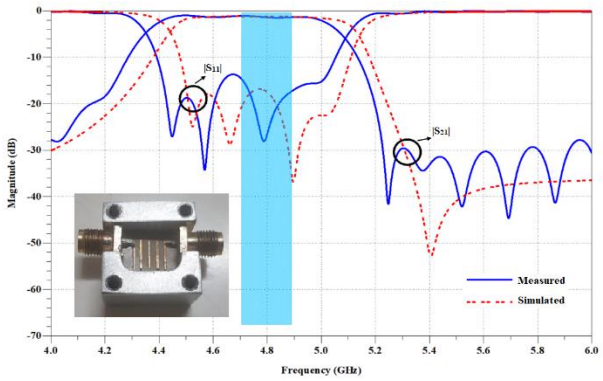


Figure 9: Measured and simulated results of the proposed BPF.

Next, the RF echo signals are received in the receiver subsystem via the receiving antenna, which is a composition of clutter and targets. The receiving antenna is the same as the transmitter antenna. The antenna output is amplified using ZX60-542LN+ Mini-Circuits (LNA). A BPF filter follows the LNA to select the proposed bandwidth. The filtered signal in the receiver subsystem and the output signal from the power splitter in the transmitter subsystem is mixed, using a wideband frequency mixer ZX05-14-S+ to convert the signal to baseband down.

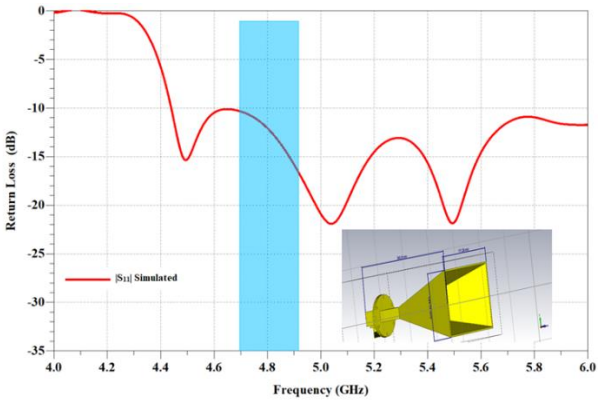


Figure 10: Simulated return loss and the 3D view of the proposed antenna.

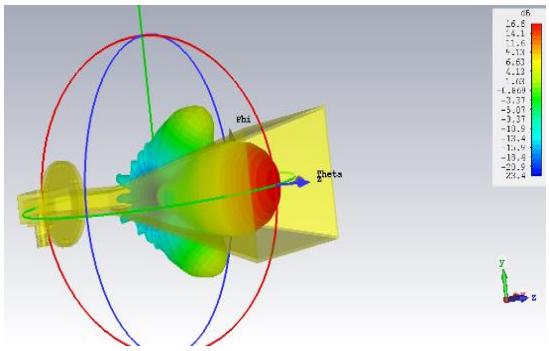


Figure 11: 3D view of the Radiation pattern.

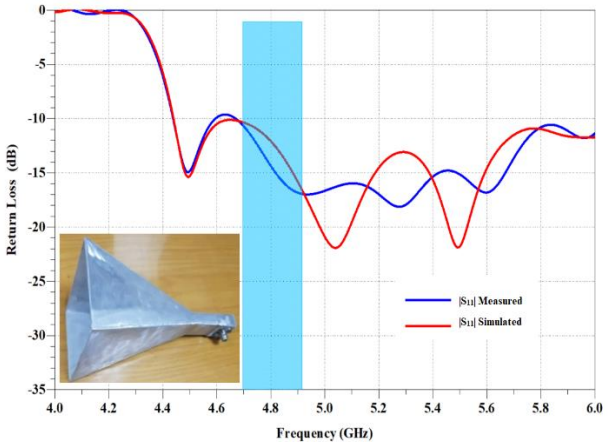


Figure 12: Measured and simulated return loss results.

3.4. Signal Processing

The output of the mixer contains two major frequency components, Sum and difference of the received signal and the reference signal. The signal of interest is the low frequency which contains the target information, Range and Doppler. A LPF is designed and implemented with a cutoff frequency of 1 MHz to filter the mixer output signal from the high frequency components, (Figure 13) shows the filter response. The filtered signal then passes through a 8-bit ADC at 2.048 MSPS sampling rate. The FPGA receives the signal for signal processing.

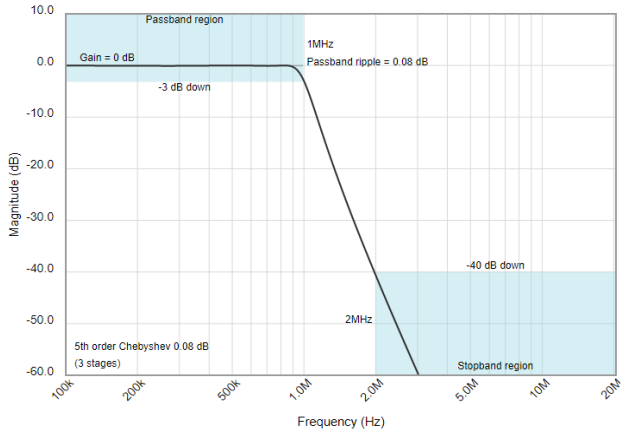


Figure 13: Frequency response of the baseband LPF.

An FPGA board of type XILINX Nexys4, Artix-7 XC7A100T FPGA board, is used for the implementation of the signal processor. A pipelined (FFT) at 2048 point is applied to the ADC

output in the fast time direction as shown in (Figure 14) to extract the frequency component corresponding to the actual range.

Figure 14.a shows that the chirp signal is received continuously through the receiver antenna and the LNA for a single moving target. Figure 14.b shows the resultant signal form the mixer output, which contains the target's information, the range and the Doppler. The amplitude of the signal increases and decreases depending on whether the target is approaching or drifting away from the radar. Figure 14.c shows the spectrum of the received signal after performing FFT, a target's frequency appears with the largest amplitude which changes with the change of the received target signal. The spectrum for each chirp then is stored in a memory Figure 14.d, so that a second FFT can be performed in the slow time for Doppler estimation as shown in Figure 14.e.

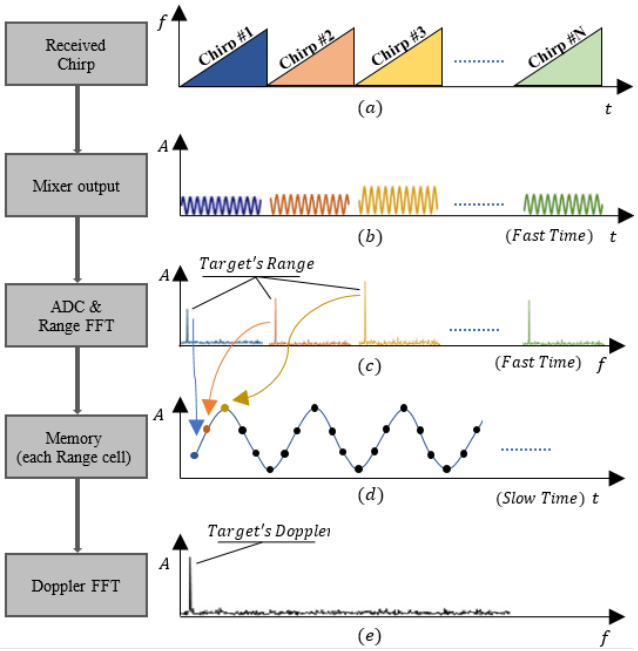


Figure 14: (a) Received chirps, and (b) Mixer output, and (c) Beat frequency in frequency domain, or spectrum of each chirp and (d) Storing target's data into memory for second FFT (slow time), and (e) Slow time FFT for Doppler estimation.

(Figure 15) shows the general block diagram of the digital signal processor. After, the Moving Target Indicator (MTI) is applied to the first FFT output, to remove the stationary targets. MTI can be enabled/disabled, it depends on the target parameters and the environmental conditions, and either one pulse or two pulse cancellers can be selected depending on the targeted speed measurement of the moving target. Then, an adaptive Constant False Alarm Rate (CFAR) is applied, and either Smallest-of-CFAR (GOCFAR) or Greatest-of-CFAR (SOCFAR) can be selected, both having the widow size adjustable (16 or 32).

A decision based on an adjustable number of successive target hits is made using a binary integrator; it can be manually turned on or off during the device run-time operation, and the required number of successive hits can be adjusted during the run-time operation as well, without affecting the radar operation.

All the parameters that can be configured, are perfectly synchronized by the synchronizer to avoid any errors due to removing/adding signal processing blocks or changing numbers within each block in real time, during run time operation. The output block can be separately sent directly to a host PC using serial communication protocol during run-time to monitor and track the received signal in each block, as well as for further analysis, testing and processing.

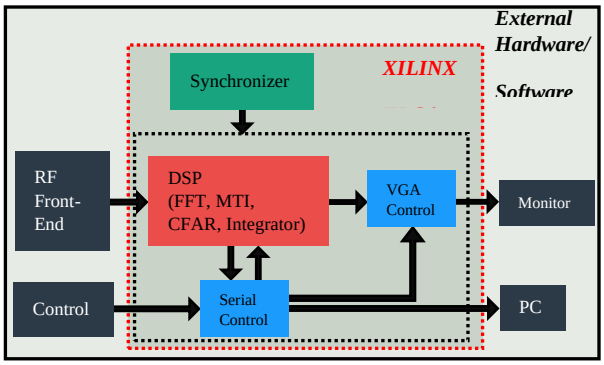


Figure 15: Developed Digital Signal Processing block diagram using XILINX FPGA.

3.5. Simulation Results

The simulation process is based on a simulated target from the MATLAB software, the simulated target then is processed in MATLAB software, the same MATLAB generated target is then processed in ModelSim software (FPGA simulator) , then the signal processing is applied, The outputs of FFT from both MATLAB and ModelSim are compared with each other, the same is applied to the MTI and CFAR. The results show that the outputs of the MATLAB and FPGA are identical. (Figure 16) shows the simulation criteria.

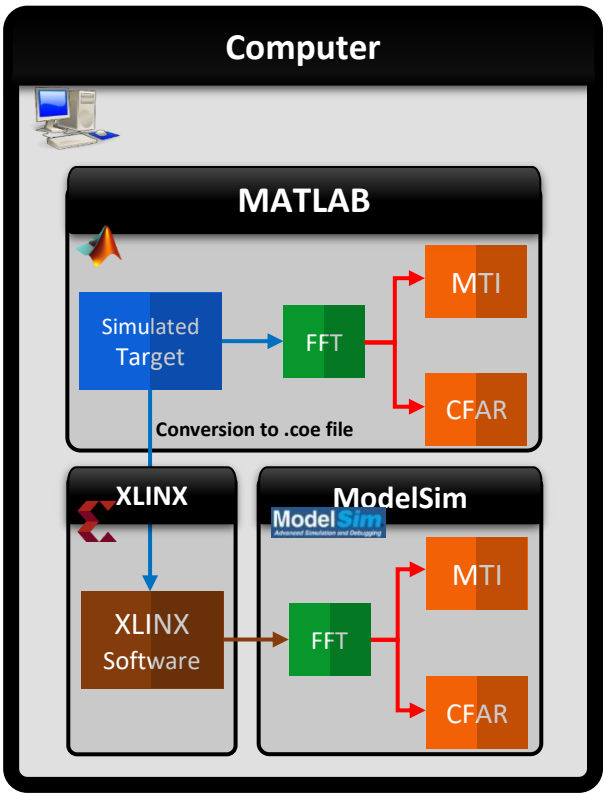


Figure 16: Developed Digital Signal Processing block diagram using XILINX FPGA.

4. Experimental Results

In this section, two experiments are presented, in the first experiment, a breathing signal is estimated from a human sitting in front of the radar device. In the second experiment, an FFT process is applied to a received signal from a human target moving back and forth within certain range from the radar, the FFT-processed data then is sent to the PC.

4.1. Breathing Detection

In this experiment (the first experiment), three measurements were taken from a human in three scenarios:

A human is sitting at a distance of 2.25m, 6m and a measurement without any target in front of the radar for results comparison.

(Figure 17) shows the breathing signal at the 3rd range cell (2.25 m) in time domain, after applying first FFT, the received breathing signal is caused by the human chest movement. After performing the 2nd FFT, a peak appears in the frequency domain which represents the breathing frequency, as shown in (Figure 18)

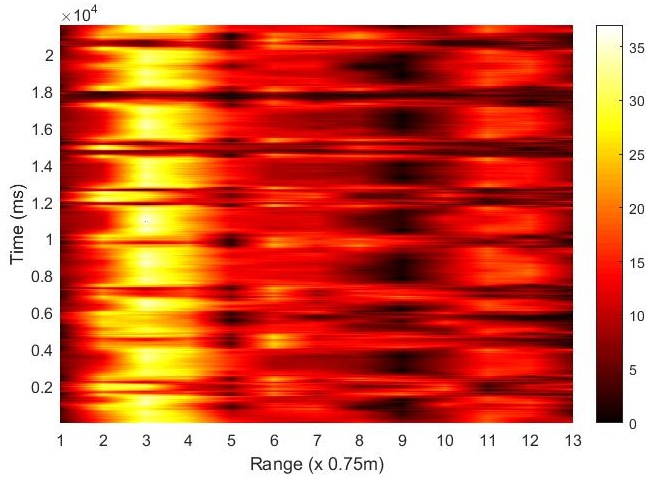


Figure 17: Time domain breathing signal at the 3rd range cell.

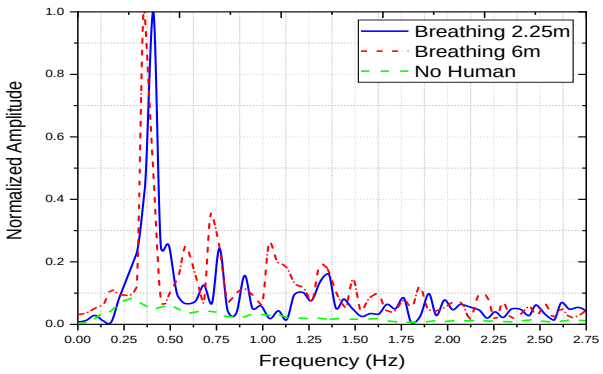


Figure 18: Breathing signal detection results of two ranges.

(Figure 18) shows that a breathing signal has been successfully extracted. Two peaks are detected at 0.4Hz and 0.37Hz detected at a distance 2.25m and 6m, respectively, which represent the breathing signal. To assure that the output peaks represent the breathing signal, not noise signal from the surrounding area or the thermal noise, a measurement was taken without a human in front of the radar, the result as in (Figure 17) green dashed line shows that the breathing frequency disappeared.

4.2. Movement Detection

In the second experiment, The radar device was placed in the free space, and there was two tanks in the left of the radar as shown in (Figure 19). The first data collected without a moving target shows a weak signal at 15 meters as the two tanks was placed on the left of the radar as shown in (Figure 20), a human is moving within 15m back and forth as shown in collected data in (Figure 21), the figure shows that the target appears distinguishable from the noise, as a result of the new RF components presented in this paper.



Figure 19: Measurement setup.

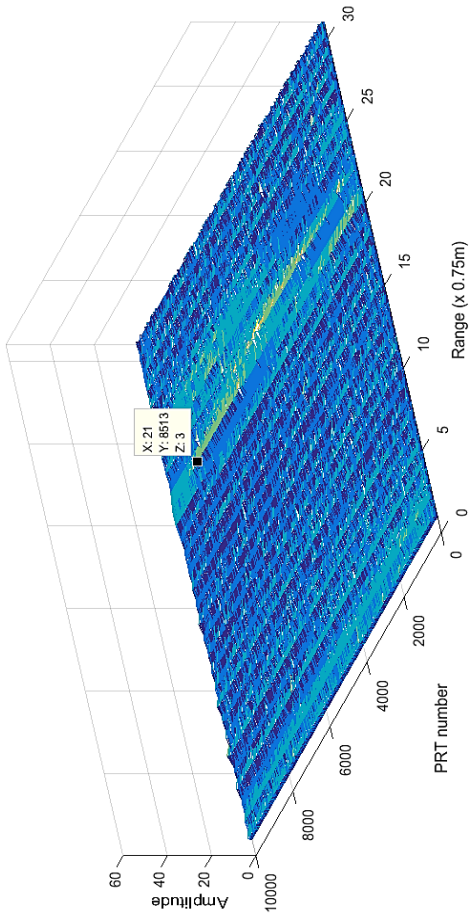


Figure 20: Data collected without moving target.

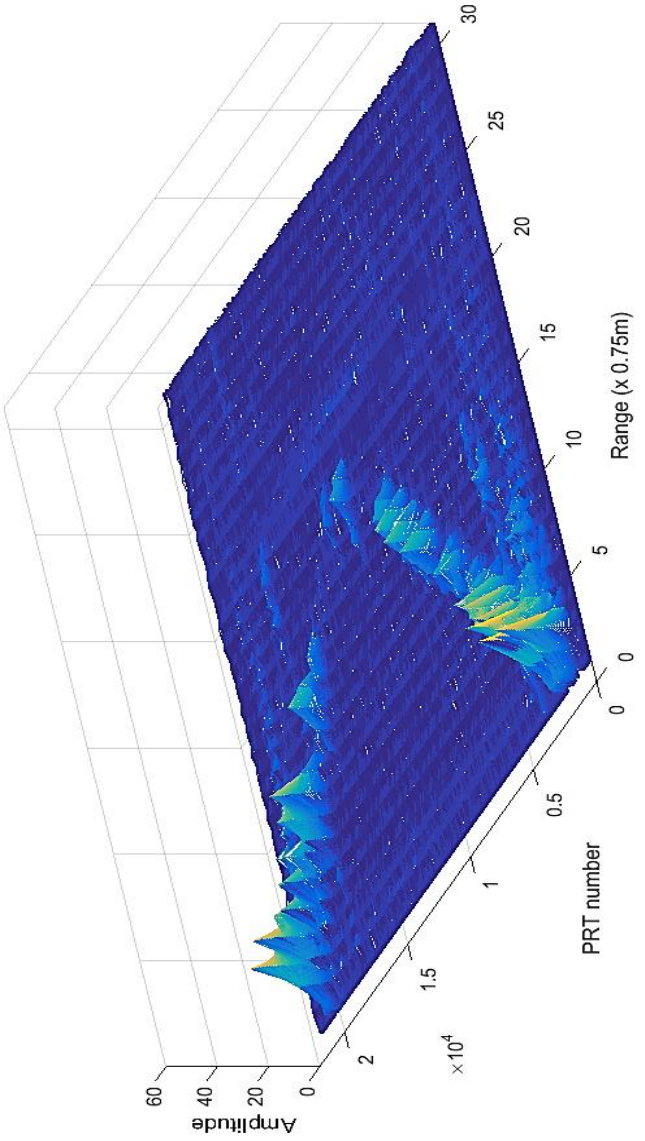


Figure 21: Data collected with the moving target.

4.3. System complexity:

In this paper, the proposed radar system contains IF components and RF components, The RF components comprises new simple RF designs yet with the best performance, while the readymade components are available in the market for commercial use.

The IF section contains the custom locally manufactured boards for this specific radar system, like ADC, LPF, Function Generator (FG) and power supply with attached battery, all were made by locally available components which made the IF section the cheapest part in the system. The FPGA is chosen carefully by calculating the needed resources for the DSP algorithm, that assures the best choice of the FPGA to avoid the high cost FPGAs.

The whole system was intended to be designed as simple as possible, to assure the availability of the required components, the reliability of usage, simplicity in maintenance, with the best performance as shown in previous sections.

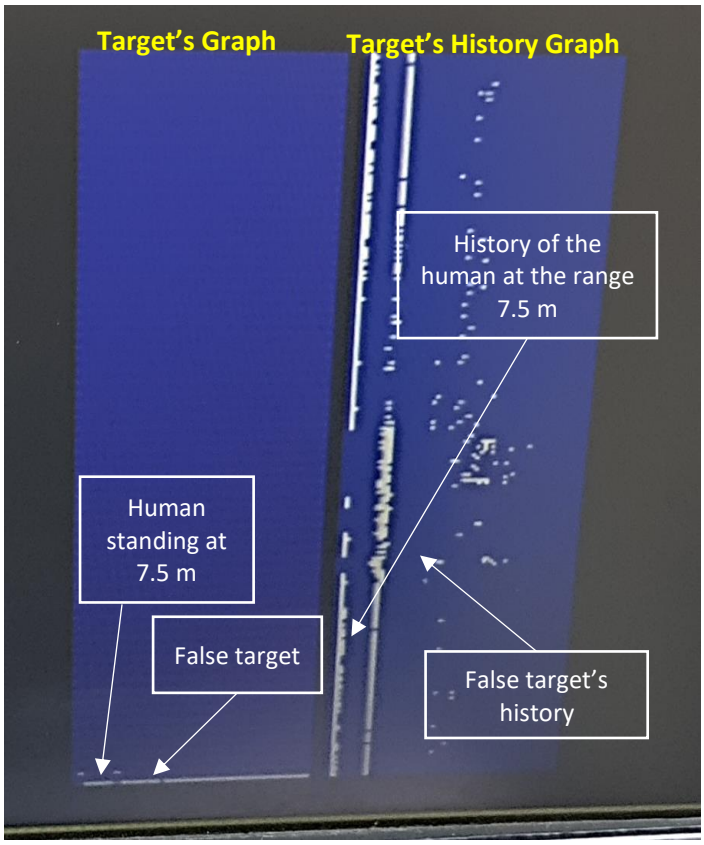


Figure 22: Real target detected on the monitor from the FPGA directly.

Table 2: Previous Work Comparison

Parameter	This Work	[23]	[24]
Start frequency	4.7 GHz	2.1 GHz	8 GHz
Bandwidth	200 MHz	12.5 MHz	4 GHz
ADC sampling rate	2.048 MSPS	25 MSPS	1 MHz
Sweep period	1 ms	1 ms	100 ms
Power Output	16 dBm	33 dBm	2 dBm
Configurability	Yes	No	No
FMCW software	No software required (Windows and Ubuntu are optional)	GNU Radio (Ubuntu)	GNU Radio (Windows)
Cost	Low	High	High

The monitoring system is a key component in this system, the system does not require a software at all, however, an optional software is designed for target monitoring. A Video Graphics Array (VGA) algorithm is implemented on FPGA, taking the advantage of the remaining FPGA resources to show the signal processing results on the VGA screen directly as shown in (Figure 22) and assures the real time detection of the targets and breathing signal. The monitor screen can show the received signal in time domain/frequency domain, breathing signal in time domain/frequency domain, CFAR target output with 5 seconds target history. Switching between all the outputs on the screen can be controlled using the configuration control screen. One

advantage of the VGA controller is that the system takes only 10 seconds to boot.

The monitoring system is a key component in this system, the system does not require a software at all, however, an optional software is designed for target monitoring. A Video Graphics Array (VGA) algorithm is implemented on FPGA, taking the advantage of the remaining FPGA resources to show the signal processing results on the VGA screen directly as shown in (Figure 22) and assures the real time detection of the targets and breathing signal. The monitor screen can show the received signal in time domain/frequency domain, breathing signal in time domain/frequency domain, CFAR target output with 5 seconds target history. Switching between all the outputs on the screen can be controlled using the configuration control screen. One advantage of the VGA controller is that the system takes only 10 seconds to boot.

4.4. Previous Work Comparison

Table 2 Shows a comparison between the proposed radar system and some other developed radar devices. The comparison shows that the proposed radar system is gaining the advantage of the configurability function, with relatively better price, in addition to that, our system does not require any software or operating system for the operation.

5. Conclusion and Future Work

In this paper, a reconfigurable radar system is proposed. The radar device is designed for human movement as well as for breathing detection. The proposed radar system is implemented using a new low-cost RF front end component and a reconfigurable radar signal processor. The radar device is aimed to operate in different situations and multiple environmental conditions. Depending on the required situations, MTI, CFAR and binary integrator can be configured for better performance. The system parameters can be reconfigured during the run time; without affecting the working efficiency. The Proposed radar system is compact and low cost compared to other radar systems with some similar parameters and operates in outdoor and indoor scenarios.

This research work will be extended by using software-defined radio (SDR), to generate the LFMCW waveform and to implement the digital receiver and DSP subsystem. SDR will help in generating different transmitted waveform configurations, decreasing the main to side-lobe level of the transmitted waveform, decreasing the overall size of the radar system, and reducing the cost. After that, many adjustments to the system can help in increasing the operating range and the breathing rate estimation accuracy. Additionally, some functions will be added to the VGA screen for ease to use, like adding grids, numbers and letters.

References

[1] M. Mabrouk, S. Rajan, M. Bolic, I. Batkin, H.R. Dajani, V.Z. Groza, "Model of human breathing reflected signal received by PN-UWB radar," 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC **2014**, 4559–4562, 2014, doi: 10.1109/EMBC.2014.6944638.

[2] S. Sharpe, Seals, J.; MacDonald, A.H.; Crowgey, S.R. Non-Contact Vital Signs Monitor. U.S. Patent **4**,958,638, 25 September 1990.

[3] J. Pan, "Medical Applications of Ultra-Wideband (UWB)," Hadley, 2007.

[4] T. R. Frieden, "The State of Aging and Health in America," Atlanta, 2013.

[5] F. Quaiyum, L. Ren, S. Nahar, F. Foroughian, A.E. Fathy, "Development of a reconfigurable low cost multi-mode radar system for contactless vital signs detection," IEEE MTT-S International Microwave Symposium Digest, **6**, 1245–1247, 2017, doi: 10.1109/MWSYM.2017.80 58832.

[6] D. Zhang, M. Kurata, T. Inaba, "FMCW radar for small displacement detection of vital signal using projection matrix method," International Journal of Antennas and Propagation, **2013**(1), 1–6, 2013, doi: 10.1155/2013/57198 6.

[7] H. Lee, B.H. Kim, J.K. Park, J.G. Yook, "A novel vital-sign sensing algorithm for multiple subjects based on 24-GHz FMCW Doppler radar," Remote Sensing, **11**(10), 2019, doi: 10.3390/rs11101237.

[8] N. Maaref, P. Millot, C. Pichot, O. Picon, "A study of UWB FM-CW radar for the detection of human beings in motion inside a building," IEEE Transactions on Geoscience and RemoteSensing, **47**(5), 1297–1300, 2009, doi: 10.1109/TG RS.2008.2010709.

[9] [9] A. Nezirovic, A.G. Yarovoy, L.P. Ligthart, "Experimental Verification of Human Being Detection Dependency on Operational UWB Frequency Band." IEEE International Conference on Ultra-Wideband, Singapore, 305-310, 2007, doi: 10.1109/ICUWB.2007.4380960.

[10] L. Anitori, A. De Jong, F. Nennie, "FMCW radar for life-sign detection," 2009 IEEE Radar Conference, 1-6, 2009, doi: 10.1109/RADAR .2009.4976934.

[11] C. Gu, G. Wang, T. Inoue, C. Li, "Doppler radar vital sign detection with random body movement cancellation based on adaptive phase compensation," 2013 IEEE MTT-S International Microwave Symposium Digest, 1-3, 2013, doi: 10.1109/MWSYM.2013.6697618.

[12] A. Ahmad, J.C. Roh, D. Wang, A. Dubey, "Vital signs monitoring of multiple people using a FMCW millimeter-wave sensor," 2018 IEEE Radar Conference (RadarConf18), **4**, 1450–1455, 2018, doi:10.1109 /RADA R.2018.8 37 8778.

[13] M. Alizadeh, G. Shaker, J.C.M. De Almeida, P.P. Morita, S. Safavi-Naeini, "Remote monitoring of human vital signs using mm-Wave FMCW Radar," IEEE Access, **7**(4), 54958–54968, 2019, doi: 10.1109/ACCESS. 2019.2912956.

[14] M. Mabrouk, S. Rajan, M. Bolic, M. Forouzanfar, H.R. Dajani, I. Batkin, "Human Breathing Rate Estimation from Radar Returns Using Harmonically Related Filters," Journal of Sensors, 2016, doi: 10.1155/2016/9891852.

[15] M. I. Skolnik, Radar Systems, New York: McGraw-Hill, 2001.

[16] M. Mabrouk, W. Mohamed, S. Aly, H.M. Hassan, "MTD Design and Implementation Using Short Coherent Processing Interval Phase Trace," Proceedings of 172nd ISERD International Conference, Cairo, Egypt ,40-44, 2019.

[17] R. Sor, J.S. Sathone, S.U. Deoghare, M.S. Sutaone, "OS-CFAR Based on Thresholding Approaches for Target Detection," Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA **2018**, 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697389.

[18] C. Xu, Y. Li, C. Ji, Y. Huang, H. Wang, Y. Xia, "An improved CFAR algorithm for target detection," International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2017 -Proceedings, 883–888, 2017, doi: 10.1109/ISPACS .2017.8266600.

[19] X. Meng, Y. He, N. Aeronautical, E. Academy, "Two Generalized Greatest of Selection CFAR Algorithms," CIE International Conference on Radar Proceedings (Cat No.01TH8559), Beijing, China, 359–362, 2001, 359-362, doi: 10.1109/ICR.2001.984700.

[20] Meng Xiangwei, Guan Jian, He You, "A generalized smallest of selection CFAR algorithm [radar signal processing]," 130–132, 2004, doi: 10.1109/radar.2003. 1278724.

[21] M.M. Abdul-atty, M. Mabrouk, S. Elramly, "Design and Implementation of a Low Cost FMCW Radar with Configurable Signal Processor for Human Movement and Breathing Detection," International Journal of Electrical, Electronics and Data Communication, **7**(9), 36–41, 2019, doi: RW.18042019.12277.

[22] X. Chen, "EM modeling of microstrip conductor losses including surface roughness effect," IEEE Microwave and Wireless Components Letters, **17** (2), 94–96, 2007, doi: 10.1109/LMWC.2006.890326.

[23] JIAXI ZHU, "LOW-COST, SOFTWARE DEFINED FMCW RADAR FOR OBSERVATIONS OF DRONES". M.Sc. thesis, Oklahoma univ., Norman, Oklahoma, 2017.

[24] A. Anghel, G. Vasile, R. Cacoveanu, C. Ioana, S. Ciochina, "Short-Range wideband FMCW radar for millimetric displacement measurements," IEEE Transactions on Geoscience and Remote Sensing, **52** (9), 5633–5642, 2014, doi: 10.1109/TGRS.2013.2291573.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1306-1311 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Basic Study of 3-D Non-Invasive Measurement of Temperature Distribution using Ultrasound Images during HIFU Heating

Ryosuke Sakakibara[*, 1], Yasuhiro Shindo[2], Kazuo Kato[3], Pak Kon Choi[4], Akira Takeuchi[5]

[1]*Graduate School of Science and Technology, Meiji University, 214-8571, Japan*

[2]*Department of Mechanical Engineering, Toyo University, 350-0815, Japan*

[3]*Department of Mechanical Engineering Informatics, Meiji University, 214-8571, Japan*

[4]*Department of Physics, Meiji University, 214-8571, Japan*

[5]*Department of Thermotherapy, Luke Hospital, 104-0045, Japan*

ARTICLE INFO

Article history:
Received: 16 September, 2020
Accepted: 01 December, 2020
Online: 16 December, 2020

Keywords:
HIFU
Ultrasound
Non-Invasive Temperature Measurement

ABSTRACT

High Intensity Focused Ultrasound (HIFU) was widely used for treating tumors non-invasively. In this treatment, ultrasound is focused on the target volume inside the human body to ablate cancerous tissues and Magnetic Resonance Imaging (MRI) is mainly used to grasp the target position and to measure the temperature distributions around the target. However, MRI is very expensive, and a large space is required.

In this paper, we presented a method for measuring the temperature distribution using an ultrasound diagnostic device, which is inexpensive and commonly used in many clinics, and actually showed the results of heating experiments on a human shaped agar phantom. The proposed method for measuring the temperature distribution around the heated target was conducted by performing image processing on two ultrasound images before and after heating. Furthermore, it was confirmed that it was possible to grasp the three-dimensional temperature distribution from the images in multiple layers. The effectiveness of the temperature distribution measurement results by the proposed method was shown by comparing the temperature measurement results with the infrared thermal camera. The error between the results was approximately 1 ℃.

It was found that the non-invasive measurement method of the three-dimensional temperature distribution around the target volume using the ultrasound images was useful for effective HIFU treatments.

1. Introduction

This paper is an extension of a previous work originally presented at the 21st edition of the International Conference on Electromagnetics in Advanced Applications (ICEAA 2019) [1].

High Intensity Focused Ultrasound (HIFU) devices are used for treating uterine fibroids, liver tumors, brain tumors, etc. as a minimal invasive treatment method [2, 3]. As shown in Figure 1, this treatment method focuses on ultrasound waves emitted from the outside of the body to a target inside the body and causes ablation necrosis of the tumor by the thermal energy generated at the focal point. This method is considered to be a treatment method that has a much smaller burden on the patient than conventional open surgery. Focused ultrasound treatment is performed at a high temperature of approximately 70 °C or higher, and therefore there is a risk of cauterizing normal tissue around the tumor. When performing this treatment, a device with MRI's guidance (MR-guided focused ultrasound surgery: MRgFUS) is commonly used. However, a large amount of introduction funds and a large space are required. In addition, this method is not considered to be suitable for cases in which there is metal inside the human body or MRI contrast media cannot be used because the inside of the body is monitored using MR images [4-6].

[*]Corresponding Author: Ryosuke Sakakibara, 1-1-1, Higashi-mita, Tama-ku, Kawasaki, Kanagawa, 214-8571, +81-44-934-7907, rsakakibara9@gmail.com

Therefore, in order to perform Focused Ultrasound Surgery (FUS) treatment, non-invasive temperature measurement using an ultrasound diagnostic imaging device that is cheaper and more convenient to move than an MRI device has been previously attempted [7-12]. In this study, the ultimate goal is to establish a method for non-invasively measuring three-dimensional temperature distribution in a living body with high resolution in subpixel units [9]. In the temperature measurement method using ultrasound images, it is a prerequisite that the used ultrasound diagnostic imaging system was designed to draw images with constant ultrasound velocity. The velocity of ultrasound waves changes depending on the temperature of the propagating tissue. Therefore, if there is a temperature change inside the tissue, the ultrasound velocity in that part locally changes, and it appears as an extremely small image displacement on the ultrasound image. The temperature could be measured by detecting this slight displacement with image analysis technology [8, 9].

In the present paper, first, the algorithm of two-dimensional temperature distribution measurement using ultrasound images is described. Next, the agar phantom was heated in a temperature-controlled water tank, the relationship between the temperature increase of the agar and the displacement of the ultrasound image was described. Next, the temperature distribution was measured from the ultrasound image inside the agar phantom when it was heated using the focused ultrasound heating device. The temperature distribution inside the agar phantom during heating was measured using two types of transducers with different focal lengths. In the previous research, we investigated the basic heating characteristics by measuring the temperature distribution using a cylindrical agar phantom [10]. In this study, we measured a wider region of temperature distributions using a human shaped agar phantom in order to confirm that the proposed temperature measurement method is effective in situations closer to clinical treatment.

From the examination of the above experimental results, the usefulness of temperature distribution measurement using an ultrasound diagnostic imaging system was clarified quantitatively.

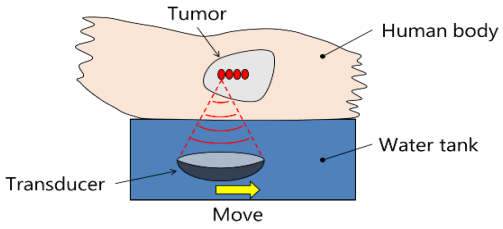


Figure 1: Illustration of Focused Ultrasound Surgery.

2. Materials and Methods

2.1. Principle of Temperature Measurement Using Ultrasound Images

As reported by Iseki et al., the temperature increases in living tissue can be obtained by solving (1) and (2) [8, 9].

$$\Delta T(x) = K_{tissue} \frac{\partial(\Delta d)}{\partial x} \quad (1)$$

$$K_{tissue} = \frac{1}{\alpha - \beta} \quad (2)$$

In these equations, ΔT is the temperature increase inside the tissue and K_{tissue} is the thermal constant that varies with the tissue. The subscript "tissue" is added to emphasize that it differs depending on the tissue. Δd is the displacement of the ultrasound image due to the temperature increase (the difference between the images before and after heating). The thermal constant K_{tissue} is expressed by (2). Here, α is the coefficient of thermal expansion of the tissue, and β is the coefficient of ultrasound velocity change due to the temperature increase inside the tissue. It is considered that the thermal expansion coefficient α and the ultrasound velocity change coefficient β change linearly in a relatively narrow temperature change region such as during focused ultrasound treatment [6].

2.2. Measurement System of Thermal Constant

Figure 2 shows the general-purpose ultrasound imaging system Nemio SSA-550A (Toshiba Medical Systems, Japan) with the attached ultrasound probe PLM-805AT (Toshiba Medical Systems, Japan) used in the experiment. The ultrasound probe used in the experiment is a linear probe as shown in Figure 2 (b). The number of elements is 128, the focus length is 20 mm, and the size of the contact part with the body is (12 mm × 67 mm). The size of the captured ultrasound image is (560 pixels × 450 pixels), which is a 256-gradation bitmap image. Here, 300 pixels were assigned to a distance of 70 mm on the captured ultrasound image.

Figure 3 shows an illustration of experimental setup for measuring thermal constant "*K*" to know the relationship between the temperature increase of the agar used as the heated object and the displacement of the ultrasound images before and after heating [8,9]. As shown in Figure 3, a cube agar phantom was fixed inside a water tank where the water temperature was constantly controlled using a thermostat and a stirring pump, and the ultrasound images inside the agar were taken from 20 ℃ to 30 ℃, because the experiment of temperature distribution measurement is in that temperature range.

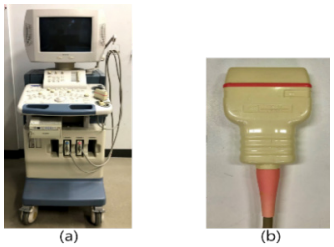


Figure 2: Ultrasound Diagnosis Device (a) Ultrasound Diagnosis Device, (b) Linear Probe.

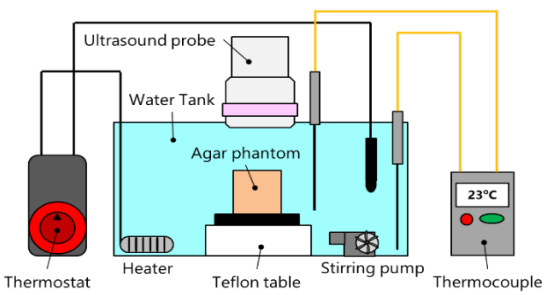


Figure 3: Illustration of Experimental Setup for Measuring Thermal Constant "*K*".

2.3. Image Analysis Algorithm

Figure 4 shows the image analysis algorithm proposed by authors for temperature distribution measurement [8, 9]. This algorithm is divided into 7 steps. First, the ultrasound images of the heated object before and after heating are input. Second, template matching is performed to measure the amount of displacement Δd between the ultrasound images before and after heating. Third, a median filter is applied to reduce the erroneous vector of displacement Δd on the ultrasound image. Fourth, since the displacement Δd distribution detected by template matching is expressed in pixel units, interpolation processing was performed to interpolate Δd distribution. Fifth, a moving average filter is applied to smooth the displacement amount of the distribution map subjected to this interpolation process. Sixth, the temperature increase ΔT during heating is calculated from the displacement Δd detected by applying the Sobel filter. Finally, a moving average filter is applied to the temperature increase distribution map to smooth the distribution. For the development of the image analysis program, general-purpose program Microsoft Visual Studio (Microsoft Japan) was used.

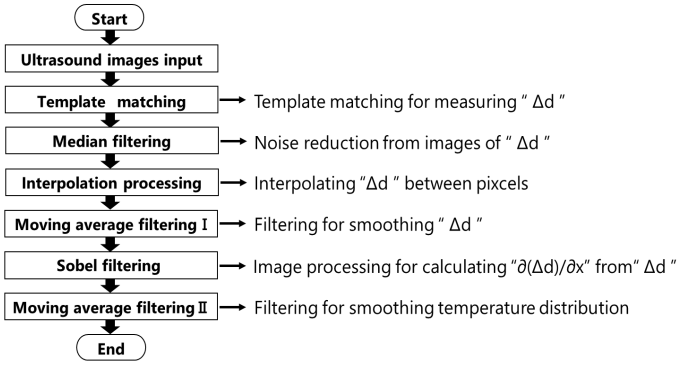


Figure 4: Flow Chart for Measuring Temperature Distributions.

2.4. Measurement System of Temperature Distributions

Figure 5 shows the device system for capturing ultrasound images before and after heating. This system mainly consists of two work sliders, an arm and a jig for fixing the ultrasound probe. As shown in Figure 5, the ultrasound probe was supported by the arm and fixed to the slider. Here, the ultrasound probe can be moved three-dimensionally by combining the two sliders. The range of motion of this slider is (0.0 to 66.0 mm), and the minimum movable interval is 0.05 mm.

Figure 6 shows an illustration of the focused ultrasound heating experiment. As shown in Figure 6, in order to measure the temperature distribution in the X-Z cross section using an infrared thermal camera, the human shaped agar phantom was previously cut at the center position. The transducer was placed underwater to focus on the position 65 mm from the bottom. Then, the ultrasound probe was moved up and down ± 15.0 mm from the center cross section every 5.0 mm along the Z direction, and ultrasound images were taken before and after heating in a total of 7 layers. These operations were performed for heating experiments with short focus and long focus transducers respectively. The heating time was 5 minutes and the heating power was 200 W. In addition, we compared the temperature distribution results by the present method using the ultrasound image and the thermal image taken by the infrared thermal camera in the central section of the human shaped agar phantom.

The focused ultrasound heating device used in the experiments was CZ901 (Mianyang Sonic Electronic, China). Two kinds of transducers were prepared for this device. Figure 7 shows the transducer used in the experiments. The short focus transducer as shown in Figure 7 (a) had an outer diameter of 170 mm, an inner diameter of 98 mm, an outer peripheral surface height of 75 mm, an inner peripheral surface height of 53 mm, and a focal length of 120 mm. The outer and inner diameters of the long focus transducer as shown in Figure 7 (b) were almost the same as those of the short focus transducer, however the inner peripheral surface height was 61 mm and focal length was 200 mm. The frequency of the focused ultrasound was 1 MHz. In the heating experiments, a frequency oscillator AWG1025F (Frequency Band: 1μHz – 25MHz, AS ONE, Japan) and a power amplifier L-400BM-HC (Frequency Band: 700kHz – 3.5MHz, JSE, Japan) were connected to the transducers to heat the human shaped agar phantom.

Figure 8 shows the human shaped agar phantom used in the experiments. The width of heating area was 160 mm and the height was 120 mm.

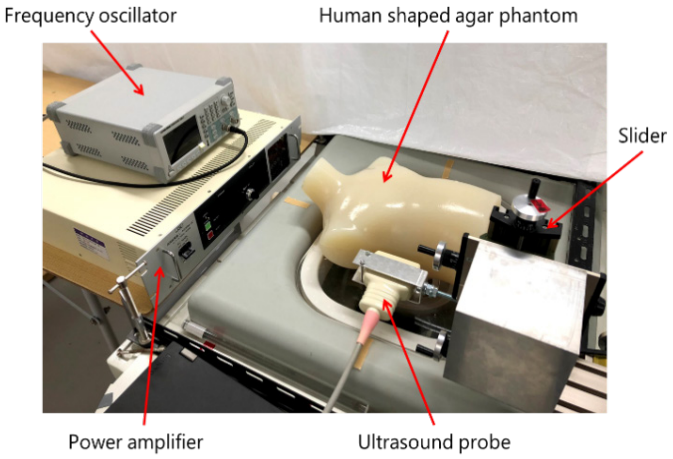


Figure 5: Experimental Setup for Capturing Ultrasound Images.

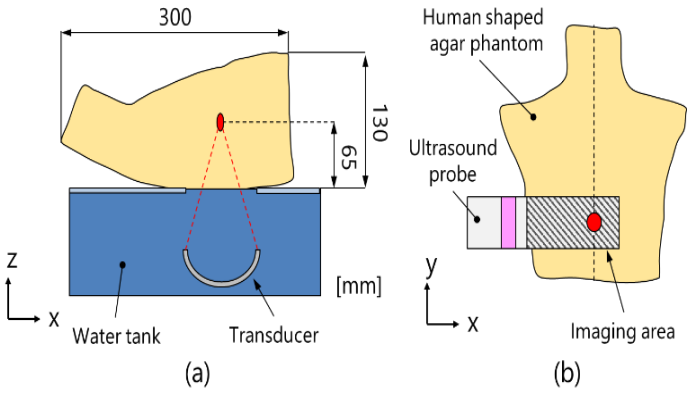


Figure 6: Illustration of Experimental Setup for Scanning Ultrasound Images. (a) Side View, (b) Top View.

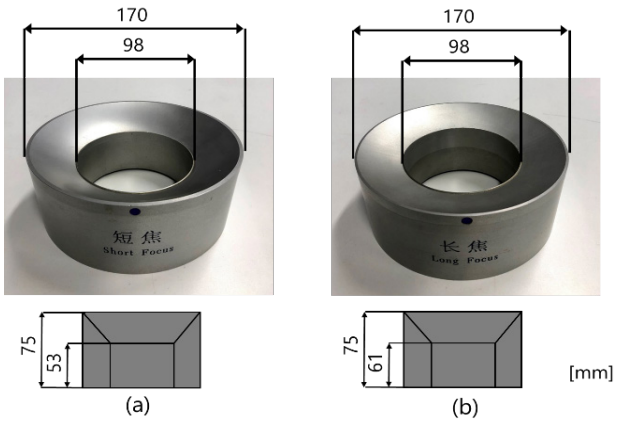


Figure 7: Transducers Used in the Experiments.
(a) Short Focal Length, (b) Long Focal Length.

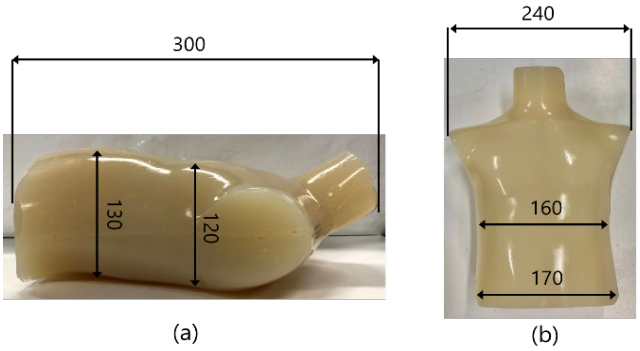


Figure 8: Human Shaped Agar Phantom.
(a) Side View, (b) Top View.

3. Results

3.1. Thermal Constant of Agar Phantom

Figure 9 shows the relationship between the temperature change (ΔT) of the agar phantom and the degree of ($\partial(\Delta d)/\partial x$) of the ultrasound image when heated as shown in Figure 3. The template size used in the image analysis as shown in Figure 4 was (13 pixels × 13 pixels). The total number of plots in Figure 9 is 45. As a result of finding the approximate straight line from these data by the method of least squares, the slope of the graph was 102.92 and the intercept was -0.5262. The thermal constant of the agar phantom used in this heating experiment was set to K_{agar} = 102.92. In addition, the coefficient of determination R^2, which is the degree of fitting of the approximate straight line, is 0.9941, which confirms that there is a high correlation.

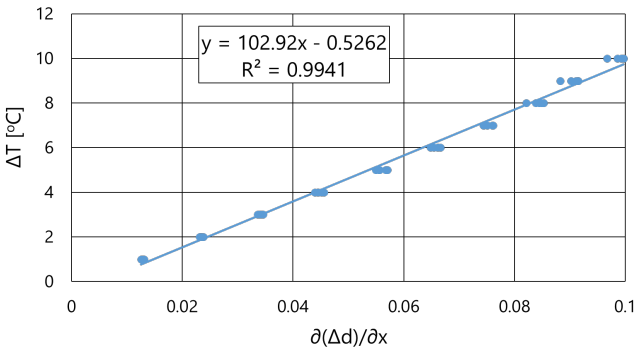


Figure 9: Relationship between ΔT and $\partial(\Delta d)/\partial x$.

3.2. 2-D Temperature Distributions

Figure 10 shows the comparison results of the temperature distribution in the central cross section of the human shaped agar phantom when heated with a short focus transducer. In the temperature distribution measurement result from the ultrasound image as shown in Figure 10 (a), the maximum temperature increase value ΔT_{max} was 10.1 °C. On the other hand, it was 10.8 °C from the infrared thermal camera as shown in Figure 10 (b). The measured temperature error by both methods was 0.7 °C, and the error rate was approximately 7 %. When comparing the absolute temperatures of both during and after heating, an error naturally occurs. However, it is useful to know the temperature distribution patterns.

Figure 11 shows the temperature in the center section of the agar phantom when the same heating was performed using the long focus transducer. In the temperature distribution measurement result from the ultrasound image as shown in Figure 11 (a), the maximum temperature increase ΔT_{max} was 7.2 °C. On the other hand, it was 5.9 °C from the image of the infrared thermal camera as shown in Figure. 11 (b). The measured temperature error was 1.3 °C, and the error rate was approximately 18 %.

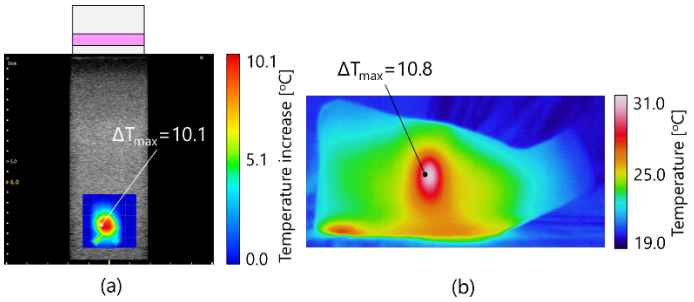


Figure 10: Comparison between Temperature Distributions inside Human Shaped Agar Phantom (Short Focal Length). (a) with Ultrasound Image, (b) with Thermal Camera

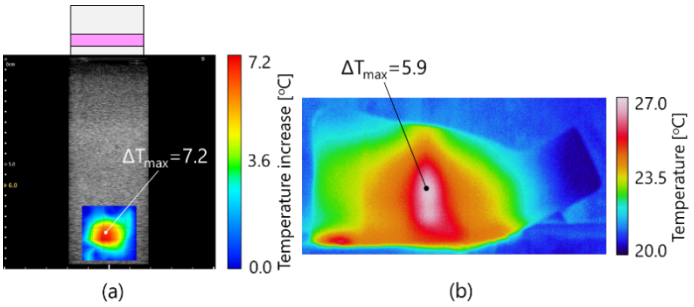


Figure 11: Comparison between Temperature Distributions inside Human Shaped Agar Phantom (Long Focal Length). (a) with Ultrasound Image, (b) with Thermal Camera.

3.3. 3-D Temperature Distributions

Figure 12 shows the two-dimensional temperature distribution results measured from the ultrasound images of the top and bottom (± 15.0 mm) taken at 5.0 mm intervals in the center section of the agar phantom. As shown in Figure 12, the two-dimensional temperature distribution results are three-dimensionally displayed by stacking them. Figure 12 (a) shows the case when using the short focus transducer. It can be confirmed that the temperature increase is the highest value in the central cross section and

decreases with increasing distance from the central cross section, and there is almost no temperature increase above ± 10.0 mm. On the other hand, Figure 12 (b) shows the case when using a long focus transducer. Also in this case, it can be seen that the temperature increase is the highest value in the central cross section. However, in the other image layers, unlike the case when the short focus transducer was used, it can be confirmed that the heating region is wide and the temperature increase value at ± 10.0 mm is similar to the value at the center cross section.

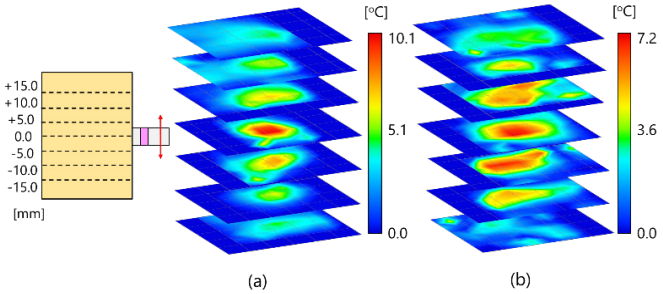


Figure 12: Temperature Distributions from Ultrasound Images. (a) Short Focal Length, (b) Long Focal Length.

4. Discussion

In order to improve the accuracy in the temperature distribution measurement method of applying ultrasound images, we have been working on, (i) taking the images before and after heating at the same position, and (ii) grasping the thermal constant K_{tissue} of the targeted object. Regarding item (i), as shown in Figure 5, the ultrasound probe could be moved three-dimensionally by using two sliders with a movement accuracy of ± 0.05 mm. Regarding item (ii), the value of the agar phantom was actually measured using the self-made device as shown in Figure 3. However, in the results shown in Figure 9, the variation of the measured 45 data was large. It is necessary to control the temperature of the object more accurately by using a higher quality thermostat.

In clinical HIFU treatments, the cauterization temperature of the tumor is mainly 70 °C - 80 °C. The human shaped agar phantom used in the HIFU experiments was prepared with the component ratio recommended by the Japan Hyperthermia Society. When the maximum temperature near the focal point, where ultrasound waves were focused, was 50 °C or higher, the agar phantom melted. Therefore, heating experiments of the agar phantom were conducted at 50 °C or lower. As shown in Figure 12 (a) and (b), heating results using transducers with different focal lengths had different maximum temperatures. Therefore, the normalized temperature (ΔT_N) shown in the following (3) was calculated, and the heating characteristics of both transducers were compared.

$$\Delta T_N = \frac{\Delta T - \Delta T_{min}}{\Delta T_{max} - \Delta T_{min}} \quad (3)$$

Here, ΔT_{max} is the maximum temperature increase value and ΔT_{min} is the minimum temperature increase value in each layer as shown in Figure 12.

Figure 13 shows the calculation results of the normalized temperature (ΔT_N). As shown in Figure 13, when a short focus transducer is used, the temperature increase at a position about ± 10 mm away from the focal point is 60 % or less of that focal point. That is, when the maximum temperature at the focal point is 80 °C, the temperature at the distance (± 10 mm) is approximately 50 °C. However, when a long focus transducer is used, the temperature at the same distance is approximately 60 °C, and the difference in temperature between them is 10 °C. From this result, careful attention is required at the boundary between the healthy tissue and the tumor tissue during HIFU treatments.

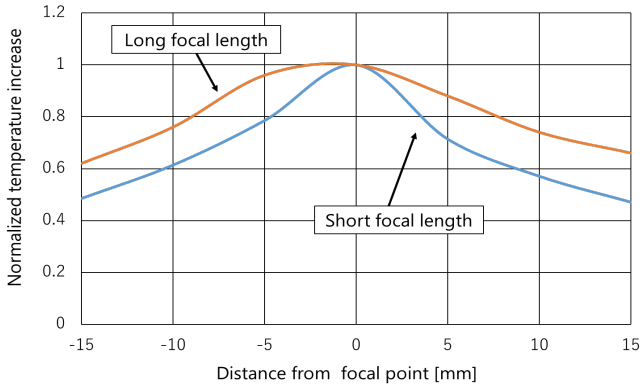


Figure 13: Normalized Temperature Profiles.

5. Conclusion

In this paper, we investigated the possibility of non-invasive measurement of the three-dimensional temperature distribution of the human shaped agar phantom during focused ultrasound heating from ultrasound images before and after heating. First, an algorithm of temperature distribution measurement using ultrasound images was presented. Second, the relationship between the temperature increase inside the heated object and the displacement of the ultrasound images before and after heating was described. Third, we measured the temperature distribution inside the agar phantom during HIFU heating using two types of transducers with different focal lengths. In the temperature measurement results, at the central cross section of the agar phantom, the error between the results of the proposed method using ultrasound images and the measurement results of the infrared thermal camera was approximately 1 °C. In addition, in order to discuss the heating characteristics of two types of transducers with different focal lengths, their three-dimensional temperature distributions were compared. Finally, we highlighted the problems in HIFU treatments using transducers with different focal lengths.

In the future, we have a plan to measure the three-dimensional temperature distribution inside heated animals and to discuss the problems of HIFU treatments.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Part of this research is supported by JSPS KAKENHI Grant Number JP18H03553.

References

[1] R. Sakakibara, K. Kato, P.K. Choi, A. Takeuchi, "Basic study of 3-D Non-invasive measurement of temperature distributions from ultrasound images

during hifu treatments," in Proceedings of the 2019 21st International Conference on Electromagnetics in Advanced Applications, ICEAA 2019, 2019, doi:10.1109/ICEAA.2019.8879194.

[2] S. Vaezy, R.W. Martin, U. Schmiedl, M. Caps, S. Taylor, K. Beach, S. Carter, P. Kaczkowski, G. Keilman, S. Helton, W. Chandler, P. Mourad, M. Rice, R. Roy, L. Crum, "Liver hemostasis using high-intensity focused ultrasound," Ultrasound in Medicine and Biology, 1997, doi:10.1016/S0301-5629(97)00143-9.

[3] G.O.N. Oosterhof, E.B. Cornel, G.A.H.J. Smits, F.M.J. Debruyne, J.A. Schalken, "Influence of high-intensity focused ultrasound on the development of metastases," European Urology, 1997, doi:10.1159/000480887.

[4] N. McDannold, Quantitative MRI-based temperature mapping based on the proton resonant frequency shift: Review of validation studies, International Journal of Hyperthermia, 2005, doi:10.1080/02656730500096073.

[5] K. Hynynen, "Feasibility of using ultrasound phased arrays for MRI monitored noninvasive surgery," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 1996, doi:10.1109/58.542049.

[6] J. Song, A. Pulkkinen, Y. Huang, K. Hynynen, "Investigation of standing-wave formation in a human skull for a clinical prototype of a large-aperture, transcranial mr-guided focused ultrasound (MRgFUS) phased array: An experimental and simulation study," IEEE Transactions on Biomedical Engineering, 2012, doi:10.1109/TBME.2011.2174057.

[7] C. Simon, P. Vanbaren, E.S. Ebbini, "Two-dimensional temperature estimation using diagnostic ultrasound," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 1998, doi:10.1109/58.710592.

[8] Y. ISEKI, K. NAKAMURA, D. ANAN, K. KATO, "Development of Ultrasound-Guided Resonant Cavity Applicator System: - Non-invasive Measurement of Temperature Distributions -," Japanese Journal of Hyperthermic Oncology, 2015, doi:10.3191/thermalmed.31.13.

[9] Y. ISEKI, D. ANAN, T. SAITO, Y. SHINDO, F. IKUTA, K. TAKAHASHI, K. KATO, "Non-Invasive Measurement of Temperature Distributions During Hyperthermia Treatments using Ultrasound B-mode Images," Thermal Medicine, 2016, doi:10.3191/thermalmed.32.17.

[10] E. MIURA, K. KATO, A. TAKEUCHI, "Basic Study of 3-D Non-Invasive Measurement of Temperature Distribution Using Ultrasound Images during FUS Heating (article in Japanese with an abstract in English)," Thermal Medicine, 2018, doi:10.3191/thermalmed.34.1.

[11] A.M. Pouch, T.W. Cary, S.M. Schultz, C.M. Sehgal, "In vivo noninvasive temperature measurement by B-mode ultrasound imaging," Journal of Ultrasound in Medicine, 2010, doi:10.7863/jum.2010.29.11.1595.

[12] M.D. Abolhassani, A. Norouzy, A. Takavar, H. Ghanaati, "Noninvasive temperature estimation using sonographic digital images," Journal of Ultrasound in Medicine, 2007, doi:10.7863/jum.2007.26.2.215.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1312-1321 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Wideband Active Switch for Electronic Warfare System Applications

Mahadev Sarkar*, Gaurav Anand, Sivakumar Ramadoss

Microwave-millimeter wave component Design Group, Product Development and Innovation Centre, Bharat Electronics Ltd. Bangalore,560013, Karnataka, India

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 23 November, 2020
Online: 16 December, 2020

Keywords:
Active SP2T Switch
Wideband SPDT
Broadband discrete switch
High speed SPDT switch
CPWG matching
Broadband equalizer
NAND Gate driver
Power Supply sequencing

ABSTRACT

Discrete component based ultra-wideband SPDT switch modelling, design and implementation challenges are reported in this article. The basic SPDT switch is planned to design using PIN diodes (bare die). Nonlinear parameters extraction steps of PIN diode from datasheets have been proposed in this article instead of using readymade nonlinear PIN diode models. Equalizers are designed as per requirement using discrete components and quarter wave transmission lines for different source and load impedances. Then all the designed components are integrated with a low noise amplifier (LNA). Extraction of PIN diodes' nonlinear parameters are carried out using diode equations along with datasheet parameters. Layout has been implemented using microstrip and CPWG techniques. Quickly responding driver is designed using transistors and logic gates. CPWG based layout achieves better isolation compared to microstrip line-based layout at highest frequencies. Measured and simulated responses are tallied and they are matching together respectively except few exceptions. Measured responses also validates the well agreement between the proposed nonlinear diode modelling and equalizer design with different input and output impedances.

1. Introduction

Modern miniaturized systems require compact, light-weight and low cost components with high yield rate [1]. Excellent ultra wideband electronic warfare (EW) receiver like RWR (Radar Warning Receiver) system requires compact, lightweight and faster SPDT (single pole double through) switch which will be used to switch between BITE and RF ports. This active SPDT switch is basically used to control the RF port isolations while examining the whole Radar in BITE (built in test) mode. As the operating bandwidth of this module is very wide, microwave devices always struggle with uncertain gain gradients and ripples. Thus, excellent class wideband receiver of an RWR system are placed far away from its distributed antenna arrangement systems and connected through a lengthy cable. Thereby, it experiences enormous gain flatness problems particularly at higher frequencies due to excessive transmission line losses. To maintain the best flatness response, it is important to compensate the gain by flattening the gain response over the operating frequency band. It is commonly carried out by using passive equalizers. This article describes various design and development concerns/ problems etc. of wideband discrete components like SPDT switch, passive equalizers and the active SPDT switch. This article also shows a detailed method of nonlinear PIN diode model parameter extraction. Then a suitable PIN diode model is made using those extracted nonlinear parameters and has been used to design the PIN diode based SPDT switch. A suitable wideband LNA (Low Noise Amplifier) has also been introduced at the output path to compensate the loss of the SPDT switch. Prior art survey is carried out over design and development procedures of SPDT switches [2-9]. Allanic et. al. proposed a new passive SPDT design using silicon substrate in [2]. It is band limited to 5GHz with unequal losses in Tx and Rx ports. Additional amplifier is required to make it active SPDT. A high-power passive SPDT design is reported in [3]. Here also an additional LNA is required to make it active switch. Yang et. al. proposed an active SPST switch in [4]. This proposed SPST is band limited to 4GHz using InGaP/GaAs HBT process. This proposed switch consists of seven transistors. This SPST (single pole single through) switch is active in nature but does not cover wide frequency band. A high isolation active absorptive SPST is reported in [5]. It is band limited to 10GHz. Values of various parameters like, minimum gain, gain flatness, isolations and VSWR are -0.4dB, ±1.5dB, 35dB and 3.6:1 max respectively. Single chip SiGe BiCMOS based active SPDT and SPST designs are reported in [6] respectively. This design is also

*Corresponding Author: Mahadev Sarkar, 8553551075, mahadevsarkar@bel.co.in

suffering from P1dB, gain, VSWR etc. Active wideband SPDT switch designs are proposed for V, D and Ku bands in [7–9] respectively. These proposed active SPDT switches are also suffering from gain, gain flatness, input output matching and isolations respectively. These reported designs are also lacking about P1dB and switching speed information except [9] (it provides only P1dB data). Present article uses a discrete LNA and integrates with PIN diode based SPDT switch. This article proposes nonlinear diode parameter extraction steps with an example using limited source of data (datasheet) and diode equations. Readymade nonlinear diode models have not been taken into consideration as vendor often do not extends support. Suitable diode models have been made using extracted nonlinear parameters and tuned up to 18GHz. Then those models have been used to design broadband switch. Equalizers are also designed separately with different input/output impedances to cope up with flatness issues. Only logic gates and discrete active components are used to design the driver where, NAND gate translates input logic to output logic control signals and transistor builds current over it. Active SPDT switch integrates separately designed SPDT, gain equalizers, LNA, attenuation pads, driver etc. Active SPDT switch was designed using microstrip lines initially, but isolation was not achieved 50dB. Then, co-planner waveguide (CPWG) approach is tried with one more extra shunt element for both the paths. This configuration achieves >50dB isolation at 18GHz. Switch isolation of microstrip line and CPWG based method is compared in the article. Switching speed of any switch is directly related to the signal throughput of the driver. Thus, how fast the driver acts switching speed will also be so faster. That is why it will be a combination of driver and switch's performance respectively [10, 11].

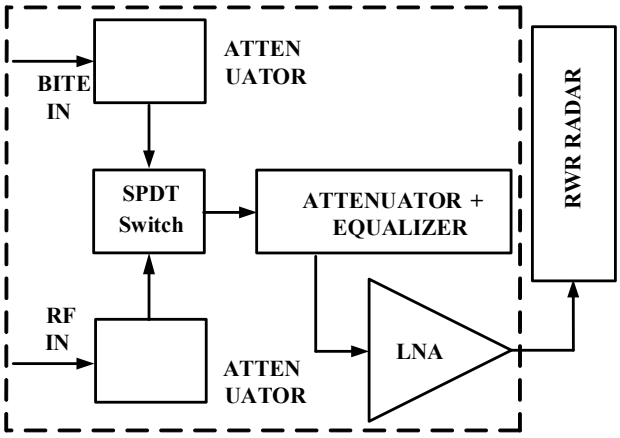


Figure 1: Block diagram of ultra-wideband high speed active SPDT Switch consists of basic SPDT, equalizers, attenuator pads and LNA

This present work advances P1dB performances, power dissipation, better input/output matching and superior packaging with reduced weight respectively from the previous work reported in [1]. Figure 1 shows the block diagram of active SPDT switch where, BITE and RF-in are the two ports of the SPDT switch. Attenuator pad/equalizer are kept at the input of the switch for better matching [1]. It increases noise figure very little as compared to the previous situation where, active switch was absent. An appropriate LNA is integrated just at the output port of the discrete SPDT switch. Equalizers are used before and after LNA to keep flatness within the specifications. Measured and simulated responses are compared in this article and analyzed the probable cause of discrepancy. An appropriate reason is explained for this discrepancy and conclude about limitations of the nonlinear diode modelling. This article has been organized as follows. In Section 2, PIN diode parameters extraction, diode modelling, SPDT switch, equalizer and driver design are discussed. In Section 3, design implementation and experimentation are shown along with simulated results. In Section 4, fabricated switch is shown and the conclusions are given in Section 5.

2. Individual Components Design

This section explains about design steps of three important components design likely, nonlinear parameter extraction of PIN diodes, modelling and comparison with measured response (plotted in the datasheets), equalizer design and driver design respectively. The major work is involved in nonlinear parameter extraction of PIN diodes from datasheet and then prepare a suitable model out of it. Here this extraction steps will be explained in details. PIN diode modelling is carried out both for the diodes used in series and shunt branches of SPDT switch. Discrete component based equalizer design explained in the next and at last NAND gate based driver design also shown in details.

2.1. *Nonlinear parameter extraction and PIN Diode Modelling*

SPDT block of the active switch shown in Figure 1, has been designed using PIN diode bare dies. Simple nonlinear model of PIN diode is costly and most often component vendors' do not support with exact nonlinear model parameter values in the datasheets. Thus it is necessary to make nonlinear model of PIN diode. It is possible to prepare nonlinear model of PIN diodes using data given in the datasheet [12]. Thus it has become designer's capability to build an acceptable model for diodes using existing data in the datasheet. Extraction of all nonlinear model parameters are extremely difficult using limited source. Approximate (mere accurate) model is possible to make using these extracted parameters by fine tuning other model parameters and that can tackle the issue [12].

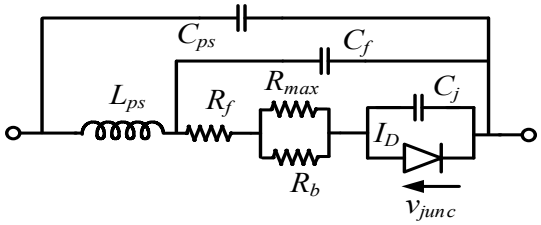


Figure 2: Classic diode model used to design circuit model of PIN diode

Adopting the above technique separate models have been prepared for shunt and series diodes respectively and compared with its measured responses. Figure 2 shows the basic diode model to be used to make equivalent PIN diode circuit model where, C_{ps}, C_f, C_j, L_{ps}, R_f, R_b, R_{max} I_D and V_{junc}, are package parasitic capacitance, beam lead parasitic capacitance, junction capacitance, package parasitic inductance, minimum resistance when diode is in forward bias, variable *i* region resistance on PIN diode, maximum resistance, diode current and voltage across the junction respectively. Equation mentioned below (in 1) describes

the current flowing through the diode where, I_s, q, n, k, T_j and I_{RBD} are, saturation current, electron charge, ideality factor, Boltzmann constant, temperature in kelvin and reverse breakdown current respectively [13].

$$I_D = I_s\left(e^{\frac{qV_{junc}}{nkT_j}} - 1\right) - I_{RBD} \quad (1)$$

and

$$I_{RBD} = I_{RBDS}e^{\frac{-q(V_{junc}+V_{RBD})}{kT_j}} \quad (2)$$

where, V_{RBD} and I_{RBDS} are reverse breakdown voltage and saturation current respectively. Datasheet does not provide and package parasitic like, C_{ps}, C_f and L_{ps} and to describe the bonding pad paracitics. Thus, it is very difficult to extract all the model parameters from manufacturers' data sheets, which only provide limited information. Fortunately, most of the critical parameters like, I_s, R_f, C_j, τ, α_c, δ etc. can be derived from the manufacturers' data sheets where, τ, α and ζ are transition time, arbitrary constant and built in potential respectively [14]. Datasheet provides the I-R and C-V characteristics of diodes. Thus, all the above parameters can be extracted using these characteristic curves.

$$\alpha_c = \frac{qV_{junc}}{nkT_j} \quad (3)$$

$$C_j = \begin{cases} C_{J0}\left(1 - \frac{V_{junc}}{\delta}\right)^{-\gamma(V_{junc})} + C_D & V_{junc} \le \mathrm{F}_c\mathrm{X}\delta \\ C_{J0}\mathrm{Xf}(V_{junc}) + C_D & V_{junc} > \mathrm{F}_c\mathrm{X}\delta \end{cases} \quad (4)$$

where C_D, C_{J0}, F_c and γ are capacitance due to the finite transit time (τ) through the *i*-region, depletion capacitance at $Vjunc = 0$, function of $Vjunc$, coefficient to divide the forward-bias, depletion capacitance into two regions and expressed by a 3rd order polynomial.

$$f(V_{junc}) = (1 - F_c)^{-(1+\gamma(V_{junc}))}\ \mathrm{X}\ \left\{1 - F_c\left(1 + \gamma(V_{junc})\right) + \gamma(V_{junc})\frac{V_{junc}}{\delta}\right\} \quad (5)$$

$$\gamma(V_{junc}) = \gamma_0 + \rho_1 V_{junc} + \rho_2 V_{junc}^2 + \rho_3 V_{junc}^3 \quad (6)$$

$$C_D = \tau\frac{\partial I_D}{\partial V_{junc}} \approx \frac{\tau I_D}{0.0256n} \text{ at } T_j = 300\mathrm{K} \quad (7)$$

The resistance in the i-region is a function of the forward current. At high forward current level, *Rb* becomes close to zero. Thereby the minimum resistance becomes *Rf*. At reverse bias condition, *Rb* is almost infinite, therefore the maximum resistance is *Rf* +*Rmax*. The resistance at low and moderate forward current level is written in (6) [14].

$$R_{Total} = \mathrm{R}_f + \frac{R_b X R_{max}}{R_b + R_{max}} \quad (8)$$

$$R_b = \begin{cases} \frac{\beta_1}{{I_D}^{\beta_2}} & I_D > 0 \\ \infty & I_D \le 0 \end{cases} \quad (9)$$

Variable resistance can be found by putting in (9) in (8) considering β_2=1, where R_m is the arbitrary RF resistance at current level I_m.

$$\beta_1 = I_m(R_m - R_f) \quad (10)$$

Table 1 shows the extracted parameters considering HPND4018 PIN diode datasheet using (1) to (10). Here few parameters are directly taken from datasheet, like forward resistance, breakdown voltage, junction capacitance, transition time etc. Few other parameters have been finalized using curve fitting technique to match the response with datasheet while simulation of model. PIN diode model simulation is performed in

Table 1: The extracted model parameters for HPND4018 Diode

Sl. No.	Parameters	Extracted Values
1	I_s	1.21X10^{-9} A
2	I_{RBDS}	1.0 X10^{-7} A
3	R_f	3.7 Ω @ 20mA
4	R_{max}	780 Ω
5	α_c	22. 464239
6	V_b	60 V
7	n	1.94
8	C_{J0}	0.0481 pF
9	τ	26 X10^{-9} S
10	τ_{rr} *(5V, 10mA)*	3.934 X10^{-9} S
11	δ	0.197 V
12	γ	0.221
13	β_1	1.257X10^{-2}
14	β_2	99.87X10^{-2}
15	ρ_1	1.61X10^{-4}
16	ρ_2	-1.613X10^{-5}
17	ρ_3	6.761X10^{-7}
18	C_p	0.095 pF
19	L_p	1.0 nH

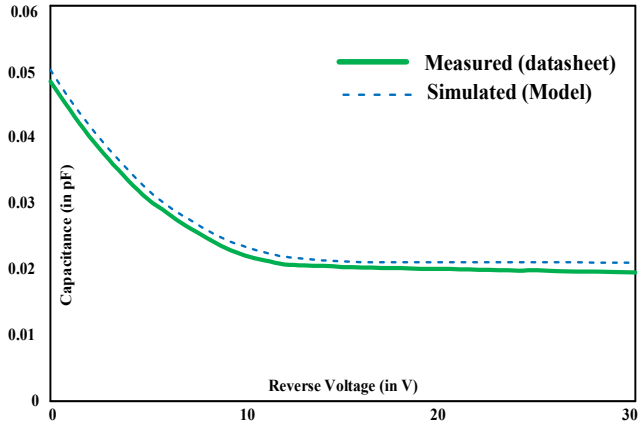


Figure 3: Comparison between modeled and measured DC C-V characteristic of HPND1084 PIN diode at 1MHz

2½D software taking parameters values from Table 1. Influence of bonding wires cannot be ignored while simulation. Thereby model is extracted including them. Tuning of the required model parameter values are carried out to match the model characteristics virtually similar as compared to the datasheet.

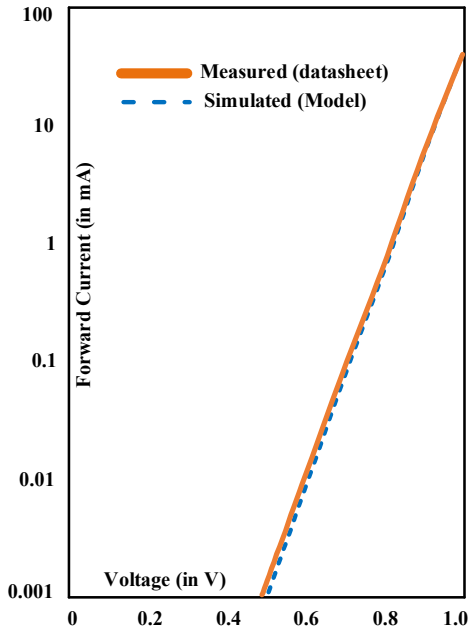


Figure 4: Comparison between modeled and measured DC I-V characteristic of HPND1084 PIN diode

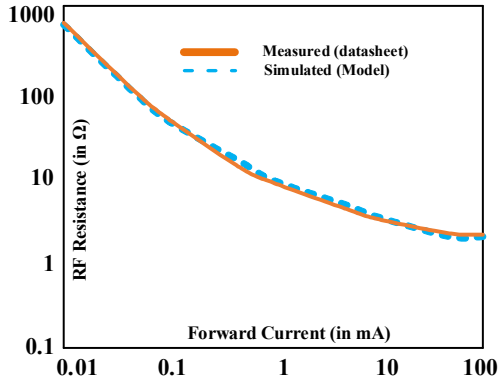


Figure 5: Comparison between modeled and measured RF resistance – Forward current at 100 MHz characteristic of HPND1084 PIN diode at 100 MHz

These models are purely made based on datasheet parameters and used for specific applications. Figure 3 shows the comparison between modeled and measured DC C-V characteristic of a HPND1084 PIN diode. Here modeled capacitance almost optimized as measured capacitance values. Figure 4 shows the comparison between modeled and measured DC I-V characteristic of the same PIN diode. Modeled current is showing a good agreement with measured current. Figure 5 shows the comparison between modeled and measured RF resistance characteristic. Here simulated resistance indicates that model parameters are optimized properly. Nonlinear model of MP6002 has also been carried out by following the above procedure. These two diode models are used for SPDT switch design, where, HPND1084 is used as series diode and MP6002 is used as shunt diode. Figure 6 shows the response of a modeled PIN diode (HPND1084) for 1– 20 GHz bandwidth. Maximum insertion loss is below 1dB at 18 GHz and it is well matched up to 18 GHz. Isolation is around 20dB at 18 GHz. The same loss characteristic is also calculated for shunt diode and kept ready for next level simulation. Here diode models are tuned several times to make it suitable and well matched with datasheet characteristics.

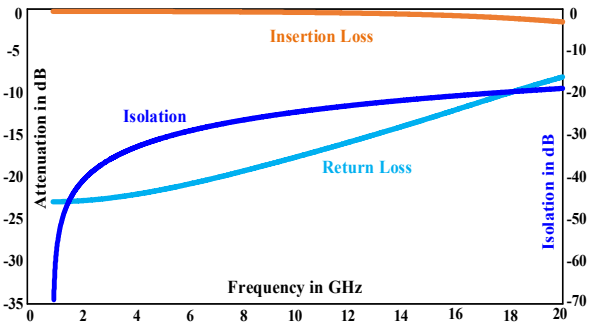


Figure 6: Simulated response of insertion loss, return loss and isolation characteristic of HPND1084 PIN diode

2.2. *SPDT Switch Design using PIN Diodes*

Simulated and verified nonlinear PIN diode models are used to design the SPDT switch where, HPND1084 is used in the series branch and MP6002 is used in the shunt branch. Schematic design

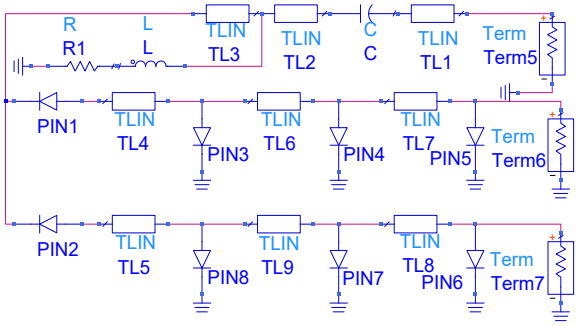


Figure 7: Schematics of basic SPDT switch using PIN diodes

of the switch is performed in 2½D software and shown in the Figure 7. Each branches of the switch having one series and three shunt diodes respectively. Here DC current return path is getting completed through inductor L and resistor R1 in the common path. Resistor R1 is used to limit the maximum current through the series diode. Bonding wires (2mil gold ribbon) have not been added in the schematics as it is already considered as parasitics in the diode modelling. The insertion loss (I_L) and isolation (I_{iso}) parameters of the SPDT switch are estimated using (11) and (12) respectively. Here, R_{fse}, R_{fsh}, Z_0, and X_C are forward resistance of series and shunt diode, characteristic impedance, and reactive impedance respectively. Tendency of the simulated isolation response is fluctuating around middle of the frequency band which will be recompensed in final layout simulation [14].

$$I_L = 10log\left[\left(1+\frac{R_{fse}}{2Z_0}\right)+\sum_{j=1}^{n}\left(\frac{Z_0+R_{fsh_j}}{2X_{C_j}}\right)\right] \quad (11)$$

where, $X_C = 1/2\pi f_L C_D$, $f_L = 1/2\pi\rho\varepsilon$, here ρ and Ɛ are the resistivity of the I region and dielectric constant of the silicon.

$$I_{iso} = 10log\left[\sum_{j=1}^{n}\left(1+\frac{Z_0}{2R_{fsh_j}}\right)^2 X\left(\frac{X_{C_j}}{2R_{fsh_j}}\right)^2 + \left(1+\frac{Z_0}{2R_{fse}}\right)^2\right] \quad (12)$$

Reverse breakdown voltage, power handling, switching speed, capacitance, forward resistance, carrier lifetime, thermal resistance etc. have been kept in mind while selecting diodes.

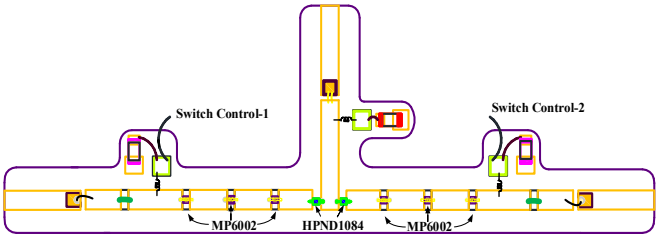


Figure 8: Proposed layout diagram of basic SPDT switch using PIN diodes

Final assembled layout of the proposed basic SPDT switch is shown in Figure 8. Here extra shunt diodes assembly option is kept in the layout so that fabrication iterations become less but, it has a disadvantage of higher insertion losses. This option will not affect at the overall performance as LNA is in the common path to compensate the extra losses. Initially, SPDT layout was carried

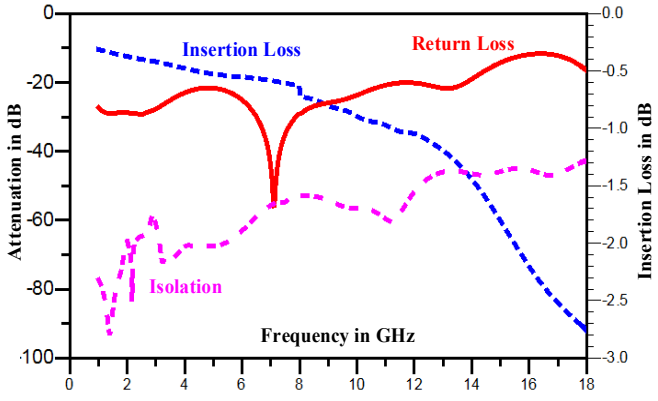


Figure 9: Simulated responses of proposed SPDT switch in 2½D software

out using microstrip lines and fabricated but, it was not meeting isolation specifications at higher frequencies (measured response is shown in the experimentation section). Thereby final layout of SPDT switch is made using CPWG lines in the first iteration. Figure 9 shows the simulated responses of the SPDT switch. Here insertion loss is about 2.8dB around 18GHz. Return loss is more than 20dB up to 14 GHz and it reduces to around 12 dB at 16–18GHz. Isolation is 50dB at 18 GHz. Now it is ready for next level integration with equalizer and LNA.

2.3. *Equalizer Design*

Equalizers are crucial parts for equalizing flatness of the module which, deteriorates with frequencies moving higher and higher. It is evident from Figure 9 that almost 2.5dB slope adjustment is must to flatten the switch response. Now there is an LNA in the common path that has its own gain slope (almost 2dB). Thus requirement specific equalizers are necessary in the various locations (common path as well as in the SPDT branches) to

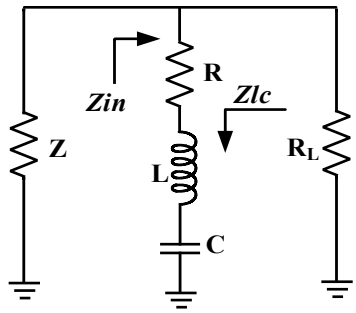


Figure 10: Schematic diagram of single section equalizer

compensate all the loss/gain slopes. Figure 10 shows elementary schematic diagram of a single resonator based equalizer [1]. This equalizer is comprising of discrete thin film resistor and a resonator of $\lambda_g/4$ length at the highest frequency. Slope of the equalizer has been finalized after analyzing the integrated switch response in the system level simulations. Thereby case to case final component twigging will be performed as per need basis. The ultimate discrete component values can be finalized at the final level of simulations to maintain tiniest gain flatness (within ±0.5dB). Driving point impedance of the R-L-C network can be calculated using (13) where, R, L and C are the shunt resistance, inductance and capacitance respectively [15-17].

$$|Z_{in}| = \sqrt{R^2 + \left|\frac{1-\omega^2 LC}{\omega C}\right|^2} \quad (13)$$

Final transfer characteristics of the equalizer containing n sections have been calculated by (14) where, Z_L, ω is the load impedance and angular frequency respectively where, $\omega = 2\pi f$ and f is the operating frequency.

$$H_{total}(\omega) = \sum_{i=1}^{i=n} \frac{2(j\omega C_i R_i - \omega^2 L_i C_i + 1)}{j\omega C_i Z_L + 2(j\omega C_i R_i - \omega^2 L_i C_i + 1)} \quad (14)$$

Although (14) has been calculated for n number of sections but the actual number of equalizer sections may be changed/decided according to flatness characteristics at the final

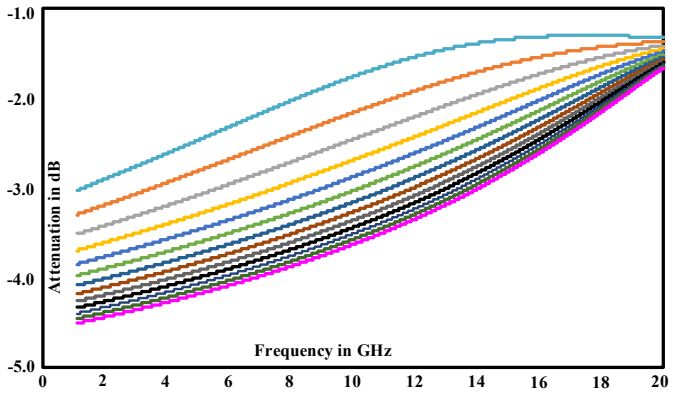


Figure 11: Insertion loss response of equalizer for different resistor values.

level of realization. Figure 11 shows insertion loss characteristics of passive equalizer varying numerous resistor (R) values where, upper and lower response curves are for lower and higher resistor values respectively. Now when it interacts with output impedance of SPDT switch as input impedance of it and input impedance of

LNA as arbitrary load then its input and output port matching also changes respectively. Figure 12 shows the input and output return loss characteristics of the equalizer for numerous R values. Here source impedance is almost 50Ω but load impedance is the input impedance of the next stage thus, input and output return loss response is not symmetric. In the system simulations various resistor values have been tuned to achieve different slopes.

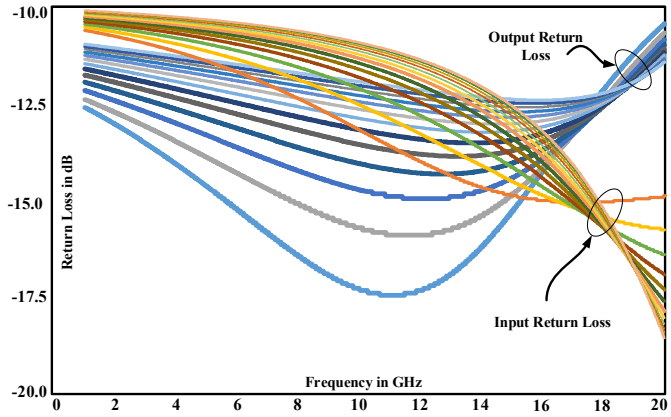


Figure 12: Input and output matching of equalizer for different resistor values

2.4. *Driver Design*

A suitable driver is designed to generate the required control signals to operate the switch and LNA (supply sequencing) from external control signals. As it is discussed that the switching speed is a combination of driver and switch's performance respectively [1]. Thereby design of appropriate driver is very crucial to achieve the best switching speed. Sequencing circuit is also required to feed supplies to the LNA. Here driver and sequencing circuit design is described in details. Logic gates, transistors, diodes etc. are used to design sequencing and driver circuit. External logic is passed through high throughput NAND gate and it passes through combination of transistors, zener diodes and resistor combinations to produce the output. Schematics level simulation of the driver is carried out using spice simulator tool. The design has also been verified by simulating the same circuit in 2½D simulator. The outcome of both the simulator tools are compared and they are very much satisfying**.** Figure 13 shows schematic diagram of a driver circuit comprises of logic gate (74LS00; dual-NAND gate) and zener diode (BGV90-C4V3) followed by transistors (current booster) and resistors in various combinations [1]. Simulated driver response is illustrated in the same Figure 13. In the Figure 13 the red color bit streams indicate the input control bits. The blue and violet color bit streams represent the output control–1 and control–2 bit streams respectively. These two output control bit streams are being used to alter the active switch RF and BITE paths respectively. It is also evident from the figure that, both the controls (control–1 and –2) are out of phase.

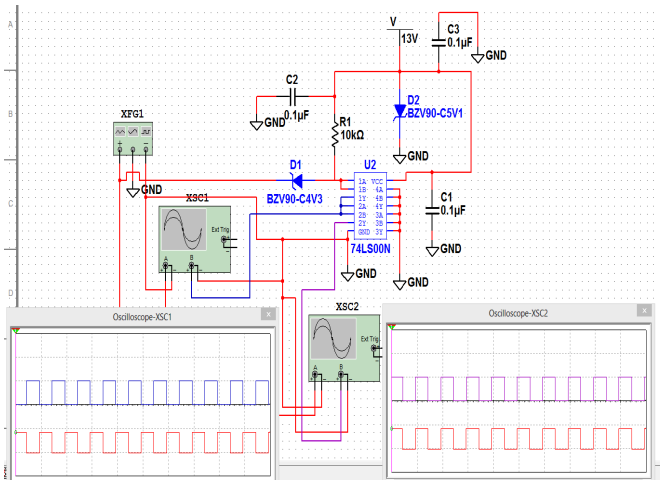


Figure 13: Schematic diagram and responses of switch driver circuit

3. Active SPDT Switch Design

Active SPDT switch design has finally been performed by employing scattering parameter (.sNp) files of all the individually designed modules. RF system analysis depends upon the precise design of all the discrete modules. Hence it will meet all the necessary specifications across all the environmental conditions over an extensive frequency spectrum. Layout is made using CPWG lines as basic SPDT switch. This technique has been adopted to rectify isolation and insertion loss characteristics of the channel respectively. The final schematic level simulation is performed using 2½ D simulator tool. Schematic diagram of the active SPDT switch for forward and reverse (isolation) channel respectively are shown in Figure 14 [1]. Tapers used to interconnect two different impedance lines are not CPWG lines also shown in this Figure 14. Figure 15 shows the simulated blue

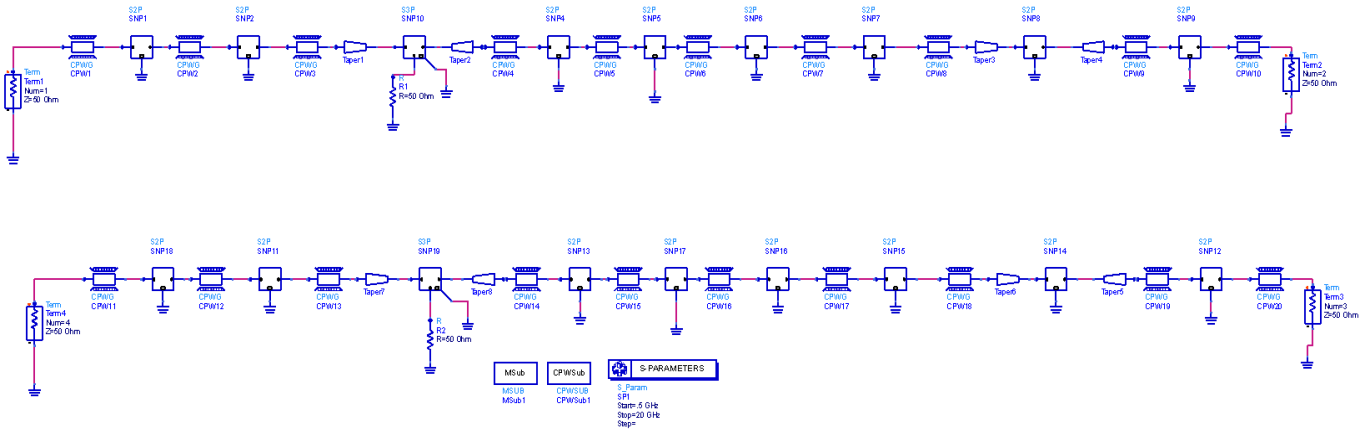


Figure 14: CPWG based schematic circuits staring from input port to output port of proposed SPDT switch for forward signal and Isolation path respectively

color insertion gain curve and saffron color isolation curves of the proposed active SPDT switch. Here the overall gain about 2 dB is achieved with flatness within ±0.5dB. More than 50 dB isolation has successfully been met at 18 GHz which is very crucial parameter. Almost 95dB isolations are also met at the lower frequencies. Appropriate LNA (die), discrete component attenuators and readily available (designed) equalizers are used in the output track of the active SPDT to match the gain flatness. Simulated VSWR response is shown in Figure 16. Budget analysis has also been performed at -7 dBm input power level and at mid-band frequency (10 GHz) using 2½D software. Parameters like isolation, dynamic range, SFDR (spurious free dynamic range) and cascaded P1 dB had been the prime concern and they have attained 54.4 dB, 64 dB, 53.24 dB and 20 .7 dBm respectively. 3D EM simulation is carried out for the layout and mechanical channels respectively. It ensures any cavity resonance is present up to 18GHz or not. Ultrathin soft substrates have been used for fabrications. Driver and control circuit is made over FR4 material. The SP2T switch and driver/control circuits are mounted on opposite sides of the same housing respectively. Switch control, sequencing and other supply connections are taken from driver circuit board to switch side through small holes, keeping in mind the EMI/EMC effects. Appropriate mechanical channeling for RF circuits are important for achieving expected responses. Microwave channel dimensions are kept narrow enough to avoid higher mode resonance (excitations), as it disturbs characteristics of the module at higher frequencies. Carrier plate are made to ensure robust ground effects at higher frequencies and very good mechanical support to the PCBs. 0.4 mm kover and copper-molly sheets having nicely nickel plated with minimum impurities have been used to make carrier plates [19].

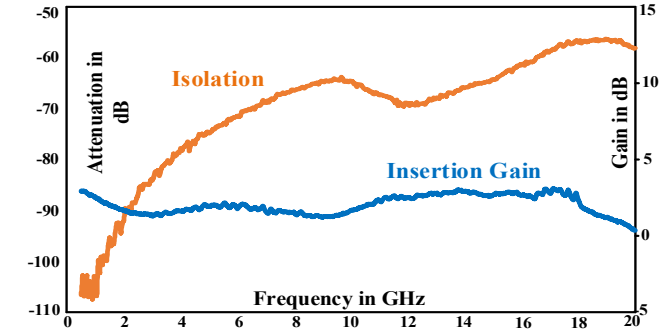


Figure 15: Gain and isolation simulation response of proposed SPDT switch in the transmission and isolation paths respectively

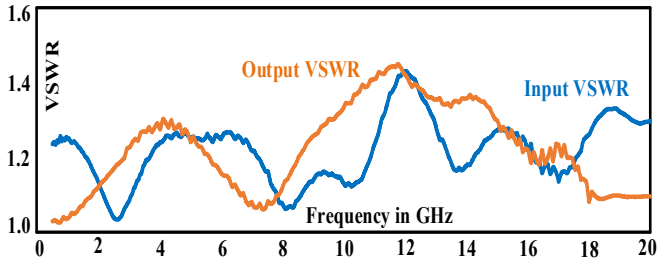


Figure 16: Input and output VSWR simulation response at input and one of the output ports of proposed SPDT switch

4. Experimentation

Active broadband SPDT switches have been developed using microstrip as well as CPWG lines respectively. Initially microstrip line based SP2T is developed and then CPWG line based switch due to noncompliance of isolation specification. The realized active SPDT switches are evaluated separately. Figure 17 shows the isolation responses of SPDT switch of microstrip line based approach. In this approach microwave channeling and air cavity sizes were not taken care precisely thereby output responses like isolation characteristics affected drastically. It is evident from the figure that isolation at 18 GHz is not meeting the requirement of 50dB. It is actually around 42dBc compared to gain plot. Gain plot of the microstrip line based SPDT is not shown as gain plot is identical to CPWG approach based SPDT. Comparison between measured and simulated responses of gain and isolation of CPWG approached based SPDT is shown in Figure 18 respectively. Here all measured and simulated responses are shown in firm and dotted lines respectively. Overall insertion gain met in this approached is 2.5dB except few frequencies. It is in good agreement with simulation responses. Each arm of this switch contains of sufficient gain equalizers to fit the gain flatness within ±0.4dB (maximum) over the entire spectrum at 27°C (Room Temperature; RT) as well as ±0.6 dB (maximum) across the wide temperature range -40°C to 71°C. Figure 18 also depicts the isolation response at RT which reached minimum 53dB (55dBc compared to gain response) in CPWG based approach.

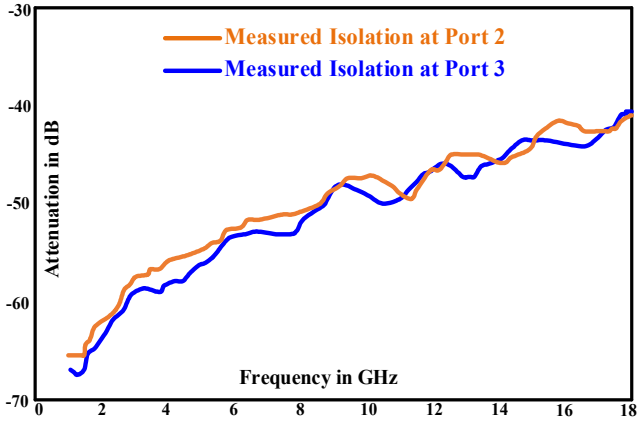


Figure 17: Measured isolation responses between common port and port-1 and port-2 of the developed SPDT switch following microstrip line approach

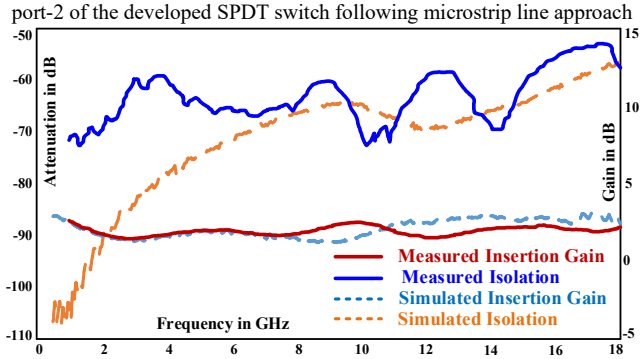


Figure 18: Comparison between measured and simulated responses of gain and isolation of common port – port-1 and common port – port-2 of the developed SPDT switch following CPWG line approach

Any input port out of two input ports of the active SPDT switch can be accessed at any instant of time according to the external logic signal. This process achieves ON/OFF input output port isolation of 55dBc at the highest frequency and more than 60dBc across 1–15.5 GHz frequency. Tested isolation response is

comparable with simulated response beyond 6GHz (6–18GHz). From 1–6GHz measured isolation is not matching to the simulation response due to more idealistic diode characteristics of the PIN diodes. Comparison between measured and simulated responses of VSWR of CPWG approached based SPDT is shown in Figure 19 respectively. The measured VSWR is 1.45:1 and 1.35:1 at the input and output ports respectively. Measured VSWR responses are in very good agreement of simulation responses. Better than 1.7:1 VSWR is measured across frequency band under ESS. About 50nS switching speed is measured with a very good rise and fall time characteristics.

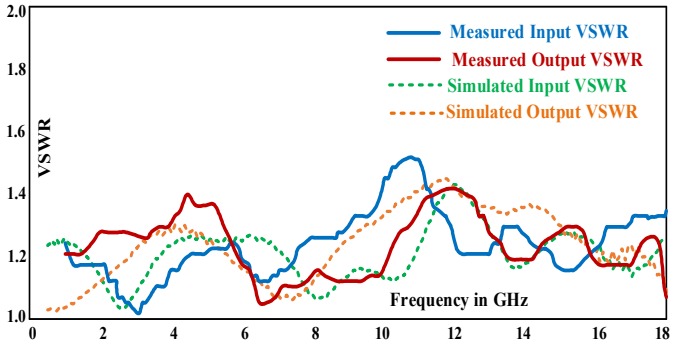


Figure 19: Comparison between measured and simulated responses of gain and isolation of common port – port-1 and common port – port-2 of the developed SPDT switch following CPWG line approach

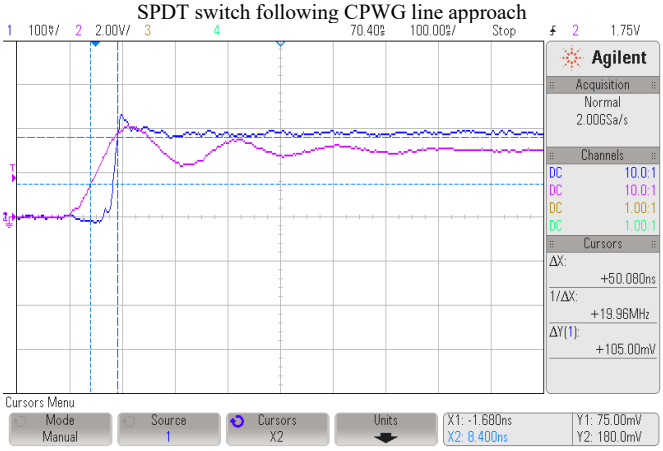


Figure 20: Response of measured switching speed at the rising edge of the developed CPWG approach based active SPDT switch

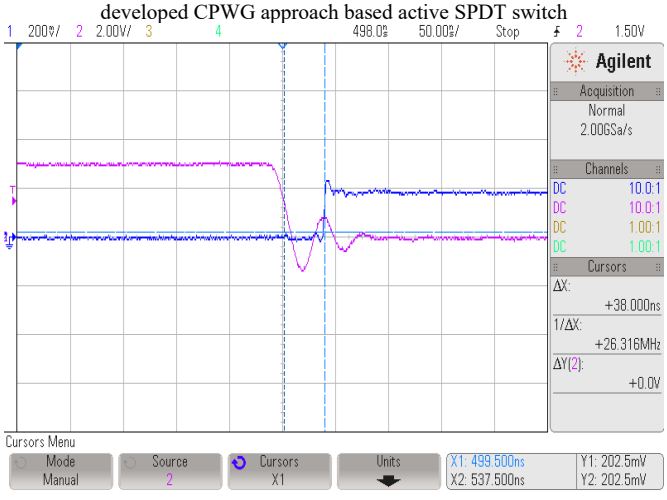


Figure 21: Response of measured switching speed at the trailing edge of the developed CPWG approach based active SPDT switch

Figure 20 and 21 describes the tested switching speed responses at rising and trailing edges against the input logic signal respectively. Here input switch control pulse is in 'volt' scale and detected output pulse in 'millivolt' scale although voltage scaling factor of input switch control and detected output pulses respectively is same (10:1). The detected output pulse will always be less than the input pulses as the input power level is fixed at -3dBm for switching speed measurement. This happens due to the absolute power level of the output pulses which is almost -1.5 – -1.0dBm (188–200mV) at 10GHz frequency and measured against a high value resistor (120Ω) is also shown in Figure 22 [18]. Thus final detected voltage after detector is very less compared to the external TTL signal. Achieved switch speed is 50nS including all RF and digital circuits. This module exhibits input P1 dB is >20 dB compared to prior work reported in [1]. This is achieved by replacing LNA with higher P1dB LNA, rectifying input/output port matching and twigging equalizer losses respectively. Developed active SPDT module also displays tremendous harmonics characteristics. Measured harmonics levels for 0dBm and 8dBm input signal for the active SPDT are 34dBc and 27dBc respectively. temperature cycling (TC) has been performed over -40° to +71° C and 3-axes random pre and post TC vibrations respectively.

Sl. no	Parameters	Specifications	Min	Max
1	Gain	2 ±2 dB	1.8 dB	2.3 dB
2	Isolation	50 dB min	54 dB	73.1 dB
3	I/P VSWR	2:1 (Max)	1.2	1.3
4	O/P VSWR	2:1 (Max)	1.2	1.5
5	P1 dB (I/P)	>10 dBm	>20 dBm	
6	2nd Harmonic	-25 dBc @ 0 dBm input -15 dBc @ 8 dBm input	-34 dBc -27 dBc	
7	Power handling	>1 watt	>1 W	
8	DC Power supply	-13V (±3%, -1%)/ 10 mA +13 (±1%, -3%)/ 50 mA	1 mA 21 mA	
9	Switching Speed	50 nS (Max)	50.08 nS	38 nS

The outcomes of environmental stress screening (ESS) of the active switch are steady. The developed module is light weight and measured weight is less than 25 grams. Figure 22 shows the

Table 3: Comparison between various parameter of recent works against present work

Ref.	Technology	f_0	BW	Gain	Isolation	Flatness	Speed	VSWR	P1dB
[2]	Sc. DDAs	5GHz	1.4 GHz	-2.38dB	43/49dB	±1.5dB	-	1.5:1	-
[3]	GaAs	30GHz	15GHz	-1.5dB	43dB	±0.35dB	4nS	1.8:1	35dBm
[4]	InGaP/GaAs HBT	22.8GHz	1.2GHz	-4.4dB	52.2dB	±3.2dB	100pS	2:1/ 3:1	-
[5]	SiGe HBT	20GHz	10GHz	3dB	<35dB	>±1.5dB	60pS	3.6:1	-
[6]	SiGe BiCMOS	22.5GHz	2.0GHz	-0.3dB	<45dB	>±1.5dB	100pS	2.3:1	0.5dBm
[7]	100nm GaAs HEMT	75.0GHz	30GHz	10dB	<15dB	>±1.0dB	-	2:1	-
[8]	100nm GaAs HEMT	145GHz	38.7GHz	3.3dB	<22dB	>±1.5dB	-	2.3:1	-
[9]	0.18µm CMOS	5.8GHz	5GHz	-1.1dB	27/40dB	>±0.3dB	-	2:1	20dBm
This work	PIN diode die	9.5GHz	17GHz	3dB	54/73dB	±0.25dB	50nS*	1.3:1 /1.5:1	>20 dBm

*switching speed has been measured including driver and others are without driver.

developed SPDT module and a classic arrangement for switching speed assessment [1]. It is clear form Figure 20 and 21, that, the switching speed is being measured between 50% of the external input control pulse and 90% or 10% of the final output pulse of the detector for rising and trailing edge respectively. Final achieved switching speed is nothing but the combination of active switch as well as doctor diode delay respectively. Thus the switching speed mentioned in the table 2 is more than the actual switching speed of the active SPDT switch. Signal delay of the detector diode has been measured and it showed almost 5nS at highest frequency. Table 3 shows a comparative study of various parameters of recent works and the present work. Switching speed shown in the above table is different from other work as additional driver delay is added to the active SPDT switch of the present work. VSWR, gain flatness, isolation etc. are excellent compared to other work listed in the table. Another significant change has happened in the DC supply circuits. Earline simple line regulators were used to produce required supply from ±13V thus power consumptions were very high (almost 1.2W) in [1]. In the present work DC-DC converters are being used to generate required supplies for active components thereby power consumptions reduce drastically (almost 270mW). Thus, it can be established that the design goal has efficiently been reached as projected in the initial calculations.

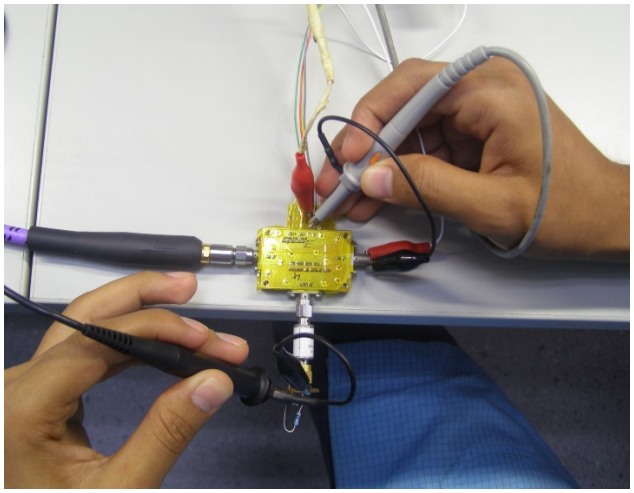


Figure 22: Measurement of switching speed by accessing input (TTL) signal to active SPDT switch and output detected pulse of the developed active SPDT switch module.

5. Conclusion

Ultra wideband active SPDT switch is designed using discrete modules like basic SPDT switch, equalizers, LNA, driver etc. Nonlinear PIN diode parameters extraction steps are shown in details using datasheet and diode equations. Nonlinear modelling is PIN diode is carried out and compared with measured responses (datasheet) for capacitance vs reverse voltage, forward current vs voltage and RF resistance vs forward current respectively. Extracted model parameters showed very good agreement with measured diode parameters. Separate modules are investigated, designed, assembled and evaluated separately. Passive equalizer with various component values are designed and finally most proper one has been selected for this active switch design. Final active switch parameters are plotted and tabulated against specification. Measured responses have been compared with simulated responses. Isolation characteristics of the developed module are not matching between simulation and measured responses from 1–5.7GHz. It happens due to limitations of PIN diode modelling. From the comparison of data, it is evident that the measured responses are in well in agreement with the simulated responses except few discrepancies. There are lots of future scopes regarding this work. LNA used in the present work has high gain which is not required as per present requirement. Thereby customized LNA can be designed to reduce power consumption further without affecting P1dB. This active SPDT switch can be compact by combining equalizers, PIN diodes, attenuators, driver and sequencing circuits as a multilayer (RF, DC and control circuit) single board solution as future work and further LNA can also be included within it using MMIC design solution as super compact light weight tiny chip.

Acknowledgment

Authors gratefully thank Mr. Manoj Jain, Head, Product Development and Innovation Centre and Mr. M.V. Rajasekhar, Director Research and Development, Bharat Electronics for their continuous inspiration to publish this work. Authors are also grateful to Microwave Super Component group, BEL for facilitating switch assembly, die bonding related processes and product design group for giving their support. They also thank all others who were involved directly or indirectly in this work.

References

[1] M. Sarkar, G. Anand, R. Sivakumar, "Light weight Active 1–18 GHz Ultra High Speed SPDT Switch for Radar Warning Receiver Systems," IEEE /MTT-S International Microwave and RF Conference (IMaRC), 1–6, 2020, doi:10.1109/IMaRC2019.9116744

[2] R. D. Allanic, L. Berre, C. Quendo, D. Chouteau, V. Grimal, D. Valente, J. Billoue, "A novel 5-GHz SPDT Switch Using Semiconductor Distributed Doped Areas," IEEE Microwave and Wireless Components Letters, **30**(4), 421-424, 2020, doi:10.1109/LMWC.2020.2978644.

[3] "24–35 GHz GaAs MMIC High Isolation SPDT Reflective PIN Switch," datasheet AP640R7–00, Alpha Industry Catalog.

[4] J. R. Yang, D. W. Kim, S. Hong "A 24 GHz Active SPST MMIC Switch with INGAP/GAAS HBTS," Microwave and Optical Technology Letters, **50**(8), 2155-2158, 2008, doi:10.1002/mop.23606

[5] I. Gresham, A. Jenkins, "A fast switching, high isolation absorptive SPST SiGe switch for 24 automotive applications", 33rd European Microwave Conference, **3**, 903–906, 2003, doi:10.1109/EUMA.2003.341108.

[6] T. M. Hancock, I. Gresham, G.M. Rebeiz, "A differential subnanosecond high-isolation absorptive active SiGe 24 GHz switch for UWB applications" Proc. IEEE Radio Frequency Integrated Circuits Symposium, 497–500, 2004, doi:10.1109/RFIC.2004.132065.

[7] A. Dyskin, S. Wagner, D. Ritter, I. Kallfass, "An Active 60–90 GHz Single Pole Double Throw Switch MMIC," Journal of Millimeter Terahz Waves, **35**, 412-417, 2014, doi:10.1007/s10762-014-0059-y.

[8] D. Muller, A. Tessmann, A. Leuther, T. Zwick, I. Kallfaaa., "Active Single Pole Double Throw Switches for D-Band Applications, IEEE/MTT-S International Microwave Symposium, 1–4, 2016, doi:10.1109/MWSYM.2016.7539969.

[9] M. Y. Chao, Z. T. Min, R. L. Chieh, K. Lin, Y. Chang, H. Wang, "Design and Analysis for a Miniature CMOS SPDT Switch Using Body-Floating Technique to Improve Power Performance," IEEE Transactions on Microwave Theory and Techniques, **54**(1), 31-39, 2006, doi:10.1109/TMTT.2005.860894.

[10] M. Sarkar, S. Jhariya, R. Sivakumar, "Correction of Nonunifirm Attenuation of Voltage Controlled Attenuator under Harsh Environment by Process Improvement" IEEE Transactions on Components, Packaging and Manufacturing Technology, **9**(8), 1632-1642, 2019, doi:10.1109/TCPMT.2019.2929093.

[11] M. Sarkar, K. L. Santha, " Designing of switch-limiter for Xband application in discrete components' topology" 9th International Radar Symposium (IRSI-2013), 1–6, 2013, https\\:radarindia.com/ProceedingdArchive/IRSI-13/13-FP-139.pdf

[12] G. Antognetti, Massobrio, Semiconductor device modelling with SPICE, McGraw-Hill, 1993.

[13] S. M. Sze, Physics of Semiconductor Devices, John Wiley & Sons, 1981.

[14] Microsemi Corporation, The PIN Diode Circuit Designers' Handbook, 1998.

[15] H. Wang, B. Yan, Z. Wang, R. Xu, "A Broadband Microwave Gain equalizer" PIER letters, **33**, 63-72, 2012, doi:10.2528/PIERL12052309.

[16] M. Sarkar, V. Kumar, R. Sivakumar, R. Manjunath, N. S. Charulu, "Performance stabilization of steep gain gradient class ultra-wideband direct conversion IFM receiver of homodyning based systems" Int. Journal of Appl. Sci. **3**(4), 27–31, 2017, issn: 2208-2182.

[17] T. F. Zhou, Y. Zhang, R. M. Xu, "Research on the millimeter wave gain equalizer," IEEE International Conference on Microwave echnology & Computational Electromagnetics (ICMTCE), 180-182, 2011, Doi:10.1109/ICMTCE.2011.5915195.

[18] M. Sarkar, V. Kumar, R. Sivakumar, S. Jhariya, B. R. Raghu, K. S. Sudheesh, "Some investigations on Some Investigations on Direct Substrate-Attachment Process for Wideband Instantaneous Frequency Measurement Receiver Development" IEEE Transactions on Components, Packaging and Manufacturing Technology, **10**(4), 704-716, 2020 doi:10.1109/TCPMT.2020.2966709.

[19] M. Sarkar, B. R. Raghu, K. S. Sudheesh, "Evaluation of Doubled-sided Substrate Attachment Technique for Round Shaped Large Metallic Housing under Airborne System Development," IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 191-194, 2020, doi:10.1109/ICCCA49541.2020.9250792.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, 1315-1328 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Evaluating the Impact of Semantic Gaps on Estimating the Similarity Using Arabic Wordnet

Mamoun Abu Helou *

Management Information System Department, Al-Istiqlal University, Jericho, 11590, Palestine

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 21 October, 2020
Online: 26 October, 2020

Keywords:
Lexical ontologies
Arabic wordnet
Semantic gaps
Semantic structure
Semantic similarity measure

ABSTRACT

Knowledge-based approach is wield used in various NLP applications. For example, to evaluate the semantic similarity between words, the semantic evidence in lexical ontologies (wordnets) is commonly used. The success of the English WordNet (EnWN) in this domain has inspired the creation of several wordnets in different languages, including the Arabic WordNet (ArWN). The English synsets have been extended to Arabic synsets through translation, which have introduced semantic gaps in ArWN structure. Therefore, compared to EnWN, ArWN has limited coverage in terms of lexical and semantic knowledge. This paper explores to what degree the richness of the wordnets' semantic structure influences the semantic evidence that can be used in wordnet-based applications, in particular the effect of filling the semantic gaps in ArWN. The paper studies the performance of applying English-based and Arabic-based similarity measures over ArWN. A set of experiments was performed by applying six path-based semantic similarity measures over Arabic benchmark dataset to investigate the usability and efficacy of the enriched structure of ArWN. The Performance measures, Person Correlation and Mean Square Error, are computed against and compared to human judgment benchmark. The obtained results demonstrate that the semantic similarity between words can be significantly improved when filling the semantic gaps. In addition, the experiment findings show that Arabic-based measures competitively perform well compared to the English-based measures. Further, ArWN enhanced structure is also available for public.

1 Introduction

In Natural Language Processing applications, a common task is to estimate the semantic similarity among words [1]. Lexical resources, such as, bilingual and multilingual dictionaries, thesauruses, lexical ontologies (wordnets), machine translation services among others, are widely used to estimate the similarity [2]. For instance, various tasks of natural language processing, knowledge engineering, and computational linguists have exploited the lexical and semantic knowledge encoded in the English WordNet (EnWN) [3, 4]; including sense disambiguation, information retrieval, text summarization, and question answering [5]–[6].

EnWN has been expanded to provide multilingual knowledge in many wordnet projects [7]–[8]. The Arabic WordNet (ArWN) [9] has extended EnWN by translating English synsets. However, English synsets that do not have translation in Arabic introduce *semantic gaps* in ArWN's semantic structure. For instance, synsets containing a single and polysemous word are difficult to determine their meaning by means of direct translation; in fact, more evidence is required to disambiguate their meaning [10]–[11]. Thus, similarity measures designed for English (i.e., English-based similarity measures) may not be effective in the same way when applied over resources in other languages; in this work we consider Arabic language.

Experiment findings in [12] showed that ArWN has limited coverage of lexical and semantic knowledge compared to EnWN. Further attempts have been made to improve the content of ArWN [9], [13]–[14]. However, resolving the semantic gaps was not considered. In [15, 16] they studied the performance of different similarity measures over ArWN. However, no explicit configuration was stated when calculating the similarity scores. Further, no explanation was given on how some semantic similarity scores were reported.

In [17], a preliminary study was conducted to examine the impact of the semantic gaps on estimating the semantic similarity scores using ArWN. They examined the impact of improving the semantic structure of ArWN on estimating the similarity between Ara-

*Corresponding Author: Mamoun Abu Helou, MIS Dept., Al-Istiqlal University, mabuhelou@pass.ps.

bic synsets. The semantic gaps were analyzed and identified. Then new synsets in Arabic were added to ArWN and mapped to their corresponding synsets in English, using interactive cross-lingual mapping approach [18]. The impact of the enriched ArWN was studied in semantic similarity experiment using only one English-based semantic similarity measure.

In this paper we extend previous work presented in [17]; a large scale experiment is conducted to further examine the degree to which wordnet-based applications can be influenced by improving their semantic structure, mainly considering ArWN. In particular, the main contributions of this work can be summarized as follow.

(i) Four settings are defined and applied over two variants of ArWN structure. In the experiment six path-based similarity measures are applied over ArWN and EnWN; including, four English-based similarity measures (Path [19], Li [2], WuP [20], and Lch [21]), and two Arabic-based similarity measures (AWSS [22], and Aldirey [16]).

(ii) Study to which extent the semantic similarity measures that are developed for Arabic-based applications can perform efficiently well compared to English-based similarity measures. A comprehensive comparison between the similarity measures over the different configurations is provided, for both EnWN and ArWN.

The similarity scores obtained from the different measures, in the different settings, are compared to a standard benchmark for Arabic word pairs obtained from the AWSS dataset [23]. Two measures, the Person Correlation and the Mean Square Error measures, are used to quantify the performance of the similarity measures. Reported values indicate the importance of the semantic evidence obtained from the enrichment process, and its significant effect on estimating the semantic similarity between words. In addition, the results show that Arabic-based measures performs competitively good compared to English-based measures.

The rest of this paper is organized as follows. Section 2 overviews related works on building wordnets, and the development of wordnet-based semantic similarity measures. Section 3 and describes the approach used to evaluate the impact of Semantic Gaps on estimating the Similarity over ArWN. Section 4 discusses experiments conducted: the benchmark dataset, the performance measures, and the obtained results. Finally Section 5 draws some conclusions and outlines future work.

2 Related works

This section provides an overview of the construction of wordnets and the ArWN contents; presents wordnet-based semantic similarity measures, which will be used in the experiment.

2.1 Wordnets overview

Wordnets, also known as lexical ontologies [24], are considered to be a resource of lexical and semantic knowledge, which organize natural language words (lexicons) into synsets. A synset is a collection of synonym words that express one meaning in a specific context (i.e., concept) [3, 25].

In wordnets, words are arranged in a lexical database. Words can have several senses, such that each sense of a given word is identified by a number and its part of speech type. For instance, the sense *village#n#2* indicates the second (#2) nominal (#*n*) sense of the word "*village*". Words are linked through lexical relations, for example, *antonym* and *synonymy* relations. When a word can have more than one meaning, it is called *polysemous word*, which can be member of several synsets. Otherwise, it is called *monosemous word*, which is a member of a single synset. For example, the word "*village*" has three noun senses as defined in EnWN; which are indicated in the following set of synsets:{{village#n#1, smalltown#n#1, settlement#n#2}, {village#n#2, hamle#n#3}, {Greenwichvillage#n#1, village#n#3}}.

Synsets are related by semantic relations. The *Hypernymy* and *Hyponymy* relations are considered to be the key semantic relations that form the semantic structure in wordnets. Hypernymy is described as the inverse of Hyponymy. For instance, in Figure 1 the synset {village#n#2, hamle#n#3} is hypernymy of the synset {settlemt#n#6}, while the synset {settlemt#n#6} is hyponymy of the synset {village#n#2, hamle#n#3}. Further, definitions (glosses) are also attached to synsets to convey their meaning. For example, the word sense *village#n#2* defined as "a settlement smaller than a town" [1].

The $HyperTree$ of a given synset (i.e, word sense) is defined as the sequence of synsets that are linked with hypernymy relations, which connect a synset with its ancestor synsets up to the root node. The function $HyperTrees(word)$ produces the set of HyperTrees which a given word belongs. Figure 1 shows an excerpt of nominal HyperTrees in English and their correspondence in Arabic [2].

EnWN has been manually produced at Princeton University over the past three decades [3, 4]. EnWN 's success in many computational language domains has inspired the development of similarly structured lexicons, for both individual and multiple languages [26], such as EuroWordNet [7], BalkaNet [27], Polylingual WordNet [8], universal wordnet [28],MultiWordNet [29], WikiNet [30],and Arabic WordNet [9].

Computational linguistics has defined the Inter-Lingual Index [7], to establish links between different wordnets which is considered to be independent of language. For instance, near-equivalence and equivalence semantic relations are used to link synsets from the individual wordnets to the Inter-Lingual Index. Wordnets for several languages have been developed under the guidance of the Global WordNet Association [3], which seeks to organize the creation and linking of wordnets. Further, the Open Multilingual WordNet project [31] offers access to open wordnets in a number of languages, which are all connected to the latest version of EnWN (v3.0) [4].

[1] Definitions can be accessed at http://wordnetweb.princeton.edu/perl/webwn.
[2] In the following, for both Arabic and English senses, the *pos* (*n*) identification is removed for readability, as they all nominal senses.
[3] http://globalwordnet.org/
[4] http://compling.hss.ntu.edu.sg/omw/.

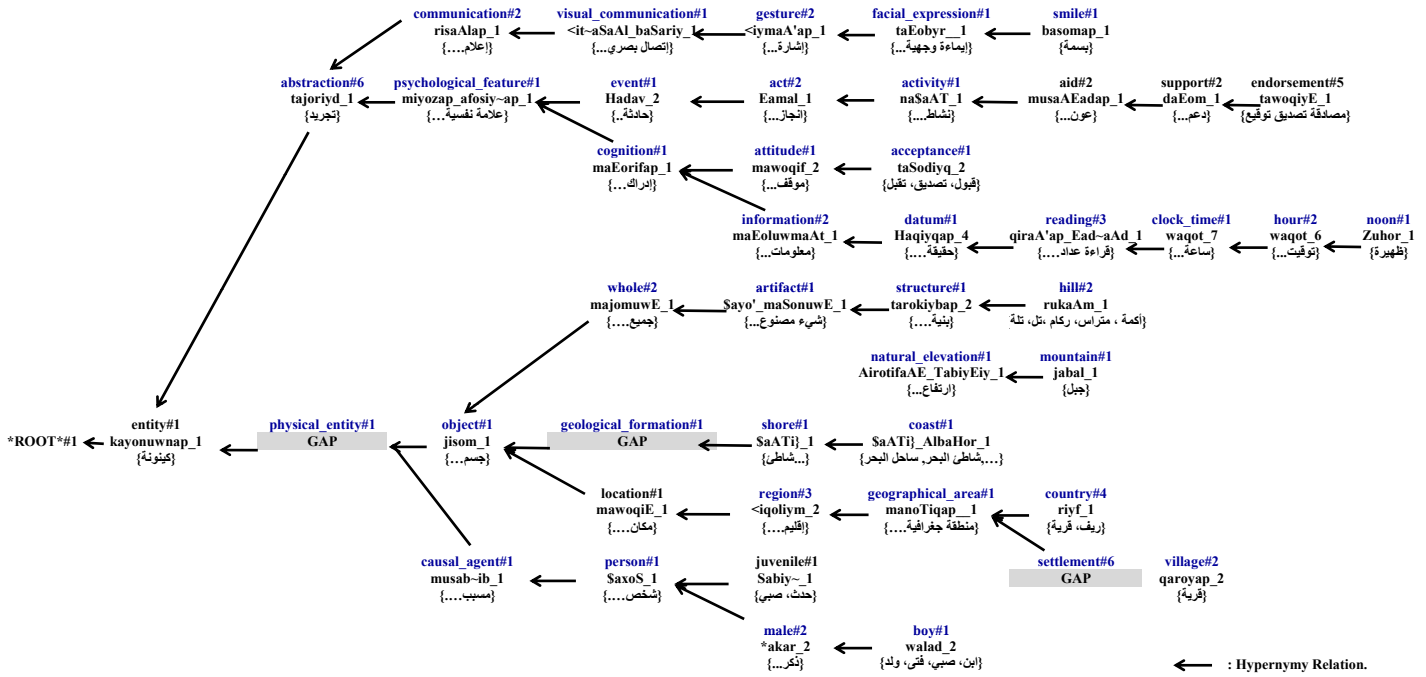


Figure 1: An excerpt of nominal HyperTrees from EnWN and its correspondences in ArWN.

2.2 *Arabic wordnet contents*

In the construction of ArWN [9], the extend method has been adopted. English Synsets have been translated into Arabic; and the structure of the EnWN (v2.0) has been inherited by ArWN . In the release of ArWN (v2.0) [5], 23,841 Arabic words, such as broken plurals, Named Entities, and roots have formed 11,296 synsets. Twenty-two types of semantic relationships have been used to connect synsets that formed 161,705 semantic links. Consequently, and in comparison with EnWN, which contains 147,306 words (117,659 synsets) [6]; one can observe that ArWN has a limited coverage in terms of semantic relations and lexicons [12].

To this end, many attempts have been made to enhance the quality of ArWN by expanding its lexical coverage [13, 9] or semantic relationships [32, 14] by different approaches. In [32] they released their work under the Lexical Markup Framework. However, the public release of ArWN ignores the synsets that are not linked to EnWN [31]. Nevertheless, synsets (semantic gaps) which are resolved in this work will be made for public [7] . In future work we plan to compile an *xml* format of ArWN enhanced structure, to enable researcher to utilize the ArWN in different applications.

2.3 *Wordnet-based similarity measures*

In linguistics, philosophy and information theory, estimating the semantic similarity between concepts is extensively studied [2, 15], which is a common and crucial task in many NLP applications, text summarization, word sense disambiguation, entailment, machine translation, among many others [33]–[6], [34, 35].

Estimating the semantic similarity between words is calculated by measuring the similarity between concepts (synsets) associated with the words [2]. Given two words, one can calculate the semantic similarity by exploiting wordnet (i.e, a lexical knowledge base). The lexical and semantic knowledge in wordnet have been used in many semantic similarity measures, which are originally designed and evaluated over EnWN (English-based measures) [36, 2].

In [15] they defined four broad categories of the similarity measures; Path-based similarity measures [2, 16, 19, 20, 21, 22]; information content similarity measures [37, 38]; feature-based similarity measures [39]; and hybrid similarity measures [40, 41]. There have been few works concerned with the similarity of Arabic; AWSS measure [22] and Aldiery measure [16]. These have mainly adapted measures from those constructed for English. In particular, Li measure [2] was adapted, which is a path-based measure that consider the depth of concepts in the HyperTrees; the distance between two compared concepts; and the depth of the least common concept (*lsc*) that subsumed two compared concepts. Noting that, these measures needs to tune weighting parameters to find the optimal values [22, 16]. In this regards, several preliminary experiments are necessary to find the best weights that provide the optimal values.

An attempt to investigate the performance of the similarity measures over ArWN was conducted in [15]. They studied the performance of seven measures; including AWSS measure [22]. All measures were applied over 40 word pairs that are selected from AWSS dataset [23], which are also considered as the benchmark dataset in this work. The experiments findings [15] showed that

[5] ArWN obtained from the AWN browser available at: AWNBrowser; or as XML file available at: Open Multilingual Wordnet
[6] EnWN-3.0 available at: https://wordnet.princeton.edu/download
[7] Full list of the Semantic Gaps available at:https://fas.alistiqlal.edu.ps/cv-2-ar.html

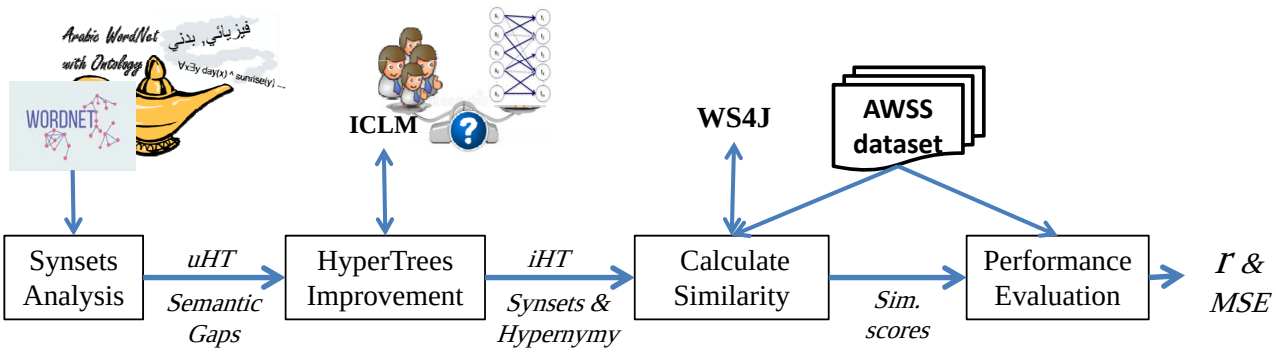


Figure 2: The adopted approach overview

WuP measure [20] has the best performance in estimating the semantic similarity between Arabic word pairs. The experiments in [16] also introduced a competitive Arabic-based similarity measures (Aldiery measure) in comparison to WuP measure.

In [17] they further studied the impact of enhancing the HyperTree over the Wup measure. This work adopted and extend their experimental configurations and examine further the impact of the enhanced semantic structure of ArWN over Six measures including English and Arabic path-based measure, further details are provided in Section 4.

Recall that, for a given concepts c_i and c_j, the function $Sim_m(c_i, c_j)$ calculates the semantic similarity between c_i and c_j, where m indicates the name of the measure. Next the description of the measures used in the experiment is given.

1. *Path* measure [19] finds the shortest path between the two concepts, by counting the number of edge (hypernymy relation) between the concepts, in order to compute the semantic similarity. Path measure which is considered as the pioneer similarity measure is defined in equation (1).

$$sim_{path}(c_i, c_j) = \frac{1}{len(c_i, c_j)} \tag{1}$$

Where the length function, $len(c_i, c_j)$, returns the length of the shortest path between c_i and c_j in the wordnet semantic hierarchy. For example, in Figure 1, $len(hill\#2, mountain\#1) = 3$, and $Sim_{path}(hill\#2, mountain\#1) = 0.333$.

2. *Wup* measure [20] calculates the similarity by computing the distance between the two concepts and the maximum depth of the least common concept (lsc) that subsumed the two concepts under evaluation. WuP measure is defined in equation (2).

$$sim_{WuP}(c_i, c_j) = \frac{2 * d(lcs(c_i, c_j))}{d(c_i) + d(c_j)} \tag{2}$$

Where $d(c_i)$ is the depth of the concept c_i using edge counting in the semantic hierarchy, $lcs(c_i, c_j)$ is the least common subsumer of c_i and cj, $d(lcs(c_i, c_j))$ is the maximum length between lcs of c_i and c_j and the *root* of the hierarchy, where $d(entity) = 1$. For example in Figure 1, $d(hill\#2) = 7$, $d(mountain\#1) = 7$, $d(lcs(hill\#2, mountain\#1) = 6$, and $Sim_{Wup}(hill\#2, mountain\#1) = 0.857$.

3. *Lch* measure [21] uses the length of the shortest path between the two concepts, and also the maximum depth of the semantic hierarchy of a given part of speech type. Lch measure is defined in equation (3).

$$sim_{Lch}(c_i, c_j) = -\log \frac{len(c_i, c_j)}{2 * maxDepth_{pos}} \tag{3}$$

Where, $maxDepth_{pos}$ is the maximum depth of the hypernymy structure for a given part of speech. For instance, $maxDepth_n$ is 20 and 15 in EnWN and ArWN, respectively. For example in Figure 1, $Sim_{Lch}(hill\#2, mountain\#1) = -log(3/2 * 20) = 2.590$.

Noting that the Lch scores reported in Section 4 are normalized into the range 0 to 1 by dividing Lch scors over 3,688, Hence, $Sim_{Lch}(hill\#2, mountain\#1) = 0.702$.

4. *Li* measure [2] computes the similarity using non-linear function, which consumes the shortest length between concepts and the minimum depth of the concepts in the semantic hierarchy. Li measure is defined in equation (4).

$$sim_{Li}(c_i, c_j) = e^{\alpha * len(c_i, c_j)} \frac{e^{\beta * d(lcs(c_i, c_j))} - e^{-\beta * d(lcs(c_i, c_j))}}{e^{\beta * d(lcs(c_i, c_j))} + e^{-\beta * d(lcs(c_i, c_j))}} \tag{4}$$

Noting that the parameters α and β need to be calculated manually for good performance. The optimal parameters are $\alpha = 0.2$ and $\beta = 0.6$ as reported in [2]. For example, $Sim_{Li}(hill\#2, mountain\#1) = 0.548$.

5. *AWSS* measure [22] is an Arabic-based measure that adapted Li measure to compute semantic similarity with modification on the depth and length computation to be proper for ArWN [23]. AWSS measure is defined in equation (5).

$$sim_{AWSS}(c_i, c_j) = e^{-\alpha * d(lcs(c_i, c_j))} * \tanh(\beta * len(c_i, c_j)) \tag{5}$$

Where the parameters α and β are the length and depth factors respectively. The optimal performance was obtained at $\alpha = 0.162$ and $\beta = 0.234$ as reported in [22]. For example in Figure 1, $len(rukaAm\_1, Jabal\_1) = 4$ and $d(lcs(rukaAm\_1, Jabal\_1)) = 8$, then $Sim_{AWSS}(rukaAm\_1, Jabal\_1) = 0.201$.

Table 1: Semantic gaps frequency distribution in ArWN for nominal synsets

Semantic Gaps	1	1	1	1	1	1	1	2	3	1	1	2	1	6	3	3	2	11	1	7	12	11	15	**88**
Freq	4525	187	100	50	48	46	36	30	24	17	15	14	12	10	9	8	7	6	5	4	3	2	1	**5493**

6. *Aldiery* measure [16] is an Arabic-based measure also adapted Li measure to compute semantic similarity with modification on the depth and length computation to be proper for ArWN. Aldiery measure is defined in equation (6).

$$sim_{Aldiery}(c_i, c_j) = \tanh(\frac{2 * d(lcs(c_i, c_j))}{d(c_i) + d(c_j)} + W(\frac{1}{len(c_i, c_j)} + \frac{\log(len(c_i, c_j))}{\log(maxDepth_{pos})})) \quad (6)$$

Where [16] defines $W = 0.5$. For example in Figure 1, $d(rukaAm\_1) = 8$, $d(Jabal\_1) = 7$, $len(rukaAm\_1, Jabal\_1) = 4$, and $d(lsc(rukaAm\_1, Jabal\_1)) = 8$, and $maxDepth_n = 15$, then $Sim_{Aldiery}(rukaAm\_1, Jabal\_1) = 0.692$.

Noting that, the similarity functions defined above consume either words, or word senses as parameters. In the first case, the similarity function returns the highest similarity score for all the possible combination of word senses for the two given words. In the second case, it returns the similarity score between the two defined senses.

In addition, the six measures defined in the equations (1,2,3,4,5, and 6) are path-based measures, this study focus on the impact of the structure without interference of other semantic evidence such as features extracted from corpuses, which depend on the quality of the used cuprous, as well as the availability of resources in Arabic.

On the other hand, Path, WuP, and Lch measures are considered as linear path-based measures, while Li measure is a non-linear path based measure. AWSS and Aldiery are also non-linear path based measures, which are derived from Li and purposely developed for Arabic.

Observe that, for the Path, Wup, and Lch measures no weights are required to be tuned. While the other measures need to find optimal value of the defined weights. The four English-based measures, as well as the two Arabic-based measures are selected because they achieved good performance against other measures [22, 16], and to compare the performance between the measures using Arabic benchmark dataset.

3 Evaluating the impact of semantic gaps on estimating the similarity

This section presents the approach that is used to evaluate the impact of enhancing the structure of ArWN on estimating the semantic similarity. Figure 2 illustrates the main phases of the approach, which are explained as follow.

1. **Synset Analysis**. In this phase the semantic gaps are identified through a comparison between the structures of ArWN (v2.0) and EnWN (v3.0). In particular, for each nominal synset in ArWN, Hypernymy relations are compared with their EnWN correspondences. The HyperTrees for each synset in ArWN is compared with its correspondence HyperTrees in EnWN. For example, Figure 1 indicates two semantic gaps in the ArWN $HyperTree(\$aATi\}\_AlbaHor\_1, ساحل)$ ={*ROOT*#1 , kayonuwnap_1, *GAP*, jisom_1, *GAP*, $aATi}_1, $aATi}_AlbaHor_1} [8], where the correspondence HyperTree in EnWN is, HyperTree(coast#1) = {*ROOT*#1, entity#1, physical entity#1, object#1, geological formation#1, shore#1, coast#1}.

   In total, [17] reported that 5,493 (69%) of the 7,960 nominal synsets in ArWN have at least one semantic gap. In particular, compared to the structure of EnWN, the semantic gaps have been resulted from the missing of 88 synsets in ArWN.

   The distribution frequency of the semantic gaps in ArWN is reported in Table 1, "Semantic Gaps" refers to the number of synsets that have the reported freq, and "Freq" indicates the number of HyperTrees that have at least one semantic gap. For instance, the first column reports an English synset ({"physical-entity#1"}) that has no correspondence in Arabic, introduces 4,525 semantic gaps in ArWN. While the 8th column indicates two synsets ({"armed-service#1",...}, and {"health-care-provider#1",...}), each introduces 30 semantic gaps in ArWN. Last column reports the totals.

2. **HyperTrees Improvement**. In this phase ICLM Web application [18] is used to fill the identified semantic gaps. ICLM is a semi-automatic matching approach that supports feedback provided by multiple users. In ICLM the number of users that are asked to perform each mapping task is estimated based on the lexical characterization of concepts under evaluation, i.e., on the estimation of the ambiguity conveyed by the concepts involved in mappings [42], with the assumption that as the selection tasks difficulties increase, more users agreement is required.

   The candidate matching of the source concepts in Arabic are automatically computed to the English target concepts using a lexical based disambiguation algorithm [43]. The study [42] recommended that combining lexical resources improves the quality of translations and provide a valuable support for candidate match retrieval in cross-lingual ontology matching problems. Accordingly, translations of the missing synsets are collected by combining lexical knowledge from different external resources. English synset translation was

[8] Represents the HyperTree of the first nominal sense for the word ساحل.

[9] https://translate.google.com/

[10] https://babelnet.org/

[11] https://www.almaany.com/ar/dict/ar-en/

Table 2: Top ten Frequent Semantic Gaps in ArWN with EnWN correspondence sysnets

	EnWN synset	Freq	Semantic gaps in ArWN
1	{physical entity}	4,525	{كيان مادي}
2	{substance}	187	{جوهر، مادة}
3	{defender, guardian, protector, shielder}	100	{مدافع ، وصي، حارس، حامي، مدافع، ولي}
4	{variety, assortment, miscellanea,...}	50	{تشكيلة،تنوع، مجموعة منوعة منوعات}
5	{aristocrat,blue blood, patrician}	48	{رفيع الجناب، نبالة، نسب أصيل، النبيل، الأرستوقراطي}
6	{formataion, geological formation}	46	{تشكيل ، التكوين الجيولوجي ، التشكلات الجيولوجية}
7	{deceiver, beguiler, cheat, cheater, ...}	36	{مخادع، مضل ، غشاش ، محتال ، دجال}
8	{armed service, service, military service}	30	{القوة المسلحة ، الخدمة العسكرية}
9	{health care provider, health professional,...}	30	{مقدم عناية صحية ، مزود الرعاية الصحية الأولية، الصحة المهنية}
10	{wrongdoer, offender}	24	{مذنب، مجرم ، ظالم ، معتد، آثم}

collected from; Google Translate[9], BabelNet[10], and Almaany dictionary[11].

The difficulties of the mapping selection tasks, that is determining the number of user which are asked to perform the task, are estimated using lexical characteristics of concepts under evaluation: Ambiguity of lexicalization, Synonym-richness, and Uncertainty in the selection Step. The mapping tasks are validated by some users based on a CAUTIOUS strategy. The task difficulty level is estimated as Low, Mid, and High level. One, three, or five users are asked to perform the Low, Mid, or High tasks, respectively.

In [17] ten users (bilingual speakers) are asked to validate the mapping tasks, that is, to fill a semantic gap in ArWN, and accordingly define new link with EnWN, hence, import the semantic relations among the concepts. The top ten frequent semantic gap are listed in Table 2. As a result 94% of the identified gaps are resolved, that is more than 98% of HyperTrees are filled in.

Observe that, some concepts are hard to resolve, and more evidences are needed. For Example, {mechanism#3}, {attache#1}, and {climber#1} synsets, which contain a single and polysemous word, are hard to determine their meaning with direct translation and no context [42], for this reason in the validation task users did not reach an agreement. Noting that, the semantic gaps for every word sense in the benchmark dataset used in the experiment are resolved.

3. **Calculate Similarity**. In this phase similarity measures defined in Section 2.3 are applied over the ArWN and EnWN using Arabic benchmark dataset (AWSS dataset [23]). Different configuration explained in Section 4.4 are applied to calculate the semantic similarity using the WS4J online application (see Section 4.1). Similarity scores are reported and passed to the next phase.

4. **Performance Evaluation**. In this phase the obtained similarity scores are compared with Human Rating benchmark [23] using two performance measures; The Person Correlation measure (r) and the Mean Squared Error (MSE). Further details are provided in the experiment Section 4.3

4 Experiment

The conducted experiment aims at studying the efficacy of the semantic evidence in ArWN. In particular, the experiment focuses on the improvement of hypernymy relations in the semantic structure of ArWN. The experiment studies the extent to which the semantic structure of ArWN affects measuring the semantic similarity between concepts. This section reports and discusses the results obtained from running a set of configurations for measuring the semantic similarity scores over ArWN and EnWN.

Next sections present the tool which is used to calculate the semantic similarity scores, the benchmark dataset, the measures used to evaluate the performance of the structure improvement, and discuss obtained results.

4.1 Similarity Measure Tools

Significant efforts are being made in developing similarity measures to consume ArWN content. For example, the Java ArWN API [12]. The application consumes Arabic words with diacritics (vocalized), whereas the benchmark dataset in this experiment contains unvocalized (without diacritics) word pairs. If Arabic words are vocalized, similar to the work done in [16, 15], then their senses will be defined in advance. The experiment's configuration *DS* (see Section 4.4) studies the performance of determining the word senses on the similarity scores.

To avoid predefined senses, in this experiment the similarity scores are obtained using the $WS4J$ online application [13]. In computing the scores, $WS4J$ uses EnWN's semantic structure (v3.0), which is used to measure the similarity scores between Arabic words. Noting that, in this experiment Arabic senses under evaluation have the same structure of their correspondence senses in English, as the semantic gaps in ArWN has been improved and linked to EnWN(v3.0). The similarity scores between the Arabic concepts are then measured using their correspondence concepts in EnWN. In addition, $WS4J$ provides the description of all HyperTree of words under evaluation. The HyperTrees which returned for EnWN are validated to obtain Arabic words' HyperTrees with semantic gaps as depicted in Figure 1. For instance, this information is necessary to measure the similarity scores in *uHT* configuration,

[12] https://sourceforge.net/projects/javasourcecodeapiarabicwordnet
[13] http://ws4jdemo.appspot.com/

details are provided in Section 4.4.

4.2 Benchmark dataset

Similar to the work performed in [15, 16], the AWSS benchmark [22] will be used in this experiment. The obtained similarity scores will be compared with Human Judgments obtained from the dataset of AWSS [23]. The AWSS dataset contains 70 nominal word pairs of Arabic, divided into three similarity levels, Low, Medium, and High; 40 word pairs are selected and used in this experiment, listed in Table 3,which are also used in [15, 16].

Table 3: Arabic word pairs benchmark dataset

NO.	Sim. level	En Word Pairs		Ar Word Pairs	HR
1	Low	Coast	Endorsement	ساحل تصديق	0.01
2	low	Noon	String	ظهر خيط	0.01
3	low	Stove	Walk	موقد مشي	0.01
4	low	Cord	midday	حبل ظهيرة	0.02
5	low	Signature	String	توقيع خيط	0.02
6	low	Boy	Endorsement	صبي تصديق	0.03
7	low	Boy	Midday	صبي ظهيرة	0.04
8	low	Smile	Village	ابتسامة قرية	0.05
9	low	Noon	Fasting	ظهر صيام	0.07
10	low	Glass	Diamond	كاس الماس	0.09
11	low	Sepulcher	Sheikh	ضريح شيخ	0.22
12	low	Countryside	Vegetable	خضار ريف	0.31
13	mid	Tumbler	Tool	قدح اداة	0.33
14	mid	Laugh	Feast	عيد ضحك	0.34
15	mid	Girl	Odalisque	فتاة جارية	0.49
16	mid	Feast	Fasting	عيد صيام	0.49
17	mid	Coach	Means	حافلة وسيلة	0.52
18	mid	Sage	Sheikh	حكيم شيخ	0.56
19	mid	Girl	Sister	فتاة اخت	0.6
20	mid	Hen	Pigeon	دجاجة حمامة	0.65
21	mid	Hill	Mountain	تل جبل	0.65
22	mid	Master	Sheikh	سيد شيخ	0.67
23	mid	Food	Vegetable	طعام خضار	0.69
24	mid	Slave	Odalisque	عبد جارية	0.71
25	mid	Run	Walk	جري مشي	0.75
26	high	Cord	String	حبل خيط	0.77
27	high	Forest	Woodland	غابة احراش	0.79
28	high	Sage	Thinker	حكيم مفكر	0.82
29	high	Journey	Travel	رحلة سفر	0.84
30	high	Gem	Diamond	جوهرة الماس	0.84
31	high	Countryside	Village	ريف قرية	0.85
32	high	Cushion	Pillow	مسند مخدة	0.85
33	high	Smile	Laugh	ابتسامة ضحك	0.87
34	high	Signature	Endorsement	توقيع تصديق	0.89
35	high	Tools	Means	اداة وسيلة	0.92
36	high	Sepulcher	Grave	ضريح قبر	0.93
37	high	Boy	Lad	صبي فتى	0.93
38	high	Wizard	Magician	ساحر مشعوذ	0.94
39	high	Coach	Bus	حافلة باص	0.95
40	high	Glass	Tumbler	كاس قدح	0.95

Noting that, some words in the dataset benchmark are not covered in ArWN. For instance, the words " موقد" stove, " ساحر" wizard, and " مشعوذ" magician are not covered in ArWN, hence, the 3rd and 38th word pairs are not covered in the experiment. While, the words "ابتسامة" smile and "جوهرة" Gem, which are also not covered in ArWN, instead the words "بسمة" and "جوهر" are used to measure the similarity scores, respectively.

4.3 Performance Measures

The obtained similarity scores are evaluated against human ratings benchmark (*HR*), which is a human judgment similarity scores of Arabic nominal word pairs obtained from the dataset of AWSS.

Two measures are used to quantify the performance of the obtained similarity scores. The Person Correlation measure (r) defines the strength of the linear relationship between the obtained similarity scores and *HR*; the Mean Squared Error (MSE) calculates the average squared difference between the similarity scores and *HR*. The best performance is indicated by a similarity measure with the smallest MSE value and r value is close to 1. While the negative r value means that the obtained scores are increase as the *HR* ratings decrease. In addition, the similarity scores are compared to the performance results reported in [15, 16], which are listed in Table 4.

Table 4: Performance measures reported in [16, 15]

#	**Measure**	r	MSE
1	WuP	0.94	0.01648
2	LCH	0.89	0.03708
3	Path	0.75	0.16038
4	LI	0.85	0.10205
5	AWSS	0.88	0.04424
6	Aldiery	0.96	0.01893

4.4 Experimental settings

Six path-based semantic similarity measures, which are defined in equations (1,2,3,4,5, and 6), will be applied over the Arabic word pairs benchmark dataset, which is described in Section 4.2. Using the following configurations, the similarity measures are applied over ArWN and EnWN to quantify the efficiency of ArWN structure enrichment:

1. *UnDefined Senses* (uDS): calculates the semantic similarity between given words without determining their senses. In this setting, which is considered as the default setting of the similarity measures, the similarity measure returns the maximum score obtained from the all possible combination of the senses of the given words.

2. *Defined Senses* (DS): calculates the semantic similarity between given words senses (i.e, sense are determined in advance). By extending the work in [17], the sense of each word pairs under evaluation is determined based on a majority vote (consensus) approach. Similar to the tasks of filling the semantic gaps [17, 18] (see Section 3), the CAUTIOUS strategy is adopted, where users are avoided to decide among word pairs that share the same words.

3. *wordnets Translation* ($wnTrans$): calculates the semantic similarity over ArWN by selecting the senses that match the

Table 5: uDs configuration over ArWN

					iHT						uHT					
NO.	Ar Word Pairs Senses		En Word Pairs Senses		WuP	LCH	Path	LI	AWSS	Aldiery	WuP	LCH	Path	LI	AWSS	Aldiery
1	$aATi}_1	taSodiyq_2	shore#1	acceptance#1	0.308	0.298	0.100	0.113	0.086	0.451	0.364	0.358	0.125	0.168	0.119	0.504
2	mu&ax~irap_1	xayoT_1	back#2	cord#4	0.706	0.436	0.167	0.301	0.335	0.808	0.667	0.436	0.167	0.300	0.312	0.781
3																
4	Hamol_1	Zuhor_1	gestation#2	midday#1	0.316	0.207	0.071	0.058	0.063	0.504	0.316	0.207	0.071	0.058	0.063	0.504
5	tawoqiyE_1	daliyl_2	endorsement#5	lead#3	0.444	0.272	0.091	0.109	0.123	0.633	0.444	0.272	0.091	0.109	0.123	0.633
6	Sabiy~_1	taSodiyq_2	juvenile#1	acceptance#1	0.308	0.298	0.100	0.113	0.086	0.451	0.333	0.326	0.111	0.138	0.102	0.475
7	Sabiy~_1	Zuhor_1	juvenile#1	midday#1	0.235	0.207	0.071	0.051	0.045	0.379	0.250	0.227	0.077	0.062	0.053	0.394
8	basomap_1	qaroyap_1	smile#1	village#1	0.375	0.272	0.091	0.105	0.102	0.553	0.375	0.272	0.091	0.105	0.102	0.553
9	Zuhor_1	Sawom_1	noon#1	fasting#1	0.364	0.188	0.067	0.049	0.065	0.574	0.364	0.188	0.067	0.049	0.065	0.574
10	kuwb_1	AlomAs_1	glass#2	diamond#2	0.353	0.248	0.083	0.086	0.087	0.535	0.267	0.248	0.083	0.076	0.062	0.411
11	maqaAm_1	ra}iyos_1	shrine#1	head#4	0.500	0.272	0.091	0.110	0.139	0.687	0.556	0.326	0.111	0.164	0.192	0.720
12	riyf_1	xuDaAr_1	country#4	vegetable#1	0.375	0.272	0.091	0.105	0.102	0.553	0.286	0.272	0.091	0.092	0.073	0.429
13	sahom_3	adaAp_2	arrow#2	instrument#1	0.857	0.922	1.000	0.819	0.826	0.943	0.842	0.546	0.250	0.449	0.499	0.874
14	<iHotifaAl_1	DaHik_2	laughter#2	celebration#2	0.824	0.546	0.250	0.449	0.485	0.864	0.824	0.546	0.250	0.449	0.485	0.864
15	fataAp_1	xaAdim_1	girl#1	retainer#2	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
16	Eiyod_1	Sawom_1	celebration#2	fasting#1	0.700	0.395	0.143	0.246	0.298	0.811	0.700	0.395	0.143	0.246	0.298	0.811
17	HaAfilap_1	wasiylap_1	coach#5	means#2	0.778	0.486	0.200	0.368	0.412	0.845	0.750	0.486	0.200	0.367	0.394	0.828
18	fayolasuwf_1	ra}iyos_1	philosopher#1	head#4	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
19	fataAp_1	>xot_1	girl#1	sister#1	0.696	0.358	0.125	0.202	0.261	0.815	0.667	0.358	0.125	0.202	0.254	0.796
20	dajaAjap_1	HamaAm_1	hen#1	pigeon#1	0.828	0.436	0.167	0.301	0.376	0.879	0.815	0.436	0.167	0.301	0.374	0.872
21	rukaAm_1	jabal_1	hill#2	mountain#1	0.533	0.358	0.125	0.199	0.201	0.692	0.500	0.395	0.143	0.233	0.195	0.649
22	say~id_1	ra}iyos_1	sir#1	head#4	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
23	TaEaAm_3	xuDaAr_1	food#2	vegetable#1	0.857	0.624	0.333	0.548	0.545	0.875	0.833	0.624	0.333	0.546	0.507	0.861
24	xaAdim_1	xaAdim_1	retainer#2	retainer#2	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
25	jaroy_1	ma$oy_1	run#7	walk#1	0.909	0.624	0.333	0.549	0.604	0.905	0.909	0.624	0.333	0.549	0.604	0.905
26	Habol_1	gazol_1	cord#1	thread#1	0.941	0.734	0.500	0.670	0.690	0.918	0.933	0.734	0.500	0.670	0.671	0.913
27	dagol_1	dagol_1	jungle#1	jungle#1	1.000	0.922	1.000	0.818	0.754	0.950	1.000	0.922	1.000	0.815	0.701	0.947
28	fayolasuwf_1	mufak~ir_1	philosopher#1	intellect#3	0.900	0.624	0.333	0.549	0.597	0.900	0.889	0.624	0.333	0.549	0.587	0.894
29	riHolap_1	safar_1	journey#1	travel#1	0.952	0.734	0.500	0.670	0.710	0.926	0.952	0.734	0.500	0.670	0.710	0.926
30	HajarN_kariym_1	AlomAs_1	gem#2	diamond#2	0.875	0.624	0.333	0.549	0.570	0.886	0.857	0.624	0.333	0.548	0.545	0.875
31	riyf_1	riyf_1	country#4	country#4	1.000	0.922	1.000	0.819	0.811	0.955	1.000	0.922	1.000	0.818	0.789	0.953
32	wisaAdap_1	wisaAdap_1	cushion#3	cushion#3	1.000	0.922	1.000	0.819	0.811	0.955	1.000	0.922	1.000	0.818	0.789	0.953
33	basomap_1	DaHik_2	smile#1	laugh#1	0.533	0.358	0.125	0.199	0.201	0.692	0.533	0.358	0.125	0.199	0.201	0.692
34	tawoqiyE_1	tawoqiyE_1	endorsement#5	endorsement#5	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
35	>adaAp_1	wasiyolap_1	tool#2	means#1	0.941	0.734	0.500	0.670	0.690	0.918	0.941	0.734	0.500	0.670	0.690	0.918
36	qabor_1	qabor_1	grave#2	grave#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955
37	Sabiy~_2	Sabiy~_2	spring chicken#1	spring chicken#1	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
38																
39	HaAfilap_1	HaAfilap_1	coach#5	coach#5	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
40	kuwb_1	kuwb_1	glass#2	glass#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955

	Performance Measures											
Sim. level	Correlation r						Correlation r					
all	0.858	0.774	0.658	0.787	0.806	0.831	0.850	0.814	0.712	0.825	0.840	0.826
low	0.060	-0.115	-0.162	-0.131	-0.074	0.155	-0.075	-0.088	-0.139	-0.135	-0.090	-0.089
mid	0.122	-0.095	-0.127	-0.077	-0.075	-0.052	0.103	0.269	0.324	0.239	0.188	0.050
high	0.152	0.314	0.393	0.265	0.300	0.177	0.171	0.314	0.393	0.267	0.345	0.197
Sim. level	MSE						MSE					
all	0.066	0.045	0.104	0.055	0.047	0.109	0.064	0.038	0.092	0.048	0.044	0.104
low	0.118	0.050	0.010	0.015	0.016	0.247	0.118	0.057	0.010	0.017	0.016	0.240
mid	0.072	0.056	0.178	0.087	0.071	0.101	0.067	0.031	0.142	0.067	0.057	0.091
high	0.020	0.030	0.109	0.056	0.050	0.009	0.020	0.030	0.109	0.056	0.053	0.008

translations defined in the benchmark dataset. In *wnTrans* the maximum similarity score is selected, such that the ArWN and the EnWN cover the Arabic word and its translation in English, respectively. Otherwise, the default setting *uDS* is applied.

4. *Upper Bound* (*UB*): calculates the semantic similarity between given words senses, such that, *UB* selects the sense pair that maximize correlation *r* values and minimize *MSE* values w.r.t the *HR* ratings (benchmark dataset). *UB* indicates the optimal scores for the considered experiment settings.

5. *Unimproved HyperTrees* (*uHT*): calculates the semantic similarity using ArWN while ignoring the structure enhancement. That is, the semantic gaps are considered in calculating the similarly scores.

6. *Improved HyperTrees* (*iHT*): calculates the semantic similarity using the enhanced structure of ArWN.

4.5 Results & Discussion

Tables 5, 6, 7, and 8 report the semantic similarity scores using six similarity measures, which resulted from applying *uDS*, *DS*, *wnTrans* and *UB* configurations over ArWN; respectively. Such that two variants, *uHT* and *iHT*, are considered. The tables also list the Arabic senses and their correspondences senses in English, which are used to provide the obtained similarity scores. Table 9 reports the semantic similarity scores that are obtained from applying *uDs*, *DS*, and *UB* configurations over EnWN. English-based

Table 6: DS configuration over ArWN

					iHT						uHT					
NO.	Ar Word Pairs Senses		En Word Pairs Senses		WuP	LCH	Path	LI	AWSS	Aldiery	WuP	LCH	Path	LI	AWSS	Aldiery
1	$aATi}_AlbaHor_1	tawoqiyE_1	coast#1	endorsement#5	0.235	0.207	0.071	0.051	0.045	0.379	0.267	0.248	0.083	0.076	0.062	0.411
2	Zuhor_1	gazol_1	noon#1	thread#1	0.200	0.154	0.059	0.028	0.028	0.343	0.211	0.170	0.063	0.034	0.033	0.354
3																
4	Habol_1	Zuhor_1	cord#1	midday#1	0.211	0.170	0.063	0.034	0.033	0.354	0.222	0.188	0.067	0.042	0.038	0.366
5	tawoqiyE_1	gazol_1	endorsement#5	thread#1	0.211	0.170	0.063	0.034	0.033	0.354	0.222	0.188	0.067	0.042	0.038	0.366
6	walad_1	tawoqiyE_1	boy#1	endorsement#5	0.235	0.207	0.071	0.051	0.045	0.379	0.250	0.227	0.077	0.062	0.053	0.394
7	walad_1	Zuhor_1	boy#1	midday#1	0.222	0.188	0.067	0.042	0.038	0.366	0.235	0.207	0.071	0.051	0.045	0.379
8	basomap_1	qaroyap_2	smile#1	village#2	0.235	0.207	0.071	0.051	0.045	0.379	0.267	0.248	0.083	0.076	0.062	0.411
9	Zuhor_1	Sawom_1	noon#1	fasting#1	0.364	0.188	0.067	0.049	0.065	0.574	0.364	0.188	0.067	0.049	0.065	0.574
10	kuwb_1	AlomAs_1	glass#2	diamond#2	0.353	0.248	0.083	0.086	0.087	0.535	0.267	0.248	0.083	0.076	0.062	0.411
11	qabor_1	ra}iyos_1	grave#2	head#4	0.444	0.272	0.091	0.109	0.123	0.633	0.375	0.272	0.091	0.105	0.102	0.553
12	riyf_1	xuDaAr_1	country#4	vegetable#1	0.375	0.272	0.091	0.105	0.102	0.553	0.286	0.272	0.091	0.092	0.073	0.429
13	kuwb_1	>adaAp_1	glass#2	tool#2	0.222	0.922	1.000	0.683	0.371	0.691	0.235	0.207	0.071	0.051	0.045	0.379
14	DaHik_2	<iHotifAl_1	laugh#1	celebration#1	0.400	0.298	0.100	0.128	0.120	0.574	0.400	0.298	0.100	0.128	0.120	0.574
15	fataAp_1	xaAdim_1	girl#1	retainer#2	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
16	<iHotifAl_1	Sawom_1	celebration#1	fasting#1	0.526	0.298	0.100	0.135	0.163	0.703	0.526	0.298	0.100	0.135	0.163	0.703
17	HaAfilap_1	wasiylap_1	coach#5	means#2	0.778	0.486	0.200	0.368	0.412	0.845	0.750	0.486	0.200	0.367	0.394	0.828
18	fayolasuwf_1	ra}iyos_1	philosopher#1	head#4	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
19	fataAp_1	>xot_1	girl#1	sister#1	0.696	0.358	0.125	0.202	0.261	0.815	0.667	0.358	0.125	0.202	0.254	0.796
20	dajaAjap_1	HamaAm_1	hen#1	pigeon#1	0.828	0.436	0.167	0.301	0.376	0.879	0.815	0.436	0.167	0.301	0.374	0.872
21	rukaAm_1	jabal_1	hill#2	mountain#1	0.533	0.358	0.125	0.199	0.201	0.692	0.500	0.395	0.143	0.233	0.195	0.649
22	say~id_1	ra}iyos_1	sir#1	head#4	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
23	TaEaAm_1	xuDaAr_1	food#1	vegetable#1	0.571	0.395	0.143	0.243	0.236	0.716	0.500	0.395	0.143	0.233	0.195	0.649
24	Eabod_1	xaAdim_1	slave#1	retainer#2	0.842	0.546	0.250	0.449	0.499	0.874	0.824	0.546	0.250	0.449	0.485	0.864
25	jaroy_1	ma$oy_1	run#7	walk#1	0.909	0.624	0.333	0.549	0.604	0.905	0.909	0.624	0.333	0.549	0.604	0.905
26	Habol_1	gazol_1	cord#1	thread#1	0.941	0.734	0.500	0.670	0.690	0.918	0.933	0.734	0.500	0.670	0.671	0.913
27	dagol_1	dagol_1	jungle#1	jungle#1	1.000	0.922	1.000	0.818	0.754	0.950	1.000	0.922	1.000	0.815	0.701	0.947
28	fayolasuwf_1	mufak~ir_1	philosopher#1	intellect#3	0.900	0.624	0.333	0.549	0.597	0.900	0.889	0.624	0.333	0.549	0.587	0.894
29	riHolap_1	safar_1	journey#1	travel#1	0.952	0.734	0.500	0.670	0.710	0.926	0.952	0.734	0.500	0.670	0.710	0.926
30	HajarN_kariym_1	AlomAs_1	gem#2	diamond#2	0.875	0.624	0.333	0.549	0.570	0.886	0.857	0.624	0.333	0.548	0.545	0.875
31	riyf_1	qaroyap_2	country#4	village#2	0.824	0.546	0.250	0.449	0.485	0.864	0.857	0.624	0.333	0.548	0.545	0.875
32	wisaAdap_1	wisaAdap_1	cushion#3	cushion#3	1.000	0.922	1.000	0.819	0.811	0.955	1.000	0.922	1.000	0.818	0.789	0.953
33	basomap_1	DaHik_2	smile#1	laugh#1	0.533	0.358	0.125	0.199	0.201	0.692	0.533	0.358	0.125	0.199	0.201	0.692
34	tawoqiyE_1	tawoqiyE_1	Endorsement#5	Endorsement#5	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
35	>adaAp_1	wasiyolap_1	tool#2	means#1	0.941	0.734	0.500	0.670	0.690	0.918	0.941	0.734	0.500	0.670	0.690	0.918
36	qabor_1	qabor_1	grave#2	grave#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955
37	muraAhiq_1	Tifol_1	adolescent#1	child#1	0.900	0.624	0.333	0.549	0.597	0.900	0.889	0.624	0.333	0.549	0.587	0.894
38																
39	HaAfilap_1	HaAfilap_1	coach#5	bus#1	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
40	kuwb_1	kuwb_1	glass#2	glass#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955

	Performance Measures											
Sim. level	Correlation r						Correlation r					
all	0.925	0.799	0.607	0.824	0.877	0.933	0.920	0.859	0.688	0.864	0.878	0.925
low	0.810	0.853	0.875	0.901	0.882	0.776	0.564	0.687	0.712	0.756	0.738	0.478
mid	0.736	-0.156	-0.389	0.018	0.463	0.636	0.698	0.770	0.702	0.739	0.695	0.686
high	0.130	0.234	0.290	0.199	0.242	0.150	0.136	0.229	0.288	0.193	0.279	0.161
Sim. level	MSE						MSE					
all	0.028	0.042	0.131	0.065	0.052	0.061	0.026	0.034	0.118	0.062	0.057	0.053
low	0.044	0.020	0.007	0.005	0.005	0.135	0.042	0.026	0.007	0.006	0.007	0.125
mid	0.024	0.060	0.207	0.104	0.076	0.055	0.022	0.033	0.176	0.099	0.088	0.042
high	0.018	0.043	0.158	0.076	0.067	0.008	0.018	0.040	0.152	0.071	0.067	0.008

similarity (Path, lch, WuP, and Li measures) scores are reported for each configuration. English senses that are used to compute the scores are also reported.

Observe that the word senses are defined differently based on the applied configuration. For example, the word boy "صبي" is selected differently w.r.t the applied configuration; in Table 5, in the *uDS* setting the selected sense is (Sabiy _1, juvenile#1) [14], in *DS* (Table 6) and *wnTrans* (Table 7) settings the selected sense is (walad_1, boy#1), and in *UB* (Table 8) setting the sense is (walad_2, boy#2). Moreover, considering *wnTrans* setting, the translation which are defined in the AWSS benchmark for 13 Arabic words that exist in 17 word pairs, does not exist in the mapping between ArWN and EnWN; the words and their translation are {signature: توقيع ; sepulcher: ضريح; sheikh: شيخ; countryside: ريف; tumbler: قدح; feast: عيد; odalisque: جارية; sage: حكيم; thinker: مفكر; pillow: مخدة; signature: توقيع; lad: فتى}. For example, the word "توقيع" has one sense in ArWN "*tawoqiyE_1*", which is mapped into the "*endorsement*#5" in EnWN, while none of the five senses for the word "*signature*" in EnWN is mapped into ArWN. Noting that 28 word pairs out of the 40 word pairs has at least one missing correspondence sense in EnWN when considering *uHT* setting, For example; similarity scores of the 21st word pairs (Hill "تل"; mountain "جبل"); which is also illustrated in

[14] Represents Arabic sense in ArWN and its correspondence in EnWN

Table 7: wnTrans configuration over ArWN

NO.	Ar Word Pairs Senses		En Word Pairs Senses		iHT						uHT					
					WuP	LCH	Path	LI	AWSS	Aldiery	WuP	LCH	Path	LI	AWSS	Aldiery
1	$aATi}_AlbaHor_1	tawoqiyE_1	coast#1	endorsement#5	0.235	0.207	0.071	0.051	0.045	0.379	0.267	0.248	0.083	0.076	0.062	0.411
2	Zuhor_1	xayoT_2	noon#1	string#9	0.353	0.248	0.083	0.086	0.087	0.535	0.353	0.248	0.083	0.086	0.087	0.535
3																
4	Habol_1	Zuhor_1	cord#1	midday#1	0.211	0.170	0.063	0.034	0.033	0.354	0.222	0.188	0.067	0.042	0.038	0.366
5	tawoqiyE_1	xayoT_2	endorsement#5	string#9	0.375	0.272	0.091	0.105	0.102	0.553	0.375	0.272	0.091	0.105	0.102	0.553
6	walad_1	tawoqiyE_1	boy#1	endorsement#5	0.235	0.207	0.071	0.051	0.045	0.379	0.250	0.227	0.077	0.062	0.053	0.394
7	walad_1	Zuhor_1	boy#1	midday#1	0.222	0.188	0.067	0.042	0.038	0.366	0.235	0.207	0.071	0.051	0.045	0.379
8	basomap_1	qaroyap_1	smile#1	village#2	0.235	0.207	0.071	0.051	0.045	0.379	0.267	0.248	0.083	0.076	0.062	0.411
9	Zuhor_1	Sawom_1	noon#1	fasting#1	0.364	0.188	0.067	0.049	0.065	0.574	0.364	0.188	0.067	0.049	0.065	0.574
10	kuwb_1	AlomAs_1	glass#2	diamond#2	0.353	0.248	0.083	0.086	0.087	0.535	0.267	0.248	0.083	0.076	0.062	0.411
11	maqaAm_1	ra}iyos_1	shrine#1	head#4	0.500	0.272	0.091	0.110	0.139	0.687	0.556	0.326	0.111	0.164	0.192	0.720
12	riyf_1	xuDaAr_1	country#4	vegetable#1	0.375	0.272	0.091	0.105	0.102	0.553	0.286	0.272	0.091	0.092	0.073	0.429
13	kuwb_1	>aadaAp_1	tool#1	glass#2	0.778	0.922	1.000	0.818	0.789	0.927	0.750	0.486	0.200	0.367	0.394	0.828
14	DaHik_2	<iHotifaAl_1	laugh#1	celebration#2	0.375	0.272	0.091	0.105	0.102	0.553	0.375	0.272	0.091	0.105	0.102	0.553
15	fataAp_1	xaAdim_1	girl#1	retainer#2	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
16	Eiyod_1	Sawom_1	day#3	fasting#1	0.316	0.207	0.071	0.058	0.063	0.504	0.316	0.207	0.071	0.058	0.063	0.504
17	HaAfilap_1	wasiylap_1	coach#5	means#2	0.778	0.486	0.200	0.368	0.412	0.845	0.750	0.486	0.200	0.367	0.394	0.828
18	fayolasuwf_1	ra}iyos_1	philosopher#1	head#4	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
19	fataAp_1	>xot_1	girl#1	sister#1	0.696	0.358	0.125	0.202	0.261	0.815	0.667	0.358	0.125	0.202	0.254	0.796
20	dajaAjap_1	HamaAm_1	hen#1	pigeon#1	0.828	0.436	0.167	0.301	0.376	0.879	0.815	0.436	0.167	0.301	0.374	0.872
21	rukaAm_1	jabal_1	hill#2	mountain#1	0.533	0.358	0.125	0.199	0.201	0.692	0.500	0.395	0.143	0.233	0.195	0.649
22	say~id_1	ra}iyos_1	sir#1	head#4	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
23	TaEaAm_1	xuDaAr_1	food#1	vegetable#1	0.571	0.395	0.143	0.243	0.236	0.716	0.500	0.395	0.143	0.233	0.195	0.649
24	xaAdim_1	xaAdim_1	retainer#2	retainer#2	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
25	jaroy_1	ma$oy_1	run#7	walk#1	0.909	0.624	0.333	0.549	0.604	0.905	0.909	0.624	0.333	0.549	0.604	0.905
26	Habol_1	xayoT_2	cord#1	string#9	0.286	0.272	0.091	0.092	0.073	0.429	0.308	0.298	0.100	0.113	0.086	0.451
27	gaAbap_1	dagol_1	forest#1	jungle#1	0.308	0.298	0.100	0.113	0.086	0.451	0.333	0.326	0.111	0.138	0.102	0.475
28	fayolasuwf_1	mufak~ir_1	philosopher#1	intellect#3	0.900	0.624	0.333	0.549	0.597	0.900	0.889	0.624	0.333	0.549	0.587	0.894
29	riHolap_1	safar_1	journey#1	travel#1	0.952	0.734	0.500	0.670	0.710	0.926	0.952	0.734	0.500	0.670	0.710	0.926
30	HajarN_kariym_1	AlomAs_1	gem#2	diamond#2	0.875	0.624	0.333	0.549	0.570	0.886	0.857	0.624	0.333	0.548	0.545	0.875
31	riyf_1	qaroyap_2	country#4	village#2	0.824	0.546	0.250	0.449	0.485	0.864	0.857	0.624	0.333	0.548	0.545	0.875
32	wisaAdap_1	wisaAdap_1	cushion#3	cushion#3	1.000	0.922	1.000	0.819	0.811	0.955	1.000	0.922	1.000	0.818	0.789	0.953
33	basomap_1	DaHik_2	smile#1	laugh#1	0.533	0.358	0.125	0.199	0.201	0.692	0.533	0.358	0.125	0.199	0.201	0.692
34	tawoqiyE_1	tawoqiyE_1	Endorsement#5	Endorsement#5	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
35	>adaAp_1	wasiyolap_1	tool#2	means#1	0.941	0.734	0.500	0.670	0.690	0.918	0.941	0.734	0.500	0.670	0.690	0.918
36	qabor_1	qabor_1	grave#2	grave#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955
37	walad_1	Sabiy~_2	boy#1	spring chicken#1	0.800	0.486	0.200	0.368	0.424	0.858	0.778	0.486	0.200	0.368	0.412	0.845
38																
39	HaAfilap_1	HaAfilap_1	coach#5	bus#1	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
40	kuwb_1	kuwb_1	glass#2	glass#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955

	Performance Measures											
Sim. level	Correlation r						Correlation r					
all	0.787	0.703	0.540	0.703	0.719	0.757	0.787	0.763	0.616	0.763	0.762	0.774
low	0.615	0.569	0.583	0.610	0.645	0.586	0.391	0.555	0.576	0.538	0.470	0.318
mid	0.453	0.067	-0.082	0.088	0.152	0.345	0.434	0.473	0.397	0.466	0.452	0.407
high	0.672	0.656	0.627	0.647	0.658	0.693	0.665	0.643	0.624	0.635	0.653	0.689
Sim. level	MSE						MSE					
all	0.051	0.061	0.155	0.097	0.089	0.082	0.050	0.051	0.140	0.087	0.084	0.076
low	0.062	0.027	0.008	0.006	0.006	0.170	0.065	0.033	0.007	0.006	0.007	0.167
mid	0.046	0.065	0.199	0.114	0.096	0.071	0.043	0.039	0.165	0.094	0.085	0.059
high	0.047	0.084	0.230	0.153	0.147	0.021	0.044	0.078	0.221	0.144	0.144	0.019

Figure 1, before the enhancement of the ArWN structure, the $HyperTree$ has two semantic gaps {*geological formation*#1} and {*physical entity*#1}; while The $HyperTree$(جبل) has one semantic gap, which is {*physical entity*#1}.

The performance measures r and MSE are reported for every configuration in the bottom ofTables 6, 5, 7, and 8; including the performance for each similarity level. Observe that, r values show that iHT achieves better performance compared to uHT. While; MSE values indicate that uHT has less difference in similarity scores than iHT, compared to HR rates. In fact, the values of MSE are strongly influenced by uHT, the semantic gaps. Noting that when $HyperTrees$ of two senses have the same semantic gaps; the lcs is reduced which decreases the similarity scores. This gives less difference in similarity scores compared to HR rates. In particular, this happens for MSE values at mid similarity level. For examples, in row 10, the $HyperTrees$ of the word pairs (Glass; Diamond) has the {*physical entity*#1} as a semantic gap. That is, $d(glass\#2) = 9$; $d(diamond\#2) = 8$ and $d(lcs(glass\#2, diamond\#2)) = 3$; while $d(kuwb\_1) = 8$; $d(AlomAs\_1) = 7$; $d(lcs(kuwb\_1, AlomAs\_1)) = 2$.

Furthermore, $wnTrans$ configuration scored the worst performance; this is due to the low Arabic word coverage. A significant finding is that, the richness of ArWN content has a high effect on the evaluation the semantic similarity between the concepts, in terms of the coverage of lexical and semantic relations.

Performance measures in [15, 16]; presented in Table 4; showed that WuP measure scored the best MSE value 0.0165 with 0.94 for r; and comparatively $Aldiery$ measure has obtained the values 0.96

Table 8: UB configuration over ArWN

					iHT						uHT					
NO.	Ar Word Pairs Senses		En Word Pairs Senses		WuP	LCH	Path	LI	AWSS	Aldiery	WuP	LCH	Path	LI	AWSS	Aldiery
1	$aATi}_AlbaHor_1	tawoqiyE_1	coast#1	endorsement#5	0.235	0.207	0.071	0.051	0.045	0.379	0.267	0.248	0.083	0.076	0.062	0.411
2	Zuhor_1	gazol_1	noon#1	thread#1	0.200	0.154	0.059	0.028	0.028	0.343	0.211	0.170	0.063	0.034	0.033	0.354
3																
4	Habol_1	Zuhor_1	cord#1	midday#1	0.211	0.170	0.063	0.034	0.033	0.354	0.222	0.188	0.067	0.042	0.038	0.366
5	tawoqiyE_1	gazol_1	endorsement#5	thread#1	0.211	0.170	0.063	0.034	0.033	0.354	0.222	0.188	0.067	0.042	0.038	0.366
6	walad_2	tawoqiyE_1	boy#2	endorsement#5	0.222	0.188	0.067	0.042	0.038	0.366	0.235	0.207	0.071	0.051	0.045	0.379
7	walad_2	Zuhor_1	boy#2	midday#1	0.211	0.170	0.063	0.034	0.033	0.354	0.222	0.188	0.067	0.042	0.038	0.366
8	basomap_1	qaroyap_2	smile#1	village#2	0.235	0.207	0.071	0.051	0.045	0.379	0.267	0.248	0.083	0.076	0.062	0.411
9	mu&ax~irap_1	Sawom_1	back#2	fasting#1	0.200	0.154	0.059	0.028	0.028	0.343	0.211	0.170	0.063	0.034	0.033	0.354
10	kuwb_1	AlomAs_1	glass#2	diamond#2	0.353	0.248	0.083	0.086	0.087	0.535	0.267	0.248	0.083	0.076	0.062	0.411
11	qabor_1	$ayox_2	grave#2	senator#1	0.400	0.227	0.077	0.073	0.089	0.601	0.333	0.227	0.077	0.070	0.074	0.519
12	riyf_1	xuDar_1	country#4	green#7	0.353	0.248	0.083	0.086	0.087	0.535	0.267	0.248	0.083	0.076	0.062	0.411
13	kuwb_1	>adaAp_1	glass#2	tool#2	0.222	0.922	1.000	0.683	0.371	0.691	0.235	0.207	0.071	0.051	0.045	0.379
14	DaHik_1	Eiyod_1	laughter#2	day#3	0.375	0.272	0.091	0.105	0.102	0.553	0.375	0.272	0.091	0.105	0.102	0.553
15	fataAp_1	xaAdim_1	girl#1	retainer#2	0.762	0.436	0.167	0.301	0.361	0.842	0.737	0.436	0.167	0.301	0.351	0.827
16	<iHotifAl_1	Sawom_1	celebration#1	fasting#1	0.526	0.298	0.100	0.135	0.163	0.703	0.526	0.298	0.100	0.135	0.163	0.703
17	HaAfilap_1	wasiylap_1	coach#5	means#2	0.778	0.486	0.200	0.368	0.412	0.845	0.750	0.486	0.200	0.367	0.394	0.828
18	fayolasuwf_1	$ayox_2	philosopher#1	senator#1	0.696	0.358	0.125	0.202	0.261	0.815	0.667	0.358	0.125	0.202	0.254	0.796
19	fataAp_1	>xot_1	girl#1	sister#1	0.696	0.358	0.125	0.202	0.261	0.815	0.667	0.358	0.125	0.202	0.254	0.796
20	dajaAjap_1	HamaAm_1	hen#1	pigeon#1	0.828	0.436	0.167	0.301	0.376	0.879	0.815	0.436	0.167	0.301	0.374	0.872
21	rukaAm_1	jabal_1	hill#2	mountain#1	0.533	0.358	0.125	0.199	0.201	0.692	0.500	0.395	0.143	0.233	0.195	0.649
22	say~id_1	$ayox_1	lord#3	graybeard#1	0.696	0.358	0.125	0.202	0.261	0.815	0.667	0.358	0.125	0.202	0.254	0.796
23	TaEaAm_3	xuDar_1	food#2	green#7	0.800	0.546	0.250	0.449	0.464	0.850	0.769	0.546	0.250	0.447	0.431	0.831
24	Eabod_1	xaAdim_1	slave#1	retainer#2	0.842	0.546	0.250	0.449	0.499	0.874	0.824	0.546	0.250	0.449	0.485	0.864
25	jaroy_1	ma$oy_1	run#7	walk#1	0.909	0.624	0.333	0.549	0.604	0.905	0.909	0.624	0.333	0.549	0.604	0.905
26	Habol_1	xayoT_1	cord#1	cord#4	0.750	0.486	0.200	0.367	0.394	0.828	0.714	0.486	0.200	0.366	0.367	0.805
27	dagol_1	dagol_1	jungle#1	jungle#1	1.000	0.922	1.000	0.818	0.754	0.950	1.000	0.922	1.000	0.815	0.701	0.947
28	fayolasuwf_1	mufak~ir_1	philosopher#1	intellect#3	0.900	0.624	0.333	0.549	0.597	0.900	0.889	0.624	0.333	0.549	0.587	0.894
29	riHolap_1	safar_1	journey#1	travel#3	0.857	0.546	0.250	0.449	0.508	0.883	0.857	0.546	0.250	0.449	0.508	0.883
30	HajarN_kariym_1	AlomAs_1	gem#2	diamond#2	0.875	0.624	0.333	0.549	0.570	0.886	0.857	0.624	0.333	0.548	0.545	0.875
31	riyf_1	qaroyap_2	country#4	village#2	0.824	0.546	0.250	0.449	0.485	0.864	0.857	0.624	0.333	0.548	0.545	0.875
32	wisaAdap_1	wisaAdap_1	cushion#3	cushion#3	1.000	0.922	1.000	0.819	0.811	0.955	1.000	0.922	1.000	0.818	0.789	0.953
33	basomap_1	DaHik_2	smile#1	laugh#1	0.533	0.358	0.125	0.199	0.201	0.692	0.533	0.358	0.125	0.199	0.201	0.692
34	tawoqiyE_1	tawoqiyE_1	endorsement#5	endorsement#5	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
35	>adaAp_1	wasiyolap_1	tool#2	means#1	0.941	0.734	0.500	0.670	0.690	0.918	0.941	0.734	0.500	0.670	0.690	0.918
36	qabor_1	qabor_1	grave#2	grave#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955
37	muraAhiq_1	Tifol_1	adolescent#1	child#1	0.900	0.624	0.333	0.549	0.597	0.900	0.889	0.624	0.333	0.549	0.587	0.894
38																
39	HaAfilap_1	HaAfilap_1	coach#5	bus#1	1.000	0.922	1.000	0.819	0.835	0.958	1.000	0.922	1.000	0.819	0.826	0.957
40	kuwb_1	kuwb_1	glass#2	glass#2	1.000	0.922	1.000	0.819	0.826	0.957	1.000	0.922	1.000	0.819	0.811	0.955

	Performance Measures											
Sim. level	Correlation r						Correlation r					
all	0.935	0.787	0.580	0.813	0.874	0.945	0.934	0.849	0.662	0.856	0.879	0.943
low	0.825	0.696	0.710	0.765	0.815	0.828	0.621	0.458	0.456	0.509	0.591	0.613
mid	0.815	-0.087	-0.358	0.093	0.537	0.739	0.806	0.792	0.734	0.766	0.773	0.788
high	0.345	0.415	0.402	0.419	0.465	0.333	0.372	0.413	0.401	0.417	0.509	0.369
Sim. level	MSE						MSE					
all	0.023	0.046	0.144	0.073	0.058	0.054	0.022	0.037	0.131	0.070	0.062	0.047
low	0.035	0.019	0.008	0.007	0.006	0.114	0.034	0.025	0.008	0.008	0.008	0.105
mid	0.022	0.060	0.205	0.105	0.074	0.055	0.018	0.033	0.175	0.099	0.084	0.041
high	0.015	0.054	0.193	0.096	0.083	0.006	0.015	0.051	0.186	0.091	0.085	0.006

and 0.0189 for r and MSE; respectively. Nevertheless, [15, 16] did not explicitly state which configuration was considered in calculating the similarity scores. For instance. in Table 8; where the UB scores indicate the best value for r is 0.945 (with 0.542 for MSE); which is obtained by *Aldiery* measure; and the best MSE value is 0.0203 (with 0.935 for r); which is obtained by WuP measure. Further, in [15, 16] semantic similarity scores were reported to be equal to zero for the word pairs in rows $1-9$, which are at the low similarity level, and the word pair in row 21 was considered as not covered ArWN, hence, this increased the r values and reduced MSE values. However, no explanation is provided.

Overall, the reported performance values show that the enhancement of the semantic structure has a strong effect on estimating the semantic similarity between the concepts. Observe that, word pairs at low and mid similarity levels gives better r values than high similarity level. While words pairs in high similarity level gives better MSE values. in other words, similarity measures obtained best coloration values when the concepts are not similar. Both ArWN and EnWN, r and MSE measures indicate that best performance is achieved when word senses are determined in advance, i.e., DS configuration. However, it is important to distinguish the approach which is used to define the sense, in this work consensus based approach is used.

In other hand; the user feedback based approach, ICLM application that adopted to fill the semantic gaps, shows its effectiveness in selecting the senses, such that scores obtained in DS are close to optimal scores achieved with upper bound setting UB. Further, Arabic-based measure *Aldiery* performs better than $AWSS$, also

Table 9: uDS, DS, and UB configuration over EnWN

	uDS						DS						UB					
NO.	En Word Pairs senses		WuP	LCH	Path	LI	En Word Pairs senses		WuP	LCH	Path	LI	En Word Pairs senses		WuP	LCH	Path	LI
1	coast#4	endorsement#2	0.632	0.436	0.125	0.202	coast#1	endorsement#1	0.286	0.350	0.091	0.092	coast#1	endorsement#5	0.235	0.285	0.071	0.051
2	noon#1	string#9	0.353	0.326	0.083	0.086	noon#1	string#9	0.353	0.326	0.083	0.086	noon#1	string#2	0.182	0.202	0.053	0.019
3	stove#2	walk#5	0.632	0.436	0.125	0.202	stove#1	walk#1	0.167	0.175	0.048	0.013	stove#1	walk#6	0.160	0.162	0.045	0.010
4	cord#2	midday#1	0.316	0.285	0.071	0.058	cord#1	midday#1	0.211	0.248	0.063	0.034	cord#3	midday#1	0.190	0.216	0.056	0.023
5	signature#5	string#7	0.737	0.514	0.167	0.301	signature#1	string#1	0.235	0.285	0.071	0.051	signature#4	string#2	0.200	0.232	0.059	0.028
6	boy#1	endorsement#1	0.286	0.350	0.091	0.092	boy#1	endorsement#5	0.235	0.285	0.071	0.051	boy#2	endorsement#5	0.222	0.266	0.067	0.042
7	boy#1	midday#1	0.222	0.266	0.067	0.042	boy#1	midday#1	0.222	0.266	0.067	0.042	boy#2	midday#1	0.211	0.248	0.063	0.034
8	smile#1	village#1	0.375	0.350	0.091	0.105	smile#1	village#1	0.375	0.350	0.091	0.105	smile#1	village#2	0.235	0.285	0.071	0.051
9	noon#1	fasting#1	0.364	0.266	0.067	0.049	noon#1	fasting#1	0.364	0.266	0.067	0.049	noon#1	fasting#1	0.364	0.266	0.067	0.049
10	glass#1	diamond#2	0.667	0.514	0.167	0.300	glass#1	diamond#1	0.353	0.326	0.083	0.086	glass#4	diamond#3	0.148	0.138	0.042	0.007
11	sepulcher#1	sheikh#1	0.476	0.326	0.083	0.090	sepulcher#1	sheikh#1	0.476	0.326	0.083	0.090	sepulcher#1	sheikh#1	0.476	0.326	0.083	0.090
12	countryside#1	vegetable#2	0.400	0.305	0.077	0.073	countryside#1	vegetable#2	0.400	0.305	0.077	0.073	countryside#1	vegetable#1	0.353	0.326	0.083	0.086
13	tumbler#2	tool#1	0.737	0.514	0.167	0.301	tumbler#2	tool#1	0.737	1.000	1.000	0.818	tumbler#1	tool#4	0.316	1.000	1.000	0.775
14	laugh#1	feast#2	0.400	0.376	0.100	0.128	laugh#1	feast#2	0.400	0.376	0.100	0.128	laugh#2	feast#1	0.333	0.305	0.077	0.070
15	girl#1	odalisque#1	0.833	0.564	0.200	0.368	girl#1	odalisque#1	0.833	0.564	0.200	0.368	girl#3	odalisque#1	0.750	0.472	0.143	0.247
16	feast#2	fasting#1	0.526	0.376	0.100	0.135	feast#2	fasting#1	0.526	0.376	0.100	0.135	feast#4	fasting#1	0.500	0.350	0.091	0.110
17	coach#5	means#2	0.778	0.564	0.200	0.368	coach#5	means#2	0.778	0.564	0.200	0.368	coach#1	means#2	0.526	0.376	0.100	0.135
18	Sage#1	Sheikh#1	0.762	0.514	0.167	0.301	Sage#1	Sheikh#1	0.762	0.514	0.167	0.301	Sage#3	Sheikh#1	0.636	0.404	0.111	0.165
19	girl#1	sister#4	0.957	0.812	0.500	0.670	girl#1	sister#1	0.696	0.436	0.125	0.202	girl#1	sister#1	0.696	0.436	0.125	0.202
20	Hen#2	pigeon#1	0.846	0.564	0.200	0.368	hen#2	pigeon#1	0.846	0.564	0.200	0.368	hen#4	pigeon#1	0.828	0.514	0.167	0.301
21	hill#1	mountain#1	0.857	0.702	0.333	0.548	hill#1	mountain#1	0.857	0.702	0.333	0.548	hill#2	mountain#1	0.667	0.404	0.111	0.165
22	master#2	Sheikh#1	0.900	0.702	0.333	0.549	master#2	Sheikh#1	0.900	0.702	0.333	0.549	master#7	Sheikh#1	0.667	0.404	0.111	0.165
23	food#2	vegetable#1	0.857	0.702	0.333	0.548	food#2	vegetable#1	0.857	0.702	0.333	0.548	food#1	vegetable#1	0.571	0.472	0.143	0.243
24	slave#1	odalisque#1	0.727	0.472	0.143	0.247	slave#1	odalisque#1	0.727	0.472	0.143	0.247	slave#2	odalisque#1	0.696	0.436	0.125	0.202
25	run#7	walk#1	0.909	0.702	0.333	0.549	run#7	walk#1	0.909	0.702	0.333	0.549	run#6	walk#1	0.750	0.472	0.143	0.247
26	cord#1	string#1	0.941	0.812	0.500	0.670	cord#1	string#1	0.941	0.812	0.500	0.670	cord#3	string#2	0.762	0.514	0.167	0.301
27	forest#2	woodland#1	1.000	1.000	1.000	0.818	forest#2	woodland#1	1.000	1.000	1.000	0.818	forest#2	woodland#1	1.000	1.000	1.000	0.818
28	Sage#1	thinker#1	0.857	0.624	0.250	0.449	Sage#1	thinker#1	0.857	0.624	0.250	0.449	Sage#1	thinker#1	0.857	0.624	0.250	0.449
29	journey#1	travel#1	0.952	0.812	0.500	0.670	journey#1	travel#1	0.952	0.812	0.500	0.670	journey#1	travel#3	0.857	0.624	0.250	0.449
30	Gem#5	diamond#1	0.952	0.812	0.500	0.670	Gem#5	diamond#1	0.952	0.812	0.500	0.670	gem#2	diamond#2	0.875	0.702	0.333	0.549
31	countryside#1	village#2	0.778	0.564	0.200	0.368	countryside#1	village#2	0.778	0.564	0.200	0.368	countryside#1	village#2	0.778	0.564	0.200	0.368
32	cushion#3	pillow#1	0.941	0.812	0.500	0.670	cushion#3	pillow#1	0.941	0.812	0.500	0.670	cushion#3	pillow#1	0.941	0.812	0.500	0.670
33	smile#1	laugh#2	0.875	0.702	0.333	0.549	smile#1	laugh#2	0.875	0.702	0.333	0.549	smile#1	laugh#2	0.875	0.702	0.333	0.549
34	signature#1	endorsement#4	0.941	0.812	0.500	0.670	signature#1	endorsement#4	0.941	0.812	0.500	0.670	signature#1	endorsement#4	0.941	0.812	0.500	0.670
35	tool#2	means#1	0.941	0.812	0.500	0.670	tool#1	means#2	0.824	0.624	0.250	0.449	tool#2	means#1	0.941	0.812	0.500	0.670
36	sepulcher#1	grave#2	0.941	0.812	0.500	0.670	sepulcher#1	grave#2	0.941	0.812	0.500	0.670	sepulcher#1	grave#2	0.941	0.812	0.500	0.670
37	boy#1	lad#2	0.952	0.812	0.500	0.670	boy#1	lad#2	0.952	0.812	0.500	0.670	boy#1	lad#2	0.952	0.812	0.500	0.670
38	wizard#2	magician#2	1.000	1.000	1.000	0.819	wizard#2	magician#2	1.000	1.000	1.000	0.819	wizard#2	magician#2	1.000	1.000	1.000	0.819
39	coach#5	bus#1	1.000	1.000	1.000	0.819	coach#5	bus#1	1.000	1.000	1.000	0.819	coach#5	bus#1	1.000	1.000	1.000	0.819
40	glass#2	tumbler#2	0.947	0.812	0.500	0.670	glass#2	tumbler#2	0.947	0.812	0.500	0.670	glass#2	tumbler#2	0.947	0.812	0.500	0.670

Performance Measures												
sim. level	correlation r				correlation r				correlation r			
all	0.856	0.851	0.726	0.865	0.949	0.832	0.623	0.845	0.965	0.801	0.565	0.796
low	-0.071	-0.247	-0.241	-0.231	0.697	0.246	0.223	0.310	0.731	0.570	0.645	0.776
mid	0.666	0.604	0.542	0.612	0.675	0.010	-0.289	0.093	0.791	-0.319	-0.488	-0.309
high	0.261	0.268	0.234	0.279	0.139	0.174	0.173	0.170	0.611	0.550	0.422	0.601
sim. level	MSE				MSE				MSE			
all	0.066	0.038	0.104	0.046	0.036	0.040	0.126	0.056	0.016	0.043	0.160	0.091
low	0.150	0.089	0.011	0.021	0.059	0.056	0.008	0.008	0.035	0.035	0.007	0.006
mid	0.055	0.013	0.125	0.051	0.046	0.044	0.174	0.081	0.010	0.068	0.250	0.174
high	0.008	0.018	0.161	0.062	0.009	0.023	0.179	0.073	0.005	0.026	0.205	0.088

Aldiery measure provided a competitive performance in comparison to WuP measures.

5 Conclusion & Future Work

Six path-based similarity measures including English and Arabic based measures are applied over ArWN and EnWN to examine the effect of the improvement of the lexical and semantic coverage on wordnet-based semantic similarity measures. wo variants uHT and iHT of ArWN structure are considered in the experiment to evaluate the impact of filling the semantic gaps on estimating the semantic similarity. The efficacy of the improved structure is examined by experiments in the context of semantic similarity. The semantic similarity scores for a benchmark dataset, human rating for 40 Arabic nominal word pairs, are calculated over ArWN and EnWN in different configurations (uDs, DS, $wnTrans$, and UB). The obtained performance values indicate the importance of the semantic evidence gained with the enrichment process; and its signification effect on estimating the semantic similarity between concepts. Moreover, when considering Arabic-based measures the experiment results showed that *Aldiery* measure performs better than $AWSS$ measure. Beside that, *Aldiery* measure has provided a competitive performance in comparison to the English-based WuP measures. Finally, the resolved semantic gaps of the new structure are made for public.

As a future direction, we plan to compile *xml* format of the new structure, and to integrate it with available ArWN resources (i,e., ArWN release available at Open Multilingual WordNet [31]). It is also interesting is to study the effect of the semantic gaps over NLP applications; for instances Question Answering similar to the work presented in [44], and word sense disambiguation [33, 35] in the context of Arabic.

References

[1] D. Jurafsky, J. H. Martin, Speech and Language Processing (2Nd Edition), Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[2] Y. Li, Z. A. Bandar, D. Mclean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. on Knowl. and Data Eng., **15**(4), 871–882, 2003, doi:10.1109/TKDE.2003.1209005.

[3] G. A. Miller, "WordNet: A Lexical Database for English," Commun. ACM, **38**(11), 39–41, 1995, doi:10.1145/219717.219748.

[4] C. Fellbaum, editor, WordNet An Electronic Lexical Database, The MIT Press, Cambridge, MA ; London, 1998.

[5] F. Christiane, H. Amanda, "When WordNet Met Ontology," in Ontology Makes Sense - Essays in honor of Nicola Guarino, 136–151, 2019, doi: 10.3233/978-1-61499-955-3-136.

[6] N. Bouhriz, F. Benabbou, E. H. B. Lahmar, "Word Sense Disambiguation Approach for Arabic Text," (IJACSA) International Journal of Advanced Computer Science and Applications, **7**(4), 2016, doi:10.14569/IJACSA.2016.070451.

[7] P. Vossen, "Eurowordnet: A Multilingual Database Of Autonomous And Language-Specific Wordnets Connected Via An Inter-Lingualindex," International Journal of Lexicography, **17**(2), 161–173, 2004, doi:10.1093/ijl/17.2.161.

[8] M. Arcan, J. P. McCrae, P. Buitelaar, "Polylingual Wordnet," CoRR, **abs/1903.01411**, 2019.

[9] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. A. Martí, W. Black, S. Elkateb, J. Kirk, P. Vossen, C. Fellbaum, "Arabic WordNet: Current State and Future Extensions," in Proceedings of the Forth International Conference on Global WordNet, 2008.

[10] G. A. Miller, W. G. Charles, "Contextual correlates of semantic similarity," Language & Cognitive Processes, **6**(1), 1–28, 1991.

[11] F. Bond, L. Morgado da Costa, M. W. Goodman, J. P. McCrae, A. Lohk, "Some Issues with Building a Multilingual Wordnet," in Proceedings of the 12th Language Resources and Evaluation Conference, 3189–3197, European Language Resources Association, Marseille, France, 2020.

[12] L. Abouenour, K. Bouzoubaa, P. Rosso, "On the evaluation and improvement of Arabic WordNet coverage and usability," Language Resources and Evaluation, **47**, 2013, doi:10.1007/s10579-013-9254-z.

[13] H. Rodríguez, D. Farwell, J. Ferreres, M. Bertran, M. Alkhalifa, A. Martí, "Arabic WordNet: Semi-automatic Extensions using Bayesian Inference," 2008.

[14] M. A. Batita, M. Zrigui, "The Enrichment of Arabic WordNet Antonym Relations," in A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, 342–353, Springer International Publishing, Cham, 2018, doi:10.1007/978-3-319-77113-7_27.

[15] N. Mohammed, D. Mohammed, "Experimental Study of Semantic Similarity Measures on Arabic WordNet," International Journal of Computer Science and Network Security, **17**(2), 2017.

[16] M. G. Aldayri, The semantic similarity measures using Arabic ontology, (Master's theses Theses and Dissertations Master). Middle East University, Jordan, 2017.

[17] M. A. Helou, "Effects of Semantic Gaps on Arabic WordNet-Based Similarity Measures," in 2019 International Conference on Innovative Computing (ICIC), 1–10, 2019, doi:10.1109/ICIC48496.2019.8966672.

[18] M. A. Helou, M. Palmonari, "Multi-user Feedback for Large-scale Cross-lingual Ontology Matching," in Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Funchal, Madeira, Portugal, November 1-3, 57–66, 2017, doi:10.5220/0006503200570066.

[19] R. Rada, H. Mili, E. Bicknell, M. Blettner, "Development and application of a metric on semantic nets," in IEEE Transactions on Systems, Man and Cybernetics, 17–30, 1989, doi:10.1109/21.24528.

[20] Z. Wu, M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 133–138, Association for Computational Linguistics, Morristown, NJ, USA, 1994.

[21] C. Leacock, M. Chodorow, "Combining local context and WordNet similarity for word sense identification," WordNet: An electronic lexical database, **49**(2), 265–283, 1998.

[22] F. A. Almarsoomi, J. D. OShea, Z. Bandar, K. Crockett, "AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity," in 2013 IEEE International Conference on Systems, Man, and Cybernetics, 504–509, 2013, doi:10.1109/SMC.2013.92.

[23] F. A, Almarsoomi, J. D, O'Shea, Z. A, Bandar, K. A, Crockett, "Arabic Word Semantic Similarity," International Journal of Cognitive and Language Sciences, **6**(10), 2497 – 2505, 2012, doi:10.5281/zenodo.1080052.

[24] G. Hirst, "Ontology and the Lexicon," in eds. S. Staab, R. Studer, editors, Handbook on Ontologies and Information Systems, Heidelberg, Springer, 2004, doi:https://doi.org/10.1007/978-3-540-24750-0_11.

[25] G. A. Miller, C. Leacock, R. Tengi, R. T. Bunker, "A semantic concordance," in Proceedings of the workshop on Human Language Technology, HLT '93, 303–308, Association for Computational Linguistics, Stroudsburg, PA, USA, 1993.

[26] H. Graeme, "Overcoming Linguistic Barriers to the Multilingual Semantic Web," in P. Buitelaar, P. Cimiano, editors, Towards the Multilingual Semantic Web, 3–14, Springer Berlin Heidelberg, 2014, doi:10.1007/978-3-662-43585-4.

[27] D. Tufis, D. Cristea, S. Stamou, "BalkaNet: Aims, Methods, Results and Perspectives. A General Overview," In: D. Tufiş (ed): Special Issue on BalkaNet. Romanian JSTI, 2004.

[28] G. de Melo, G. Weikum, "Towards a universal wordnet by learning from combined evidence," in D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, J. J. Lin, editors, CIKM, 513–522, ACM, 2009, doi:10.1145/1645953.1646020.

[29] E. Pianta, L. Bentivogli, C. Girardi, "MultiWordNet: developing an aligned multilingual database," in Proceedings of the 1st Inter. Global Wordnet Conference, 2002.

[30] V. Nastase, M. Strube, B. Boerschinger, C. Zirn, A. Elghafari, "WikiNet: A Very Large Scale Multi-Lingual Concept Network," in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias, editors, LREC, European Language Resources Association, 2010.

[31] F. Bond, R. Foster, "Linking and Extending an Open Multilingual Wordnet," in ACL (1), 1352–1362, The Association for Computer Linguistics, 2013.

[32] M. M. Boudabous, N. Chaâben Kammoun, N. Khedher, L. H. Belguith, F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns," in 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 1–6, 2013.

[33] E. Bilel, "Arabic word sense disambiguation: a review," Artif. Intell. Rev., **52**(4), 2475–2532, 2019, doi:10.1007/s10462-018-9622-6.

[34] A. Budanitsky, G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," Comput. Linguist., **32**(1), 13–47, 2006, doi: 10.1162/coli.2006.32.1.13.

[35] N. Bouhriz, F. Benabbou, E. H. Benlahmar, Y. Zidoun, "Arabic Word Semantic similarity: a survey," in 1st International Conference On Research In Applied Mathematics And Computer, Casablanca, Morroco, 2019.

[36] H. Rubenstein, J. B. Goodenough, "Contextual correlates of synonymy," Commun. ACM, **8**(10), 627–633, 1965, doi:10.1145/365628.365657.

[37] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence, 448–453, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995.

[38] D. Lin, "An Information-Theoretic Definition of Similarity," in J. W. Shavlik, J. W. Shavlik, editors, ICML, 296–304, Morgan Kaufmann, 1998.

[39] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, 24–26, ACM, New York, NY, USA, 1986, doi:https://doi.org/10.1145/318723.318728.

[40] Z. Zhou, Y. Wang, J. Gu, "New model of semantic similarity measuring in wordnet," in 3rd International Conference on Intelligent System and Knowledge Engineering, volume 1, 256–261, 2008, doi:10.1109/ISKE.2008.4730937.

[41] M. A. Helou, A. Abid, "Semantic Measures based on Wordnet using Multiple Information Sources," in KDIR 2010 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Valencia, Spain, October 25-28, 2010, 500–503, 2010.

[42] M. A. Helou, M. Palmonari, M. Jarrar, "Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping," J. Artif. Intell. Res., **55**, 165–208, 2016, doi:10.1613/jair.4789.

[43] M. A. Helou, M. Palmonari, "Cross-lingual lexical matching with word translation and local similarity optimization," in Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS 2015, Vienna, Austria, September 15-17, 97–104, 2015, doi:10.1145/2814864.2814888.

[44] Y. Regragui, L. Abouenour, F. Krieche, K. Bouzoubaa, P. Rosso, "Arabic WordNet: New Content and New Applications," in In Proceeding of the 8th Global Wordnet Conference (GWN 2016), 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1322-1327 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Evaluation of Simple Space Interpolation Methods for the Depth of Precipitation: Application for Boyacá, Colombia

Pedro Mauricio Acosta Castellanos[*,1], Alejandra Castro Ortegon[2], Hugo Fernando Guerrero Sierra[3]

[1]*Faculty of Civil Engineering, Santo Tomas University, Tunja, 150003, Colombia*

[2]*Faculty of Environmental Engineering, Santo Tomas University, Tunja, 150003, Colombia*

[3]*Faculty of Economics and Social Sciences, La Salle University, Bogotá, 110231, Colombia*

ARTICLE INFO

Article history:
Received: 04 September, 2020
Accepted: 01 December, 2020
Online: 16 December, 2020

Keywords:
Regionalization
Interpolation
Precipitation
Hydrology

ABSTRACT

Interpolation tools are used daily in hydrology and climatology. With the purpose to regionalize spot´s registration parameters, such as depth of precipitation, temperature, humidity, among others. The accuracy of these methods is not fully validated. This research presents a comparative study between the most used interpolation methods for the regionalization of hydrological and climatological parameters. Comparative analysis of spatial interpolation; it was carried out using the IDW, Kriging and Spline methods, for this the ARGIS software was used, because its widespread and widespread use. The depth of precipitation was considered as a parameter for the comparison. Accuracy was determined by cross validation. Also, for the coefficient of determination and the comparison for visual errors such as "bull's eyes". The research was spatially restricted to a politically delimited region; Boyacá, Colombia, South America. The best method for spatial interpolation in this case was shown to be Spline.

1. Introduction

Rainfall records are of great importance, because from this information can be obtained among other analyses of flows, design storms and changes in rainfall behavior. The records and their analysis usually come from a network of meteorological stations, each corresponding to a specific geographical point [1]. With this information it is possible to estimate and design structures for water control, economic evaluation of flood protection projects, planning and management of land use, water quality control, among others [2].

Because rainfall information represents the record of a single geographical point, it is common for information to be researched and developed for interpolation in the form of regionalization maps [3]. This information is more common than rainfall, due to its simplicity in the form of record, in both cases can be obtained from calculation treatments and analysis rainfall runoff models and obtain important information such as the curves intensity, duration and frequency [4]. In this sense, regionalization can be defined as the transfer of data or spatial information from one station to another [5]. The transfer can range from data characteristics to hydrological models, [6]. A factor which determines a better condition in regionalisation is homogeneity which essentially refers to the geographical proximity or contiguity and hydrological similarity of meteorological stations [7].

Geographical dispersion in measurement networks means that there are gaps in the spatiality of information, specifically in areas without station coverage [8]. It is evident that the same applies to rainy seasons, for this reason, it is usual that different investigations are presented that seek to find the best method of regionalization and interpolation of hydrological data in order to present a viable alternative for obtaining data in areas where there is no availability. This may influence better and more accurate hydrological models, especially in small and medium catchment areas [9].

This paper presents the evaluation of three interpolation methods; Interpolation with Weighted Inverse Distance (IDW), Interpolation by Gaussian Regression Processes (Kriging) and Segmental Interpolation (Spline), applied to the regionalization of rainfall data, looking for the best method for current and future research, determining the degree of reliability of each method. The interpolation and evaluation were carried out for the rainfall data of monthly and annual records, the application of the study was

[*]Corresponding Author: P. M. Acosta Castellanos, pedro.acosta@usantoto.edu.co

limited to the geographical area comprised by the entire department of Boyacá, Colombia. These three methods or interpolation techniques were evaluated because they are the most commonly used by consultancies and research [10].

2. Methodology

2.1. Interpolation

This research seeks to validate or give support to what is used daily, for this reason the ordinary kriging and the circular model were used because within the wide spectrum of the kriging method, these are the two most used, it is important to note that they can make and create the empirical models for semivariograms but the vast majority of references and even companies like ARGIS indicate that the ordinary kriging and the circular model are the most used [11–19].

Due to the fact that rainfall information is of a point type due to the need for a spatial range and not a point range, interpolation tools are often used, which can define regions with similar behavior [20]. The commonly used methods are Inverse Distance Weighting or IDW, Kriging and Spline, in this sense there are few comparative studies that seek to define the most suitable spatial interpolation method for climatic or environmental variables such as precipitation [21]. Spatial interpolation methods can be defined in four categories: regression method or model and trend surface, local methods, geostatistical methods and mixed methods [9], Only local methods (IDW, Spline) and a known geostadistic method Kriging were evaluated for this research. Likewise, only monthly and annual rainfall depth record values were interpolated for the region of Boyacá, Colombia, because it has a spatially acceptable coverage of rainfall stations, which has a direct bearing on the use of interpolation methods since it is necessary to have a regional coverage to carry out interpolation, since the techniques preside over the variable of interest in a specific place taking values from the surrounding region, resulting in homogeneity [22]. The project will be carried out in the department of Boyacá, Colombia, which has 123 municipalities and 116 rainfall stations of which 60 were useful, since they fulfilled a range of 20 years of historical records counting from 1998 to 2018.

2.2. IDW method

The IDW interpolation method is widely used because of its simplicity [23]. Which is based on the fact that the unknown or sought value of a point has the greatest influence of the nearby control points and the degree of influence or weight of them [24], this implies that it is directly proportional to the inverse of the distance between the points posed [25]. When defining a high power, greater emphasis will be on the nearest points and the result of the surface will be more detailed, but less smoothed. By specifying a lower power, greater importance will be given to the more distant points and therefore the resulting surface will be smoother. equation (1) represents the method

$$Z = \frac{\sum_{i=1}^{N} W_i Z_i}{\sum_{i=1}^{N} W_i} = \frac{\sum_{i=1}^{N}\left[\frac{Z_i}{D_i^P}\right]}{\sum_{i=1}^{N}\left[\frac{1}{D_i^P}\right]} \quad (1)$$

where Z is the interpolated value for a point with the unknown observed value, Wi is the weighting function that determines the importance and relationship of each control point Zi, while Zi is the observed value of the control point i (i=1,...,n), which is the closest zone to the interpolated point and n is the total number of points used in interpolation [26]. The surfaces obtained by this method tend to be symmetrical over the points and reflect the mean value of the dispersion points in the regions between the points. In this method it is possible to control the importance of known points over interpolated values based on distance.

2.3. Kriging method

The method of interpolation Kriging, has a large number of variants that have been developed in recent times, the common is the Kriging Ordinary (KO). We have in this method that the observed or not observed search points are calculated by the linear weighted average of the observed or existing points [23], as shown in equation (2).

$$\widehat{Z_0} = \sum_{i=1}^{n} \lambda_i Z_i \quad (2)$$

where $\widehat{Z_0}$ represents the point searched or not observed, Zi means the attribute of the value of an observed point i, λi is the weight or significance of Zi.

The KO method presents the best linear prediction for points not observed, this means that in theory the value sought or point not observed is equal or very similar to the true value and the variance in the error is minimal [27].

2.4. Spline method

This interpolation method determines the unobserved points using a mathematical function that minimizes the general curvature of the surface, which has the particularity of being a smoothed surface. Many tests performed with Spline show their accuracy which means that the interpolated regions pass through the observed points [28]. There are two types of Splines mostly diffused, the Regularized and Stress, the first refers to softened regions or surfaces and the second creates less smoothed surfaces that are closer to the observed values, both cases result from the general equation of the equation method (3).

$$Z_{(x,y)} = T_{(x,y)} + \sum_{i=1}^{n} \lambda_i Z_{ri} \quad (3)$$

where Z is the value sought or point not observed, n, the number of points, λi coefficient of the system of linear equations, Zi is the distance of the point (x,y) to the point I [29].. T and R, are defined according to whether it is regularized or tension.

2.5. Hydrological Information

The information used for the assessment of interpolation methods is of the monthly and annual rainfall type. We reviewed the records of 116 stations located at the site of the study (Boyacá, Colombia), of which 60 had continuous records between 15 and 20 years. It should be recalled that rainy seasons record only the precipitation depth measured in (mm). The place where the research was applied is located in Colombia, South America, in the department of Boyacá, with an area of 23189 km², which represents 2.03% of the Colombian territory [30]. It is a region with three main hydrographic basins on the rivers Magdalena, Arauca and Meta, in the central area of Colombia, the figure 1, shows the area of study.

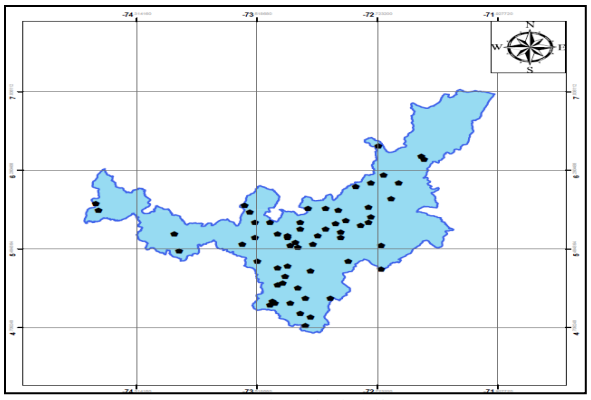


Figure 1: Location of rainfall seasons

3. Results and discussion

3.1. Validation of the methods evaluated.

The validation of the interpolated points using the three methods was performed by cross-validation, technique used to evaluate the results of a statistical analysis and to ensure that they are independent of the partitioning of the data obtained in this case from rainfall stations. The estimated value must be similar to the actual value, that is the measured value, so that the error committed by the model, that is, the difference between the two, is as low as possible [31].

The difference indicates how out-of-date the precipitation depth data calculated by the software are, when compared with the actual or measured information presented in the study area, that can be analyzed by the coefficient of determination or R2.

3.2. Cross-validation

Cross-validation is a way of predicting the fit of a model to a hypothetical test data set when the explicit test data set is not available. The cross-validation process gives a precise approximation of performance and error prediction [32]. Cross-validation in this study was performed by interpolating the totality of the rainfall stations and subtracting one weather station at a time, interpolating again, to verify by crossing the interpolated information against the actual or measured of each meteorological station. In other words, the interpolated data at the points where the measured data is held are verified and the actual or measured values and the interpolated values are compared.

3.3. Comparison of interpolation methods

Raster maps generated by a GIS were used in order to identify which interpolation method presents better results from different points of analysis; cross validation, coefficient of determination (R2) and the visual aspect, in the latter was sought to identify the error called "bull's eye". Based on the measured and calculated differences, statistical analyses were performed for the maximum annual and maximum monthly precipitation depth. The mean, median, minimum and standard deviation were compared, although it should be noted that greater importance was given to the statistical parameter of the mean, since it was considered to be the most relevant parameter in this case.

3.4. Interpolation of Annual precipitation depth.

It was observed that the Kriging technique has the least variation in the statistical parameter of the mean (54.70), of data standardization, but has the greatest dispersion of data with respect to the measured data (standard deviation), in addition it does not present an anisotropic behavior with spherical distribution behavior normalizing thus adjusting the plateau, data tolerance and bandwidth as shown in Figure 2.

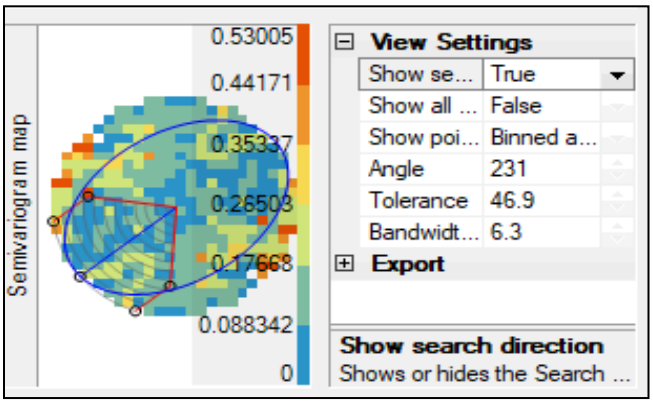


Figure 2: Data adjustment

In the standardization of data, a circular or spherical model was used in the semivariogram, adjusting the data to an exponential trend Figure 3, thus obtaining the highest reliability to perform the interpolation of IDW, ordinary kriging and spline as shown in Figure 4, 5, and 6 for 2008 year.

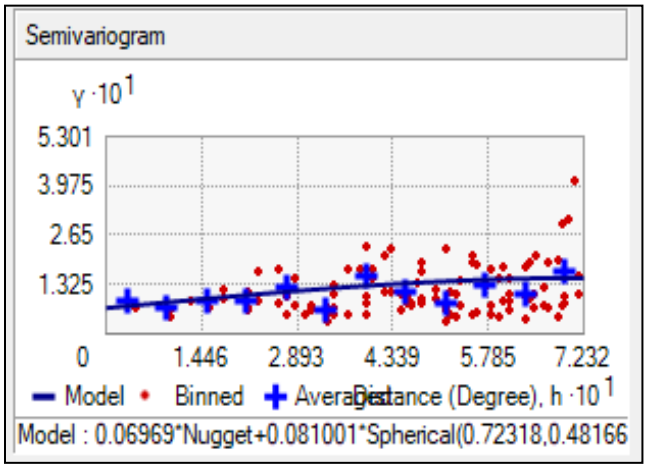


Figure 3: Semivariogram

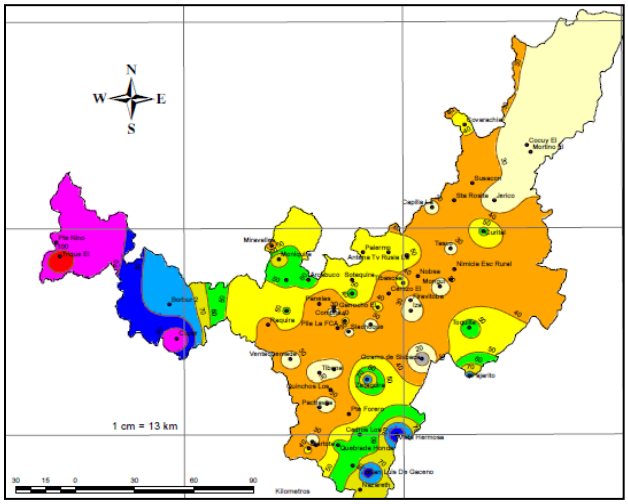


Figure 4: Interpolation by IDW, january 2008

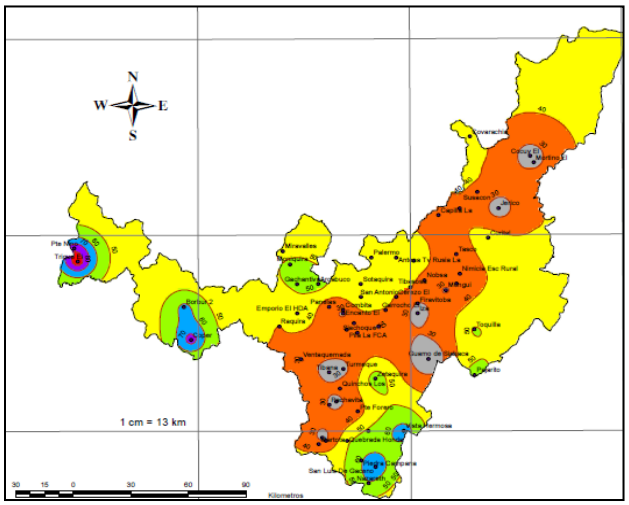


Figure 5: Interpolation by Kriging, january 2008

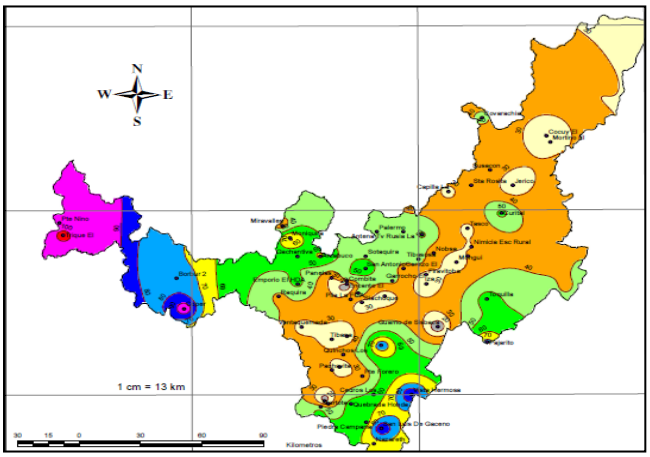


Figure 6: Interpolation by Spline, january 2008

On the other hand, the Spline technique also shows good performance in terms of average (54.66), but less than the Kriging value. The Spline method presented less dispersion than Kriging. Similarly, it was found that the IDW technique shows the greatest difference between the measured data and the interpolated data, in Table 1; the results of the analysis are evident.

Table 1: Statistical analysis of annual precipitation depth for 2008

Statistics	*Data measured*	*IDW*	*Spline*	*Kriging*
Mean	56.81	53.02	54.66	54.70
Median	47.1	49.04	48.40	51.04
Estándar deviation	30.51	20.12	20.99	16.26
Minimun	10.50	25.10	27.40	34.12

For the analysis of the interpolation of annual precipitation, it was determined that the kriging technique is the best fit in the statistical analysis of data, since the mean with the Kriging technique has a more appropriate behavior than the other two evaluated techniques.

3.5. Interpolation of monthly precipitation depth.

In the case of the interpolation of the monthly precipitation depth data, the Spline technique presented better results in terms of the statistical parameters of the mean and median, compared with the other techniques, although with great closeness between each method. In the second place, the Kriging technique is found in Table 2; the results of the analysis are evident.

Table 2: Statistical results of methods evaluated against mesuared monthy values, 2008

Statistics	Data measured	IDW	Spline	Kriging
Mean	66.31	62.22	63.95	63.86
Median	55.15	54.60	55.14	53.17
Estándar deviation	32.04	22.58	24.75	25.20
Minimun	28.80	35.13	31.70	34.44

3.6. Análisis de contornos.

Raster maps because they represent the space in a regular array of pixels, where the representation of the elements is made by pointing out the existence or not of data within each pixel [33], in the case of the regionalization of precipitation depth facilitates the visual behavior of the three interpolation techniques used in this investigation. Where we visually compared which of the three techniques has better contours and a more detailed surface in areas where there are no weather stations. One of the characteristics that is intended to show both in presence and absence is the typical error of "ox eyes" where interpolation techniques usually create concentration points or form islands around interpolated or measuring points [34], in this case the rainfall seasons. This anomaly is evident that with the IDW method where concentric circles are presented, as can be seen in Figure 1, this anomaly was presented as a visual trend a large percentage of the generated maps.

Ordinary kriging was the technique that presented the best results, since visually it presents different contours of precipitation depth, which is more accommodated to the recorded precipitation, consistent with the cross validation performed. In addition to performing the ordinary kriging, an additional raster is also interpolated to validate the interpolation of standard prediction errors, where the less data is those that may reliability and the larger data are those that should be resampled, as shown in figure 7.

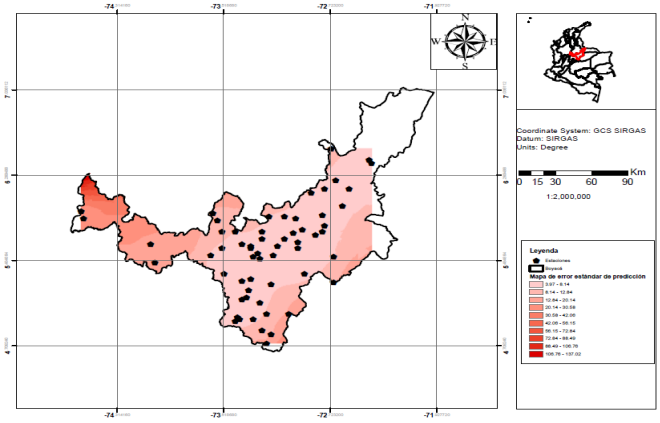


Figure 7: Standard Prediction Error Map.

3.7. Coefficient of determination R2

Raster maps because they represent the space in a regular array of pixels, where the representation of the elements is made

by pointing out the existence or not of data within each pixel [33], in the case of the regionalization of precipitation depth facilitates the visual behaviour of the three interpolation techniques used in this investigation.

3.8. Measurement of R2 in precipitation interpolation Annual

In this sense the method of interpolation of Kriging presents better correlation in the interpolation with depth of precipitation, the coefficient of determination is close to 1, with respect to the other models (Idw and Spline), thus yielding a high degree of reliability for choosing this model and carrying out the interpolation with which the regionalization of precipitation depth can be generated, due to the nature of this method it can be observed that it has a better behaviour with a higher data density, since analyzing interpolation with depth of monthly precipitation shows that the so-called spline model presents a greater correlation between its data, this allows choosing it to generate monthly regionalization.

Table 3: R2, para los métodos evaluados.

Interpolation Depth of Precipitation Max. Annual		Interpolation Precipitation Depth Max. Monthly	
Method	R2	Method	R2
Kriging	0,5762	Kriging	0,6717
Spline	0,5484	Spline	0,6861
IDW	0,5308	IDW	0,655

4. Conclusions

It was identified that visually the methods do not present an obvious difference, but in the case of the IDW technique generates the so-called portholes. For the regionalization of the maximum monthly precipitation depth it was possible to identify that the Spline method has a better behavior compared to the real data, since the R2, had a higher value (0.6861) compared to the other two methods evaluated.

In the case of the maximum monthly depth, the method that had a greater difference compared to the real data was the IDW, added to the visual errors (portholes) it is not advisable to use this method to regionalize hydrological data. Based on the analysis of both visual and cross-validation results, it can be stated that the kriging technique has a high degree of reliability when performing regionalizations with annual precipitation information.

In addition to the comparative results of the methods evaluated, the resulting information can be used as input for the determination of hydrological models in areas or locations that do not have their own information or measurement. In other words, they do not have a hydrological station. It could be shown that it is possible to use regionalized information with a high degree of confidence provided that it comes from data interpolated by the Spline method or the Kriging method

Twenty-four raster maps were generated, with maximum monthly and maximum annual rainfall depths, where twelve of these maps belong to the Spline technique which generates greater reliability with the maximum monthly values, and the remaining twelve were generated by the Kriging technique for maximum annual precipitation depth values. Maps generated by the IDW technique were discarded. It is evident that having a greater number of stations in optimum condition could make a more precise regionalization since interpolation would be more accurate, as well as having support stations would minimize the effect of "bull's eye".

References

[1] T.S. Kokkonen, A.J. Jakeman, P.C. Young, H.J. Koivusalo, «Predicting daily flows in ungauged catchments: Model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina», Hydrological Processes, **17**(11), 2219-2238, 2003, doi:10.1002/hyp.1329.

[2] M. Jessen, T. Einfalt, A. Stoffer, B. Mehlig, «Analysis of heavy rainfall events in North Rhine-Westphalia with radar and raingauge data», Atmospheric Research, **77**(1-4 SPEC. ISS.), 337-346, 2005, doi:10.1016/j.atmosres.2004.11.031.

[3] J.M. Mirás-Avalos, R.M. Mestas-Valero, P. Sande-Fouz, A. Paz-González, «Consistency analysis of pluviometric information in Galicia (NW Spain)», Atmospheric Research, 2009, doi:10.1016/j.atmosres.2009.05.011.

[4] G. Blöschl, M. Sivapalan, «Scale issues in hydrological modelling: A review», Hydrological Processes, **9**(3-4), 251-290, 1995, doi:10.1002/hyp.3360090305.

[5] Problems of rainfall-runoff modelling in arid and semiarid regions, oct. 2020, doi:10.1080/02626668809491261.

[6] R.J. Nathan, T.A. McMahon, «Evaluation of automated techniques for base flow and recession analyses», Water Resources Research, **26**(7), 1465-1473, 1990, doi:10.1029/WR026i007p01465.

[7] D.A. Post, A.J. Jakeman, «Predicting the daily streamflow of ungauged catchments in S.E. Australia by regionalising the parameters of a lumped conceptual rainfall-runoff model», Ecological Modelling, **123**(2-3), 91-104, 1999, doi:10.1016/S0304-3800(99)00125-8.

[8] A. Agarwal, R. Maheswaran, V. Sehgal, R. Khosa, B. Sivakumar, C. Bernhofer, «Hydrologic regionalization using wavelet-based multiscale entropy method», Journal of Hydrology, **538**(538), 22-32, 2016, doi:10.1016/j.jhydrol.2016.03.023.

[9] C.C.F. Plouffe, C. Robertson, L. Chandrapala, «Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of Sri Lanka», Environmental Modelling and Software, **67**, 57-71, 2015, doi:10.1016/j.envsoft.2015.01.011.

[10] Y. Gyasi-Agyei, G. Pegram, «Interpolation of daily rainfall networks using simulated radar fields for realistic hydrological modelling of spatial rain field ensembles», Journal of Hydrology, **519**(PA), 777-791, 2014, doi:10.1016/j.jhydrol.2014.08.006.

[11] M.A. Oliver, R. Webster, «Kriging: A method of interpolation for geographical information systems», International Journal of Geographical Information Systems, **4**(3), 313-332, 1990, doi:10.1080/02693799008941549.

[12] Cómo funciona Kriging—Nuevas características y mejoras que se han agregado en esta versión. | Documentación, oct. 2020.

[13] A. Mazzella, A. Mazzella, «The importance of the model choice for experimental Semivariogram modeling and its consequence in evaluation process», Journal of Engineering (United Kingdom), **2013**, 2013, doi:10.1155/2013/960105.

[14] I.K. Tsanis, M.A. Gad, «A GIS precipitation method for analysis of storm kinematics», Environmental Modelling and Software, **16**(3), 273-281, 2001, doi:10.1016/S1364-8152(00)00068-2.

[15] D. Koutsoyiannis, «Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation», Hydrological Sciences Journal, **49**(4), 575-590, 2004, doi:10.1623/hysj.49.4.575.54430.

[16] E. Rojas, B. Arce, A. Peña, F. Boshell, M. Ayarza, «Cuantificación e interpolación de tendencias locales de temperatura y precipitación en zonas alto andinas de Cundinamarca y Boyacá (Colombia)», Corpoica Ciencia y Tecnología Agropecuaria, **11**(2), 173, 2010, doi:10.21930/rcta.vol11_num2_art:209.

[17] M. Sivapalan, G. Blöschl, «Transformation of point rainfall to areal rainfall: Intensity-duration-frequency curves», Journal of Hydrology, **204**(1-4), 150-167, 1998, doi:10.1016/S0022-1694(97)00117-0.

[18] R.J. Moreano Viteri, SISTEMA DE INFORMACION PARA LA INTERPOLACION ESPACIAL Y TEMPORAL DE DATOS SOBRE EL TIEMPO ATMOSFERICO Y EL CLIMA DEL ECUADOR , QUITO/ EPN/ 2008, 2008.

[19] Ana Carolina SantosAndrés Vargas-LunaAndrés Vargas-LunaNelson

Obregón-NeiraNelson Obregón-NeiraEDER CÁRDENAS, «Análisis de la distribución e interpolación espacial de las lluvias en Bogotá, Colombia», Dyna, 151-159, 2010.

[20] G. Gong, S. Mattevada, S.E. O'Bryant, «Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas», Environmental Research, **130**, 59-69, 2014, doi:10.1016/j.envres.2013.12.005.

[21] D. Zimmerman, C. Pavlik, A. Ruggles, M.P. Armstrong, «An experimental comparison of ordinary and universal kriging and inverse distance weighting», Mathematical Geology, **31**(4), 375-390, 1999, doi:10.1023/A:1007586507433.

[22] S.M. Vicente-Serrano, M.A. Saz-Sánchez, J.M. Cuadrat, «Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): Application to annual precipitation and temperature», Climate Research, **24**(2), 161-180, 2003, doi:10.3354/cr024161.

[23] F. Huang, D. Liu, X. Tan, J. Wang, Y. Chen, B. He, «Explorations of the implementation of a parallel IDW interpolation algorithm in a Linux cluster-based parallel GIS», Computers and Geosciences, **37**(4), 426-434, 2011, doi:10.1016/j.cageo.2010.05.024.

[24] P.M. Bartier, C.P. Keller, «Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW)», Computers and Geosciences, **22**(7), 795-799, 1996, doi:10.1016/0098-3004(96)00021-0.

[25] M.P. Armstrong, R.J. Marciano, «Local interpolation using a distributed parallel supercomputer», International Journal of Geographical Information Systems, **10**(6), 713-729, 1996, doi:10.1080/02693799608902106.

[26] Q. Huang, C. Yang, «Optimizing grid computing configuration and scheduling for geospatial analysis: An example with interpolating DEM», Computers and Geosciences, **37**(2), 165-176, 2011, doi:10.1016/j.cageo.2010.05.015.

[27] H. Hu, H. Shu, «An improved coarse-grained parallel algorithm for computational acceleration of ordinary Kriging interpolation», Computers and Geosciences, **78**, 44-52, 2015, doi:10.1016/j.cageo.2015.02.011.

[28] N.C.-N.Y.A.W.-I. Publicatiorq, undefined 1993, «Statistics for spatial data. revised edition».

[29] M.F. Hutchinson, «A new procedure for gridding elevation and stream line data with automatic removal of spurious pits», Journal of Hydrology, **106**(3-4), 211-232, 1989, doi:10.1016/0022-1694(89)90073-5.

[30] R. Franke, «Smooth interpolation of scattered data by local thin plate splines», Computers and Mathematics with Applications, **8**(4), 273-281, 1982, doi:10.1016/0898-1221(82)90009-8.

[31] J.A. Becerra Oviedo, L.F. Sánchez Mazorca, P.M. Acosta Castellano, J.L. Díaz Arévalo, «Regionalización de curvas IDF para el uso de modelos hidrometeorológicos en la Sabana Occidental del departamento de Cundinamarca», Ingeniería y Región, **14**(2), 143, 2016, doi:10.25054/22161325.701.

[32] P.M. Acosta Castellanos, L.X. Sierra APonte, «Evaluación de métodos de construcción de curvas IDF a partir de distribuciones de probabilidad y parámetros de ajuste», REVISTA FACULTAD DE INGENIERÍA, **22**(35), 25, 2013, doi:10.19053/01211129.2512.

[33] Colombia: Diagnóstico departamental de situación de DDHH y DIH - Boyacá - Colombia | ReliefWeb, oct. 2020.

[34] T Izquierdo., Álvaro Márquez, «Estudio comparativo de diferentes métodos de interpolación para la realización de mapas de precipitación en la isla de La Gomera (Islas Canarias): Evaluación de la incorporación de la altura - Dialnet», Geogaceta, 131-134, 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1328-1334 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

SH-CNN: Shearlet Convolutional Neural Network for Gender Classification

Chaymae Ziani*, Abdelalim Sadiq

MISC Laboratory, Faculty of Sciences, Ibn Tofail University, Kenitra, 14000, Morocco

ARTICLE INFO

Article history:
Received: 04 September, 2020
Accepted: 01 December, 2020
Online: 16 December, 2020

Keywords:
CNN
Shearlet transform
Gender prediction
Features extraction

ABSTRACT

Gender detection and age estimation become an active research area and a very important field today, wish has been widely used in various applications including them: biometrics, social network, Targeted advertising, access control, human-computer interaction, electronic customer, etc. The need to further improve the recognition or classification rate keeps increasing day after day. In this paper, we explore how deep learning techniques can help in the classification of gender from human face images and moreover raise the recognition rate. We propose in this contribution an approach called SH-CNN based on Discrete Shearlet Transform (DST) as a first step of feature extraction layer, and Deep Convolutional Neural Network (DCNN) as a second automatic feature extraction layer and also a classification step. The idea behind our contribution is to generate trough DST several features (in different decomposition and orientation) of an image. These features of each image will be the input of the DCNN, to enrich the training step, and so, improve the recognition rate. The obtained results have shown that the proposed approach import a significant enhancement.

1. Introduction

This paper is an extension of work originally presented in The Third International Conference on Intelligent Computing in Data Sciences ICDS2019 [1].

Automatic age and gender classification of humans through their image faces is a fundamental task in computer vision. It is an important research focus this last years, and plays a key role in a large variety of applications. In this last years, The Convolutional Neural Networks (CNNs) model has shown high level classification capacity, and mainly in cases of face images taken in real-world conditions. The performance of this model result in the ensured high level feature extraction and classification, through the used layers.

The Feature Extraction Function (FEF) is the key factor that determines the efficiency of a Face Recognition System and considered as one of the important parts. Several feature extraction function have been proposed in this domain of face feature extraction, such as, wavelets [2], SIFT [3] and other [4], [5]. Until today, the need for more efficient feature extraction function increases, seen the use of face gender recognition system in sensitive and important area. The Shearlet transform as feature extraction function allows generating more sensitive and significant features [6]-[8].

The field of learning has undergone a revolution with the appearance of CNN because the CNN can extract features of different complexities. Unlike the learning technique, convolutional neural networks learn the features of each image automatically, and that's where their strength lies. Today, the most successful models for image classification are those based on CNN. Since the model is based on learning, so we will enrich this phase with features extracting by DST. The Discrete Shearlet Transform (DST) is used in our paper as a first feature extraction layer for our proposed SH-CNN. The idea is to extract more significant features of the input face, to further enhance the gender recognition rate.

2. Related work

There is already a rich literature on the subject of classification by age and gender in an attempt to further improve the recognition rate. We briefly reviewed the age and gender classification studies that use neural network or Convolutional neural network. The authors in [9] present a structure including CNN and Extreme Learning Machine (ELM), where the CNN is used to extract the features from the input images and ELM, to classify the interposed results, and to combine the co-action of the two classifiers, to perform age and gender classification. In [10], the authors

*Corresponding Author: Chaymae Ziani, Email: chema.zin@gmail.com, chaymae.ziani@uit.ac.ma

implement a CNN of 3 convolutional layers and 3 fully connected layers for the tasks of estimating age and gender, it is trained exclusively on Audience. For the forecast, they do several crops around the face in different sizes. The final prediction is the average prediction value for all of these crops. They achieve an 86.8% rate of good classifications on the audience test images. A Deep gender classification is performed in [11], which is based on AdaBoost-fusion of isolated facial features and foggy faces. They applies an approach based on local features of the face by separating it into several parts (mouths, eyes, nose). They also include the discriminating features surrounding the face while blurring it. These images are then used to train several CNNs followed by an Adaboost classifier grouping the classification scores of the CNNs. This method yielded 91.75% on LFW and 83.06% on Adience using Gallagher as training data. In [12] the authors address the estimations problem with a classification based model which considered the age value as an independent etiquette and a separate class, and propose a deep convolutional neural network model learns the pertinent informative age and gender characterizations from the image pixel. A Male/female identification method from very low resolution face images by using neural network have proposed in [13]. In [14], the authors proposed a method named the full component CNN, which holds the full faces and the facial component networks, to obtain human age/gender classification through the facial image data. For age and gender classification, the authors in [15] propose a method applying deep learning based learning, the feature extractor HOG is applying on the input image on the network. The structure of the network used in this study is GoogLeNet. In [16], the authors present a review and a study of the advanced research technique for gender classification, which can be useful for researchers who work on gender classification. In [17], a convolutional neural network is used for the estimation of age and gender of a person from the facial image, and the PCA is applied to reduce the dimensions of the extracted features.

Among the research works countable on fingers that use CNN and shearlet transform, we find a contribution in [18], where the authors use an algorithm based on CNN and Shearlet transform, to classify the patient having Alzheimer's disease and early or late mild cognitive impairment using the florbetapir PET (positron emission tomography) amyloid imaging data. The authors had concatenated shearlet coefficients into the convolution layers on the CNN.

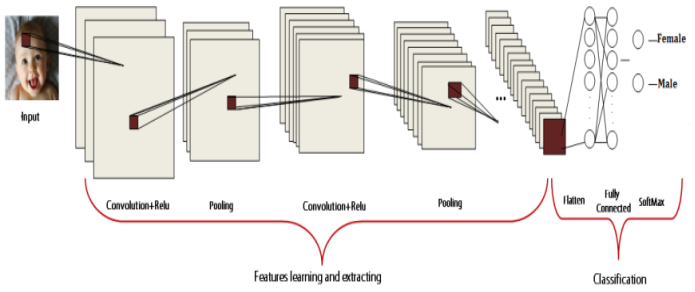


Figure 1: Example of a DCNN architecture for gender prediction

In our knowledge, this is the first time that Shearlets are used as first pretreatment, in order to enrich the training phase of the DCNN. The idea is to add as first treatment a DST, to exploit the calculated features extracted through decompositions and orientations. These operations allow us to generate from one input image, several feature views. Based on these generated features, we enrich the training phase, by injecting to DCNN the image and its feature views (the DST decompositions and orientations). The proposed approach will allow us to enhance the gender recognition, seen that the training phase is improved.

3. Deep Convolutional neural network and shearlet transform

3.1. DCNN

Convolutional Neural Networks [19] was proposed mainly to treat 2D shapes variability. They have been used widely in computer vision and they showed the best results and outperformed all other technique. Now, DCNN has become a major research orientation for researchers in the domain of computer vision, especially for image classification.

DCNN is powerful model, which is capable to capture effective information due to his architecture (Figure 1). DCNN architecture is divided into 2 steps, the first step is learning and extracting features and the second step is classification. There are four types of layers for a CNN: the convolutional, the pooling, the ReLU correction, and the fully-connected.. These components work together to learn a dense feature representation of an input.

- The convolution layer (Figure 2) is the important component of CNN, and always constitutes their first layer. Its goal is to identify features in the images received as input. To do this, convolution filtering is applied, by sliding window that represents feature, and by computing the convolution product of each image part and feature. Then, a feature is viewed as a filter. An activation map is obtained for each pair (image, filter) which indicates where the features are located in the image, when the value elevated, the more the corresponding place in the image looks like the feature. And then, these feature maps are normalized (with an activation function) [20].
- The ReLU correction layer used as an activation function, it is defined by ReLU (x) = max (0, x) [20], to replace all negative values with zeros

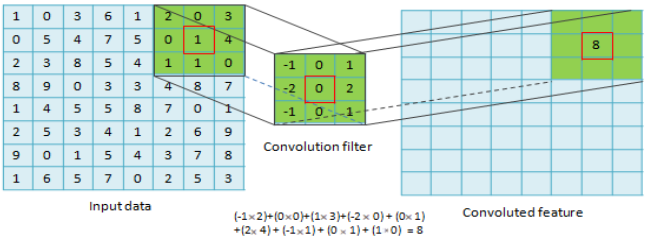


Figure 2: Convolution filter applied on input image

- The pooling layer (Figure 3): feature maps are injected as input to this layer. The principal of pooling is to reduce image size without losing the essential characteristics. In general, the pooling layer is nested in two convolution layers, and aims to enhance network-learing, by minimizing over-calculations [20].
- As last layer of the network. The fully-connected layer [20] has as input a vector, on which it applies linear combination

and activation function, and so, generated output vector. The returned vector by this layer to classify the input image represents the possible output classes of the classification system. Each value of the vector represents the corresponding probability of the input image to a class. The input matrix resulting from the previous layer contains an activation map, which represents the position of feature in the image. According to the position of feature in the image, the fully-connected layer can calculate the corresponding probability to each output classes. In general, for the DCNN classification systems, the softmax method is often used at the end of the network [21].

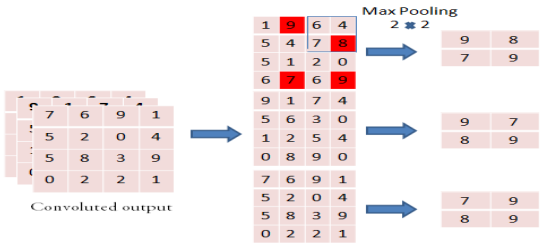


Figure 3: Example of a max pooling operation [22]

3.2. Shearlet transform

A multidimensional version of wavelet transform were invented in 2006 [23] named Shearlet, that encode anisotropic features in multivariate problem classes. The limitation of wavelets is the incapacity of capturing anisotropic features. Therefore, this transform provided optimal sparse representation for images with edges, surpassing the wavelet.

The successful use of shearlets in many image processing applications, especially face detection, feature extraction and denoising [24]-[26], is due to their properties: parabolic scaling, shearing, and translation. Expressly, the Continuous Shearlet Transform (CST) [27] is defined as:

$$f \longmapsto SH_\psi f(a,s,t) = \langle f, \psi_{a,s,t,} \rangle, \quad f \in L^2(\mathbb{R}^2)\ , (a,s,t) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}^2 \tag{1}$$

where,

$$\psi_{a,s,t(x)} = \mid det M_{a,s} \mid^{-\frac{1}{2}} \psi\left(M_{a,s}^{-1}(x-1)\right) \tag{2}$$

The parameter of this system calls as wavelets with composite dilations.

And $$M_{a,s} = \begin{pmatrix} a & s \\ 0 & \sqrt{a} \end{pmatrix}. \tag{3}$$

The matrix $M_{a,s}$ is factorized as B_sA_a, where $B_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$ is a shearing matrix and $A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$ is a parabolic scaling matrix (or anisotropic dilation matrix).

The Shearlets transform function is based on three variables (Figure 4): a is the scale parameter measuring the resolution level, s is the shear parameter measuring the directionality and t is the translation parameter measuring the position.

To get a frame (or Parseval Frame) for $L^2(R)$, a CST sampling is applied on a suitable discrete set of the three parameters: scale, shear, and translation. A discrete shearlets can be generated by sampling those three parameters:

$$a_j = 2^j (j \in \mathbb{z}), \qquad s_{j,k} = k a_j^{\frac{1}{2}} = k 2^{\frac{j}{2}} (k \in \mathbb{z})\ ,$$
$$\text{and} \qquad t_{j,k,m} = D_{a_j, s_{j,k}} (m \in \mathbb{z}^2) \tag{4}$$

For $\psi \in L^2(\mathbb{R}^2)$, the Discrete Shearlet Transform of $f \in L^2(\mathbb{R}^2)$ is the mapping defined by:

$$f \longmapsto SH_\psi f(j,k,m) = \langle f, \psi_{j,k,m} \rangle,$$
$$f \in L^2(\mathbb{R}^2)\ , (j,k,m) \in \mathbb{z} \times \mathbb{z} \times \mathbb{z}^2 \tag{5}$$

Where (6) is the discrete Shearlets function (DST)

$$\psi_{j,k,m,}(x) = 2^{\frac{3j}{2}} \psi(B_k A^j x - m), j \geq 0, -2^j \leq k \leq 2^j - 1, \quad m \in \mathbb{z}^2 \tag{6}$$

and $M_{a,s}$ is discretized as $M_{j,k} = B_k A^j$. Thus, SH_ψ maps the function f to the coefficients $SH_\psi f(j,k,m)$ correlated with the indexes of scale j, the orientation k, and the position m.

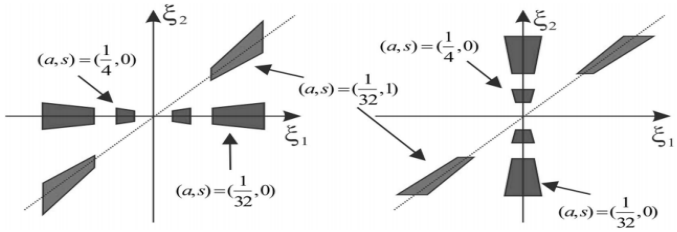
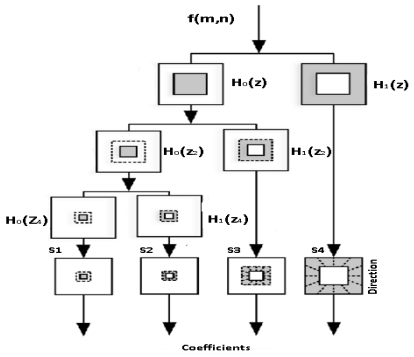


Figure 4: The horizontal and vertical shearlets Frequency support [28]



The shearlet transformation is a decomposition function based on scales and direction set. As shown in Figure 5, the decomposition is done by 4 scales levels denoted S1, S2, S3, and S4 with 6 directional filtered coefficients. At first, repetitively, a decomposition of the input image is done to a low-pass image and a high-pass image, based on a decomposition of two-channel non-subsampled [30], [31]. The used filters are: $H_0(z)$ for low pass filter function, and $H_1(z)$ for high pass filter function. The decomposition steps are: first, a frequency transformation domain applied on the high pass image, using two dimensional Fourier transform. Secondly, a cartesian coordinate system with six directions is used, to generate six directional subbands from the frequency domain. For the low pass image, a decomposition of two-channel non-subsampled is applied again, and the same

decomposition steps are repeated. The shearlet coefficients are generated as result of this transformation [29].

4. Proposed Approach

The idea which motivated us to propose this solution is the objective of enriching the learning phase to improve gender recognition. To enrich the DCNN learning input sample. We proposed to use as the first preprocessing of the CNN, a layer which uses the DST. The latter allows us to generate from a CNN learning input image, several views made up of the features generated from the decompositions and orientations of the DST. At this point, in the learning phase, the input will be represented not just by the images but also by their feature views. In this way, we will increase the input sample from the same number of starting images. Such a technique will be very useful in the case where the image sample in the training phase is small.

In our proposed approach, the training and the testing phase of the system performs essential operations. Each operation contributes to ensure a high recognition level. Figure 6 and Figure 7 showed these operations:

- Face detection
- Preprocessing
- Features extraction using DST
- Automatic features extraction and classification

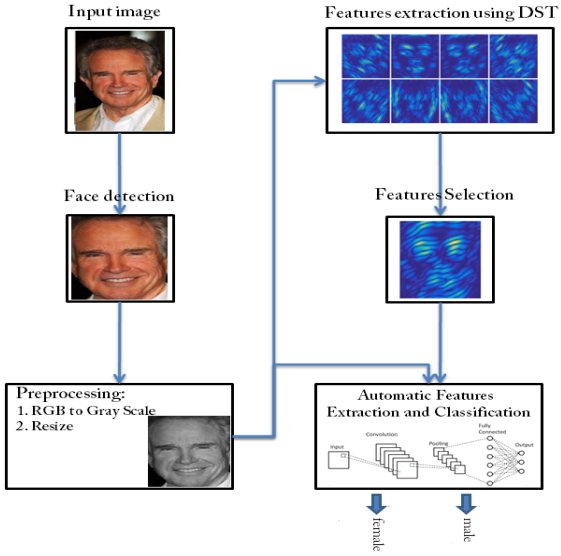


Figure 6: Proposed SH-CNN approach (training phase)

In our approach, haar-cascades [32] are used for face detection in order to eliminate any unusable information in the image. As a result, we will focus just on faces.

4.2. Preprocessing

At this level, two main treatments are applied on each face photo. The first treatment is to convert the photo from a RGB color level to a Gray level. The second treatment is to reduce the size of the faces photos. These steps reduce the calculation time in extracting features, and also in CNN layers.

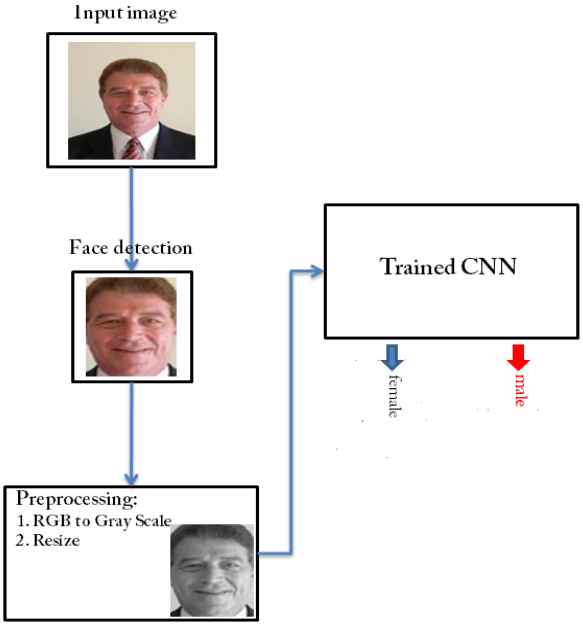


Figure 7: Proposed SH-CNN approach (testing phase)

The Feature Extraction Function (FEF) is a key factor that determines the efficiency of gender Recognition System. DST as a FEF is applied in each face image; the images will decompose along a number of directions and scales to extract important facial features. At the end, for each image we get a set of images in different orientations. That will enrich the CNN's learning step and facilitate the classification phase. And that is the goal for combining it with the CNN. This operation is used only in training phase, and so, we will not increase the processing time in the testing phase.

4.4. Automatic features extraction and classification (CNN)

The idea is very simple. The trainable system consists of a series of modules, each one representing a processing step. Each module is trainable, with adjustable parameters similar to the weights of linear classifiers. The system is trained from end to end: in each instance, all parameters of all modules are adjusted to bring the output produced by the system closer to the desired output.

To enrich the training phase, we inject as an input in our CNN the face and the corresponding coefficients calculated using DST. This operation allows the CNN to adapt more the weights and to increase the learning step.

5. Experiment results

This section presents the experimental results of our proposed approach. Our method is implemented using Python as a high-level programming language, with the Tensorflow [33] (an OSS framework for machine learning on decentralized data) and Keras(is a deep learning API) [34] libraries for CNN learning and classification.

5.1. UTKFace dataset

UTKFace dataset [35] is composed from over 20,000 face images (e.g. Figure 8), which are tagged by age and gender. The

images of this dataset present multiple facial expressions. Certain images are with high and low resolution, things that allow testing our system well. The UTKFace dataset is frequently used in gender/age classification and face detection, etc.

The images size is 200×200 pixels. We have resized all the face with the dimension of 100×100 pixels.



5.2. Operations of extracting features and classification

In the previous section, we presented our approach with successive operations. In this section, we present more details on the proposed SH-CNN approach for automatic gender recognition.

The DST ensures a good representation level of images with edges. At the features extraction step, the pre-treated images were decomposed into 4 levels (Table 1), by varying shear parameter from 2^0 to 2^3, and scale parameter from 0 to 3, and hence we obtained shearlet coefficients on many directions (Figure 9).As we said before these coefficients will be used later as input of a CNN. We repeat this procedure for all faces of the training database.

Table 1: Used DST parameters.

Decomposition levels	Directions	Shearing filters size
Level 1	4	32×32
Level 2	8	32×32
Level 3	16	16×16
Level 4	16	16×16

In our proposed approach, and during all experiments for gender classification, we used a CNN composed from three convolutional layers and one fully-connected layer. Gender classification on the face dataset requires distinguishing between two classes, that the reason why we choose a smaller network.

The input layer is a face image of size 100 × 100 pixels. The three convolutional layers are respectively of 96@7 × 7, 256@5 × 5, and 384@3 × 3 feature maps. After each one of these convolution layer, a ReLU function and a max pooling layer (the max value of 3×3) are applied. To generate the needed 512 neurons for the fully connected layer, a flatten layer is used. The next layer is a fully connected, that maps to the final gender classes. As a last step, a soft-max layer is applied to attribute a corresponding probability for each class. The class having the probability max value represents the closest prediction

The used criteria in this experiment results take into consideration our need in terms of features extraction part and CNN classification part. For the DST feature extraction part, we have used four level of decomposition with directions (4, 8, 16, and 16). In general in the existing works, and as the recommended values, the used decomposition number is approximatively 3, 4 or 5 levels, and for directions we found (4, 8, 16, and 32) [6]-[8], [26]. In our contribution we have used four level with directions (4, 8, 16, 16), which have clearly shown the added value of our contribution. For the CNN classification part, we have used three convolutional layers with a kernel size of 7x7, 5x5 and 3x3, in order to reduce the training time. And for the fully connected layer, we used 512 neurons, since using larger values in case of 100x100 input image size will not improve the results, contrary, it will increase the calculation time.

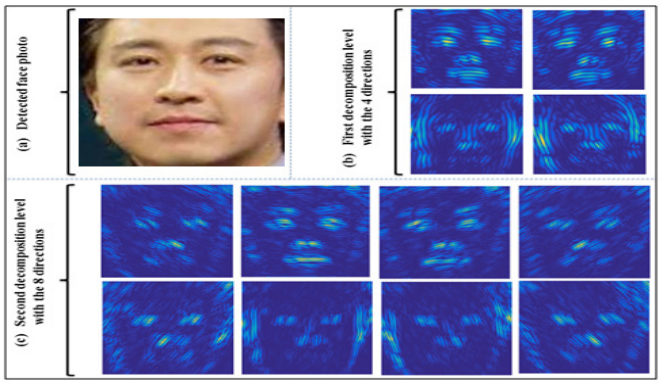


Figure 9: (a) Detected face photo (b) 1st decomposition level with 4 orientations applied on the detected face (c) 2nd decomposition level with 8 orientations applied on the detected face.

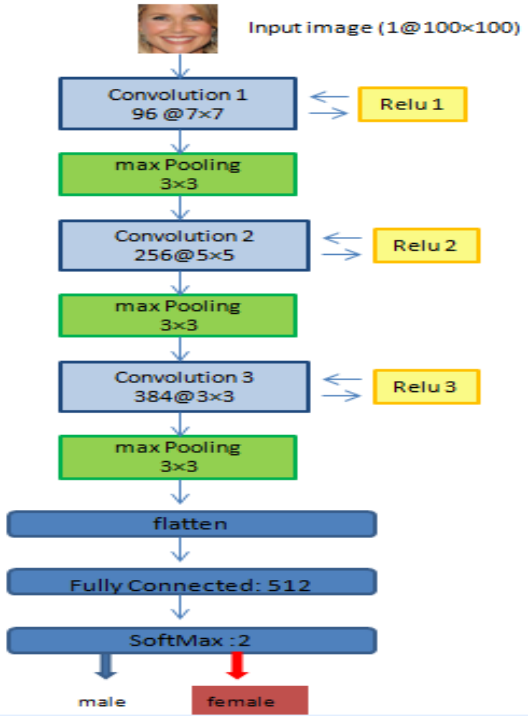


Figure 10: Proposed CNN architecture

5.3. Results and discussion

Regarding the comparison phase of the proposed approach (SH-CNN), we carried out a direct comparison study with the CNN, without seeing the other models that already exist, proposed as part of the improvement of the gender recognition, and based on CNNs. It is seen that our first objective was to test the added value of our SH-CNN compared to CNN. Add to this that the proposed preprocessing layer concatenated with the CNN can be concatenated with all the proposed models based on the CNN as

well. Because the add layer plays the role of an enrichment phase of the sample of the learning part of the CNN. From this explanation, we understand that the comparison must be made with CNN directly.

The obtained results in terms of accuracy rate and loss rate (see Figure 11 and 12), shown clearly the added value of our proposed approach. The generated coefficients for each face using DST, enrich the training phase as expected, and so, increase the recognition rate. In this simulation, we have used just one generated coefficient of DST as input with the associated face. If we use more coefficient as input to enrich the training phase, we may obtain more efficient recognition system, and this will be our coming research work. Also, in our simulation, we have used just 3 epochs, to show the capacity of our approach to increase more the accuracy rate within a few epochs.

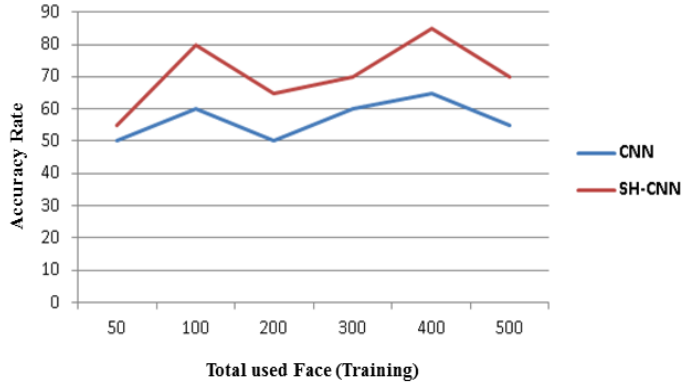


Figure 11: The obtained Accuracy Rate (CNN vs SH-CNN)

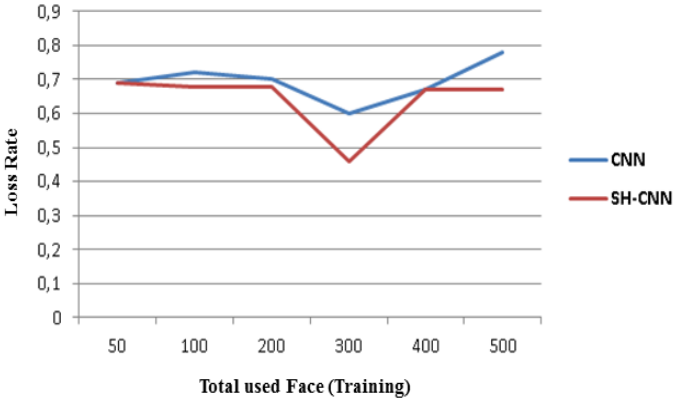


Figure 12: The obtained Loss Rate (CNN vs SH-CNN)

6. Conclusion

The need to increase the gender recognition rate has pushed us to search lower-cost solutions (processing time, dataset size, etc.). In this paper, we have proposed an approach based on CNN and especially on DST as a first layer. The idea behind this solution is to enrich more the training phase, by using the generated coefficients of each image using DST. The obtained results have shown the clear added value of this approach. The ensured accuracy rate of our system is also related to the capacity of DST to extract the significant features of each image.

In our future work, we will add also the age as output of our proposed system. We will also use high and low values of epoch, and explore the capacity of our proposal in others databases (FERET, etc.).

References

[1] C. Ziani, A. Sadiq, "Shearlet Convolutional Neural Network Approach for Age and Gender recognition," in 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), 1-4, 2019, doi: 10.1109/ICDS47004.2019.8942359.

[2] K. Ramesha, K. B. Raja, "Face Recognition System Using Discrete Wavelet Transform and Fast PCA," in Information Technology and Mobile Communication, **147**, 13-18, 2011, https://doi.org/10.1007/978-3-642-20573-6_3.

[3] P. G. Ambhore , L. Bijole, "Face Sketch to Photo Matching Using LFDA,". International Journal of Science and Research (IJSR), 3(4), 11-14, 2014,

[4] J .Qian, J. Yang, "A Novel Feature Extraction Method for Face Recognition under Different Lighting Conditions," in Chinese Conference on Biometric Recognition (CCBR), **7098**, 17-24 , 2011, https://doi.org/10.1007/978-3-642-25449-9_3.

[5] T. T. Do, T. H. Le, " Facial Feature Extraction Using Geometric Feature and Independent Component Analysis," In Knowledge Acquisition: Approaches, Algorithms and Applications, **5465**, 231-241, 2009, https://doi.org/10.1007/978-3-642-01715-5_20.

[6] Z. Zeng, J. Hu, "Face recognition based on shearlets transform and principle component analysis," in 2013 5th International Conference on Intelligent Networking and Collaborative Systems, 697-701, 2013, doi: 10.1109/INCoS.2013.134.

[7] A. Danti, K. M. Poornima, "Face Reconition using Shearlets," in 2012 IEEE 7th International Conference on Industrial and Information Systems, 1-6, 2012, doi: 10.1109/ICIInfS.2012.6304796..

[8] X. Sun, X. Wu, and Y. Wei, "Face Recognition Based on Shearlet Transform and Fast ICA," in 2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control, 832–835, 2014, doi: 10.1109/IMCCC.2014.175.

[9] M. Duan, K. Li, C. Yang, K. Li, "A hybrid deep learning cnn–elm for age and gender classification" , Neurocomputing, **275**, 448–461, 2018, https://doi.org/10.1016/j.neucom.2017.08.062

[10] G. Levi , T. Hassncer, "Age and gender classification using convolutional neural networks," in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 34-42, 2015, doi: 10.1109/CVPRW.2015.7301352.

[11] A. Mahmoud ,A. Abdelhamed, "AFIF4: Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces," Journal of Visual Communication and Image Representation, **62**, 77-86, 2019 , https://doi.org/10.1016/j.jvcir.2019.05.001.

[12] O. Agbo-Ajala, S. Viriri, " Age Group and Gender Classification of Unconstrained Faces," In Advances in Visual Computing, **11844**, 418-429 , 2019 https://doi.org/10.1007/978-3-030-33720-9_32.

[14] C. Huang, Y. Chen, R. Lin , C. J. Kuo, "Age/gender classification with whole-component convolutional neural networks (WC-CNN)," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1282-1285, 2017, doi: 10.1109/APSIPA.2017.8282221.

[16] C. P. Divate, S. Z. Ali, "Study of Different Bio-Metric Based Gender Classification Systems," in 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 347-353, 2018, doi: 10.1109/ICIRCA.2018.8597340.

[17] B. Agrawal, M. Dixit, "Age Estimation and Gender Prediction Using Convolutional Neural Network," In: Intelligent Computing Applications for Sustainable Real-World Systems ,Proceedings in Adaptation, Learning and Optimization, **13**, 163-175, 2020, https://doi.org/10.1007/978-3-030-44758-8_15.

[18] E. Jabason, M. O. Ahmad, M. N. S. Swamy, "Shearlet based Stacked Convolutional Network for Multiclass Diagnosis of Alzheimer's Disease using the Florbetapir PET Amyloid Imaging Data," in 2018 16th IEEE International New Circuits and Systems Conference (NEWCAS), 344-347, 2018, doi: 10.1109/NEWCAS.2018.8585550.

[19] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning

applied to document recognition," in Proceedings of the IEEE, **86**(11), 2278-2324, 1998, doi: 10.1109/5.726791.

[20] R. Waseem , Z. Wang. "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," Neural Computation, **29**(9), 2352-2449, 2017, https://doi.org/10.1162/neco_a_00990

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun.ACM, **60**(6), 84–90, 2017, https://doi.org/10.1145/3065386

[22] "Machine Learning, Max Pooling With Color Images", https://dev.to/sandeepbalachandran/machine-learning-max-pooling-with-color-images-2i21, 2020

[23] G. Kanghui, G. Kutyniok, D. Labate, "Sparse multidimensional representations using anisotropic dilation and shear operators," Wavelets and Splines , Nashboro Press, Nashville, 189–201, 2006.

[24] G.R. Easley, D. Labate, "Critically sampled wavelets with composite dilations," IEEE Trans. Image Process, **21**(2), 550-561, 2012, 10.1109/TIP.2011.2164415.

[25] G.R. Easley, D. Labate ,V. Patel, "Directional multiscale processing of images using wavelets with composite dilations," J. Math. Imag. Vision, **48**, 13-44, 2014, https://doi.org/10.1007/s10851-012-0385-4.

[26] S. Yi, D.Labate, G.R. Easley, H. Krim, "A Shearlet approach to Edge Analysis and Detection," IEEE Trans. Image Process, **18**(5), 929–941, 2009, doi: 10.1109/TIP.2009.2013082.

[27] G. Kutyniok , D. Labate, "Resolution of the wavefront set using continuous shearlets," Transactions of the American Mathematical Society, **361**(5), 2719-2754, 2009, http://www.jstor.org/stable/40302874.

[28] C.Liu, D.Wang, J.Sun, T.Wang, "Crossline-direction reconstruction of multi-component seismic data with shearlet sparsity constraint," Journal of Geophysics and Engineering, **15**(5), 1929–1942, 2018, https://doi.org/10.1088/1742-2140/aac097.

[29] F.Yuan, L. Po , M.Liu, X. Xu, W. Jian., K. Wong, K. Cheung , "Shearlet Based Video Fingerprint for Content-Based Copy Detection," Journal of Signal and Information Processing, **7**, 84-97. doi: 10.4236/jsip.2016.72010.

[30] J.P.Zhou, A.L. Cunha, M.N. Do, "Nonsubsampled Contourlet Transform Construction and Application Enhancement," IEEE International Conference on Image Processing, 1-469, 2005, doi: 10.1109/ICIP.2005.1529789.

[31] G.Easley, D.Labate, W.Q.Lim , "Sparse Directional Image Representations Using the Discrete Shearlet Transform," Applied and Computational Harmonic Analysis, **25**, 25-46, 2007, http://dx.doi.org/10.1016/j.acha.2007.09.003

[32] P. Viola, M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", IEEE Computer Society, **1**, 511-518, 2001, doi: 10.1109/CVPR.2001.990517.

[33] "keras", https://keras.io/, 2020

[34] "Tenserflow", https://www.tensorflow.org/federated, 2020

[35] "UTKface database", https://susanqq.github.io/UTKFace/, 2020

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1422-1427 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Handling Priority Data in Smart Transportation System by using Support Vector Machine Algorithm

Sara Ftaimi*, Tomader Mazri

Laboratory of Advanced Systems Engineering, Ibn Tofail science university, Kenitra, 14020, Morocco

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 04 December, 2020
Online: 21 December, 2020

Keywords:
Vehicular Adhoc network
Autonomous vehicle
Intelligent transportation system
Smart city
Big Data
Machine learning algorithms
Support Vector Machine

ABSTRACT

In an intelligent transportation system (ITS), time is a big challenge since processing a huge amount of data in a short time is very difficult, especially when the processed data is heterogeneous, consisting of a mixture of emergency data, normal data, and noise. In an ITS, an ambulance is one of the priority vehicles, and the data sent by the ambulance to the infrastructure and other vehicles must be treated first because if the ambulance does not receive a status of road from the infrastructure in time, it could take the wrong road or takes a road where there is a traffic jam has a high chance of arriving late, which could put the patient's life in danger.
Prioritizing treatment of this type of data has become paramount and vital in such cases. This paper proposes modifying the big data process to include handling the intelligent transport system's urgent data type. We will add a classification step that allows us to classify the data according to the priority degree. We use the SVM algorithm of machine learning because it has given good results concerning data classification.

1. Introduction

Nowadays, it seems that every city is trying to become a smart city. To make a city smarter, life necessities (water management, waste management, transport, and security) and technology need to be transformed into smart services. A simple definition of a smart city is an environment in which citizens can live smoothly and easily by developing their various facilities and making them smart [1].

In smart cities, sensors will be placed all around the city, and they will generate continuous data. The analysis of such data will be very helpful for the city with the potential to be smarter. As we will have many sensors installed in the city, we will have a very large amount of data generated called Big DATA, and big data is a very large amount of data and data that can come from different sources in different shapes and forms. In the intelligent transportation that will be our case study, data will be generated from sensors placed in the city and sensors placed in the car (traffic light, Lidar camera, etc.). Those sensors will generate a mass of data that will be analyzed to help the car to become autonomous on the one hand and on the other hand to make the ride a joyful experience in addition to reducing the rate of accidents on the road.

As mentioned before, the data will be generated from different sensors located in the city and the sensors placed in the vehicle. In addition to these data, we will also have another type of data circulating in the network, namely the data exchanged between the Road-Side Units and between the latter and the vehicles and between the vehicles and the network data. So the challenge will be which data should be processed first, which data should have a higher priority; we will answer these questions in the present article, which is an extended version of the work presented in [1] where we propose to make a change on big data analysis process to handle priority data.

In the first section of this paper, we will briefly review the intelligent transportation system we will define the autonomous vehicle, explore the sensors embedded in it. We will see the Vehicular Adhoc Network (VANET) and the type of communication in VANET network. In section three, we will define big data in smart transportation. We will apply the principle of classification of transportation data to see different types of data collected by sensors. In the fourth section, we will explain our contribution by explaining the case study and the different steps of the proposed process. We will also define the machine learning algorithm used. In the final section, we will present the experimental result, and we conclude our work.

2. Autonomous vehicle

An autonomous vehicle is defined as a vehicle with the capacity to detect the surrounding environment to drive without human intervention. The autonomous vehicle depends on sensors, a Cloud platform, advanced algorithms, Machine learning systems,

*Corresponding Author: Sara Ftaimi, Email: saraftaimi@gmail.com

and efficient CPUs to operate algorithms. Thanks to the sensors integrated into the vehicle as shown in figure 1, it can generate a map of its environment. The camera and the LiDar are the eyes of the vehicle. They detect traffic lights, obstacles, traffic signs; all this data is processed by an algorithm which allows reading the traffic lights and the traffic signs, to avoid obstacles and to trace the path of the vehicle, and finally to send instructions to the accelerators to accelerate or to brake depending on the existence of obstacles in the path.

2.1. Sensors

Car manufacturers have integrated several sensors into the vehicle, as shown in figure 1 so that it becomes autonomous and can move around safely by detecting obstacles and reading signal lights. For a vehicle to be able to see and understand its environment, at least eight cameras are placed in the vehicle in addition to the LiDar to detect obstacles as well as RADAR, GPS to determine the location of vehicles over and above authors sensors.

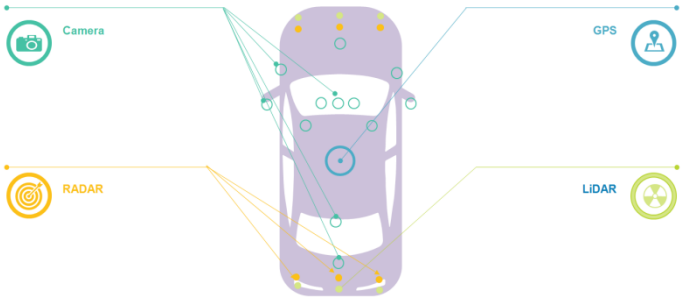


Figure 1: sensors in autonomous vehicle

2.2. VANET network

Vehicle Ad-hoc Networks (VANETs) emerged as one of the most relevant scientific research areas as they focused on supporting intelligent navigation and safe driving. It represents a specific type of MANET network that allows inter-node communication and organizes its communication system without any dependency on any other infrastructure. VANET is an intelligent building block of transport systems since vehicles interact with each other and infrastructure [2]. The high mobility of the nodes is one of the most relevant features of the VANET network vehicles are moving continuously in many directions and at varying speeds. The topology of the vanet network changes rapidly due to the fact that the vehicle can join or leave a cluster of vehicles in a very short time [3].Owing to the two above-mentioned features, combined with climate change and traffic congestion, vehicles can frequently become disconnected from the network. VANET has no energy or storage problems [4–6], unlike MANET networks, where energy constraint is a challenge. Besides having an unlimited network size since it has no geographical limits.The vehicles are referred to be communication nodes in VANET. To make communication possible between vehicles, the vehicle must be equipped with an On-Board Unit (OBU), and along the road, we must install the Roadside Unit (RSU).

The RSU is installed along the road based on the IEEE 802.11p radio technology to establish dedicated short-range communication (DSRC). It allows establishing local connectivity to the moving vehicle. It can also be used to communicate with other network devices in other infrastructure networks [7].

OBU is a device installed onboard the vehicle it allows to make a connection with the RSU, and with other OBUs via a wireless link IEEE 802.11p [8] The OBU gathers data from the sensors in the car, namely the Global Positioning System (GPS), the Event Data Recorder (EDR) and obtains the input power from the vehicle's battery [7].

2.3. Communication in Vanet

In the vanet network, Vehicles communicate with each other and with the infrastructure via three modes of communication

V2V: In V2V communication mode, a vehicle can directly communicate with other vehicles within the same range; for this mode, the architecture is decentralized. It is very efficient to transmit information on road safety services, but due to the changing speed and mobility of the nodes [7, 9] permanent connections between vehicles cannot be guaranteed.

V2I: In this mode of communication, the exchanges of information are between the vehicle and the infrastructure (RSU, antenna, satellite). The V2I mode offers a better use of the common resources,besides the connection status, which is more stable in this mode than V2V communication [7, 9, 10].

V2X means communication between vehicle and X where X can be a vehicle, a pedestrian, infrastructure, network, etc... In this diverse ecosystem, vehicles will communicate with other vehicles, with infrastructures such as traffic lights, pedestrians using smartphones, and data centers via cellular networks using short-range wireless signals [9].

3. Big data in smart transportation

In a smart city, an enormous amount of data will be generated by an intelligent transportation system (ITS). An autonomous vehicle driving around the city generates more than 25 gigabytes of data per hour. The sensors integrated into the vehicle (camera, GPS, Lidar, etc.) continuously generate information that is processed and sometimes sent to other vehicles or the infrastructure. Besides this data, there is also the data generated by the road infrastructure: RSUs (Road-Side Unit), traffic signs, traffic lights, and parking spaces.

All transport components are nowadays equipped with intelligent sensors that generate and send data continuously, e.g. a sensor is installed in a parking space and periodically sends the status of the space either it is empty or not. In addition to this data, there is also the data from the VANET network used for communication. [11, 12].

Figure2 resumes the data in a smart transportation system (ITS). Cars have to manage much larger amounts of data than ever before. Processing this data quickly and transparently is now and will continue to be one of the main challenges for car manufacturers in the future.

4. Taxonomy of data in a smart transportation system

Before classifying the data, we must know the sources of the data. In fact, in an intelligent transport system, the data generated and collected may come from a variety of sources:

- Data shared between vehicles.
- Data shared between the vehicle and the infrastructure
- The environmental data (climate)
- The data generated by the sensors placed on board the vehicles

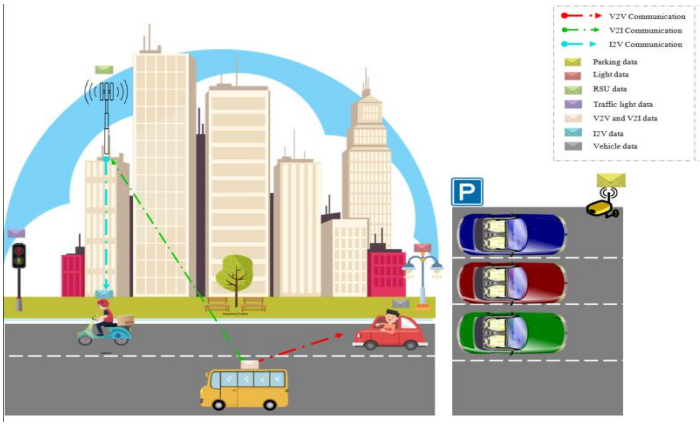


Figure 2: Communication and data type in ITS

Based on data collected from an intelligent transportation system and the taxonomy carried out in [1, 13], the data in an intelligent transportation system can be classified into three categories: the APP-level data, network-level data, and environmental data as shown in figure 3. APP-level data is the data generated by sensors installed in the ITS equipment (traffic light, etc) and the sensors integrated into the vehicles. The Network-level data is the data generated by the network that manages the communication between vehicles and road infrastructure besides the vehicles themselves (VANET network). The last type of data is environmental data.

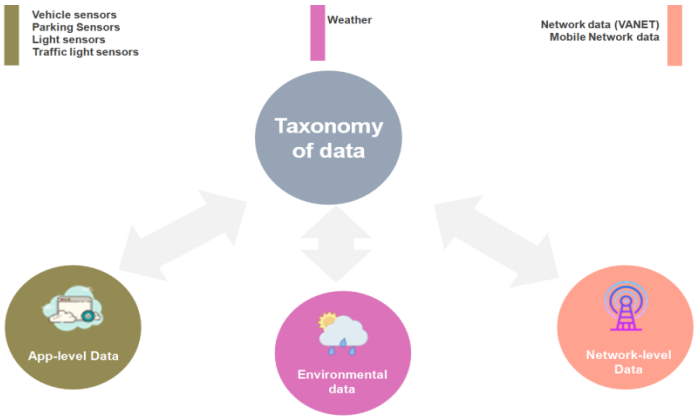


Figure 3: Taxonomy of data in ITS

5. Proposed Method

As the volume of data generated by the various intelligent transport equipment is large, processing has become vital. For a vehicle to move safely, it is necessary to collect, analyze, and transmit the information in short-time. When processing this data, the problem is to determine which data has the highest priority and needs to be processed first. To solve this problem, we propose to make a change in the Big data analysis process. Usually, the process of big data consists of the following steps: data collection, cleaning, analysis, and, finally comes the step of data interpretation[14]. However, to identify the data with the highest priority, we have added a step before the cleaning step, which will be the ordering and classifying step, and we will detail in the next sections.

5.1. Case study

In the intelligent transport system, we have several types of data that are assigned a priority. We are interested in the data shared between vehicles and RSUs, and our case study is as follows:

These RSUs will receive a large amount of data; amongst this data are the critical data that is of a higher priority than others. For instance, in the case of an accident on the road, or a fire engine trying to reach the scene of the incident, the messages sent by these vehicles have a higher priority than those sent by other vehicles, as shown in figure 4. In an intelligent transportation system, it must be able to broadcast information messages to alert road users about the accident on the road and advise them to change or turn back, as well as be able to alert the hospital through notifications, so it will send an ambulance if there is injury. The challenge here is the large number of messages received by the RSUs. The question we ask ourselves in this situation is how can we process this data and come out with the ordered data from the most critical to the least important. Figure 4 represents our case study;

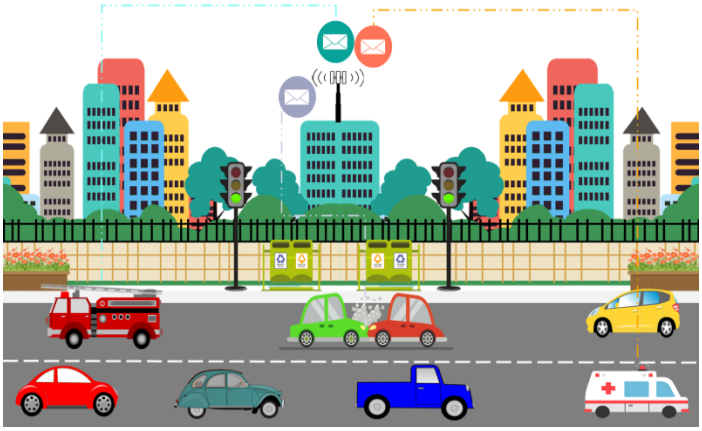


Figure 4: Case Study emergency data

We have three vehicles that have priority over the other vehicles, which are the fire engine, the accident vehicle and finally the ambulance. The messages received by these vehicles must be processed first to avoid causing traffic problems and other more serious consequences, such as loss of life.

5.2. Critical Data

To process the priority data, we must first determine the type of vehicles given priority [15]. In our case study, we have already listed three of them, and we add to them the police vehicles. Therefore, to process the priority data, we propose the following process.

First of all, it is necessary to collect the data received by the different vehicles and then apply a classification algorithm that allows us to determine the classes to which they belong. For this purpose, we have two levels of priority that are shown in figure five.

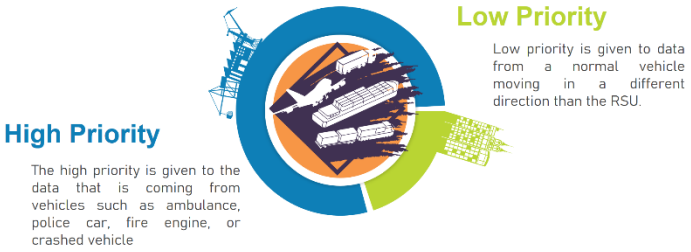


Figure 5: Levels of data priority in ITS

After the data classification step, we will have the possibility to know the data with priority, to process them first. The same process can be applied to the vehicles that will receive the data from other vehicles and the RSU data, and the data generated by the different sensors inside the vehicle. In the case of a vehicle traveling through the city at a given speed, if an accident occurs on its way, the priority must be given to the information of the change of path and the information of the change of the geolocalization map, whereas in the case of the ambulance, the alert must be given so that it passes first and requests the other vehicles in its area to clear the way.

5.3. Data packet

The communication that takes place between the infrastructure and the vehicle is communication managed by the RSU and the OBU. In an intelligent transportation system, the RSU is installed at the roadside to provide a local connection to the vehicle in motion, although it can also be used to communicate with other network devices in other infrastructure networks [7]. RSU uses short-range communication (DSRC), which is based on IEEE 802.11p radio technology. The RSU communicates to the OBU continuously. The main function of the OBU is to connect to the RSU or another OBU via the IEEE 802.11p wireless link [8]. Besides, the OBU gathers information from sensors embedded in the vehicle (e.g. Global Positioning System (GPS), Event Data Recorder (EDR)) and gets input power from the battery vehicles [16]. The packet from OBU to RSU contains the following fields OBU_ID, TIME_STAMP, POSITION, SPEED, DIR, ACC, VEHICLE_TYPE, DUTY_FLAG [15]. The table 1 represents an explanation of the fields.

Table 1: fields of the packet send between the OBU and the RSU

Fields	Explanation
OBU_ID	Contains the vehicle ID
TIME_STAMP	It is the time when the frame was sent.
POSITION	Is the position of the vehicle the value f position is gathered from the GPS sensor
SPEED	Is the speed of the vehicle
DIR	It is the direction of the vehicle.
ACC	Acceleration
VEHICLE_TYPE	Is the type of vehicle (Ambulance, fire engine, etc ...)
DUTY_FLAG	Contains information if there is an accident or the vehicle is on duty

The values of the fields above are calculated from GPS sensors that allow them to obtain directly the position, the speed, and the time. We can deduce the value of DIR from the GPS, the other values that identify the vehicle such as the id and the vehicle type are gathered from the latter.

5.4. Proposed algorithm

5.4.1. Support Vector Machine (SVM)

The SVM is a machine learning algorithm that belongs to the family of supervised learning algorithms. SVM can solve regression and classification problems besides outlier detection. The SVM algorithm can be used to classify linear and non-linear data. Therefore, it can carry out classification in a non-linear way using the kernel trick [3, 17].The SVM can be used to detect network intrusion [18] and spoofing attacks [19]. In [20], the author have used security messages to identify classes and types of attacks using SVM and KNN algorithms. In [20], the author noticed that KNN works better than SVM in some cases. When SVM has a reduction in performance, it is when the training base of both algorithms contains attacks that had extreme levels of misconduct and only a few attacks that were common errors. Also, judging the accuracy of recall, they found that SVM and KNN yield better results in some types of attacks, and SVM outperforms KNN in some attacks. In [21], the author have conducted an experiment on Support Vector Machine (SVM) and Feed Forward Neural Network (FFNN) to design intelligent IDSs that can detect two types of attacks, which are grey hole attacks and Rushing Attacks [21]. It is noted that FFNN-based IDSs could better detect malicious vehicles with lower false-negative alarm rates than SVM-based IDSs. However, the performance of SVM is much faster than FFNN. The choice of SVM algorithms was based on the comparative study conducted in [3] found that SVM is an algorithm that can handle multidimensional data and does not need much memory, and can detect outliers because the outlier detection is very important since the outliers can make false the classification model. Also, SVM can be used for both the separable and non-linearly separable data and work very well with a higher dimension. Besides, by using SVM we can replace the feature engineering, which is a very hard step for exploring data by kernelization.

5.5. Kernels

The SVM algorithm works by using a kernel technique to classify non-linear data, it maps data from the lower dimension data to the higher one and separates them with an optimal hyperplane. Therefore the kernel trick can help to transform non-linear problems into linear separable problems[17].

There are three types of kernel functions that are commonly used: linear kernel, Polynomial kernel, and Radial Basis Function Kernel (RBF). To choose which one to use, the SVM algorithm uses cross-validation to test several combinations of the kernel function and picks the one that gives the best results.

$$K(\bar{x}) = \begin{matrix} 1 & \text{if} \|\bar{x}\| \leq 1 \\ 0 & \text{otherwise} \end{matrix} \quad (1)$$

- Linear kernel: The linear kernel is commonly used to classify text and is also utilized in situations where the data is linearly separable. This means that the data may be separated with only one line. A linear kernel is generally faster to train due to the only optimization of the C regularization parameter.
- Polynomial Kernel(poly): The polynomial kernel is widely used for processing natural language (NLP). The standard degree is d = 2 (quadratic), since higher degrees over-adjust NLP problems.

- Radial Basis Function Kernel (RBF): The radial basis function kernel is the most used in SVM. This is because it has a localized and finite response along the whole x-axis. as well as being used when we have no previous knowledge of the data.

6. Experimental results

Our research results and their explanations, are presented in this section. First, we explain the methodology of our experiment and discuss the experimental result of using the SVM algorithm.

6.1. Methodology

Before starting the cleaning step, we will apply a classification algorithm to determine the type of data e.g, whether it is high-priority data or not.

The proposed process is schematically depicted in figuresix. Firstly, vehicles send every second packet, which is described in section 5.3. Each packet is read by the Machine Learning algorithm, which is the support vector machine (SVM) algorithm that we have chosen for this purpose. It has to classify data into two types see section 5.2. The classification is based on three criteria: the type of vehicle (e.g. ambulance, fire engine, police vehicle, accident vehicle, normal vehicle), the vehicle's direction, and finally, the duty flag field. If the type of vehicle is normal and the value of the duty flag mentions it as not crashed, then the algorithm will classify the data as not priority; otherwise, if the type of vehicle is an ambulance, fire engine, or police car, the algorithm must check the direction of the vehicle either it is the same as the one of the RSU, then the SVM classifies the data as a high priority. Once classified, the data can follow the standard BIG data process described in the figure six.

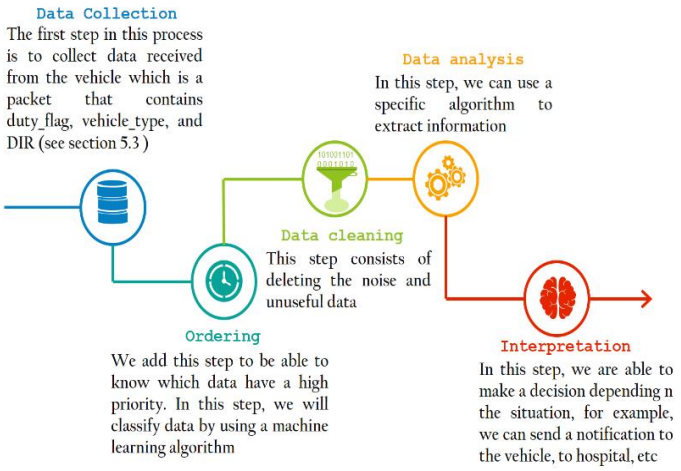


Figure 6: Proposed process

The urgent and priority data is a very challenging type of data in every domain, especially in the transportation system. If we do not treat them urgently, we can cause big damage and sometimes lose a life if the treatment was delayed. Our paper aims to handle this type of data. We propose an SVM algorithm that will read the packet sent to the vehicles or the RSUs and detect which packet has the priority to be treated first. Before starting the classification, we show in figure 7 the distribution of some vehicles based on their speed captured by the GPS sensor. We capture 100 packets sent from vehicles to RSU, and we use it to test our SVM algorithms. The SVM algorithm was applied to this database to classify the data into two types: priority and non-priority. We focused on three types of kernel poly, linear et RBF. The SVM algorithm uses two parameters to build the model, which are the gamma and the C value. In fact, the two variables, gamma, and C could change the classification model. So if the gamma value is very high, the algorithm will consider the closest points to build the limits and aims at it. We can fall into overfitting when C's value is very high and into underfitting if it is small.For this purpose, we applied a function that takes several input values for the C and Gamma parameters to find the optimal values that give the best classification results and high accuracy. The result was resumed in Table 2, where we represent C and Gamma's value that gives an accuracy that is equal to 0.90.

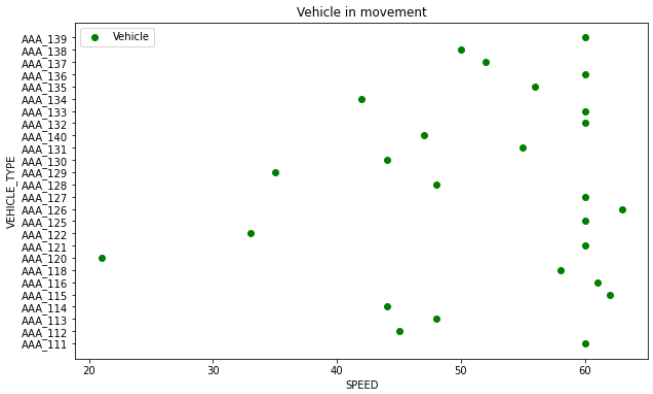


Figure 7: Distribution of vehicle based on speed

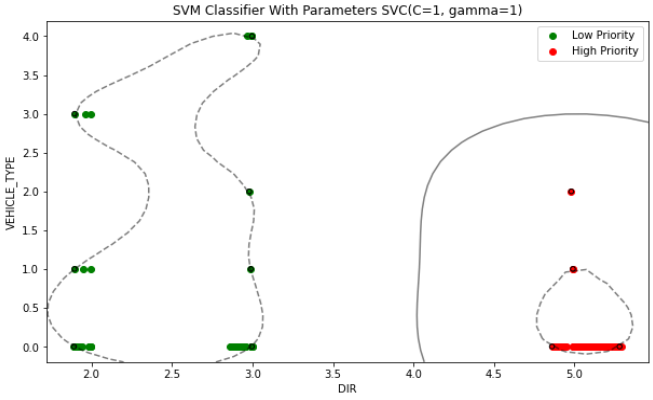


Figure 8 : SVM classifier with RBF kernel

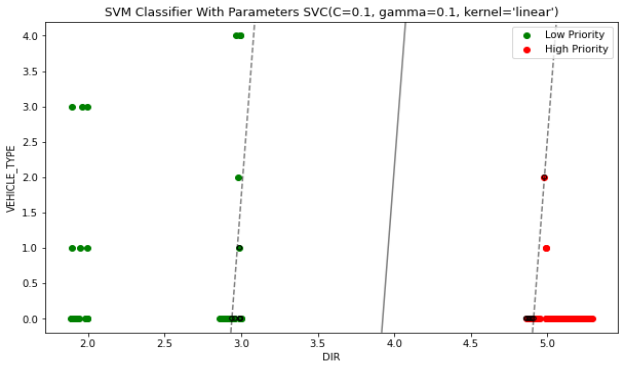


Figure 9: SVM classifier with Linear kernel

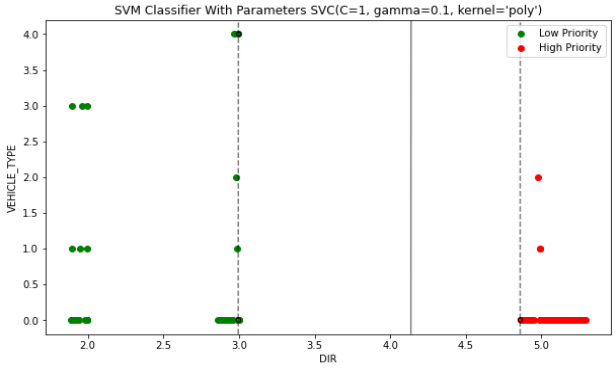


Figure 10: SVM classifier with polynomial kernel

Table 2: The best value of gamma and C

Gamma	C	accuracy	kernel
0.1	1	0.90	poly
0.1	0.1	0.9	linear
1	1	0.9	RBF

The figure 8,9 and 10 represents the classification results with each kernel. Figure 8 represents a two-dimensional classification problem. We plot the priority and the normal packet using two colors, red and green. There is a linear decision boundary for this dataset, but in figure 8, we can see how an RBF kernel can automatically decide a non-linear classification. In figure 9, we have three hyperplanes, and all of them separate the classes well. We can see that the margin between the continuous line and the data point is high compared to the dotted line. Therefore, we name the hyperplane that is represented by a continuous line as the good hyperplane. The reason for choosing the hyperplane with a higher margin is robustness, and if we choose the hyperplane with a low margin, then there is a high risk of wrong classification. The polynomial kernel is used with degree two, and he also gave a result close to the one with a linear kernel.

7. Conclusion

This paper proposes an SVM algorithm for classification problems in the context of the Vanet network. The contributions of this paper are as follows. First, we propose to add an additional step to the big data processing, which is the ordering step, and we have explained its role in identifying the data with high priority. This additional step can help to improve the cleaning step because by ordering data, we can remove the ones with low priority. It can also be employed to eliminate unused data namely the data of vehicle that is sent to the RSU and that possess a different direction from the latter. In the second part of this paper, we apply a machine learning algorithm to classify data. We select the SVM algorithm because of its high performance as a classifier, and we choose three kernels, which are the poly, linear, and the RBF. Since C and gamma's value can make a big change in the construction of the classifier, we apply a function that takes a different value of C and gamma for each kernel and return the value that gives the best accuracy. The paper aims to propose a process that can handle priority and urgent data and treat them at the right time. We propose adding a classification step that will allow us to classify data into two types: normal data and priority data. By using SVM, we achieve a good result. However, In future work, we will use the same proposed process, and we will apply the neural network to find the priority data. We will compare the SVM results with the neural network results. we chose to compare the neural network with SVM because, based on the comparative study conducted in [3], the neural network also gives good results in handling outliers, and also it is a very fast algorithm and can also handle multidimensional data and does not need much memory.

References

[1] S. Ftaimi, T. Mazeri, "The role of big data in a connected environment Case study: smart city," in 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), IEEE, Agadir, Morocco: 1–7, 2019, doi:10.1109/ICCSRE.2019.8807501.

[2] J. Lloret Mauri, ed., Cognitive networks: applications and deployments, CRC Press/Taylor & Francis Group, Boca Raton, 2015.

[3] S. Ftaimi, T. Mazri, "A comparative study of Machine learning algorithms for VANET networks," in Proceedings of the 3rd International Conference on Networking, Information Systems & Security, ACM, Marrakech Morocco: 1–8, 2020, doi:10.1145/3386723.3387829.

[4] R.G. Engoulou, M. Bellaïche, S. Pierre, A. Quintero, "VANET security surveys," Computer Communications, **44**, 1–13, 2014, doi:10.1016/j.comcom.2014.02.020.

[5] D. Dinesh, M. Deshmukh, "Challenges in Vehicle Ad Hoc Network (VANET)," International Journal of Engineering Technology, **2**(7), 14, 2014.

[6] M.N. Mejri, J. Ben-Othman, M. Hamdi, "Survey on VANET security challenges and possible cryptographic solutions," Vehicular Communications, **1**(2), 53–66, 2014, doi:10.1016/j.vehcom.2014.05.001.

[7] S. Al-Sultan, M.M. Al-Doori, A.H. Al-Bayatti, H. Zedan, "A comprehensive survey on vehicular Ad Hoc network," Journal of Network and Computer Applications, 37, 380–392, 2014, doi:10.1016/j.jnca.2013.02.036.

[8] IEEE P1609.0/D5, September 2012: IEEE Draft Guide for Wireless Access in Vehicular Environments (WAVE) - Architecture., IEEE, Place of publication not identified, 2012.

[9] M.S. Sheikh, J. Liang, "A Comprehensive Survey on VANET Security Services in Traffic Management System," Wireless Communications and Mobile Computing, 2019, 1–23, 2019, doi:10.1155/2019/2423915.

[10] M.S. Sheikh, J. Liang, W. Wang, "Security and Privacy in Vehicular Ad Hoc Network and Vehicle Cloud Computing: A Survey," Wireless Communications and Mobile Computing, 2020, 1–25, 2020, doi:10.1155/2020/5129620.

[11] L. Zhu, F.R. Yu, Y. Wang, B. Ning, T. Tang, "Big Data Analytics in Intelligent Transportation Systems: A Survey," IEEE Transactions on Intelligent Transportation Systems, **20**(1), 383–398, 2019, doi:10.1109/TITS.2018.2815678.

[12] S. Kaffash, A.T. Nguyen, J. Zhu, "Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis," International Journal of Production Economics, **231**, 107868, 2021, doi:10.1016/j.ijpe.2020.107868.

[13] X. Cheng, L. Fang, L. Yang, S. Cui, Mobile Big Data, Springer International Publishing, Cham, 2018, doi:10.1007/978-3-319-96116-3.

[14] M. Chen, S. Mao, Y. Liu, "Big Data: A Survey," Mobile Networks and Applications, **19**(2), 171–209, 2014, doi:10.1007/s11036-013-0489-0.

[15] W.-H. Lee, C.-Y. Chiu, "Design and Implementation of a Smart Traffic Signal Control System for Smart City Applications," Sensors, **20**(2), 508, 2020, doi:10.3390/s20020508.

[16] S. Al-Sultan, M.M. Al-Doori, A.H. Al-Bayatti, H. Zedan, "A comprehensive survey on vehicular Ad Hoc network," Journal of Network and Computer Applications, 37, 380–392, 2014, doi:10.1016/j.jnca.2013.02.036.

[17] P. Walsh, "Support Vector Machine Learning for ECG Classification," 10.

[18] M.A. Alsheikh, S. Lin, D. Niyato, H.-P. Tan, "Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications," IEEE Communications Surveys & Tutorials, **16**(4), 1996–2018, 2014, doi:10.1109/COMST.2014.2320099.

[19] M. Ozay, I. Esnaola, F.T. Yarman Vural, S.R. Kulkarni, H.V. Poor, "Machine Learning Methods for Attack Detection in the Smart Grid," IEEE Transactions on Neural Networks and Learning Systems, **27**(8), 1773–1786, 2016, doi:10.1109/TNNLS.2015.2404803.

[20] S. So, P. Sharma, J. Petit, "Integrating Plausibility Checks and Machine Learning for Misbehavior Detection in VANET," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, Orlando, FL: 564–571, 2018, doi:10.1109/ICMLA.2018.00091.

[21] K.M. Ali Alheeti, A. Gruebler, K.D. McDonald-Maier, "On the detection of grey hole and rushing attacks in self-driving vehicular networks," in 2015 7th Computer Science and Electronic Engineering Conference (CEEC), IEEE, Colchester: 231–236, 2015, doi:10.1109/CEEC.2015.7332730.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1428-1435 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Fabrication and Optimization of High Frequency ZnO Transducers for Both Longitudinal and Shear Emission: Application of Viscosity Measurement using Ultrasound

Hatem Dahmani[*,1], Ibrahim Zaaroura[2], Abbas Salhab[1,3], Pierre Campistron[1], Julien Carlier[1], Malika Toubal[1], Souad Harmand[2], Vincent Thomy[4], Marc Neyens[3], Bertrand Nongaillard[1]

[1]*Univ. Polytechnique Hauts-de-France, CNRS, Univ. Lille, ISEN, Centrale Lille, UMR 8520 – IEMN – Institut d'Électronique de Microélectronique et de Nanotechnologie, DOAE – Département d'Opto-Acousto-Électronique, F-59313 Valenciennes, France*

[2]*Univ. Polytechnique Hauts-de-France, CNRS, UMR 2201- LAMIH – Laboratoire d'Automatique de Mécanique et d'Informatique Industrielles et Humaines, F-59313 Valenciennes, France*

[3]*STMicroelectronics Crolles2, 850 rue Jean Monnet 38921 Crolles, France*

[4]*Univ. Lille, Centrale Lille, UVHC, ISEN, LIA LICS/LEMAC—IEMN UMR CNRS 8520, F-59000 Lille, France*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 01 December, 2020
Online: 21 December, 2020

Keywords:
ZnO Transducer
High frequency
Shear and Longitudinal waves
Matching network
C-axis inclined ZnO
Volume and Shear viscosities

ABSTRACT

This paper covers the study of high-frequency (~ 1 GHz) ZnO piezoelectric transducer integrated on a silicon substrate able to generate both compressional and shear acoustic waves. First, to promote the longitudinal mode, an electrical matching of the transducer in this high-frequency range is effectuated. Second, to promote shear waves, new deposition conditions were applied, giving thin zinc oxide films of inclined c-axis. The RF microprobe was used to validate the transducer design and to conduct the viscosity measurements. Thus, the shear and the volume viscosity of a water droplet were measured.

1. Introduction

This paper is an extension of work originally presented in the International Conference on Design & Test of integrated micro & nano-System (DTS 2019) [1]. Piezoelectric sensors have many applications in different areas, like the medical, manufacturing, and research fields [2-5]. One of these applications is the piezoelectric ZnO transducer, which has been well known for about 50 years, mainly for its high-frequency applications. [6]. Compared to monocrystalline lithium niobate (LiNbO3) [7], ZnO has lower piezoelectric performances (coupling coefficients), but it is easier to be deposited with the classical sputtering process. The thickness of the deposited ZnO can be from 10s nm to several microns. Additionally, it is possible to lead to the right (C-axis) crystallite orientation for longitudinal acoustic wave emission with a relatively short process duration.

Prior publications have demonstrated that depending on the distance between the sample's axis and the sputtering target axis, the orientation of crystallites cannot be homogeneous over a large substrate surface [8]. The inclination of the C-axis can be modified by controlling this distance, which slightly increases the proportion of shear acoustic waves at the expense of compressional waves.

Only a few applications report using a transducer capable of generating longitudinal and shear acoustic waves simultaneously [9, 10]. The first step in this article is to present the manufacturing, the characterization and the electrical modeling of ZnO piezoelectric transducers that generate both shear and longitudinal waves in the solid. Then, an electrical matching was used in order to increase the longitudinal mode.

The deposition of layers with appropriate orientation for shear excitation has been mentioned in several articles using different methods and substrates [11-13]. Thus, to increase the shear mode, one of these methods, mainly based on the addition of a blind

[*]Corresponding Author: Hatem Dahmani, Email: hatem.issat342@gmail.com

positioned between the sputtering target and substrate, was used. The growth of ZnO films with inclined orientation makes it possible to generate shear waves.

Finally, to demonstrate the usability of this transducer**,** both longitudinal acoustic reflection and shear coefficients at the solid/liquid interface were used to characterize the viscous fluid (a sessile water droplet) properties from its shear and volume viscosity.

2. Manufacturing, characterization and modeling of ZnO transducers

2.1. The manufacturing steps of the ZnO transducers

The piezoelectric material used is ZnO. It is deposited between two metal electrodes. The deposition configuration is shown in Figure 1, and it is fabricated by a standard thin-film technology: deposition and lift-off.

The choice of ZnO as a piezoelectric material can be related to its ease of preparation and its increased scope in high-frequency applications. One can site, for example, the acoustic microscopy applications in 100s MHz to 1s GHz frequency range [14,15].

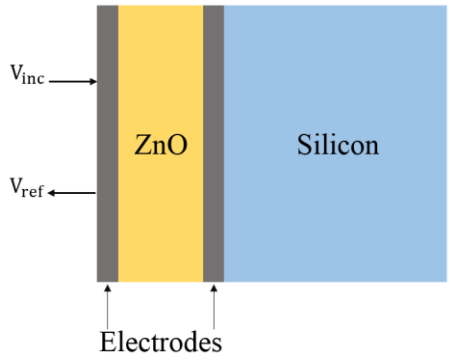


Figure 1: Multilayer Diagram of a Piezoelectric Transducer.

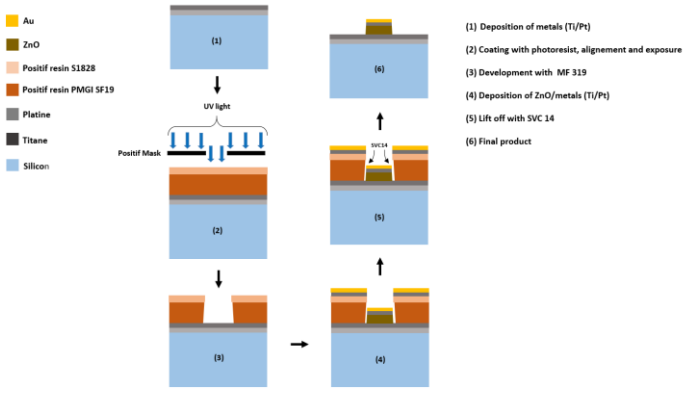
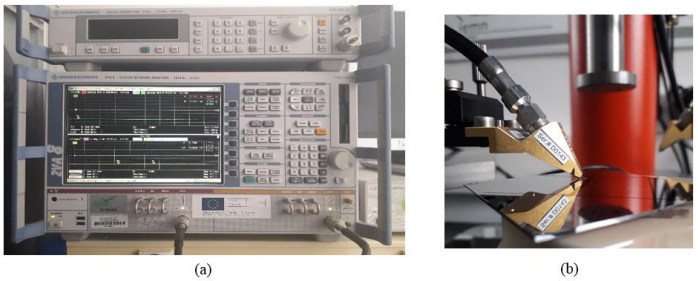


Figure 2: Piezoelectric transducer fabrication steps.

The ZnO piezoelectric transducer's fabrication starts with the deposition of 10 nm of a titanium layer by cathode sputtering on the surface of a 400 µm thick silicon wafer, followed by deposition of 80 nm of platinum. These two first layers will act as the lower electrode. A photolithography process is then performed to make the transducers pattern, where a thick 8 µm layer of PMGI SF19 (Microchem Corporation) is dispensed by spin coating on the lower electrode. Further, a layer of 1.5 µm of S1828 (Shipley Corporation) is appended. After ultraviolet exposure using a mask, a commercial developer MF 319 (Microposit) was used to obtain a pattern of different diameters in the PMGI layer. After that, a 2 µm thin ZnO layer was deposed. Finally, a 10 nm of titanium adhesion layer followed by 400 nm of gold were deposited. The gold layer is used to enable soldering of the contact electrode to a gold wire to connect the transducer to an external circuit. The last step is the lift-off process, where the UV exposed resin is removed to get the transducers pattern. The transducers have a diameter of 150 um. The microfabrication process is shown in Figure 2.

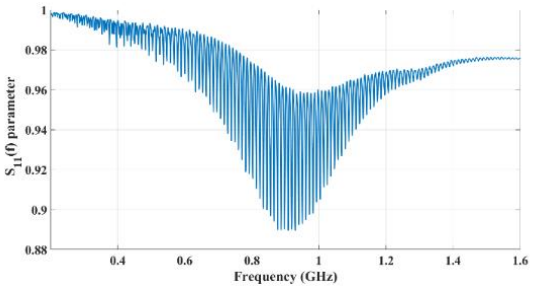
2.2. Characterization of the transducers

The ZnO transducers' characterization is carried out by a Rohde & Schwarz ZVA8 Vector Network Analyzer (VNA) as shown in Figure 3.a. The electrical contact is made using a radio-frequency measuring tip Z-probe, GS type (pitch = 1.25 mm) connected by a coaxial cable to the VNA (Figure 3.b). Since we are working within a high-frequency range in a Gigahertz order, a calibration must be done before each measurement to ensure reliable results. The silicon sample is fixed by suction (using a vacuum pump) on a circular plate.

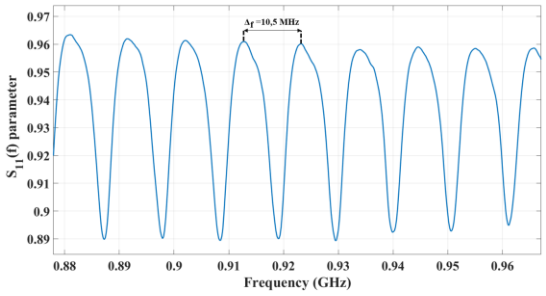


(a) (b)

Figure 3: Image of (a) Rohde & Schwarz ZVA8 Vector Network Analyzer (VNA), (b) tip Z-probe.



(a)



(b)

Figure 4: (a) Parameter S_{11} measured with the network analyzer, (b) Zoom around the longitudinal acoustic wave resonance.

The S_{11} parameter as a function of frequency, shown in Figure 4.a, is the output measured on a transducer located on the edge of the substrate. At this position, the crystal orientation of the ZnO is not perfectly vertical, thus, the generation of acoustic shear waves is clearly observed. This parameter given by Equation 1 is composed of two contributions [16,17]. The first one comes from the direct reflection of the electric wave on the transducer (S_{11}^{el}. The second is a result of all the acoustic wave reflections on the opposite side of the substrate. These acoustic reflections are then converted into electrical signal by the transducer.

$$S_{11}(f) = S_{11}^{el} + K_{11}S_{11}^{ac} \quad (1)$$

The coefficient K_{11} is the electro-acoustic-electrical coupling coefficient of the piezoelectric material used to convert energy from electric into acoustic for the emission then from acoustic into electrical for the reception.

The reflection transfer function of our system is represented by S_{11}. To obtain the temporal response, an inverse Fourier Transform is performed (Figure 5). In the time domain, the electrical term response appears shortly after the transducer's excitation while the acoustic term arises later.

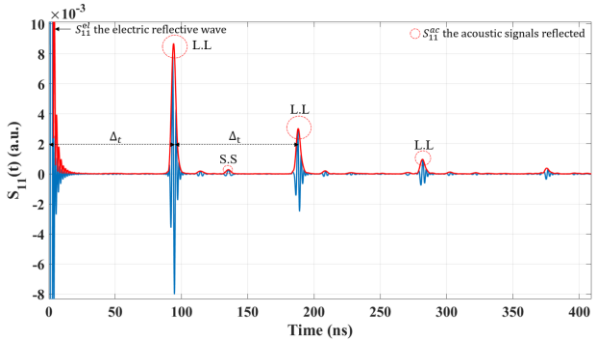


Figure 5: Impulse response $S_{11}(t)$ for f = [0.2 - 1.6 GHz].

The response shown in Figure 4 exhibits a rapid oscillation of $S_{11}(f)$ having a peak oscillation around the central frequency (f = 0.9 GHz). This is due to the fact that for each wavelength emitted in the substrate of a given thickness, the superposition of the emitted and received waves occurs with a variable phase shift. Thus, the total vibration amplitude acquired by the transducer (amplitude of the sum of the emitted and received waves) is greater for the case when the two waves (incident and reflected) are in phase and lower when they are in opposite phases (phase cancellation effect). This phenomenon is portrayed by the SWR standing wave ratio (between the highest and the lowest vibration levels). The frequency of these periodic interference ripples in $S_{11}(f)$ is $\Delta_f = c_{Si}/2e_{Si}$, where c_{Si} and e_{Si} are respectively the celerity of wave and the thickness of the silicon layer (400 µm) (Figure 4.b). The time-domain response corresponding to Figure 4.a is shown below in Figure 5. The multiple echoes are seen, the first one corresponding to the S_{11}^{el}. This electrical echo is directly due to the electrical mismatch between the transducer and the external circuit (50-Ohm load impedance of the VNA). The minor S_{11}^{el} delay corresponds to the electromagnetic wave propagation time in the cable portion between the network analyzer and the RF microprobe. The other echoes correspond to the different acoustic signals (S_{11}^{ac}) traveling back and forth in the silicon layer. They are separated by periods $\Delta_t = 1/\Delta_f = 94$ ns defined by the double Si wafer thickness's propagation time.

It is observed that the transducer emits both longitudinal and shear waves, which propagate in the silicon substrate at different speeds. The waves are denoted LL and SS for longitudinal/longitudinal or shear/shear propagation, respectively. The coupling of these 2 waves is very weak as the reflecting surface of the silicon is smooth and parallel to the transducer.

2.3. Transducer design using Auld's model

For checking the manufacturing quality of the transducers, we can model their impedance Zel and compare it to experiments. The piezoelectric layers being very thin compared to their lateral dimension (2 µm thick vs. 150 µm diameter), we can model them by a 1D model, like the Auld model [18].

As longitudinal and shear waves are emitted, we will use a modified Auld model that includes the two modes of emission. For that, it is necessary to consider the different mechanical impedances Z of the transducer's constituting layers. The analytical formulation for Zel of a transducer rigidly bounded to a semi-infinite solid is given by Equation 2; the mechanical impedances $Z^S_{(Air/Au/Ti)}$ and $Z^L_{(Air/Au/Ti)}$ in Equations 3 and 4 are hardly significant or negligible relatively to the other terms [16].

$$Z_{el} = \frac{1}{jC_0w} + \frac{K_S^2}{C_0w}A + \frac{K_L^2}{C_0w}B \quad (2)$$

where,

$$A = \frac{Z^S_{ZnO}}{\theta_S}\frac{2Z^S_{ZnO}[\cos(\theta_S)-1]+j\left[Z^S_{(Air/Au/Ti)}+Z^S_{(Si/Ti/Pt)}\right]\sin(\theta_S)}{Z^S_{ZnO}\left[Z^S_{(Air/Au/Ti)}+Z^S_{(Si/Ti/Pt)}\right]\cos(\theta_S)+j\left[(Z^S_{ZnO})^2+Z^S_{Air/Au/Ti)}.Z^S_{(Si/Ti/Pt)}\right]\sin(\theta_S)}$$

and

$$B = \frac{Z^L_{ZnO}}{\theta_L}\frac{2Z^L_{ZnO}[\cos(\theta_L)-1]+j\left[Z^L_{(Air/Au/Ti)}+Z^L_{(Si/Ti/Pt)}\right]\sin(\theta_L)}{Z^L_{ZnO}\left[Z^L_{(Air/Au/Ti)}+Z^L_{(Si/Ti/Pt)}\right]\cos(\theta_L)+j\left[(Z^L_{ZnO})^2+Z^L_{(Air/Au/Ti)}.Z^L_{(Si/Ti/Pt)}\right]\sin(\theta_L)}$$

L and S indices stand to the longitudinal and shear waves. $C_0 = \varepsilon_{ZnO}.S_{ZnO}/e_{ZnO}$ is the static capacitance of the ZnO transducer, ε_{ZnO} the dielectric permittivity of ZnO, e_{ZnO} and SZnO are the thickness and the active surface of the ZnO layer. K^2 is the square of the electromechanical coupling coefficient of the ZnO. $\theta = k_{ZnO}.e_{ZnO} = 2\pi f\,(e_{ZnO}/c_{ZnO})$ with k_{ZnO} the wave number in ZnO and c_{ZnO} the longitudinal or transverse wave's velocity the ZnO.

From an experimental point of view, the S-parameter is linked to the transducer impedance as follows:

$$S_{11} = \frac{Z_{el}-Z_0}{Z_{el}+Z_0} \quad (3)$$

Z_0 is the characteristic impedance of the network analyzer equal to 50 Ω.

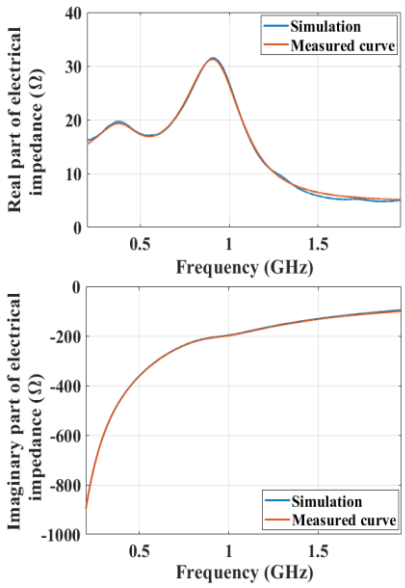


Figure 7: Simulated (blue curve) and measured (red curve) electrical impedance Z_{el} of the 150 µm diameter ZnO transducer.

The 150 µm diameter ZnO transducer's simulation results are shown in Figure 7 by the solid blue curves. The real part of Z_{el} (Figure 7.a) exhibits two peaks occurring at the resonance frequencies of the transverse (f_S= 0.38 GHz) and longitudinal (f_L= 0.9 GHz) waves.

A good fit between the simulation and the experimentation can be seen in Figure 7. So, the Auld model can be used to design a transducer for which the real part is close to 50 Ω by adjusting the thicknesses of the thin layers and the surface.

The manufacturing process is well calibrated which allows us to know the thicknesses of the different materials quite precisely. We were able, during various works, to measure the acoustic properties of ours deposited layers [17]. We therefore know the parameters on which the model depends, except the electro-mechanical coupling coefficients of the transducers which depend on the unknow orientation of the ZnO crystallites. We get them by an optimization routine using Nelder-Mead's algorithm. In the simulation, K_s and K_L were set to 0.10 and 0.23, respectively. The result of optimization for K_L is close to the theoretical one obtained for a perfect crystallite c-axis alignment (K_L = 0.27, according to [7]).

As shown in Figure 8, the pure electrical model we use represents the 2 modes of vibration of the transducer by the association of 2 RLC circuits shunted by the static capacitance C_0 . From a physical point of view, the capacity C_0 reflects the resonators electromagnetic behavior, while the two parallel resonant circuits exhibit the resonators acoustic behavior.

The electrical parameters of the circuit are determined by using the electrical impedance thanks to the measurement of the S_{11} parameter. As for the evaluation of the coupling coefficients, we obtained the electrical parameters by an optimization procedure.

The electrical parameters of the circuit are determined by using the S_{11} parameter:

$C_0 = 0.82$ pF, $L_L = 641$ nH, $C_L = 0.047$ pF, $R_L = 1645$ Ω, $L_S = 11.87$ µH, $C_S = 0.012$ pF, $R_S = 24$ kΩ.

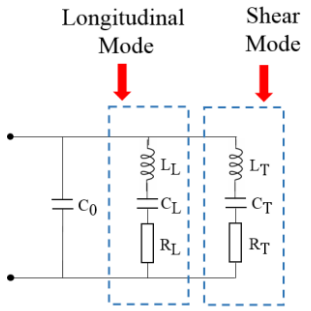


Figure 8: The equivalent electric circuit of the ZnO transducer.

3. Transducer optimization

3.1. Increasing the longitudinal acoustic waves emission by impedance matching

In any kind of application, we need to favor either one or both types of the generated waves. In the present study, regarding the matching network, promoting the longitudinal waves is sought [19].

The concept of electrical matching is to transfer the maximum power from the source to the transducer or, in other words, to eliminate electrical reflections on the transducer in a chosen frequency range. As piezoelectricity is a reciprocal phenomenon, the electrical matching network also acts as an acoustical matching layer. So, the electrical matching reduces the reflection of the acoustical waves returning on the piezoelectric layer, which increases the acoustical transmission coefficient at the piezoelectric device. Simultaneously, levels of both acoustic waves propagating in the device and electrical measurement signal increase.

There are two steps for the electrical impedance matching. The first one consists of connecting the transducer with a gold wire (25 µm of diameter). The second step uses an L-section impedance matching [20], consisting mainly of two reactive elements (capacitor and inductor). The capacity is used to adjust the real part of the impedance to 50 Ω, while the coil reduces the imaginary part. A capacitance has a significant loss effect at high-frequency, causing critical insertion losses [21]. For this reason, a chip inductor with a value of 8.6 nH was put in series with the load (transducer connected by the gold wire) at $f_{L,match}$ = 1.12 GHz. The real part of the load at this frequency is equal to 50 Ω.

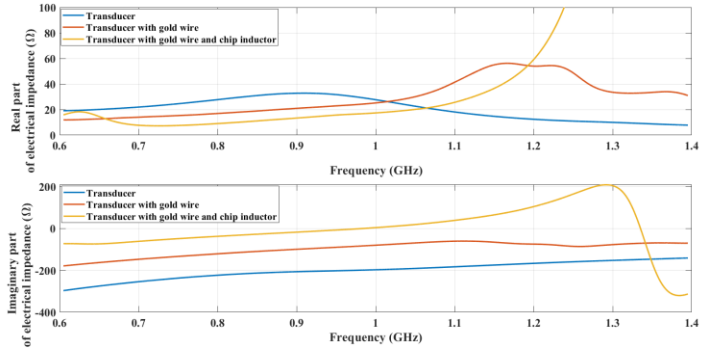


Figure 9: The real and imaginary parts of the electrical impedance of the ZnO transducer (blue curve), the transducer with gold wire (red curve) and transducer with gold wire and chip inductor (yellow curve).

It is clearly seen that the electrical response has been modified (Figure 9). This modification can be related to the losses and the parasite impedances of bonding wire and packaging. So, it has become complicated to choose a filter based on the equivalent electric circuit of the transducer, such as the Butterworth or Chebychev filters [22, 23, 24]. Therefore, the simplest solution is to use the transducer impedance's real and imaginary parts, including the wire.

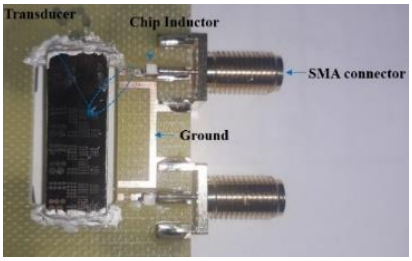


Figure 10: Image of the different elements of the circuit.

The expected result is the gain in longitudinal wave amplitude, which will make it easy to connect the transducer to the VNA or a generator via an SMA connector, as shown in Figure 10.

To verify that the electrical adaptation works as expected, we compare the insertion loss of the transducer before and after the adaptation, through an SMA connector. Afterwards, the responses in terms of received amplitude level will be compared. The electrical insertion losses L_e was calculated by the following equation:

$$L_e = -10 \log \left(1-\left|S_{11}^{el}\right|^2\right) \quad (4)$$

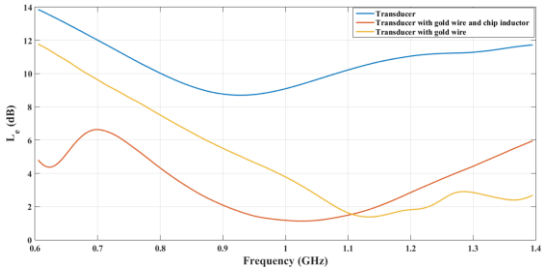


Figure 11: Measured electrical insertion losses.

As shown in Figure 11, the electrical adaptation has reduced the insertion losses. The matched transducer exhibits an amplitude gain of 8 dB and a bandwidth enlarged to 250 MHz.

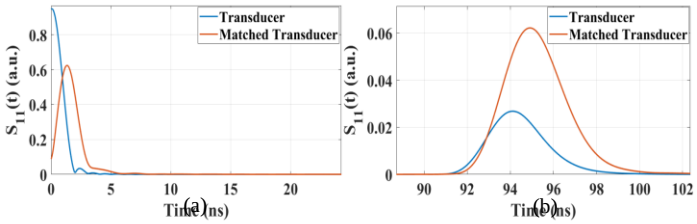


Figure 12: Impulse response $S_{11}(t)$, (a) the electrical reflection S_{11}^{el}, (b) the acoustic response S_{11}^{ac}.

The matching network's direct effect can be represented in Figure 12, showing the echoes before (blue curve) and after (red curve) the electrical matching. The attenuation of the reflected electrical signal is notably due to the conversion from electrical into acoustic energy. This conversion can be seen by the simultaneous amplitude decrease of the electrical echo S_{11}^{el} (Figure 12.a) and the increase of the first acoustical echo S_{11}^{ac} (Figure 12.b). However, this result only applies to $f_{L,match}$.

Although the electrical adaptation improves the transducer's acoustic sensitivity, the electrical reflection does not completely disappear because the use of an inductor is only at a single frequency $f_{L,match}$ and not over the bandwidth of the transducer.

3.2. Increasing the shear acoustic waves emission by deposition of ZnO inclined c-axis by R.F. magnetron sputtering

It appears difficult to use a matching network for increasing the level of shear waves as they are too weak. Previous measurements have been made with a transducer from a misaligned area (with respect to the ZnO target of the silicon wafer). By increasing that effect, it is possible to increase the shear level.

The ZnO thin film fabrication process was modified to balance the amplitude of generated longitudinal and shear waves by changing the material crystalline orientation. The film was deposited on three wafers with the following stack: Pt (80 nm) / Ti (10 nm) / Si (400 µm) by R.F. magnetron sputtering technique. The substrate was fixed in the chamber, whose total pressure was 0.02 mBar. A rectangular cover (the blind) is positioned between the target and the substrate, making it possible to obtain oblique incidence of the particles in order to locally influence the growth of the ZnO thin film (Figure 13). The blind was made of aluminum, had a height of 30 mm and a thickness of 1 mm. It was placed in the center of the substrate. The deposition rate of the ZnO films varied from 15 to 25 nm/min.

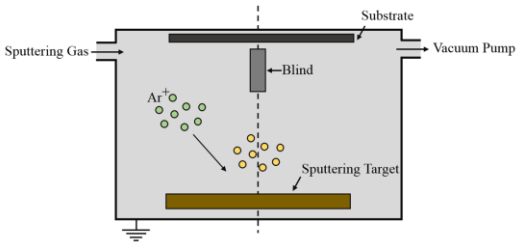


Figure 13: R.f. magnetron sputtering system with additional blind between substrate and target.

A cross-section SEM image of the ZnO deposited on the silicon substrate is shown in Figure 14. The inclination of the crystallite orientation is approximately 14.5° to the surface normal can be observed. So, shear and longitudinal waves can be excited. However, the inclination of the crystal orientation is strongly dependent on the distance between the blind and the transducer (in the substrate plane); therefore, the c-axis inclination will be higher in an area close to the blind or close to the edge of the substrate.

Figure 15.a shows the resonance diagram of the piezoelectric transducer; two distinguished downward peaks can be seen, the first one in the frequency range [0.35 - 1 GHz] which corresponds to critical transversal waves response compared to Figure 4 and the other peak in the range [1 - 2 GHz] corresponding to longitudinal

waves response. Two oscillations with different frequencies Δf = {S: 7.4 MHz, L: 10.5 MHz}, are presented in Figure 15.b. Each of them presents a different wave mode. This result directly reflects the transducers' ability sputtered with the inclined ZnO c-axis to generate both shear and longitudinal waves.

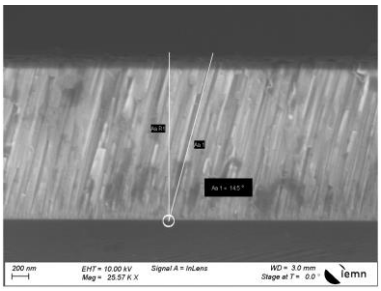
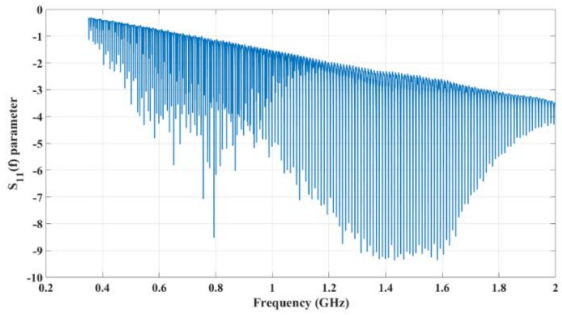
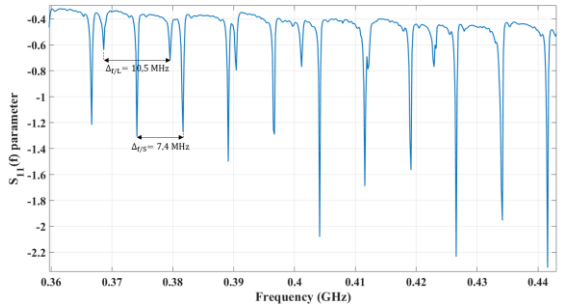


Figure 14: Cross-sectional SEM image of ZnO film with c-axis inclined of 14.5°.



(a)



(b)

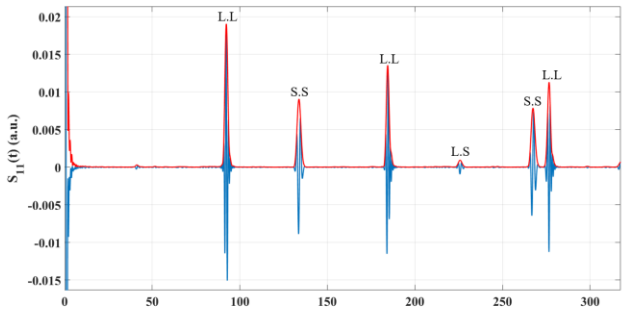
Figure 15: Impulse response $S_{11}(f)$, (a) for f=[0.35 - 2 GHz], (b) Zoom around the longitudinal and the shear acoustic waves resonance.

Figure 15 (b), as we expect each of them are close to the theoretical ones {Cs/2e = 5843/800e-6 = 7,3 MHz and Cl/2e = 8432/800e-6 = 10,54 MHz}.

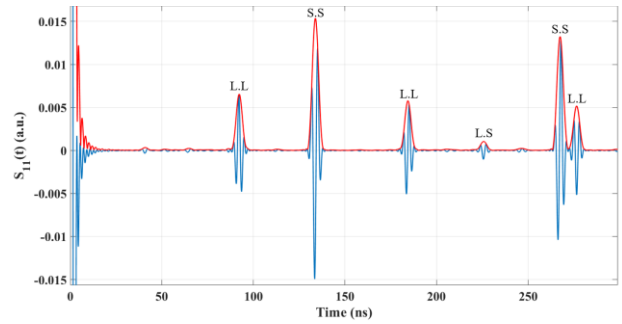
The time-domain representation is used below to check the level of both signals. Figure 16 shows the inverse Fourier transform of the S11 parameter. For [0.35 - 2 GHz], two different acoustic reflections LL, and SS, are the responses from the generated acoustic waves, with LL being a purely longitudinal response with longitudinal wave emitted and the longitudinal wave reflected. The longitudinal speed of sound in silicon (8432 m/s) is greater than the transversal one (5832 m/s); thus, the time of reception of the signal for the longitudinal waves should be the fastest (the first reflection on Figure. 16.a corresponds to LL). The second impulse SS is the purely transversal reflected signal with the time delay being significantly greater than LL. For a bandwidth of 1.65 GHz, the ratio of the shear wave to the longitudinal wave is 0.5.

We decreased the frequency range to [0.35 - 1 GHz] (Figure 16.b), the shear to shear (SS) reflection amplitude increases, and the longitudinal to longitudinal (LL) amplitude decreases, which what was expected to be seen since the lower frequency range favors the transversal waves (Figure 16.a). The ratio of shear to longitudinal amplitudes increases to 2.6.

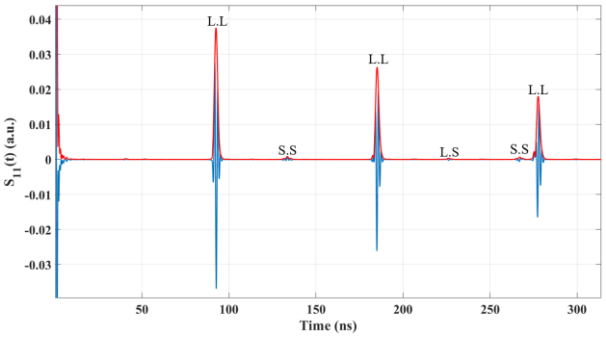
For f = [1 - 2 GHz] (Figure 16.c), the result is the profound decrease in the SS reflection amplitudes and the profound increase in the LL amplitudes, which is what to expect since the high frequencies favor the longitudinal waves. We now have a ratio of shear to longitudinal amplitudes decreased to 0.01.



(a)



(b)



(c)

Figure 16: Impulse response $S_{11}(t)$, the time delays $\Delta_{t/L} = 1/\Delta_{f/L} = 95$ ns and $\Delta_{t/S} = 1/\Delta_{f/S} = 135$ ns between them corresponds to the propagation delays of the longitudinal and transverse wave travelling twice the thickness of the wafer. (a) f = [0.35-2 GHz], (b) f = [0.35-1 GHz] and (c) f = [1-2 GHz].

Levels are weak, but all signals are out of the noise and can be used for measurements. If necessary, they can be increased using a convenient matching network, as demonstrated in part 1. Thus, we propose a method allowing the emission of both longitudinal and shear waves using only one fixed fabrication process.

However, the c-axis tilt is not the same on the wafer surface, so the transducers cannot generate the same wave level for each frequency range, so the ratio of shear to longitudinal amplitudes changes.

4. Application: Measurement of the viscosity of a liquid using both longitudinal and shear waves generated by the ZnO transducer

There are a vast number of applications that utilize physical parameters in relation to acoustic waves-material interactions. An application to measuring the fluid viscosity is presented below to demonstrate the usability of the proposed transducer design. Water is used as the reference fluid whose acoustic properties are given in [25, 26].

The total viscosity of water can be extracted from the shear and longitudinal waves generated at the liquid/ solid interface [27,28]. Our transducers can generate these two types of waves. We used the evaporation of 1 µL droplet of water to determine the two kinds of viscosities (namely shear and volume viscosities). The dynamic shear viscosity was extracted using the shear wave, while the volume viscosity from both waves.

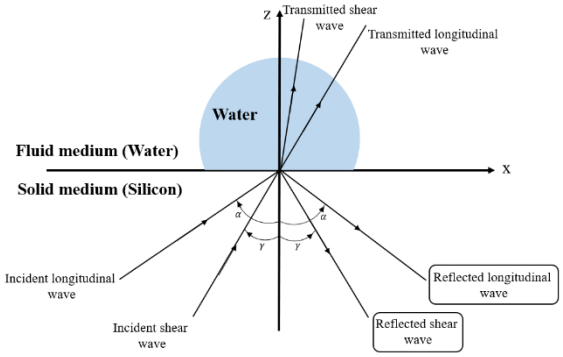


Figure 17: Incident acoustic waves interaction at droplet/silicon interface.

Figure 17 presents the direct contact of the acoustic waves to the water droplet/silicon interfaces. The acoustic waves reflect at the interface and generate two components, i.e., longitudinal and shear reflected waves.

Several mechanical properties at the silicon/water interface influence the measurement, and these are also strongly dependent on the temperature. The reflection coefficient at the solid/fluid interface is given by equation 5, Z_0 is the mechanical impedance of the silicon, considered purely elastic, so Z_0 is purely real [29], while water presents viscoelastic properties. As a consequence, its acoustic impedance Z_1^* is in a complex form.

$$r^* = \frac{Z_1^* - Z_0}{Z_1^* + Z_0} \tag{5}$$

where $Z = \rho C$, ρ is the density of the medium, and C is the acoustic velocity of the wave propagating inside of the medium. The reflection coefficients are determined for both longitudinal and shear reflected waves. So, the complex form of the shear reflection coefficient ($r_{SS}{}^*$) and the longitudinal reflection coefficient ($r_{LL}{}^*$) are shown in the Equations 6 and 7. These coefficients give access to the modulus and the phase angle of the two reflections acoustic wave at the interface.

$$r_{SS}^* = \frac{Z_w^{S*} - Z_{si}^S}{Z_w^{S*} + Z_{si}^S} = \frac{\rho_w C_S^* - Z_{si}^S}{\rho_w C_S^* + Z_{si}^S} \tag{6}$$

where,

$$C_S^* = \omega/k_S^* \quad \text{and} \quad k_S^* = \omega/C_S - i\alpha_S = \omega/C_S - i\sqrt{\frac{\rho_w \omega}{2\eta}}$$

$$r_{LL}^* = \frac{Z_w^{L*} - Z_{si}^L}{Z_w^{L*} + Z_{si}^L} = \frac{\rho_w C_L^* - Z_{si}^L}{\rho_w C_L^* + Z_{si}^L} \tag{7}$$

where,

$$C_L^* = \omega/k_L^* \text{ and } k_L^* = \omega/C_L - i\frac{\omega^2}{2\rho_w C_L^3}\left[\mu + \frac{4}{3}\eta\right]$$

where C_L^*, C_S^* are the complex longitudinal and shear velocity of water, ρ_w is the density, k_L^*, k_S^* the complex longitudinal and shear wave number, ω (rad/s) the angular velocity. η and μ are the shear and volume viscosity, respectively.

C_L and C_S are the effective longitudinal and shear phase velocities in water respectively, α_S is the shear attenuation. The kinetics of the shear and volume viscosity is shown in Figure 18. A sessile droplet of water evaporating at room temperature of 25°C is used. It is clearly seen that the consistency of the theoretical and experimental results.

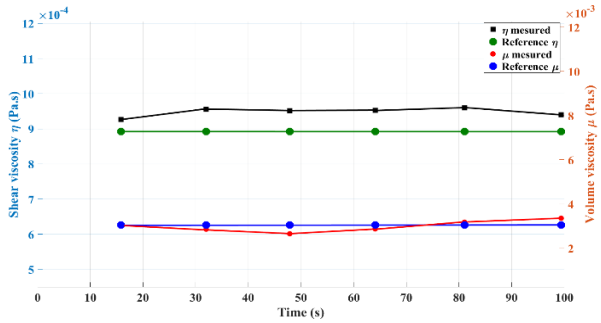


Figure 18: Kinetics of the shear and volume viscosity of a water droplet at T=25°C compared to the reference values.

It was essential to improve the shear waves emission to accurately measure the shear viscosity values for different complex fluids such as nanofluids. These nanofluids contain nanoparticles in suspension in a matrix fluid; that is why it is important to have a high sensitivity of these particles that settled above the surface.

5. Conclusion

This article is the extended version of a paper published in IEEE Xplore. We have studied the fabrication and the optimization of a ZnO high-frequency transducer able to generate both longitudinal and shear waves. To model the transducer's electrical response, we used the Auld model; we noticed a good agreement between electrical impedance measurements and simulations, allowing its use in the design of transducers. To

improve the sensitivity, we have used an electrical circuit including a chip inductor and an SMA connector so that it could be used without the requirements of an external RF microprobe. This matching circuit had been experimented with for the longitudinal mode and allowed us to gain ~ 8 dB on the electrical insertion losses.

In order to increase the shear wave level, the growth of ZnO films with tilting of the crystallite axis has been achieved using a blind of 10 mm that was positioned between the target and the substrate during the R.F. magnetron sputtering. The inclination of the crystal orientation was approximately 14.5°.

Both shear and longitudinal waves can be excited with this multi-mode transducer; the wave type can be selected by adjusting the emission frequency.

As to proof of the usability of this kind of measurement, an application has been proposed. We have exploited the simultaneous measurements of the two types of waves to extract both shear and volume viscosities of the water. Comparisons between the experimental and theoretical results show a good agreement.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was partially supported by the French Renatech network. A part of the cost has also been supported by STMicroelectronics and Hauts-de-France region. The authors wish to thank all the members of the cleanroom at IEMN- UMR CNRS 8520. They are also grateful to the professor Bertrand Nongaillard for his encouragement and support in this work. Finally, the authors also thank the laboratory engineers, Dr. Malika Toubal, and Dr. Nikolay Smagin, for their help in preparing the article.

References

[1] H. Dahmani, I. Zaaroura, P. Campistron, J. Carlier, M. Toubal, B. Nongaillard, "Study of a dual longitudinal and shear acoustic wave ZnO transducer in the gigahertz frequency range: Application to viscosity measurement and first step for device integration," in IEEE International Conference on Design and Test of Integrated Micro and Nano-Systems, DTS 2019, 2019, doi:10.1109/DTSS.2019.8914906.

[2] F.L. Di Scalea, H. Matt, I. Bartoli, A. Srivastava, G. Park, C. Farrar, "The fundamental response of piezoelectric guided-wave sensors and applications to damage and impact location," in Proceedings of the 3rd European Workshop - Structural Health Monitoring 2006, 1271-1280, 2006.

[3] C. Caliendo, "Longitudinal modes along thin piezoelectric waveguides for liquid sensing applications," Sensors (Switzerland), **15**(6), 12841-12856, 2015, doi:10.3390/s150612841.

[4] L.P. Suwal, R. Kuwano, "Disk shaped piezo-ceramic transducer for P and S wave measurement in a laboratory soil specimen," Soils and Foundations, **53**(4), 510-524, 2013, doi:10.1016/j.sandf.2013.06.004.

[5] F.A. Alexandre, P.R. Aguiar, R. Götz, M.A.A. Viera, T.G. Lopes, E.C. Bianchi, "A novel ultrasound technique based on piezoelectric diaphragms applied to material removal monitoring in the grinding process," Sensors (Switzerland), **19**(18), 2019, doi:10.3390/s19183932.

[6] J. de Klerk, "Thin film zinc oxide transducers for use in microwave devices," Ultrasonics, **8**(3), 159-164, 1970, doi:10.1016/0041-624X(70)90614-1.

[7] D. Royer, E. Dieulesaint, Elastic Waves in Solids I: Free and Guided Propagation, translated by D. P. Morgan, 2000.

[8] M. Link, Study and realization of shear wave mode solidly mounted film bulk acoustic resonators (FBAR) made of c-axis inclined zinc oxide (ZnO) thin films: application as gravimetric sensors in liquid environments, Ph. D Thesis, Henri Poincaré University, 2006.

[9] I. Zaaroura, M. Toubal, H. Reda, J. Carlier, S. Harmand, R. Boukherroub, A. Fasquelle, B. Nongaillard, "Evaporation of nanofluid sessile drops: Infrared and acoustic methods to track the dynamic deposition of copper oxide nanoparticles," International Journal of Heat and Mass Transfer, **127**, 1168-1177, 2018, doi:10.1016/j.ijheatmasstransfer.2018.07.102.

[10] C. Virgilio, Caractérisation du mouillage de surfaces micro / nanostructurées par méthode acoustique haute fréquence : application aux traitements humides dans l'industrie de la microélectronique Ph. D Thesis, Valenciennes and Hainaut-Cambresis University, 2017.

[11] M. Link, M. Schreiter, J. Weber, R. Primig, D. Pitzer, R. Gabl, "Solidly mounted ZnO shear mode film bulk acoustic resonators for sensing applications in liquids," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, **53**(2), 492-496, 2006, doi:10.1109/TUFFC.2006.1593389.

[12] S. Bensmaine, L. Le Brizoual, O. Elmazria, J.J. Fundenberger, B. Benyoucef, "Deposition of ZnO inclined c-axis on silicon and diamond by r.f. magnetron sputtering," in Physica Status Solidi (A) Applications and Materials Science, **204**(9), 3091-3095, 2007, doi:10.1002/pssa.200776322.

[13] M. Link, M. Schreiter, J. Weber, D. Pitzer, R. Primig, M.B. Assouar, O. Elmazria, "C-axis inclined ZnO films deposited by reactive sputtering using an additional blind for shear BAW devices," in Proceedings - IEEE Ultrasonics Symposium, **1**, 202-205, 2005, doi:10.1109/ULTSYM.2005.1602831.

[14] K. Yamada, H. Shimizu, M. Minakata, "PLANAR-STRUCTURE FOCUSING LENS FOR OPERATION AT 200 MHZ AND ITS APPLICATION TO THE REFLECTION-MODE ACOUSTIC MICROSCOPE.," in Ultrasonics Symposium Proceedings, 745-748, 1986, doi:10.1109/ultsym.1986.198834.

[15] B.T. Khuri-Yakub, "Scanning acoustic microscopy," Ultrasonics, **31**(5), 361-372, 1993, doi:10.1016/0041-624X(93)90070-G.

[16] Y. Deblock, P. Campistron, M. Lippert, C. Bruneel, "Electrical characterization of plate piezoelectric transducers bonded to a finite substrate," The Journal of the Acoustical Society of America, **111**(6), 2681-2685, 2002, doi:10.1121/1.1416904.

[17] P. Campistron, J. Carlier, N. Saad, J. Gao, M. Toubal, L. Dupont, G. Nassar, B. Nongaillard, "High frequency ultrasound, a tool for elastic properties measurement of thin films fabricated on silicon," in Advanced Materials Research, **324**, 277-281, 2011, doi:10.4028/www.scientific.net/AMR.324.277.

[18] B.A. Auld, R.E. Green, Acoustic Fields and Waves in Solids: Two Volumes, Physics Today, 1974.

[19] J. Gao, J. Carlier, S. Wang, P. Campistron, D. Callens, S. Guo, X. Zhao, B. Nongaillard, "Lab-on-a-chip for high frequency acoustic characterization," Sensors and Actuators, B: Chemical, **177**, 753-760, 2013, doi:10.1016/j.snb.2012.11.037.

[20] D.M. Pozar, Microwave Engineering, 4th Edition, 2012.

[21] R. Fiore, "Capacitors in broadband applications," Applied Microwave and Wireless, **13**(5), 40, 2001.

[22] R.N. Thurston, "Effect of Electrical and Mechanical Terminating Resistances on Loss and Bandwidth According to the Conventional Equivalent Circuit of a Piezoelectric Transducer," IRE Transactions on Ultrasonic Engineering, **7**(1), 16-25, 1960, doi:10.1109/T-PGUE.1960.29256.

[23] D.C. Youla, A new theory of broadband matching, IEEE Transactions on Circuit Theory, 1964.

[24] W.-K. Chen, The Circuits and Filters Handbook, 2003.

[25] L. Korson, W. Drost-Hansen, F.J. Millero, "Viscosity of water at various temperatures," Journal of Physical Chemistry, **73**(1), 34-39, 1969, doi:10.1021/j100721a006.

[26] R.A. Horne, D.S. Johnson, "The viscosity of water under pressure," Journal of Physical Chemistry, **70**(7), 2182-2190, 1966, doi:10.1021/j100879a018.

[27] J. Holmes, G. Parker, W. Povey, "Temperature dependence of bulk viscosity in water using acoustic spectroscopy," in Journal of Physics: Conference Series, **269**(1), 2011, doi:10.1088/1742-6596/269/1/012011.

[28] A.J. Zuckerwar, R.L. Ash, "Variational approach to the volume viscosity of fluids," Physics of Fluids, **18**(4), 2006, doi:10.1063/1.2180780.

[29] M. Trempa, C. Reimann, J. Friedrich, G. Müller, D. Oriwol, "Mono-crystalline growth in directional solidification of silicon with different orientation and splitting of seed crystals," Journal of Crystal Growth, **351**(1), 131-140, 2012, doi:10.1016/j.jcrysgro.2012.04.035.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1436-1457 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Parameter Estimation for Industrial Robot Manipulators Using an Improved Particle Swarm Optimization Algorithm with Gaussian Mutation and Archived Elite Learning

Abubakar Umar[1], Zhanqun Shi[1, *], Lin Zheng[2], Alhadi Khlil[1], Zulfiqar Ibrahim Bibi Farouk[3]

[1]*School of Mechanical Engineering, Hebei University of Technology, Tianjin, 300401, China*

[2]*School of Mechanical and Electrical Engineering, Dalian Minzu University, Dalian, 116600, China*

[3]*School of Mechanical Engineering, Tianjin University, Tianjin, 300350, China*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 01 December, 2020
Online: 21 December, 2020

Keywords:
Robotics
Dynamic model
Swarm intelligence

ABSTRACT

This work presents the analysis and formulation for optimizing the dynamic model and parameter estimation of all the six joints of a 6DOF industrial robot manipulator by utilizing swarm intelligence to optimize two excitation trajectories for the first three links at the arm and the last three links at the wrist of the robot manipulator. Numerical techniques were used to reduce the observation matrix to a minimum linear combination of parameters, thereby maximizing the identifiable parameters, and the Linear Least Square method was used for parameter identification. An improved particle swarm optimization algorithm with mutation and archived elite learning was proposed for solving the dynamic optimization problem of the industrial robotic manipulator. The basic parameters of the algorithm have been optimized for robotic manipulator analysis. The proposed algorithm is computationally economical while completely dominating other Evolutionary algorithms in solving robot optimization problems. The algorithm was further used to analyze 36 benchmark functions and produced competitive results.

1. Introduction

This paper is an extension of the work originally presented in the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications [1], where a novel mutating Particle swarm optimization (MuPSO) based solution was proposed for analyzing the inverse kinematic problem of multi-degree-of-freedom robot manipulator. It was observed that the basic parameters of the Particle swarm optimization (PSO) algorithm are incapable of analyzing robot optimization problems, and it was established that the PSO parameters need to be modified for robot analysis. Further investigations were published in [2] where various parameters were used to analyse the inverse kinematic problem of four popular robot configurations and a new range of parameters for solving robot optimization problems was established, thereby an optimized mutating PSO algorithm for robot analysis was developed. If the algorithm runs into stagnation while solving a robot optimization problem, the mutation function is used to generate a completely new swarm and all the information regarding the previous swarm is lost. It is keen to note that if a particle from the previous swarm is retained, that particle becomes the global best particle for the new swarm, making the entire swarm converge again at the previous solution. It was observed that it would be beneficial to save previously non-dominated solutions in an archive for future reference. This work therefore seeks to further modify the mutating PSO algorithm by implementing it with an archived elite learning system, and using the new eMuPSO algorithm for dynamic parameter estimation of a 6 Degree-of-freedom (DOF) industrial manipulator and finally compare the improved algorithm with other state of the art algorithms in analyzing 36 benchmark functions. The rest of the paper is structured as follows; section 2 presents previous literature, section 3 introduces the new Mutating PSO algorithm, section 4 presents the results of parameter estimation of robot manipulators, and compares the new PSO algorithm with state of the art swarm-based algorithms on thirty-six benchmark functions and finally section 5 concludes the findings.

2. Literature

In industrial robotics, a dynamic model determines the actuator force required to achieve the desired joint configuration, coupling

*Corresponding Author: Zhanqun Shi, Hebei University of Technology, Tianjin-China, Tel.: +86-2260438217, z_shi@hebut.edu.cn

effects and non-linearity of the system introduces dynamic uncertainties. These errors degrade the robot's performance. Therefore, re-calibrating robot parameters are necessary for efficient robot analysis and control, that can only be achieved experimentally. In dynamic model estimation, obtaining experimental data requires that firstly, a linearized model of the robot manipulator is developed, then a trajectory that excites the robot's rigid body dynamics is developed, afterwards input torque and output joint configuration data is experimentally obtained and finally the dynamic parameters are estimated. The accuracy of the estimation method is largely dependent on the chosen trajectory and measuring accuracy. The actuator torque can be measured from input electric currents while optic encoders are incorporated for measuring joint output conditions. The recent improvement in technology makes available encoders and reduction gears that are capable of preventing cumulative errors and minimizing gearbox backlash. Therefore, carefully selecting an exciting trajectory is paramount in dynamic model estimation. Reference [3] observed that the excitation trajectory can be optimized by minimizing the condition number of the observation matrix. The observation matrix is a function of only the joint variables; it enables the solution of nonlinear systems to be presented as linear. The condition number of the observation matrix is a measure of noise immunity, a value closer to unity results in a better signal to noise ratio (SNR). The presence of unidentifiable parameters results in a high condition number of the observation matrix. References [4, 5] proposed techniques for categorizing dynamic parameters into identifiable parameters, unidentifiable parameters, and identifiable parameters in linear combinations. These works suggested that eliminating unidentifiable parameters from the observation matrix enhances estimation accuracy. In [6, 7] a set of Minimal Linear Combinations (MLC) of identifiable dynamic parameters from the total amount of parameters to be identified was generated. Reference [8] proposed formulating the periodic Finite Fourier series (FFS) as an excitation trajectory. FFS inhibits noise interference, if the identification experiment is repeated several times, averaging the measured data in the time domain improves SNR. Reference [9] used a modified Fourier series (MFS), where the FFS is implemented on a circular trajectory in the Cartesian space such that the initial and final conditions of the joints are the same, permitting continuity in the optimization experiment without necessarily stopping and restarting the manipulator. The particle swarm optimization algorithm (PSO) is a stochastic metaheuristic swarm-based evolutionary algorithm (EA), each particle in the swarm is attracted towards the global best particle's position and its personal best position denoted as (Gbest and Pbest). The velocity and position of canonical PSO are updated according to (1) and (2).

$$X_i^d(t+1) = X_i^d(t) + V_i^d(t+1)\,, \tag{1}$$

$$V_i^d(t+1) = wV_i^{\dim}(t) + c_1 r_1\left(Pbest_i^{\dim}(t) - X_i^{\dim}(t)\right)\cdots + c_2 r_2\left(Gbest^{\dim}(t) - X_i^{\dim}(t)\right)\,, \tag{2}$$

where i = 1, 2, 3…$nPop$ denotes for the index of each particle, $nPop$ is the swarm size, and dim = 1,2…dof denotes for the dimensionality of the solution space, dof is the degrees of freedom of the robot. V_i^{dim} and X_i^{dim} stand for the position and velocity vectors of the ith particle, respectively. $V_i^{dim} = [v_i^1, v_i^2, \ldots, v_i^{dof}]$, $X_i^{dim} = [x_i^1, x_i^2, \ldots, x_i^{dof}]$. w is the inertia weight. c_1 and c_2 are cognitive and social learning coefficients. r_1 and r_2 are two uniformly distributed random numbers within the range of [0,1]. Over the years, PSO has undergone various modifications either to improve its internal dynamics (enhance convergence and exploitation characteristics) or to satisfy the requirement of a specific real-world optimization problem or both.

2.1. PSO with Improved Dynamics

An index based ring topology PSO was proposed by [10] for solving multimodal multi-objective optimization problems having more than one Pareto optimal solution corresponding to the same objective function. Reference [11] studied the impact of communication topology in PSO, the star topology (Gbest) where each particle communicates with every other particle, was found better in analyzing unimodal separable problems while the ring topology (LBest) where each particle communicates with only its two neighbor particles, is preferred for multimodal, non-separable and composite functions. The ring topology is more capable of jumping between optimum basins and it was also found to never converge at a local optimum, but with a considerably higher computational cost. Memetic algorithms (MA) are hybrid EAs that are a fusion of global and local search techniques. EAs are usually employed for global search while unconstrained optimizers are employed for local search. Adaptive MAs have been successfully combined with other EAs but not PSO, [12] proposed a co-evolutionary memetic PSO for solving multi-objective optimization problems. Feature selection (FS) is a combinatorial problem for large dimension data processing which requires large memory and high computational cost, [13] proposed a variable-length PSO for FS. The Proposed algorithm is flexible and simplifies large data analysis, and can jump out of local optima while narrowing the search space. The cooperative search strategy which prevents particle from being trapped in local optima was merged with PSO by [14] for unconstrained optimization, where cooperative multiple swarms were used to improve the convergence and efficiency of the canonical PSO.

2.2. Problem Oriented PSO

Medicine and medical healthcare are amongst the top beneficiaries of artificial intelligent swarm algorithms. Health care services are shifting from inpatients to outpatients. Reference [15] showed that accident and emergency centers in hospitals are currently being optimized to ensure 98% of patients get the required attention within 24 hours. The increasing demand for primary health care outlets has led to a requirement to optimize the operational efficiency of primary healthcare centers. Grouping the patients into an appropriate category optimizes time and manpower whereas patient no shows after grouping reduces operational efficiency of clinics. In [16], using PSO to improve the accuracy of traditional clustering techniques was proposed for analyzing real-time patient attendance data and group them into clusters, while [15] compared PSO with opposition based learning and self-adaptive cohort intelligence for identifying significant features capable of predicting no shows. Gene selection from microarray data for cancer diagnosis and testing also involve large data sets and are computationally expensive, support vector machines algorithm is a fast and efficient classification model

which requires its parameters tuned to satisfy specific problems. Reference [17] compared the performance of PSO-SVM and Memetic SVM for classification and feature selection of cancer cells. Alzheimer's disease is a primary stage of dementia which is a type of memory disorder that affects the brain. The structural changes of the brain's internal regions are the most commonly measured in diagnosing AD. Reference [18] compared the accuracy of PSO and other metaheuristic algorithms in segmenting the brain sub-regions. Air pollution is a major contributor to human mortality and a potential danger to the environment and ecological system [19]. Thermal power plants such as coal, petroleum, and natural gas are the main sources of air pollution, mercury contamination has been identified to be the most acute air pollutant produced by power plants [20]. In [21], a novel combination of modified genetic algorithm and the improved particle swarm optimization was proposed for minimizing the fuel cost of generating plants and emission simultaneously, by managing and controlling the integration of renewable energy and thermal power production, where an optimal generation plan for maximizing system efficiency can be achieved. Reference [22] proposed using an adaptive neuro-fuzzy inference system with PSO (ANFIS-PSO) for predicting mercury emission in power plants. Reference [23] sought to minimize the construction cost of reinforced concrete retaining walls (RCRW). These structures include bridges, railways, dams, etc., capable of withstanding the pressure resulting from the difference in level by an embankment, excavation, or natural processes. PSO was used to determine the optimum solution between popular techniques. In rural areas where the utility grid is unavailable, renewable energy is an attractive alternative for water pumping applications. The power-voltage curve of a photovoltaic cell was found to have multiple maximum power points under partial shading conditions, making the solar tracking mechanism unstable. In [24], the performance of PSO was compared with salp swarm algorithm for maximum power point tracking of solar panels. Electric motors are most popular for electrical-mechanical power conversion and employed in several industrial applications including robotics, because of their simplicity, durability, and low maintenance cost. They usually require a controller for high performance and efficiency. The electrical parameters of the electrical motor are very essential to design, performance assessment, and feasibility of the control technique, any difference in the actual motor parameters adversely affects the system performance. Reference [25] proposed using PSO for estimating the electrical parameters of induction motors.

3. Mutating PSO with Elite Archive Learning

The proposed PSO is equipped with a modified set of parameters and governing equations. The mutation function is used to generate a new swarm around the vicinity of the most promising particles when the algorithm runs into stagnation. The operation of the proposed mutating PSO with elite learning (eMuPSO) is divided into two, during the early stages, the algorithm searches the solution space for the minimum solution, then when it stagnates, elite solutions are being recalled from the archive and merged with the best solutions of the current run to generate a new population. The new population would be created either through Random or Gaussian mutation depending on the application. It was found that a mix of both was suitable for kinematic analysis while the Gaussian mutation was suitable for dynamic analysis.

3.1. Modified Parameters

It was established in [1] and [2] that the basic parameters of canonical PSO are not capable of ensuring converging solutions for robot optimization problems especially when the DOF is greater than three. The basic parameters of the mutating PSO were modified to satisfy the requirement for solving robot optimization problems. This was achieved by testing the performance of various PSO parameters on four popular robot configurations, and a relationship between the inertia weight and the social learning coefficient was derived. More details on modifying PSO parameters can be found in [2]. A non-linearly decreasing inertia weight is implemented with values between (2.1 – 0.6), the cognitive learning coefficient is constantly at 2.24, while the social learning coefficient is non-linearly increasing between (1.8 – 3.9). The equations for updating w and c2 are given in (3) and (4).

$$w_{iter} = w_{initial} * n^{iter} , \quad (3)$$

$$c_{2iter} = c_{2initial} / m^{iter} , \quad (4)$$

where *iter* is the iteration number, and *n* and *m* are coefficients. The coefficients *n* and *m* can be determined by setting *iter* to the maximum value. Robot configurations have complex dynamics described by highly coupled and non-linear, second-order differential equations, there are possibilities of numerous strong local minimizers in the solution space.

3.2. Mutation

In [26] it was shown that structural bias is a characteristic of swarm algorithms that tends to confine the algorithm to a region of the total search area. Mutation is a valid technique for overcoming such constraints. It was established in [2] that even the most intelligent swarm-based solutions are capable of running into stagnation, random mutation was used to push the algorithm out of stagnation and enhance convergence. The Gaussian mutation was used in [27] to improve swarm performance. Reference [28] also used a Gaussian mutation throw point strategy to redistribute the swarm at sparse positions. In the proposed algorithm, when the swarm stagnates at a local minimum solution, a Gaussian mutation is used to push the algorithm out of stagnation and redistribute the swarm to enhance unbiased coverage of the search space. The radius of the swarm is monitored according to (5). When the swarm radius becomes too small, it signifies that the swarm has lost diversity and might be stagnating, then the mutation function is initiated. The condition in (6) states that if there is no change in the swarm radius or when the change in the swarm radius is negligible, then the counter *g* increases sequentially otherwise the counter is reset to zero. A second condition in (7) terminates the algorithm when the minimum solution is achieved or the swarm is mutated when *g* is greater than a threshold value *G*.

Assuming there are *k* elite particles, the entire swarm population would be divided into *k+1* parts and each *k* part would be replaced by off-springs of the kth elite particle, while the *k+1* part consists of particles generated from a random distribution, thereby maintaining diversity in the swarm. such that if x^{dim} is the

dimth dimension of the swarm, then all the elements in the *dim*th dimension shall mutate according to (8).

$$rad_{iter} = GlobalBestCost_{iter} - mean(BestCost)_{iter} , \quad (5)$$

$$g = \begin{cases} g+1, & if, (rad_{iter-1} - rad_{iter}) = 0 \\ 0 , & Else \end{cases} , \quad (6)$$

$$f(MuPSO) = \begin{cases} end & if \quad 1e^{-8} \\ mutate & if \quad g \geq G \end{cases} , \quad (7)$$

$$^{\dim} x_{mutate} = {}^{\dim} x_{elite} + gaussian(\sigma) , \quad (8)$$

where *rad* is the radius of the swarm, *iter* refers to the current iteration, $^{dim}x_{mutate}$ and x_{elite} are the mutated off-spring and the elite solution in the *dim*th dimension of the swarm, and σ is a variable between [0-1] multiplied by size of the search space.

3.3. Elite Archive

When the algorithm is searching for the best solution, the experiences of the best particle is transferred to the next iteration through the global best information, but at the beginning of subsequent generation or after mutation, all the information regarding the best particle from the previous generation is lost. This inspired the introduction of an archive where the experiences of the elite non-dominated solutions are saved and can be called in the future to enhance the search. Reference [29] applied a dynamic archive maintenance strategy to improve the diversity of solutions in multi-objective particle swarm optimization. In [30] an external archive was employed to preserve the non-dominated solutions visited by the particles to enable evolutionary search strategies to exchange useful information among them, and [31] used a grid-based approach for the archiving process and ε-dominance method to update the archive, which helps the algorithm to increase the diversity of solutions. Reference [32] used the cooperative archive to exploit the valuable information of the current swarm and archive. Information about the elite particles from dynamic sub swarms was used in [33] to improve the following sub-swarm, while [32] introduced a new velocity updating technique that explores the external archive of non-dominated solutions in the current swarm. In this article, an elite archive learning is used to refine the solution in the final stages of the algorithm. Elites from previous searches are saved as global elites, while elites of the current search are saved as local elites. When the algorithm stagnates, a combination of both elite vectors is used to generate a new population, and the global elites are updated with new solutions from the current local solutions

3.4. Robot Dynamics

Industrial robot manipulator

A 6 DOF industrial robot manipulator was analyzed, the structure of the robot manipulator is shown in Figure 1 and its D-H parameters are given in Table 1. A homogeneous transformation matrix describes the state (orientation and position coordinates) of a robot link with respect to the previous link in the Cartesian coordinate space and can be used to describe the state of the end-effector (tool) of the manipulator with respect to the base (global coordinate) frame. If all the joint parameters are known (D-H Parameters), the homogeneous matrix of each successive pair of frames for the forward and inverse directions can be obtained from (9). The robot manipulator's end-effector coordinates would be a product of post multiplication as shown in (10).

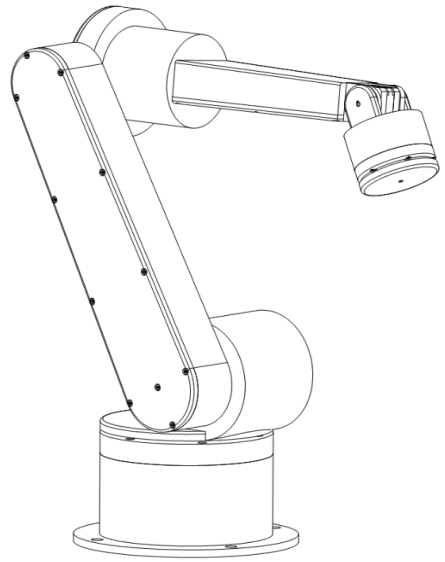


Figure 1: A 6 degree of freedom industrial robot manipulator

Table 1: D-H parameters of 6 degree of freedom industrial robot manipulator

Joint	Link	Off-set	Joint	Off-set
1	300	320	(-165 : +165)	-pi/2
2	700	0	(-110 : +110)	0
3	0	0	(-110 : +70)	-pi/2
4	0	200+497.5	(-160 : +160)	pi/2
5	0	0	(-120 : +120)	-pi/2
6	0	97.5+30	(-242 : +242)	0

$$T_{\dim}^{\dim-1} = \begin{bmatrix} c\theta_{\dim} & s\theta_{\dim}c\alpha_{\dim-1} & s\theta_{\dim}s\alpha_{\dim-1} & a_{\dim}c\theta_{\dim} \\ -s\theta_{\dim} & c\theta_{\dim}c\alpha_{\dim-1} & c\theta_{\dim}s\alpha_{\dim-1} & a_{\dim}s\theta_{\dim} \\ 0 & -s\alpha_{\dim-1} & c\alpha_{\dim-1} & d_{\dim} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (9)$$

$$T_4^0 = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t_{21} & t_{22} & t_{23} & t_{24} \\ t_{31} & t_{32} & t_{33} & t_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} = T_1^0 T_2^1 T_3^2 T_4^3 \quad (10)$$

where $c\theta$ and $s\theta$ refers to $cos\theta$ and $sin\theta$ respectively.

Dynamic parameter estimation

The Linear-least-square (LLS) method together with an optimized excitation trajectory was used for dynamic parameter estimation of the robot manipulator. LLS requires that the model equations are linear with respect to dynamic parameters and it is

sensitive to noise in the measured data. [34] observed that the actuator forces of an industrial manipulator are linear functions of dynamic parameters, [4] inferred that reformulating the Newton-Euler dynamic model such that the link inertia tensors are expressed about the link coordinate frames instead of the center of the links' mass, would result in a linearized Newton-Euler formulation. Equations (11-14) presents a summarized Newton-Euler formulation in the form of observation matrix. For detailed information on linearizing the Newton-Euler equation, see [4] and [34].

$$\Phi = \left(\Omega^T \Omega\right)^{-1} \Omega^T * \Gamma \ , \tag{11}$$

$$\Gamma = \left(\left(\Omega^T \Omega\right)^{-1} \Omega^T\right)^{inv} \Phi \ , \tag{12}$$

$$\Gamma = \begin{vmatrix} \tau_{\dim} \\ \vdots \\ \tau_{dof} \end{vmatrix} = \begin{vmatrix} \omega_{\dim}\left(q,\dot{q},\ddot{q}\right) \\ \vdots \\ \omega_{dof}\left(q,\dot{q},\ddot{q}\right) \end{vmatrix} * \begin{vmatrix} \phi_{\dim} \\ \vdots \\ \phi_{dof} \end{vmatrix} , \tag{13}$$

$$\phi_{\dim} = \begin{bmatrix} mm & mc_x & mc_y & mc_z & I_{XX} & I_{YY} & I_{ZZ} & I_{XY} & I_{XZ} & I_{YZ} \end{bmatrix}^T , \tag{14}$$

$$fitness = cond(\Omega) + \frac{1}{\varphi_{\min}(y)} + \varphi_{\max} y \ , \tag{15}$$

where Γ is the compound torque for the manipulator comprising of individual joint torques τ, the compound observation matrix Ω comprises of individual joint vectors $\omega(q, \dot{q}, \ddot{q})$, and is a function of the joint position, velocity, and acceleration only. ϕ is a vector of unknown robot parameters for each link that require to be identified, it includes link mass *mm*, first moments $[mc_x\ mc_y\ mc_z]^T$ along the x-y-z axis, and six inertia tensor I_{ij}, these compose the compound vector of unknown parameters Φ. The intelligent swarm-based algorithm would be required to minimize the condition number of the observation matrix while maximizing the smallest singular value and minimizing the largest singular value. Therefore, the fitness function would be as described in (15).

Exciting trajectory

A modified Fourier series (MFS) formulation was used to define the excitation trajectory with a cubic polynomial. The equation of motion for the trajectory of the first three joints are given in (16-18), a circular trajectory path was implemented so as to allow continuous trajectory tracking of the robot during experimentation. The trajectory is to be optimized by applying the eMuPSO algorithm to minimizing the condition number of the observation matrix. The trajectory is implemented in Cartesian space, and joint positions are determined from inverse kinematic solutions. The wrist of the manipulator has a z-x-z orientation, therefore the excitation trajectory along the *x* and *y* axis can be derived independently from (16) and (17) while that of the *z* axis is estimated from (19), where *d6* is the offset length of the sixth link of manipulator (see DH parameters) and t_{14} is the position coordinate of transformation matrix T_4^0 from the *Base* to *Link-04*.

$$Q_x(it) = d0_j^{it} + \sum_{j=1}^{N} \begin{pmatrix} \cos(P_{it}) * a0_j^{it} + \dots \\ \frac{b0_j^{it}}{freq * j} \sin(freq * j * tym) - \dots \\ \frac{c0_j^{it}}{freq * j} \cos(freq * j * tym) \end{pmatrix} , \tag{16}$$

$$Q_y(it) = e0_j^{it} + \sum_{j=1}^{N} \begin{pmatrix} \sin(P_{it}) * a0_j^{it} + \dots \\ \frac{b0_j^{it}}{freq * j} \sin(freq * j * tym) - \dots \\ \frac{c0_j^{it}}{freq * j} \cos(freq * j * tym) \end{pmatrix} , \tag{17}$$

$$Q_{z,arm}(it) = f0_j^{it} + \sum_{j=1}^{N} \begin{pmatrix} \frac{b0_j^{it}}{freq * j} \sin(freq * j * tym) - \dots \\ \frac{c0_j^{it}}{freq * j} \cos(freq * j * tym) \end{pmatrix} , \tag{18}$$

$$Q_{z,wrist}(it) = f0_j^{it} + \sqrt{d6^2 - \left(t_{14} - Q_x(it)\right)^2} \ , \tag{19}$$

$$Vector(it) = \begin{bmatrix} vec_{11} \\ vec_{21} \\ vec_{31} \end{bmatrix} = \begin{bmatrix} t_{14} \\ t_{24} \\ t_{34} \end{bmatrix} - \begin{bmatrix} Q_x(it) \\ Q_y(it) \\ Q_z(it) \end{bmatrix} , \tag{20}$$

$$\theta_4 = a\tan\left(vec_{31}, vec_{21}\right) , \tag{21}$$

$$\theta_5 = a\tan\left(norm\left(\left[0, vec_{21}, vec_{31}\right]^T\right), vec_{21}\right) , \tag{22}$$

$$\theta_6 = a\tan\left(\left(vec_{31}(it) - vec_{31}(it-1)\right),\ \left(vec_{21}(it) - vec_{21}(it-1)\right)\right) , \tag{23}$$

$$\left.\begin{array}{l} \min_{a_j, b_j, c_j} \ cond(\Omega) \quad such \quad that; \\ \quad \{s(Q(it))\} \subset S \\ if \quad Q_x = [q_{x1}, q_{x2}, \cdots q_{xk}] \\ \quad 0 \le Q_x(it) \le 2\pi \\ \quad 0 \le \dot{Q}_x(it) \le 0 \\ \quad 0 \le \ddot{Q}_x(it) \le 0 \end{array}\right\} , \tag{24}$$

where *Q* describes the trajectory in Cartesian space such that $[Q_x, Q_y, Q_z]^T$ are the *x-y-z* coordinates of *Q*, *P* is the polynomial, the variables *a0* is the radius, *b0* and *c0* are the coefficients of the MFS while the variables *d0*, *e0*, and *f0* are the x-y-z position coordinates of the excited trajectory. $j = 1,2 \dots N$ denotes the harmonics of the MFS, while the index of time segment/period is denoted as *it*; *freq* is the fundamental frequency and *tym* is the period. The joints 4, 5, and 6 coincident at the joint 5, described by the transformation matrix T_4^0 in (10). The inverse kinematic solution of these joints can be estimated given the orientation matrix, but in this case only the position vector is known (i.e. $[Qx, Qy, Qz]^T$). Therefore, the rotation angles of the joints at the wrist would be estimated from

(20-23), and the orientation matrix is evaluated by substituting θ in (9) then inverse kinematics can be subsequently evaluated. To achieve continuity along the tool-path, the initial and final conditions of the trajectory must be the same as described in the conditions in (24). The initial and final values for the circular trajectory *P* are set as *0* and *2π* radians, 3 harmonics for the coefficients of MFS were used (N=3), and the fundamental frequency is *0.6283* with a period of *10* seconds.

4. Results

The eMuPSO algorithm was first implemented together with the LLS for optimizing the observation matrix and dynamic parameter estimation of all the six links of the 6DOF robot manipulator. The ideal parameters of the manipulator are given in (25-32). Then the proposed algorithm was finally implemented in evaluating thirty-six benchmark functions including twenty-four variable-dimension benchmark functions and twelve constant-dimension benchmark functions. The variable dimension benchmark functions consist twelve functions each, of unimodal and multi-modal roles (a full description of all benchmark function used are given in the appendix).

The results were compared with the Standard PSO (PSO) [35] and four other swarm-based EAs including the Whale Optimization Algorithm (WOA) [36], Grey Wolf Optimization (GWO) [37], Grass-hopper Optimization Algorithm (GOA) [38], and the Differential Evolution (DE) algorithms [39].

4.1. Dynamic Parameter Estimation

Generating initial elite particles

Parameter estimation was in two stages; initial elite particles were first generated which were then used to generate the swarm for parameter estimation. This reduces the probability for the algorithm to run into stagnation. A swarm of ten particles was first generated for each variable *(a0, b0, ..., f0)*. From inverse kinematics, the joint conditions (position, velocity and acceleration) corresponding to the generated trajectory of each set of particle is estimated and compared with the stated constraints in (24). Trajectories that do not meet the constraints are flagged and a penalty was introduced.

The fitness function for this swarm is a measure of the degree of compliance of the generated trajectory with the stated robot constraints. If a trajectory is found to completely satisfy these constraints, the variables corresponding to such trajectory is elevated to elite status and a new particle is generated to replaces that particle, else, the entire swarm's velocity and position is updated in accordance with PSO. Elite particles should be generated within 5-10 iteration if the minimum and maximum limits of the variables are carefully selected. This algorithm is terminated when 5 elite particles have been generated. The variable limits are set at ±10. While the variables *a* and *b* span between negative and positive limits, others must always be positive.

$$\mathrm{Mm} = \begin{bmatrix} \mathrm{mm1} \\ \mathrm{mm2} \\ \mathrm{mm3} \\ \mathrm{mm4} \\ \mathrm{mm5} \\ \mathrm{mm6} \end{bmatrix} = \begin{bmatrix} 5.029912009 \\ 4.530794148 \\ 3.056783430 \\ 1.815211338 \\ 0.990530741 \\ 0.186899621 \end{bmatrix}, \tag{25}$$

$$\mathrm{Mc} = \begin{bmatrix} \mathrm{mcx1} & \mathrm{mcy1} & \mathrm{mcz1} \\ \mathrm{mcx2} & \mathrm{mcy2} & \mathrm{mcz2} \\ \mathrm{mcx3} & \mathrm{mcy3} & \mathrm{mcz3} \\ \mathrm{mcx4} & \mathrm{mcy4} & \mathrm{mcz4} \\ \mathrm{mcx5} & \mathrm{mcy5} & \mathrm{mcz5} \\ \mathrm{mcx6} & \mathrm{mcy6} & \mathrm{mcz6} \end{bmatrix} \cdots = \begin{bmatrix} 4.3956 & -15.8576 & 10.1804 \\ -44.1094 & 93.9465 & 183.7506 \\ 0.37299 & 0.001086 & 100.6463 \\ -0.01376 & 12.03674 & 18.1007 \\ -0.00300 & -0.00352 & -21.2584 \\ 0.004296 & 0.000028941 & 4.9163 \end{bmatrix}, \tag{26}$$

$$\mathrm{I}_1 = \begin{bmatrix} \mathrm{Ixx1} & \mathit{Ixy1} & \mathit{Ixz1} \\ \mathit{Ixy1} & \mathit{Iyy1} & \mathit{Iyz1} \\ \mathit{Ixz1} & \mathit{Iyz1} & \mathit{Izz1} \end{bmatrix} \cdots = \begin{bmatrix} 20975.4835 & 0 & 0 \\ 0 & 13651.6206 & 0 \\ 0 & 0 & 19154.2091 \end{bmatrix}, \tag{27}$$

$$\mathrm{I}_2 = \begin{bmatrix} \mathrm{Ixx2} & \mathit{Ixy2} & \mathit{Ixz2} \\ \mathit{Ixy2} & \mathit{Iyy2} & \mathit{Iyz2} \\ \mathit{Ixz2} & \mathit{Iyz2} & \mathit{Izz2} \end{bmatrix} \cdots = \begin{bmatrix} 103901.5055 & 0 & 0 \\ 0 & 79173.3684 & 0 \\ 0 & 0 & 30809.7396 \end{bmatrix}, \tag{28}$$

$$\mathrm{I}_3 = \begin{bmatrix} \mathrm{Ixx3} & \mathit{Ixy3} & \mathit{Ixz3} \\ \mathit{Ixy3} & \mathit{Iyy3} & \mathit{Iyz3} \\ \mathit{Ixz3} & \mathit{Iyz3} & \mathit{Izz3} \end{bmatrix} \cdots = \begin{bmatrix} 9855.9631 & 0 & 0 \\ 0 & 9759.3916 & 0 \\ 0 & 0 & 7360.3570 \end{bmatrix}, \tag{29}$$

$$\mathrm{I}_4 = \begin{bmatrix} \mathrm{Ixx4} & \mathit{Ixy4} & \mathit{Ixz4} \\ \mathit{Ixy4} & \mathit{Iyy4} & \mathit{Iyz4} \\ \mathit{Ixz4} & \mathit{Iyz4} & \mathit{Izz4} \end{bmatrix} \cdots = \begin{bmatrix} 16365.8046 & 0 & 0 \\ 0 & 1616.0497 & 0 \\ 0 & 0 & 17081.9923 \end{bmatrix}, \tag{30}$$

$$I_5 = \begin{bmatrix} Ixx5 & Ixy5 & Ixz5 \\ Ixy5 & Iyy5 & Iyz5 \\ Ixz5 & Iyz5 & Izz5 \end{bmatrix} \cdots = \begin{bmatrix} 1058.2078 & 0 & 0 \\ 0 & 996.4668 & 0 \\ 0 & 0 & 1033.5495 \end{bmatrix}, \quad (31)$$

$$I_6 = \begin{bmatrix} Ixx6 & Ixy6 & Ixz6 \\ Ixy6 & Iyy6 & Iyz6 \\ Ixz6 & Iyz6 & Izz6 \end{bmatrix} \cdots = \begin{bmatrix} 103.8927 & 0 & 0 \\ 0 & 103.8903 & 0 \\ 0 & 0 & 204.5344 \end{bmatrix}, \quad (32)$$

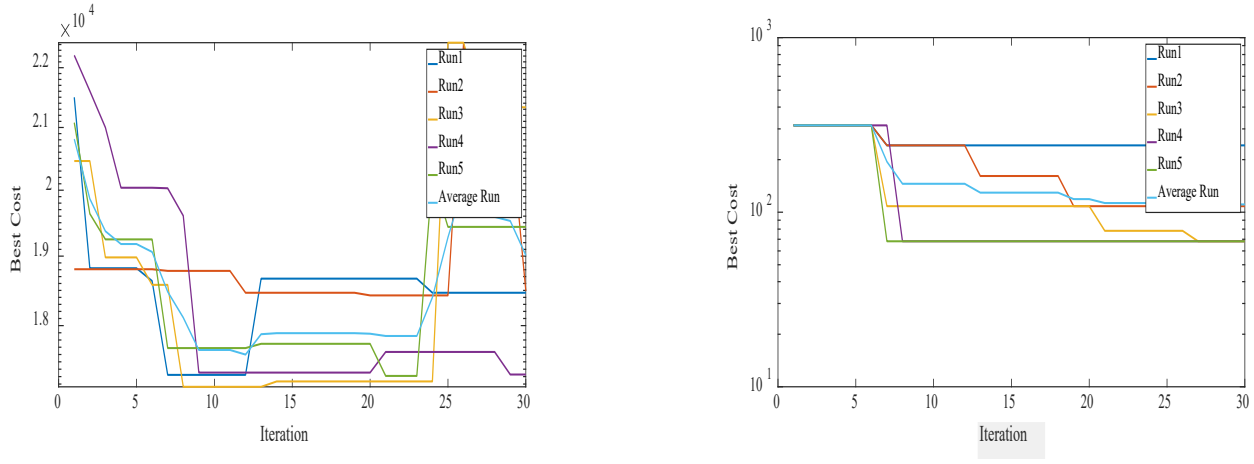
Optimizing the excitation trajectory

A swarm of 20 particles is generated from the previously generated *initial elite particles* through Gaussian mutation. Similar to previous procedures, the boundary conditions are set, joint conditions corresponding to each generated trajectory is estimated, again penalties are introduced to make trajectories that do not satisfy the constraints in (24) less likely to be selected as exemplars. The fitness function for this swarm is evaluated according to (15). The swarm of the algorithm is mutated at 10 iterations after stagnation (i.e. G=10).

Parameter estimation

The dynamic parameters of a 6DOF manipulator was estimated in two stages, the last three links corresponding to the wrist were first estimated followed by the first three links corresponding to the arm of the manipulator. A numeric technique for reducing the observation matrix to a minimum linear combination of identifiable parameters was implemented as described in [40] where three empty matrices were created corresponding to *identifiable* parameters, *unidentifiable* parameters, and *identifiable parameters in linear combinations*. Understanding that the rank of the observation matrix is equal to the number of identifiable parameters, the norm of each column of the observation matrix is determined independently. The columns with zero norms represent coefficients of *unidentifiable* parameters and are moved to the matrix of *unidentifiable* parameters, else it is added to the matrix of *identifiable* parameters. Note that the initial rank of the empty matrix of *identifiable* parameters is zero, if the newly added column does not increase the rank of the matrix of *identifiable* parameters, it is moved to the matrix of *identifiable parameters in linear combination*. The eMuPSO was used to select the trajectory that best optimizes the observation matrix. Removing *unidentifiable* parameters from the observation matrix increases estimation accuracy. In the first estimation stage (wrist), out of the 30 unknown parameters at the wrist, 15 were independently identifiable, 7 were identifiable in linear combination, while 8 were unidentifiable. Out of the 7 parameters that were identifiable in linear combination, it was observed that for most elite particles, 5 were identifiable with lower errors. The 15 independently identifiable parameters include *Iyy4*, *Ixz4*, *mx5*, *Ixx5*, *Iyy5*, *Izz5*, *Ixy5*, *Ixz5*, *Iyz5*, *mcx6*, *mcy6*, *Iyy6*, *Izz6*, *Ixz6*, and *Iyz6*. The 7 parameters identifiable in linear combination include *mcy4*, *mcy5*, *mcz5*, *mm6*, *mcz6*, *Ixx6*, and *Iyz6*. It was found that *Ixx6* and *Iyz6* were identified with very large errors for most elite particles. In the second estimation stage (arm), 20 parameters were independently identifiable, 2 were identifiable in linear combinations while 8 were unidentifiable. All parameters were identified with little errors. The 22 identified parameters include *mcy1*, *Iyy1*, *mm2*, *mcx2*, *mcy2*, *mcz2*, *Iyy2*, *Izz2*, *Ixz2*, *Iyz2*, *mm3*, *mcx3*, *mcy3*, *mcz3*, *Ixx3*, *Iyy3*, *Izz3*, *Ixy3*, *Ixz3*, *Iyz3*, *Ixz1*, and *Ixx2*.

Figure 2.0 (a) and (b) show the convergence plot at the wrist and at the arm of the manipulator respectively. The eMuPSO was populated with twenty particles in thirty iterations and five runs. The Figure 2.0 shows all five runs, the sixth being the average of all five runs. It would be observed that in the third run of Figure 2.0 (a), the algorithm converges to the global minimum solution in eight iterations, then it is mutated in the 14th iteration where it stagnates until it is mutated again in the 24th iteration. The best solutions are stored in the archive before mutation. The Figure 2.0 (b) show less agitation of the algorithm signifying less mutation and simpler analysis.



(a) (b)

Figure 2: Convergence plot of the observation matrices (a) at the wrist, and (b) at the arm.

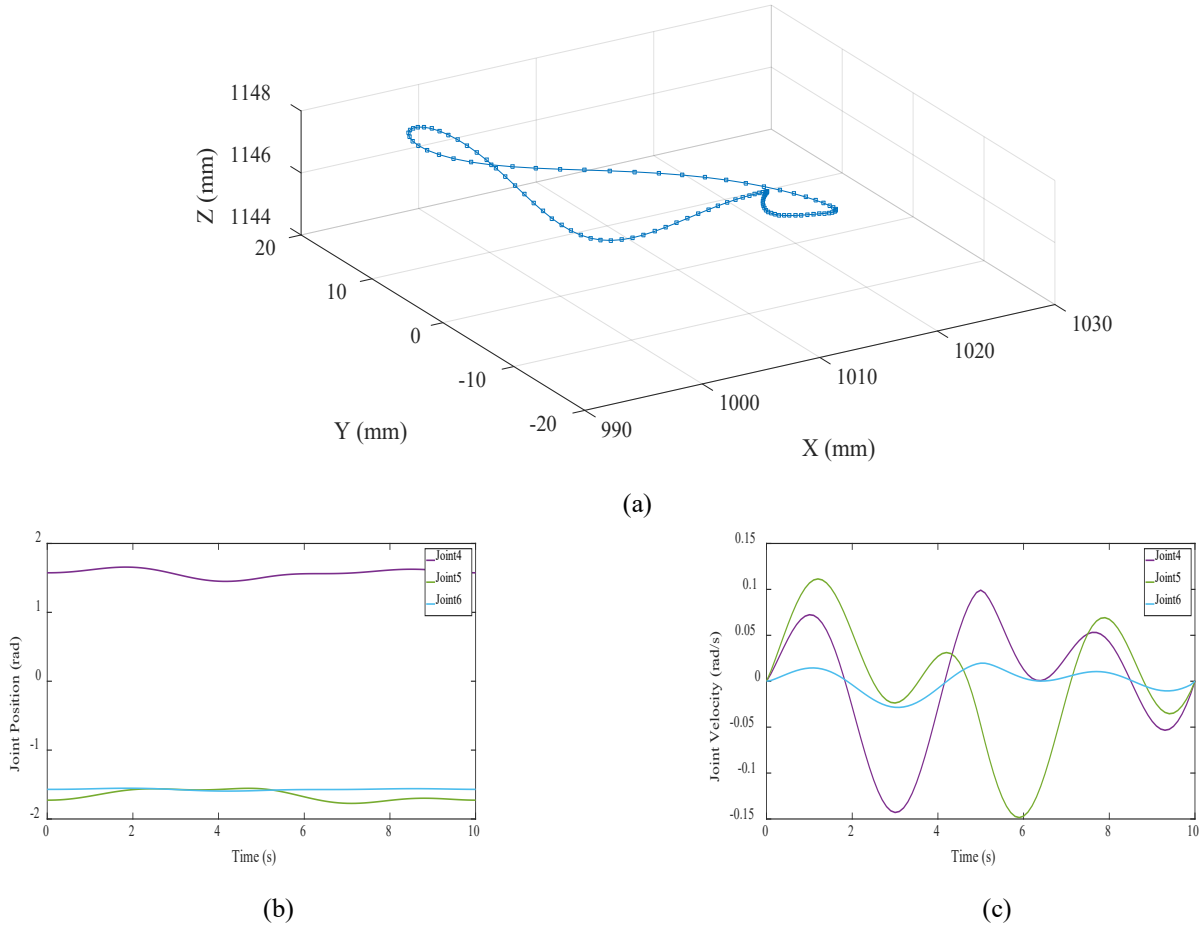
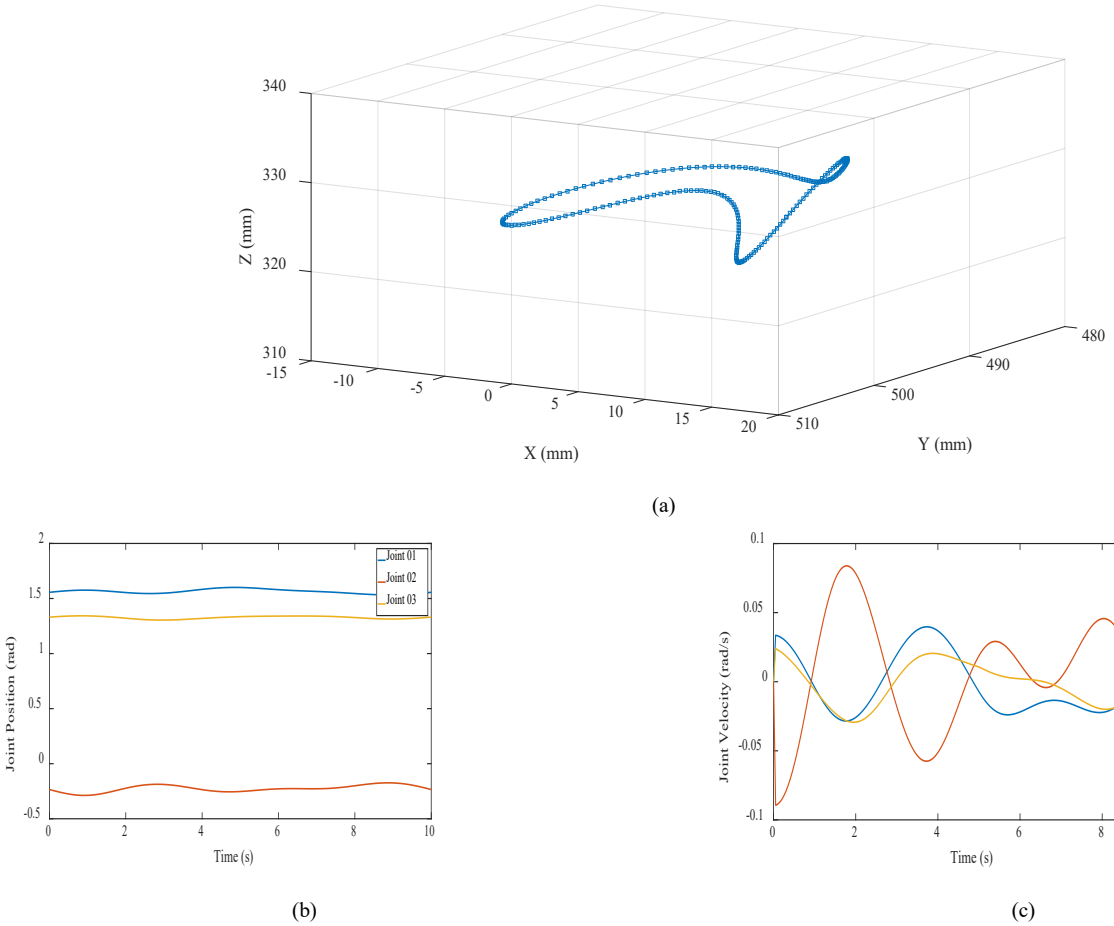


Figure 3: Excitation trajectory for the links at the wrist of manipulator. (a) Trajectory in Cartesian space (b) Joint displacement (c) Joint velocities

Table 2: Dynamic parameter estimation for the last three joints of a 6 DOF industrial robot

Parameter	Elite Particles (wrist)					
	1	2	3	4	5	6
a0(1)	-3.8711	-8.7566	-6.1829	-4.824	-6.7566	-4.4276
a0(2)	9.8782	1.1156	5.7223	-2.4838	3.1156	4.0066
a0(3)	0.42207	4.5867	-4.0733	-3.0035	3.748	-5.7368
b0(1)	2.1907	2.8328	4.9482	2.6421	4.7319	6.1336
b0(2)	-3.5264	-7.3354	-2.4158	-1.0079	-5.3354	5.1722
b0(3)	-2.293	-7.4121	0.31497	-3.5628	-5.4121	0.91808
c0	6.5802	5.4755	6.423	6.8105	5.0595	2.9898
d0	6.5802	3.1097	6.423	0.85837	5.0595	2.9898
e0	0	0	0	0	0	0
f0	t_{34}	t_{34}	t_{34}	t_{34}	t_{34}	t_{34}
fitness	418.354	401.380	357.993	351.894	349.889	297.128
condition number	154.591	152.878	156.902	157.041	155.063	154.087
Error(20)	10.8120	188.679	14.5859	99.1558	86.8486	25.3959
Error(22)	2220.44	16789.5	529.495	4568.45	8777.48	44.3582



(a) (b) (c)

Figure 4: Excitation trajectory for the links at the arm of manipulator. (a) Trajectory in Cartesian space (b) Joint displacement (c) Joint velocities

Table 3: Dynamic parameter estimation for the first three joints of a 6 DOF industrial robot

Parameter	Elite Particles (Arm)					
	1	2	3	4	5	6
a0(1)	-1.2742	5.2376	-3.2742	-2.8265	0.72577	-3.2742
a0(2)	-6.5754	6.2129	-4.5754	-6.5754	-8.5754	-8.5754
a0(3)	-9.0638	8.2042	-7.0638	-9.0638	-6.8925	-6.7108
b0(1)	-0.4967	6.2173	1.5033	-0.4967	1.5033	-2.4967
b0(2)	4.5411	-2.6712	2.6293	2.9774	2.5411	6.5411
b0(3)	-1.6233	-7.73	-0.74646	1.2535	-3.6233	0.37674
c0	3.0295	1.4518	1.0295	1.7199	1.0295	1.0295
d0	0	0	0	0	0	0
e0	173.11	101.03	184.34	187.21	188.04	166.14
f0	5.7001	6.0889	1.0295	1.7199	1.0295	1.0295
fitness	293.392	199.588	145.3132	103.235	70.9066	64.7651
condition number	53.5757	44.7731	26.41273	17.8764	13.4862	11.8133
Error(20)	0.00270	0.00286	0.035218	0.01955	0.03211	0.00822
Error(22)	0.00460	0.00766	0.074774	0.03507	0.08359	0.01544

Tables 2 and 3 presents the parameters of the non-dominated elite solutions for the joints at the arm and wrist of the manipulator, including the three harmonics each for *a0* and *b0*, with corresponding fitness values, condition number of the observation matrix and cumulative errors for the best 20 identified parameters and for all 22 identified parameters. Figure 3.0 (a) presents the resultant trajectory in Cartesian space for the 5th elite particle at the robot's wrist, showing a somewhat circular continuous trajectory with fluctuations that are typical of FFS. Figures 3.0 (b) and (c) show the position and velocity of the three joints at the robot's wrist in joint space, while Figure 4.0 shows the equivalent for the 1st elite particle at the robot's arm. From the results presented in Table 2 (at the wrist), it would be observed that while the *6th elite particle* has the lowest fitness value and the lowest cumulative estimation error for all the 22 identified parameters, the *2nd elite particle* presents the trajectory with the lowest condition number, and the *1st elite particle* presents the lowest cumulative estimation error for the 20 best identified parameters. Likewise, in Table 3 (at the arm), although all the 22 identifiable parameters were identified with higher accuracy, yet it would also be observed that the *6th elite particle* has the lowest fitness and presents the trajectory with the lowest condition number, but the *1st elite particle* presents the lowest cumulative estimation errors for both the 20 best identified parameter and the 22 identifiable parameters. This supports the observation of [41] that the best solution does not guarantee the global minimum solution because the dynamic optimization problem is not convex.

4.2. Benchmark Function Analysis

A total of thirty-six benchmark functions (BF) were evaluated, and the results presented in Tables 4-6. Reference [2] showed that accuracy and convergence time are paramount in robot analysis, the average and standard deviations alone do not completely describe the results, and some solutions with good averages are incapable of finding the global minimum solution (inaccurate). Therefore, the minimum values are also presented in the tables. The solution that best minimizes the function is presented in bolded font. From the Table 4, in the unimodal BF analysis, all the algorithms satisfactorily found the minimized solution for the function f6; eMuPSO, PSO, and DE equally minimized the function f7 while GWO and WOA equally minimized the function f11. WOA and GWO had the best performance, producing the best minimizing solutions for five out of twelve unimodal BF each. WOA dominated other algorithms in f1, f3, and f12 functions, tied with GWO in f11 and tied with all other metaheuristics in f6. While GWO dominated other metaheuristics in f8, f9, and f10. Therefore, WOA and GWO dominated other algorithms in three unimodal BF, tied in two BF and were dominated in seven BF each as elaborated in Table 7. The PSO dominated other metaheuristics in f4 and f5, while eMuPSO was dominant in f2. Both tied with other metaheuristics in f6 and f7 and were dominated in eight and nine unimodal BF respectively. From the Table 5, in the multimodal BF analysis, it can be observed that the eMuPSO and PSO tied in f1, f4, and f12; PSO, WOA, and GWO tied in f2; WOA and GWO tied in f3; eMuPSO, PSO and GOA tied in f5; all the metaheuristics except GWO and DE tied in f6, eMuPSO and DE tied in f9, and all the metaheuristics except GOA tied in f11. eMuPSO had the best performance producing the best minimum results on eight out of twelve multimodal BF, followed by PSO, WOA, and GWO producing the best minimum solutions for seven, five, and four multimodal BF respectively. eMuPSO dominated other metaheuristics on f10 and tied with other metaheuristics on seven multimodal BF. PSO did not dominate other metaheuristics on any of the multimodal BF, but tied on seven and was dominated in five BF, WOA was dominant on f7, tied with other metaheuristics on 4 Multimodal BF and was dominated in 7 BF. GWO was dominant on f8, tied with other metaheuristics on 3 BF, and was dominated on 8 BF. From the Table 6, in the constant dimension BF analysis, eMuPSO and PSO tied in producing the best minimizing solution for f1 and f6, all metaheuristics except GOA and GWO tied on f2; eMuPSO, PSO and DE tied on f3, f5, f7, f8, f9, f11. All the metaheuristics except WOA and GWO tied on f4 and f12. eMuPSO once again had the best performance producing the best minimum solutions for all twelve BF, dominated other metaheuristics in f10, and tied in eleven BF, followed by PSO and DE producing the best-minimized solution for eleven and nine BF respectively. Compared to other algorithms, WOA and GWO performed well in the unimodal BF analysis but performed poorly in the constant-dimension BF analysis. After numerous experiments, at various dimension sizes, it can be suggested that the WOA algorithm is convenient for analyzing optimization problems with high dimensionality (especially unimodal roles), but may not necessarily maintain its lead when the dimensionality is low, while the proposed eMuPSO is convenient for analyzing optimization problems with lower dimensionality. Compared to other EAs, the eMuPSO and PSO algorithms performed averagely in the unimodal BF evaluations but were excellent in analyzing multimodal and constant-dimension BF. DE performed poorly in the unimodal and multimodal BF, performed well in the constant-dimension BF analysis.

The Wilcoxon signed-rank test was performed at a 5% significance level to compare the results of each run and decide on the significance of the results, and the p-values are reported in Table 8. In the statistical analysis, the best algorithm in each test function is the algorithm with the best average value, the best algorithm is compared with other algorithms independently. For instance, if eMuPSO is the best algorithm, pairwise comparison is done between eMuPSO/PSO, eMuPSO/GOA, and so on. Observe that since the best algorithm cannot be compared with itself, N/A with bold font size has been assigned to it. From Table 8, in the unimodal statistical analysis (F1-F12) WOA algorithm was most significant in six out of twelve BF as elaborate in Table 9. SPSO and GWO were significant in four functions each, while DE and eMuPSO were significant in three and two of the functions respectively. In the multimodal statistical analysis (F13-F24), the eMuPSO and DE were most significant in six out of the twelve BF each, followed by WOA with five BF, while SPSO and GWO were most significant in 3 of the BF each. In the constant-dimension statistical analysis (F25-F36), the eMuPSO and DE remained most significant for ten out of the twelve BF each, followed by the SPSO with 8 BF, GOA was most significant in 2 of the BF, while others were significant in only one each of the benchmarks functions.

Table 4: Unimodal benchmark function analysis.

Unimodal Benchmark Functions		**Evolutionary Algorithms**					
		eMuPSO	PSO	WOA	GWO	GOA	DE
F1: Sphere	*Min*	2.00E-12	4.90E-44	**0**	2.85E-240	2.25E-06	0.001903
	Ave	9.46E-08	5.44E-42	0	1.91E-232	8.66E-05	0.031038
	StDev	3.13E-07	2.23E-41	0	0	9.17E-05	0.039287
F2: Beale	*Min*	**1.67E-08**	7.70E-08	0.000176	6.42E-05	0.000149	3.147E-08
	Ave	0.0178315	0.015488	0.026001	0.005977	0.035998	1.43E-07
	StDev	0.0308013	0.029336	0.044659	0.0154745	0.038128	2.13E-07
F3: Matyas	*Min*	1.69E-12	2.12E-24	**0**	1.33E-143	1.42E-06	3.50E-07
	Ave	3.42E-09	1.07E-18	0	1.57E-120	1.55E-05	6.23E-06
	StDev	7.91E-09	5.33E-18	0	4.10E-120	1.51E-05	5.08E-06
F4: Rosenbrook	*Min*	10.167783	**0.087952**	14.80708	15.116454	13.29022	6.665086
	Ave	28.023866	6.257364	15.61825	16.217369	258.643	28.08251
	StDev	32.409539	3.514322	0.417166	0.6497502	480.0779	31.98284
F5: Step 2	*Min*	3.91E-10	**0**	1.38E-05	1.91E-07	7.31E-07	0.004025
	Ave	8.03E-08	0	4.99E-05	0.1749605	0.000109	0.047807
	StDev	1.77E-07	0	3.02E-05	0.1875981	0.000145	0.055443
F6: Scahffer 2	*Min*	**0**	0	0	0	0	0
	Ave	0	0	0	0	0	0
	StDev	0	0	0	0	0	0
F7: Scahffer 3	*Min*	**-3.37E-4**	-3.37E-4	-3.37E-4	-3.37E-4	-3.37E-4	-3.37E-4
	Ave	-3.37E-4	-3.37E-4	-3.36E-4	-3.37E-4	-3.37E-4	-3.37E-4
	StDev	1.65E-19	1.65E-19	3.00E-06	3.45E-08	1.08E-09	1.65E-19
F8: Scwefel 1.2	*Min*	0.0003765	1.92E-11	0.317571	**7.14E-99**	1.759081	0.003903
	Ave	0.0071479	4.45E-09	207.2391	4.23E-82	33.02237	0.079942
	StDev	0.0069211	6.37E-09	408.9857	2.15E-81	28.64997	0.090924
F9: Scwefel 2.21	*Min*	0.370036	9.76E-10	1.62E-09	**8.56E-66**	0.018632	0.179791
	Ave	1.0249904	1.39E-08	4.32931	1.20E-62	1.084041	2.315132
	StDev	0.5007591	1.17E-08	9.450776	4.98E-62	1.537078	2.441406
F10: Scwefel 2.22	*Min*	0.0027291	5.00E-26	**2.0800E-322**	4.67E-136	0.018623	0.077971
	Ave	0.0326953	2.66E-24	2.7900E-321	1.61E-133	0.192329	1.517522
	StDev	0.0266489	5.88E-24	0	5.42E-133	0.578804	2.606056
F11: De Jong 4	*Min*	2.35E-22	1.53E-79	**0**	0	1.40E-13	5.83E-10
	Ave	3.14E-14	4.80E-73	0	0	2.75E-09	0.117002
	StDev	9.95E-14	1.73E-72	0	0	3.50E-09	0.640849
F12: Axis Parallel	*Min*	4.79E-12	2.48E-46	**0**	2.61E-241	3.93E-05	6.28E-05
	Ave	2.12E-08	2.71E-43	0	2.61E-235	0.001067	0.000993
	StDev	6.13E-08	9.27E-43	0	0	0.001414	0.001253

Table 5: Multimodal benchmark function analysis.

Multimodal Benchmark Functions		**Evolutionary Algorithms**					
		eMuPSO	PSO	WOA	GWO	GOA	DE
F13: Easom	*Min*	**0**	0	3.26E-15	7.77E-12	2.11E-16	NaN
	Ave	0	0	0.023333	1.00E-09	0.086667	NaN
	StDev	0	0	0.043018	9.16E-10	0.034575	NaN
F14: Griewank	*Min*	2.00E-07	**0**	0	0	0.000577	0.011976
	Ave	0.0343134	0.015987	0.001742	0.0001222	0.04842	0.147971
	StDev	0.0330935	0.017756	0.009541	0.0006694	0.06238	0.143891
F15: Rastrigin	*Min*	8.9546265	7.959672	**0**	0	31.83894	9.732789
	Ave	17.378784	16.01883	0	0	67.56469	36.96119
	StDev	5.8526738	5.267084	0	0	29.12814	22.58322
F16: Pen Holder	*Min*	**-9.04E-08**	-9.04E-08	2.87E-07	0.0393752	0.026031	NaN
	Ave	0.0168779	0.001703	0.029721	0.0652148	0.082208	NaN
	StDev	0.0251307	0.004706	0.02437	0.0179396	0.034506	NaN
F17: Test-tube Holder	*Min*	**-2.04E-08**	-2.04E-08	-2.04E-08	-2.04E-08	-2.04E-08	-2.040E-08
	Ave	-2.04E-08	-2.04E-08	0.000198	-1.63E-08	0.002207	-1.50E-05
	StDev	0	0	0.000604	4.80E-09	0.00908	3.01E-05
F18: Egg Holder	*Min*	**0**	0	0	1.03E-10	0	NaN
	Ave	69.652218	52.92821	1.14E-13	25.398919	149.1868	NaN
	StDev	77.530967	74.41267	1.16E-13	50.110379	130.2925	NaN
F19: Damavandi	*Min*	2	2	**0.009163**	2.0000051	2	2.000002
	Ave	2	2	0.278923	2.2578664	2.000002	2.000018
	StDev	3.27E-10	0	1.197583	1.412306	3.83E-06	1.32E-05
F20: Cross-Leg Table	*Min*	0.0037107	0.049876	0.099255	**0**	0.09978	0.0150
	Ave	0.0603736	0.084285	0.099835	0.0832343	0.099862	0.011415
	StDev	0.0271336	0.019337	0.000111	0.0378599	2.45E-05	0.006958
F21: Bucking 4	*Min*	**0**	1.47E-90	2.20E-11	1.58E-08	3.80E-08	0
	Ave	0	5.03E-79	1.17E-06	1.15E-06	0.00028	0
	StDev	0	2.35E-78	1.28E-06	1.26E-06	0.000495	0
F22: Bucking 6	*Min*	**0.0023858**	0.003192	0.003289	0.092071	0.0065	0.013404
	Ave	0.1010663	0.119555	0.124073	0.152622	0.098243	0.097272
	StDev	0.0585834	0.074511	0.08727	0.0926189	0.066986	0.055002
F23: Cosine	*Min*	**0**	0	0	0	1.94E-15	0
	Ave	0.0084647	0	0	0	0.08306	0
	StDev	0.0080534	0	0	0	0.05875	0
F24: Modified Rosenbrook	*Min*	**3.4040243**	3.404024	3.404024	3.4040243	3.404024	3.404024
	Ave	4.8692154	5.80161	3.404024	6.8672033	7.000974	3.404024
	StDev	1.9585576	1.991086	4.00E-08	1.3815921	1.219484	1.88E-15

Table 6: Constant-dimension benchmark function analysis.

Constant-Dimension Benchmark Functions		Evolutionary Algorithms					
		eMuPSO	SPSO	WOA	GWO	GOA	DE
F25: Paviani	*Min*	**-2.34E-06**	-2.34E-06	4.52E-06	1.91E-05	-2.34E-06	NaN
	Ave	-2.34E-06	-2.34E-06	0.000137	4.28E-05	-2.34E-06	NaN
	StDev	5.70E-15	3.54E-15	0.00013	2.09E-05	2.03E-11	NaN
F26: Modified Ackley	*Min*	**0**	0	0	1.53E-10	1.78E-15	0
	Ave	0	0	2.01E-10	4.45E-09	3.00E-14	0
	StDev	0	0	3.77E-10	3.79E-09	2.01E-14	0
F27: Camel 6 Hump	*Min*	**2.22E-16**	2.22E-16	1.11E-15	1.11E-13	4.44E-16	2.22E-16
	Ave	2.22E-16	2.22E-16	5.03E-13	3.93E-10	8.44E-16	2.22E-16
	StDev	0	0	1.25E-12	4.48E-10	3.37E-16	0
F28: Branin Rcos	*Min*	**-2.26E-11**	-2.26E-11	-2.26E-11	1.84E-10	-2.26E-11	-2.26E-11
	Ave	-2.26E-11	-2.26E-11	6.93E-09	1.94E-08	-2.26E-11	-2.26E-11
	StDev	0	0	1.32E-08	2.29E-08	4.85E-15	0
F29: Hartmann 3	*Min*	**-1.78E-05**	-1.78E-05	-1.78E-05	-1.78E-05	-1.78E-05	-1.78E-05
	Ave	-1.78E-05	-1.78E-05	0.000528	0.0005009	0.051517	-1.78E-05
	StDev	0	0	0.001523	0.0017303	0.196121	0
F30: Hartmann 6	*Min*	**0**	0	9.46E-07	4.68E-08	6.22E-15	4.44E-16
	Ave	0.0435941	0.047557	0.041273	0.056637	0.043647	0.11493
	StDev	0.0582734	0.059241	0.059382	0.0745266	0.058344	0.021707
F31: Shekel 5	*Min*	**0**	0	2.36E-06	2.51E-06	6.04E-14	0
	Ave	1.5943441	2.514812	0.81061	0.6767119	4.505148	0
	StDev	2.7615053	3.018129	2.487387	1.7547512	3.369024	0
F32: Shekel 7	*Min*	**-20.80588**	-20.80588	-20.80588	-20.80588	-20.80588	-20.80588
	Ave	-19.2602	-19.6892	-20.62807	-20.80587	-17.28696	-20.80588
	StDev	2.9070217	2.583984	0.970316	1.00E-05	3.854556	3.61E-15
F33: Shekel 10	*Min*	**-2.20E-07**	-2.20E-07	2.50E-06	1.27E-06	-2.20E-07	-2.20E-07
	Ave	0.4489944	1.336099	0.455673	1.37E-05	4.623238	-2.20E-07
	StDev	1.7463952	2.782463	1.750715	6.81E-06	3.885799	5.42E-16
F34: Perm	*Min*	**1.00E-30**	1.81E-26	0.067631	0.0001552	0.00023	0.0049
	Ave	0.042594	0.03135	2.678428	1.3509178	0.046792	0.006286
	StDev	0.095434	0.089466	3.288481	2.9794594	0.093625	0.007091
F35: Goldstein Price	*Min*	**-8.08E-14**	-8.08E-14	5.46E-11	5.70E-10	-7.02E-14	-8.08E-14
	Ave	-7.87E-14	-7.86E-14	5.70E-07	1.04E-06	2.7	-7.93E-14
	StDev	9.85E-16	8.92E-16	1.13E-06	9.80E-07	14.78851	8.83E-16
F36: Langerman 5	*Min*	**0.3033287**	0.303329	0.694475	0.3033287	0.303329	0.303329
	Ave	0.5761576	0.61325	1.021407	0.5757514	0.742518	0.329675
	StDev	0.3298947	0.33584	0.186392	0.3026761	0.321734	0.10027

Table 7: Summary of benchmark function evaluations.

Summary of Benchmark Function Analysis		Evolutionary Algorithms					
		eMuPSO	PSO	WOA	GWO	GOA	DE
Minimum function Evaluations	*Unimodal*	1/2/9	2/2/8	3/2/7	3/2/7	0/1/11	0/2/10
	Multimodal	1/7/4	0/7/5	1/4/7	1/3/8	0/3/10	0/3/9
	Constant Dim.	1/11/0	0/11/1	0/1/11	0/0/12	0/2/10	0/9/3
	Total	3/20/13	2/20/14	4/7/25	4/5/27	0/6/30	0/14/22

Table 8: p-Values obtained from Wilcoxon signed rank pairwise test.

Wilcoxon signed rank pairwise test (Benchmark Functions)	**Evolutionary Algorithms (P_val)**					
	eMuPSO	SPSO	WOA	GWO	GOA	DE
F1: Sphere	1.73E-06	1.73E-06	NA	1.73E-06	1.73E-06	1.73E-06
F2: Beale	NA	1.73E-06	1.73E-06	1.73E-06	1.73E-06	1
F3: Matyas	1.73E-06	1.73E-06	NA	1.73E-06	1.73E-06	1.73E-06
F4: Rosenbrook	1.92E-06	NA	1.73E-06	1.73E-06	1.73E-06	1.92E-06
F5: Step 2	1.73E-06	NA	1.73E-06	1.73E-06	1.73E-06	1.73E-06
F6: Scahffer 2	NA	1	1	1	1	1
F7: Scahffer 3	NA	1	1.73E-06	1.73E-06	1.73E-06	1
F8: Scwefel 1.2	1.73E-06	1.73E-06	1.73E-06	NA	1.73E-06	1.73E-06
F9: Scwefel 2.21	1.73E-06	1.73E-06	1.73E-06	NA	1.73E-06	1.73E-06
F10: Scwefel 2.22	1.73E-06	1.73E-06	NA	1.73E-06	1.73E-06	1.73E-06
F11: De Jong 4	1.73E-06	1.73E-06	NA	1	1.73E-06	1.73E-06
F12: Axis Parallel	1.73E-06	1.73E-06	NA	1.73E-06	1.73E-06	1.73E-06
F13: Easom	NA	1	1.67E-06	1.73E-06	1.98E-07	NaN
F14: Grievant	1.73E-06	5.94E-05	1	1	1.73E-06	1.73E-06
F15: Rastrigin	1.73E-06	1.73E-06	NA	1	1.73E-06	1.73E-06
F16: Pen Holder	NA	1.01E-06	1.73E-06	1.73E-06	1.73E-06	NaN
F17: Test-tube Holder	0.015625	0.015625	1.73E-06	1.73E-06	5.06E-06	1
F18: Egg Holder	NA	0.000375	0.014795	0.002585	4.73E-06	NaN
F19: Damavandi	3.11E-05	3.11E-05	NA	2.84E-05	3.11E-05	3.11E-05
F20: Cross-Leg Table	1.92E-06	1.73E-06	1.73E-06	7.69E-06	1.73E-06	1
F21: Bucking 4	1.73E-06	1.73E-06	1.73E-06	1.73E-06	1.73E-06	1
F22: Bucking 6	NA	0.19861	0.14704	0.010444	0.53044	0.734325
F23: Cosine	NA	1	1	1	1.73E-06	1
F24: Modify Rosenbrook	NA	3.38E-05	1.73E-06	1.73E-06	1.72E-06	1
F25: Paviani	1.73E-06	1	1.73E-06	1.73E-06	1.73E-06	NaN
F26: Modified Ackley	NA	1	2.56E-06	1.73E-06	1.73E-06	1
F27: Camel 6 Hump	NA	1	1.73E-06	1.73E-06	1.50E-06	1
F28: Branin Rcos	NA	1	1.73E-06	1.73E-06	0.007813	1
F29: Hartmann 3	NA	1	1.73E-06	1.73E-06	1.73E-06	1
F30: Hartmann 6	0.877403	0.106394	1	0.544006	0.082206	0.000616
F31: Shekel 5	0.25	0.000244	1.73E-06	1.73E-06	1.73E-06	1
F32: Shekel 7	NA	0.0625	1.73E-06	1.73E-06	1.71E-06	1
F33: Shekel 10	0.375	0.007813	1.73E-06	1.73E-06	1.73E-06	1
F34: Perm	0.24519	0.236936	1.73E-06	0.000388	0.093676	1

F35: Goldstein Price	0.000509	0.004651	1.73E-06	1.73E-06	1.73E-06	1
F36: Langerman 5	0.165027	0.025637	1.73E-06	0.001382	0.000148	1

Table 9: Summary of Wilcoxon signed rank pairwise test evaluations.

Summary of Wilcoxon signed rank pairwise test		**Evolutionary Algorithms**					
		eMuPSO	PSO	WOA	GWO	GOA	DE
Wilcoxon Signed Rank test	*Unimodal*	0/2/10	2/2/8	4/2/6	2/2/8	0/1/11	1/2/9
	Multimodal	2/4/6	0/3/9	1/4/7	0/3/9	0/1/11	2/4/6
	Constant Dim.	0/10/2	1/7/4	0/1/11	0/1/11	0/2/10	1/9/2
	Total	2/16/18	3/12/21	5/7/24	2/6/28	0/4/32	4/15/17

4.3. Convergence Analysis

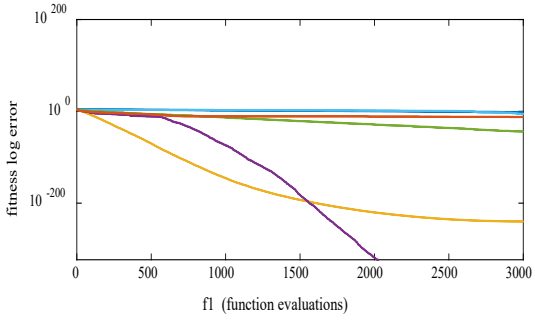
In Figure 5, the unimodal convergence plot analysis, it can be observed that eMuPSO was dominant in f2, tied with other metaheuristics in f6 and f7, produced the second-best minimum solution for f4 and f5, and the third-best minimum solution for f8. A large overshoot can also be observed in the results of f7. In Figure 6, the multimodal convergence plots, eMuPSO dominated others in f10 and tied in f1, f4, f5, f6, f9, f10, f11. eMuPSO produced the second-best minimized solution in f8. All the algorithms except WOA stagnated at the local minimum in f7. In Figure 7, the Constant-dimension benchmark function convergence plots, eMuPSO best minimized all twelve BMF dominating others in f10. Overshoot can be observed in the results of eMuPSO in f25, f28, f29, f33, and f35. Similarly, an overshoot can also be observed in the results of WOA in f28 and f29. In the convergence plots presented, it can be seen that the GOA and GWO produced the slowest converging results, this is most evident in f13, f16, f18, f22, and most of the constant-dimension benchmark functions. This suggests good exploration capabilities of the swarms. Whereas eMuPSO, PSO, and DE produced the fastest converging results. In robot analysis speed and accuracy are required, therefore the best algorithm has to make a balance between fast convergence speed and good exploration tendencies.

5. Conclusion

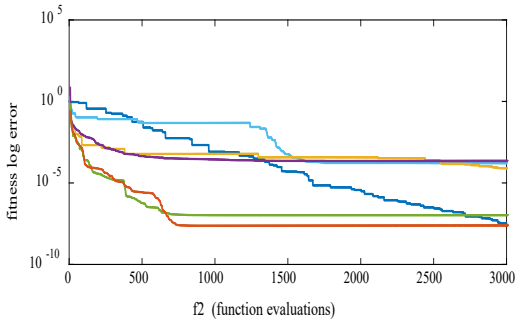
An enhanced PSO algorithm was proposed with a Gaussian mutation and an archive elite learning (eMuPSO), the parameters and dynamic update technique were optimized for robot analysis. The parameter identification problem of robot manipulators is generally computationally demanding, therefore finding a computationally efficient solution is important for which swarm-based algorithms are suitable. In Dynamic parameter estimation, excitation trajectory was optimized using the proposed eMuPSO, it allowed the swarm to be populated with fewer particles thereby requiring less computational time. The mutation function allowed a wholesome search of the solution space. The LLS method was subsequently used to estimate the dynamic parameters of a 6DOF industrial robot manipulator where it was seen that estimation errors are larger at the wrist of the manipulator. Parameter estimation accuracy is better at the arm with less dexterity, while the total estimation accuracy for the 6DOF manipulator is deteriorated by the cumulative estimation errors at the wrist with increased dexterity. This suggests that larger DOF manipulators would be suitable for manufacturing, agriculture and other industrial applications, but may not be suitable for high end application that are keen on accuracy like medical operations, airplane manufacture, etc. The eMuPSO particularly gives a scientist the advantage of choosing from the set of elite particles, the solution that best solves the desired problem. The eMuPSO is therefore capable of analyzing both kinematic and dynamic robot problems of robot manipulators. Further experimentation showed that the landscape of the solution space for robot optimization problems are non-linear, non-separable and multi-modal with multiple possibilities of sub-optimal solutions (local minimizers) some of which are very steep. When a solution approaches the vicinity of these steep local minimizers, it easily falls into its trap and it is usually very difficult to break out of such stagnation without artificial perturbation, resulting to the failure of most evolutionary algorithms in analyzing robot optimization problems. The performance of the eMuPSO can be related to its origins. The parameters of the eMuPSO was derived experimentally by comparing the performance of various robot manipulator configurations as described in [2]. The robot optimization problem is a non-linear and multi-modal problem with dimension size usually less than twenty. Robot manipulator problems are generally computational demanding, therefore a fast convergence time and high accuracy are important. The eMuPSO best showed its veracity in analyzing multimodal and constant dimension BF. PSO and DE performed well in constant dimension BF. It would also be observed that the proposed eMuPSO cannot drill a solution below 1E-30, this is also suitable for robot analysis because the solution to robot optimization problems is usually in the range of 1E-10.

For a successful dynamic parameter estimation analysis, the following should be carefully observed.

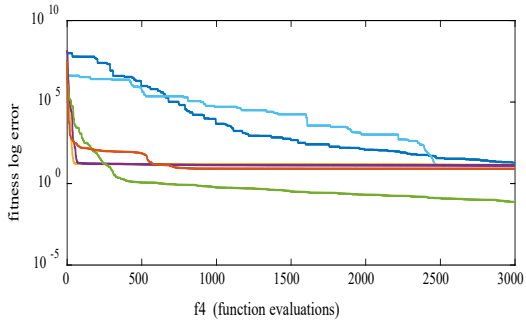
- It is recommended to keep either *d0* or *e0* equal to zero to avoid singularity so that the excitation trajectory is limited two quadrants on the x-y plane.
- It would be keen to differentiate between dynamic and kinematic DH-parameters. When determining the position coordinated of the robot manipulator, the sum of lengths of the 3rd and 4th links make up *d4*, while during dynamic analysis this results in actuator torque errors. Likewise, the length *d6*.



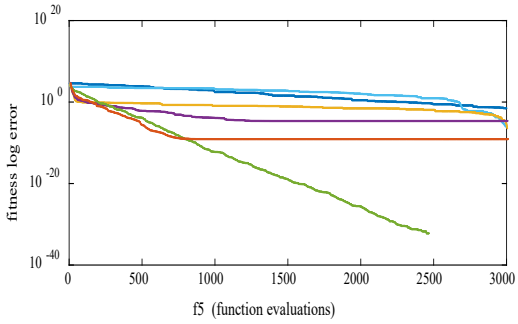
(a) Sphere



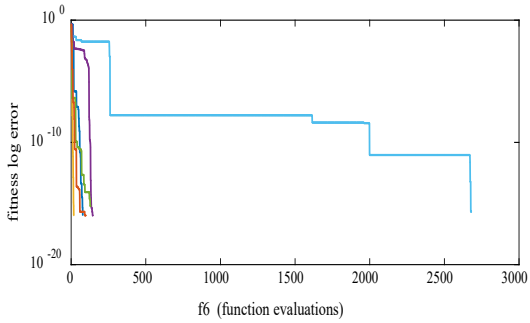
(b) Beale



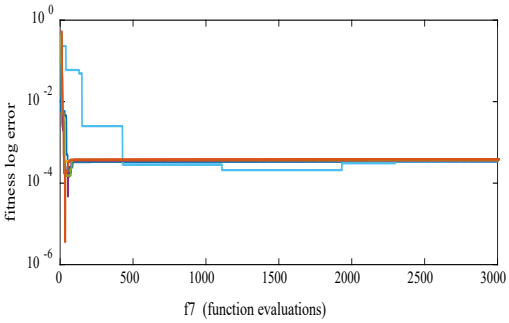
(c) Rosenbrook



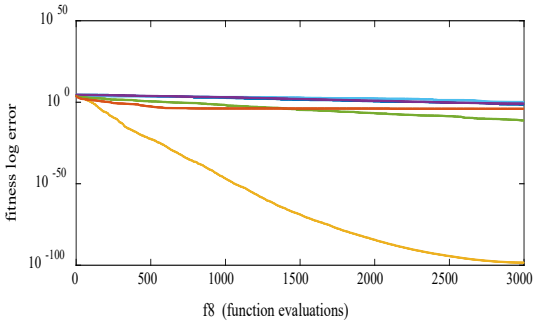
(d) Step 2



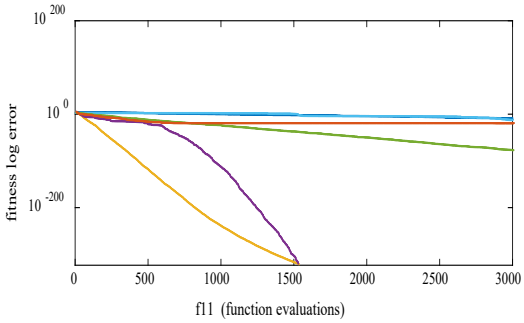
(e) Schaffer 2



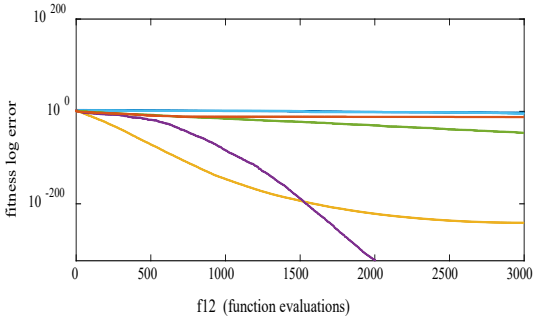
(f) Schaffer 3



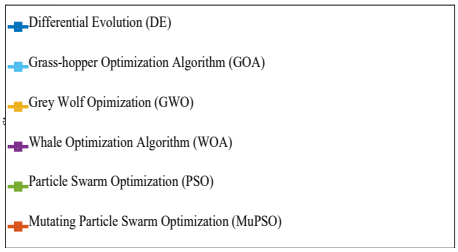
(g) Scwefel 1.2



(h) De Jong 4



(i) Axis Parallel



(j) Legend

Figure 5: Convergence analysis for unimodal benchmark functions.

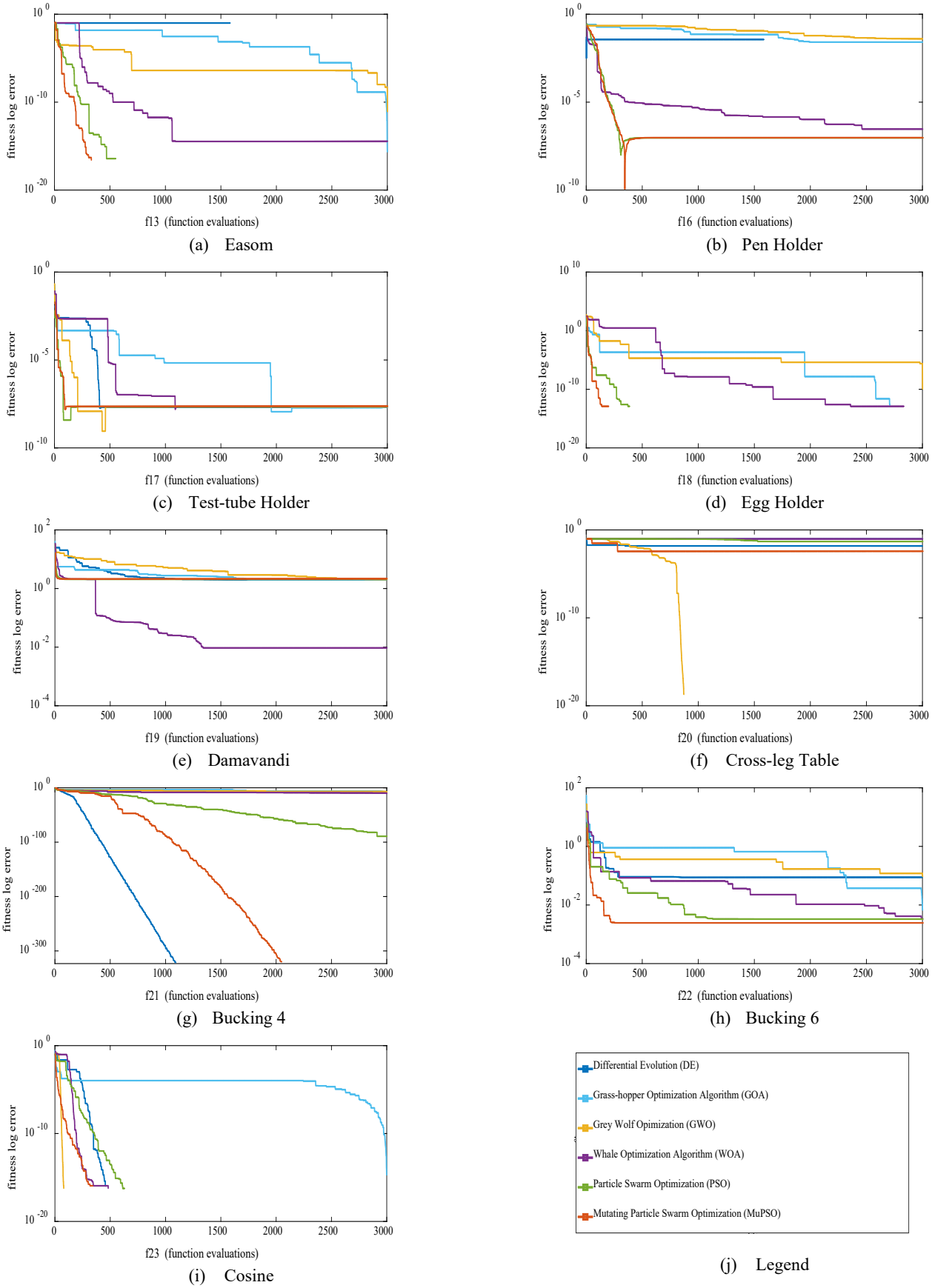
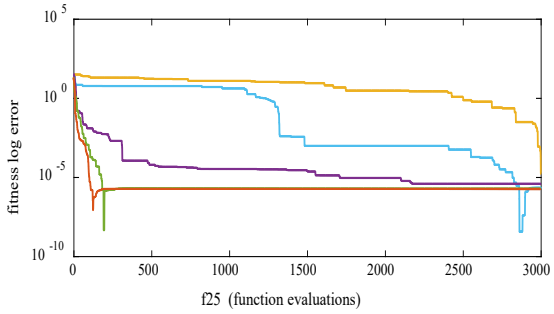
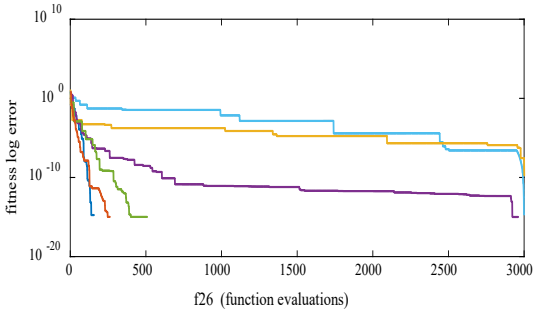


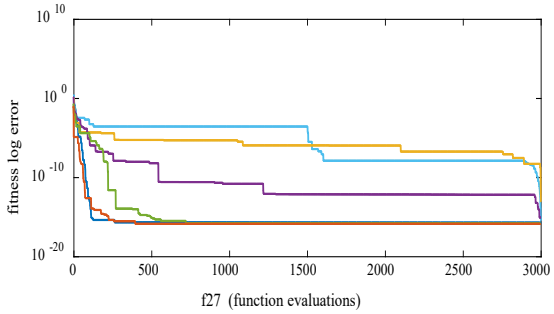
Figure 6: Convergence analysis for multimodal benchmark functions.



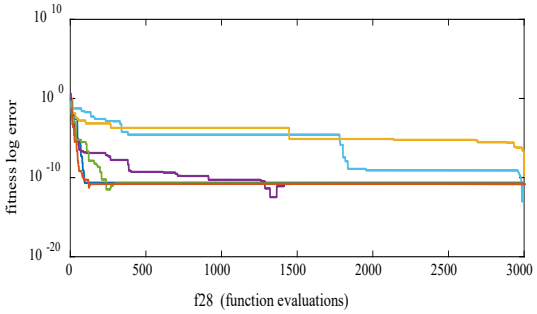
(a) Paviani



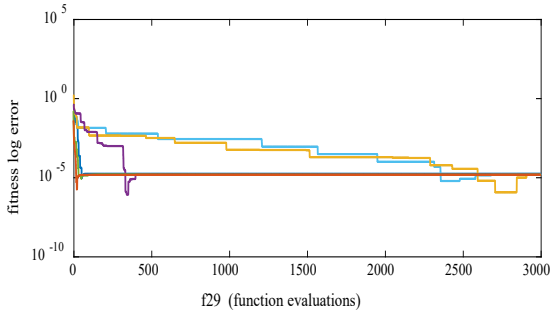
(b) Modified Ackley



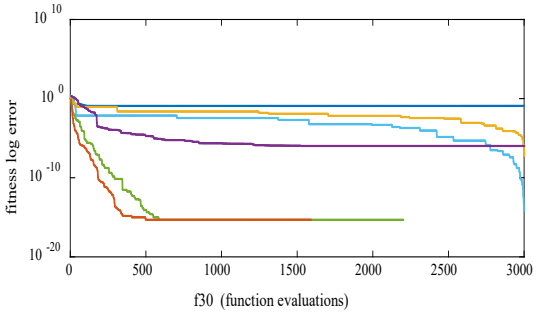
(c) Camel 6 Hump



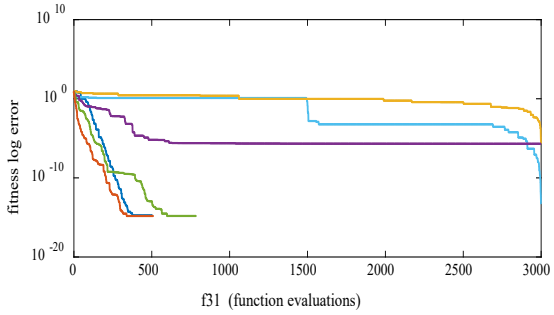
(d) Branin R cos



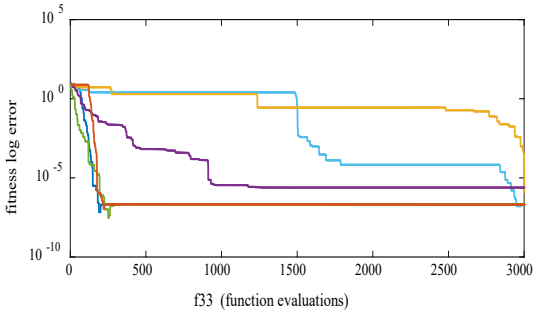
(e) Hartmann 3



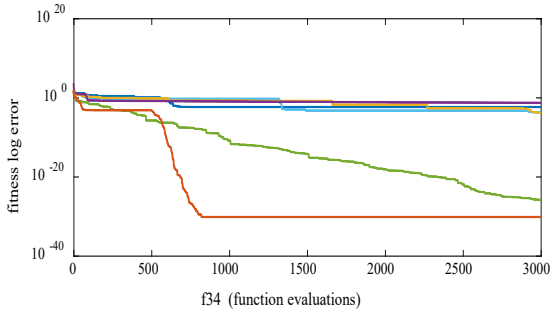
(f) Hartmann 6



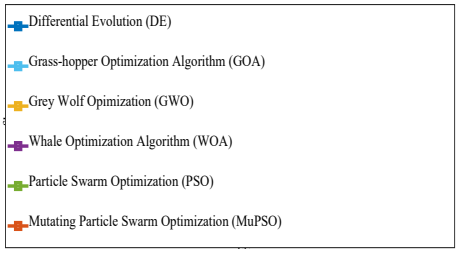
(g) Shekel 5



(h) Shekel 10



(i) Perm



(j) Legend

Figure 7: Convergence analysis for constant-dimension benchmark functions.

- When determining the 6th joint angle from the excitation trajectory vector with (27), the initial joint position is unknown, so it is recommended to assume the initial joint position at the first segment (*@ it = 1*) is equal to the joint position at the last time segment to reduce the overshoot in velocity.

Funding

This research was funded by the National Natural Science Foundation of China, grant number 51875166.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I would like to first acknowledge my late mum whose unflinching support kept me going against all odds and who unfortunately is no more with us to witness my achievements, my lovely wife and children who were patient with me through these trying times and finally to Prof. Shi for giving me the opportunity and finding me worthy.

References

[1] A. Umar, Z. Shi, W. Wang, Z.I.B. Farouk, "A Novel Mutating PSO Based Solution for Inverse Kinematic Analysis of Multi Degree-Of-Freedom Robot Manipulators," Proceedings of 2019 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2019, 459–463, 2019, doi:10.1109/ICAICA.2019.8873449.

[2] A. Umar, Z. Shi, A. Khlil, Z.I.B. Farouk, "Developing a New Robust Swarm-Based Algorithm for Robot Analysis," Mathematics, **8**(2), 158, 2020, doi:10.3390/math8020158.

[3] B. Armstrong, "On Finding Exciting Trajectories for Identification Experiments Involving Systems with Nonlinear Dynamics," The International Journal of Robotics Research, **8**(6), 28–48, 1989, doi:10.1177/027836498900800603.

[4] P.K. Khosla, "Categorization of Parameters in the Dynamic Robot Model," IEEE Transactions on Robotics and Automation, **5**(3), 261–268, 1989, doi:10.1109/70.34762.

[5] K. Bouabaz, Q. Zhu, "Improved numerical technique for industrial robots model reduction and identification," in 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), IEEE, Hefei: 1032–1038, 2016, doi:10.1109/ICIEA.2016.7603734.

[6] S.K. Lin, C.J. Fang, "Efficient formulations for the manipulator inertia matrix in terms of minimal linear combinations of inertia parameters," Journal of Robotic Systems, **16**(12), 679–695, 1999, doi:10.1002/(SICI)1097-4563(199912)16:12<679::AID-ROB2>3.0.CO;2-Y.

[7] S.K. Lin, "Minimal Linear Combinations of the Inertia Parameters of a Manipulator," IEEE Transactions on Robotics and Automation, **11**(3), 360–373, 1995, doi:10.1109/70.388778.

[8] J. Swevers, C. Ganseman, J. De Schutter, H. Van Brussel, "Experimental robot identification using optimised periodic trajectories," Mechanical Systems and Signal Processing, **10**(5), 561–577, 1996, doi:10.1006/mssp.1996.0039.

[9] W. Wu, S. Zhu, X. Wang, H. Liu, "Closed-loop dynamic parameter identification of robot manipulators using modified fourier series," International Journal of Advanced Robotic Systems, **9**, 2012, doi:10.5772/45818.

[10] C. Yue, B. Qu, J. Liang, "A Multiobjective Particle Swarm Optimizer Using Ring Topology for Solving Multimodal Multiobjective Problems," IEEE Transactions on Evolutionary Computation, **22**(5), 805–817, 2018, doi:10.1109/TEVC.2017.2754271.

[11] T. Blackwell, J. Kennedy, "Impact of Communication Topology in Particle Swarm Optimization," IEEE Transactions on Evolutionary Computation, **23**(4), 689–702, 2019, doi:10.1109/TEVC.2018.2880894.

[12] U. Bartoccini, A. Carpi, V. Poggioni, V. Santucci, "Memes evolution in a memetic variant of particle swarm optimization," Mathematics, **7**(5), 1–13, 2019, doi:10.3390/math7050423.

[13] B. Tran, B. Xue, M. Zhang, "Variable-Length Particle Swarm Optimization for Feature Selection on High-Dimensional Classification," IEEE Transactions on Evolutionary Computation, **23**(3), 473–487, 2019, doi:10.1109/TEVC.2018.2869405.

[14] S.K.S. Fan, C.H. Jen, "An enhanced partial search to particle swarm optimization for unconstrained optimization," Mathematics, **7**(4), 2019, doi:10.3390/math7040357.

[15] M. Aladeemy, L. Adwan, A. Booth, M.T. Khasawneh, S. Poranki, "New feature selection methods based on opposition-based learning and self-adaptive cohort intelligence for predicting patient no-shows," Applied Soft Computing Journal, **86**, 105866, 2020, doi:10.1016/j.asoc.2019.105866.

[16] W. Liu, Z. Wang, X. Liu, N. Zeng, D. Bell, "A Novel Particle Swarm Optimization Approach for Patient Clustering from Emergency Departments," IEEE Transactions on Evolutionary Computation, **23**(4), 632–644, 2019, doi:10.1109/TEVC.2018.2878536.

[17] S.K. Baliarsingh, W. Ding, S. Vipsita, S. Bakshi, "A memetic algorithm using emperor penguin and social engineering optimization for medical data classification," Applied Soft Computing Journal, **85**, 105773, 2019, doi:10.1016/j.asoc.2019.105773.

[18] D. Chitradevi, S. Prabha, "Analysis of brain sub regions using optimization techniques and deep learning method in Alzheimer disease," Applied Soft Computing Journal, **86**, 105857, 2020, doi:10.1016/j.asoc.2019.105857.

[19] H. Paramesh, "Air Pollution and Allergic Airway Diseases: Social Determinantsand Sustainability in the Control and Prevention," Indian Journal of Pediatrics, **85**(4), 284–294, 2018, doi:10.1007/s12098-017-2538-3.

[20] S. Zhao, D. Pudasainee, Y. Duan, R. Gupta, M. Liu, J. Lu, "A review on mercury in coal combustion process: Content and occurrence forms in coal, transformation, sampling methods, emission and control technologies," Progress in Energy and Combustion Science, **73**, 26–64, 2019, doi:10.1016/j.pecs.2019.02.001.

[21] A. Goudarzi, Y. Li, J. Xiang, "A hybrid non-linear time-varying double-weighted particle swarm optimization for solving non-convex combined environmental economic dispatch problem," Applied Soft Computing Journal, **86**, 105894, 2020, doi:10.1016/j.asoc.2019.105894.

[22] S. Shamshirband, M. Hadipoor, A. Baghban, A. Mosavi, J. Bukor, A.R. Várkonyi-Kóczy, "Developing an ANFIS-PSO model to predict mercury emissions in combustion flue gases," Mathematics, **7**(10), 1–16, 2019, doi:10.3390/math7100965.

[23] N. Moayyeri, S. Gharehbaghi, V. Plevris, "Cost-based optimum design of reinforced concrete retaining walls considering different methods of bearing capacity computation," Mathematics, **7**(12), 2019, doi:10.3390/MATH7121232.

[24] J. Arfaoui, H. Rezk, M. Al-Dhaifallah, F. Elyes, M. Abdelkader, "Numerical performance evaluation of solar photovoltaic water pumping system under partial shading condition using modern optimization," Mathematics, **7**(11), 1–19, 2019, doi:10.3390/math7111123.

[25] H. Rezk, A.A. Elghany, M. Al-Dhaifallah, A.H.M. El Sayed, M.N. Ibrahim, "Numerical estimation and experimental verification of optimal parameter identification based on modern optimization of a three phase induction motor," Mathematics, **7**(12), 2019, doi:10.3390/MATH7121135.

[26] A. V. Kononova, D.W. Corne, P. De Wilde, V. Shneer, F. Caraffini, "Structural bias in population-based algorithms," Information Sciences, **298**, 468–490, 2015, doi:10.1016/j.ins.2014.11.035.

[27] H. Liang, K. Zhang, J. You, H. Yu, "Multi-objective Gaussian particle swarm algorithm optimization based on niche sorting for actuator design," Advances in Mechanical Engineering, **7**(12), 1–7, 2015, doi:10.1177/1687814015620326.

[28] Y. Sun, Y. Gao, "A multi-objective particle swarm optimization algorithm based on Gaussian mutation and an improved learning strategy," Mathematics, **7**(2), 2019, doi:10.3390/math7020148.

[29] S. Cheng, H. Zhan, Z. Shu, "An innovative hybrid multi-objective particle swarm optimization with or without constraints handling," Applied Soft Computing Journal, **47**, 370–388, 2016, doi:10.1016/j.asoc.2016.06.012.

[30] Q. Lin, J. Li, Z. Du, J. Chen, Z. Ming, "A novel multi-objective particle swarm optimization with multiple search strategies," European Journal of Operational Research, **247**(3), 732–744, 2015, doi:10.1016/j.ejor.2015.06.071.

[31] V.K. Patel, V.J. Savsani, "A multi-objective improved teaching-learning based optimization algorithm (MO-ITLBO)," Information Sciences, **357**(June), 182–200, 2016, doi:10.1016/j.ins.2014.05.049.

[32] A. Lin, W. Sun, H. Yu, G. Wu, H. Tang, "Adaptive comprehensive learning particle swarm optimization with cooperative archive," Applied Soft Computing Journal, **77**, 533–546, 2019, doi:10.1016/j.asoc.2019.01.047.

[33] X. Xia, Y. Tang, B. Wei, L. Gui, "Dynamic multi-swarm particle swarm

optimization based on elite learning," IEEE Access, **7**, 184849–184865, 2019, doi:10.1109/ACCESS.2019.2960890.
[34] C.G. Atkeson, C.H. An, J.M. Hollerbach, "Estimation of Inertial Parameters of Manipulator Loads and Links," The International Journal of Robotics Research, **5**(3), 101–119, 1986, doi:doi.org/10.1177/027836498600500306.
[35] J. Kennedy, R. Eberhart, "Particle Swarm Optimization," in Proceedings of the 1995 IEEE International Conference on Neural Networks. Part 1 (of 6), IEEE: 1942–1948, 1995.
[36] S. Mirjalili, A. Lewis, "The Whale Optimization Algorithm," Advances in Engineering Software, **95**, 51–67, 2016, doi:10.1016/j.advengsoft.2016.01.008.
[37] S. Mirjalili, S.M. Mirjalili, A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, **69**, 46–61, 2014, doi:10.1016/j.advengsoft.2013.12.007.
[38] S. Saremi, S. Mirjalili, A. Lewis, "Grasshopper Optimisation Algorithm: Theory and application," Advances in Engineering Software, **105**, 30–47, 2017, doi:10.1016/j.advengsoft.2017.01.004.
[39] R. Storn, "Differential evolution design of an IIR-filter," Proceedings of the IEEE Conference on Evolutionary Computation, 268–273, 1996, doi:10.1109/icec.1996.542373.
[40] K. Radkhah, D. Kulic, E. Croft, "Dynamic Parameter Identification for the CRS A460 robot," in IEEE International Conference on Intelligent Robots and Systems, IEEE, San Diego, CA, USA: 3842–3847, 2007, doi:10.1109/IROS.2007.4399314.
[41] D. Kostić, B. de Jager, M. Steinbuch, R. Hensen, "Modeling and identification for high-performance robot control: An RRR-robotic arm case study," IEEE Transactions on Control Systems Technology, **12**(6), 904–919, 2004, doi:10.1109/TCST.2004.833641.

Appendix A1: Unimodal benchmark functions.

Unimodal functions	**Test function**	**Dim**	**Range**	**f_{min}**
F1: Sphere	$f_1(x)=\sum_{i=1}^{n} x_i^2$	30	-100:100	0
F2: Beale	$f_2(x)=\left(1.5-x_i+x_i x_{i+1}\right)^2+\left(2.25-x_i+x_i x_{i+1}^2\right)^2+\left(2.625-x_i+x_i x_{i+1}^3\right)^2$	30	-4.5:4.5	0
F3: Matyas	$f_3(x)=0.26\left(x_i^2+x_{i+1}^2\right)0.48 x_i x_{i+1}$	30	-10:10	0
F4: Rosenbrook	$f_4(x)=\sum_{i=1}^{n}\left[100\left(x_{i+1}-x_i^2\right)^2+\left(x_i-1\right)^2\right]$	30	-30:30	0
F5: Step2	$f_5(x)=\sum_{i=1}^{n}\left(\left\lvert x_i+0.5\right\rvert\right)^2$	30	-100:100	0
F6: Scahffer 2	$f_6(x)=0.5+\frac{\sin^2\left(x_1^2-x_2^2\right)^2-0.5}{1+0.001\left(x_1^2-x_2^2\right)^2}$	30	-100:100	0
F7: Scahffer 3	$f_7(x)=0.5+\frac{\sin^2\left(\cos\left\lvert x_1^2-x_2^2\right\rvert\right)^2-0.5}{1+0.001\left(x_1^2-x_2^2\right)^2}$	30	-100:100	0.0015
F8: Schwefel 1.2	$f_8(x)=\sum_{i=1}^{D}\left(\sum_{j=1}^{D} x_j\right)^2$	30	-100:100	0
F9: Schwefel 2.21	$f_9(x)=\max_{1\le i\le D}\left\lvert x_i\right\rvert$	30	-100:100	0
F10: Schwefel 2.22	$f_{10}(x)=\sum_{i=1}^{D}\left\lvert x_i\right\rvert+\prod_{i=1}^{n}\left\lvert x_i\right\rvert$	30	-10:10	0
F11: De Jong	$f_{11}(x)=\sum_{i=1}^{D} i x_i^4$	30	-5.12:5.12	0
F12: Axis Parallel	$f_{12}(x)=\sum_{i=1}^{D} i x_i^2$	30	-5.12:5.12	0

Appendix A2: Multimodal benchmark functions.

Multimodal functions	Test function	Dim	Range	f_{min}
F13: Easom	$f_{13}(x)=-\cos(x_i)\cos(x_{i+1})\exp\left(-(x_i-\pi)^2-(x_2-\pi)^2\right)$	30	-100 : 100	-1
F14: Griewank	$f_{14}(x)=\sum_{i=1}^{n}\frac{x_i^2}{4000}-\prod_{i=1}^{n}\frac{\cos(x_i)}{\sqrt{i}}+1$	30	-600:600	0
F15: Rastrigin	$f_{15}(x)=\sum_{i=1}^{n}\left[x_i^2-10\cos(2\pi x_i)+10\right]$	30	-5.12:5.12	0
F16: Test-tube Holder	$f_{17}(x)=-4\left\lvert\exp\left(\left\lvert\cos\left(\frac{x_i^2+x_{i+1}^2}{200}\right)\right\rvert\sin(x_i)\cos(x_{i+1})\right)\right\rvert$	30	-10:10	-10.87
F17: Egg Holder	$f_{18}(x)=\sum_{i=1}^{n-1}x_{i+1}+47\sin\left(\sqrt{\left\lvert\frac{x_i}{2}+x_{i+1}\right\rvert}\right)-x_i\sin\left(\sqrt{\lvert x_i-x_{i+1}-47\rvert}\right)$	30	-512:512	959.64
F18: Damavandi	$f_{19}(x)=\left(\sum_{i=1}^{n}\left(1-\left\lvert\frac{[\sin(\pi x_i-2)\sin(\pi x_{i+1}-2)]}{\pi^2(x_i-2)(x_{i+1}+1)-2}\right\rvert\right)^5 * \left(2+(x_i-7)^2+2(x_{i+1}-7)^2\right)\right)/n$	30	0:14	0
F19: Cross-leg Table	$f_{20}(x)=\sum_{i=1}^{n}\frac{-1}{\left(\left\lvert\exp\left(\left\lvert 100-\frac{\sqrt{x_i^2+x_{i+1}^2}}{\pi}\right\rvert\right)\sin(x_i)\sin(x_{i+1})\right\rvert+1\right)^{0.1}}$	30	-10:10	0
F20: Bucking 4	$f_{21}(x)=100x_{i+1}^2+0.01\lVert x_i+10\rVert$	30	-15:-5 -3:3	0
F21: Bucking 6	$f_{22}(x)=100\sqrt{\lVert x_{i+1}-0.01x_i^2\rVert}+0.01\lVert x_i+10\rVert$	30	-15:-5 -3:3	0
F22: Cosine	$f_{23}(x)=-0.1\sum_{i=1}^{n}\cos(5\pi x_i)-\sum_{i=1}^{n}x_i^2$	30	-1:1	0.4
F23: Modified Rosenbrook	$f_{25}(x)=74+100(x_{i+1}-x_i^2)+(1-x)^2-400\exp\frac{-(x_i-1)^2+(x_{i+1}+1)^2}{0.1}$	30	-2:2	0
F24: Schwefel 2.26	$f_{26}(x)=-\frac{1}{D}\sum_{i=1}^{D}x_i\sin\sqrt{\lvert x_i\rvert}$	30	-500:500	-418.98

Appendix A3: Constant-dimension benchmark functions.

Constant-Dimension functions	Test function	Dim	Range	f_{min}
F25: Paviani	$f_{27}(x)=\sum_{i=1}^{10}\left[(\ln(x_i-2))^2+(\ln(10-x_i))^2\right]-\left(\prod_{i=1}^{10}x_i\right)^{0.2}$	10	2.00:10	-45.778

F26: Modified Ackley	$f_{28}(x)=\sum_{i}^{D}\left(\exp^{-0.2}\sqrt{x_i^2+x_{i+1}^2}+3\left(\cos(2x_i)+\sin(2x_{i+1})\right)\right)$	2	-35:35	-3.9172
F27: Camel 6 Hump	$f_{29}(x)=\left(4-2.1x_i^2+\frac{x_i^4}{3}\right)x_i^2+x_ix_{i+1}+\left(4x_{i+1}^2-4\right)x_{i+1}^2$	2	-5:5	-1.0316
F28: Branin Rcos	$f_{30}(x)=\left(x_{i+1}-\frac{5.1x_i^2}{4\pi^2}+\frac{5x_i}{\pi}-6\right)^2+10\left(1-\frac{1}{8\pi}\right)$ $*\cos(x_i)\cos(x_{i+1})\ln\left(x_i^2+x_{i+1}^2+1\right)+10$	2	-5:10 0:15	0.3978
F29: Hartman 3	$f_{31}(x)=-\sum_{i=1}^{4}c_i\exp\left[-\sum_{j=1}^{3}a_{ij}\left(x_j-p_{ij}\right)^2\right]$	3	1:3	-3.8627
F30: Hartman 6	$f_{32}(x)=-\sum_{i=1}^{4}c_i\exp\left[-\sum_{j=1}^{6}a_{ij}\left(x_j-p_{ij}\right)^2\right]$	6	0:1	-3.3223
F31: Shekel 5	$f_{33}(x)=-\sum_{i=1}^{5}\frac{1}{\sum_{j=1}^{4}\left(x_j-a_{ij}\right)^2+c_i}$	4	0:10	-10.149
F32: Shekel 7	$f_{34}(x)=-\sum_{i=1}^{7}\frac{1}{\sum_{j=1}^{4}\left(x_j-a_{ij}\right)^2+c_i}$	4	0:10	-10.399
F33: Shekel 10	$f_{35}(x)=-\sum_{i=1}^{10}\frac{1}{\sum_{j=1}^{4}\left(x_j-a_{ij}\right)^2+c_i}$	4	0:10	-10.531
F34: Perm	$f_{36}(x)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(j^i+0.5\right)\left(\frac{x_j}{j}\right)^{i-1}\right]^2$	4	-4:4	0
F35: Goldstein Price	$f_{37}(x)=$ $\left[1+\left(x_i+x_{i+1}+1\right)^2\left(19-14x_i+3x_i^2-14x_{i+1}+6x_ix_{i+1}+3x_{i+1}^2\right)\right]*$ $\left[30+\left(2x_i-3x_{i+1}\right)^2\left(18-32x_i+12x_i^2+48x_{i+1}+36x_ix_{i+1}+27x_2^2\right.\right.$	2	-2:2	3
F36: Langerman 5	$f_{38}(x)=-\sum_{i=1}^{m}c_i\exp\left(-\frac{1}{\pi}\right)\sum_{j=1}^{D}\left(x_j-a_{ij}\right)^2\cos\left(\pi\sum_{j=1}^{D}\left(x_j-a_{ij}\right)^2\right)$	10	0:10	-1.4

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1537-1544 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Software-Defined Network Approach for The Best Hospital Localization Against Coronavirus (COVID-19)

Bilal Babayigit[1], Eda Nur Hascokadar[2,*]

[1]*Department of Computer Engineering, Erciyes University, Kayseri, 38039, Turkey*

[2]*Graduate School of Natural and Applied Sciences, Erciyes University, Kayseri, 38039, Turkey*

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 04 December, 2020
Online: 25 December, 2020

Keywords:
Software-Defined Network
Multi-Controller Placement
Genetic Algorithm
COVID-19
Pandemic Hospital

ABSTRACT

Traditional networks have difficulty in meeting the technological developments and the continuous increase in the size of data to be processed. Software-Defined Network (SDN) approach has emerged as an alternative to traditional networks. SDN separates the control and data planes from each other and manages the network over the control plane with flexibility and cost advantages. In networks with large data flow, SDN with multiple controllers will be useful to manage the network because a single controller may cause network interruptions and data loss. After deciding on the use of multiple controllers in the SDN approach, problems are encountered with the number of controllers that should be used and the placement of these controllers. Due to the increase in the coronavirus epidemic that has affected the whole world in recent months, the need for pandemic hospitals has increased. However, considering the establishment costs of pandemic hospitals, it will not be possible to establish these hospitals in every desired location in the impact area. For this reason, positioning pandemic hospitals at more strategic points will provide these hospitals with a wider coverage area and more functionality at lower costs. In this paper, a genetic algorithm-based SDN is presented for pandemic hospital localization. The number of coronavirus infected people of each city in Turkey is taken into consideration as well as the city distances in the ULAKNET data set within the Topology Zoo database. For the best localization of pandemic hospitals, the Dijkstra algorithm is used to best cover the cities where the coronavirus epidemic is at a minimum distance from the cities. In this paper, the number of controllers is set as 10, and the experimental results are given with maps and graphics.

1. Introduction

This paper is an extension of work originally presented in 2019 4th International Conference on Computer Science and Engineering (UBMK) [1]. In this study, firstly, the coronavirus coefficients of the cities on a certain date are added to the data file of [1] in addition to the coordinates of each city in the Ulaknet data set within the Topology Zoo database. Secondly, the fitness function values are obtained by dividing the coronavirus coefficients of the cities in the selected solution set by the arithmetic averages, as well as the algorithm used to find the minimum distances of the incoming solution clusters to other cities. In this way, while developing random solutions during Genetic Algorithm (GA) generations, solution clusters at the minimum distance to cities are found based on Dijkstra Algorithm (DA), and the solution set with cities with a high number of coronaviruses become the optimum solution set. The aim is to find the locations of pandemic hospitals that are close to the cities on the map, have a high coronavirus coefficient, and best cover cities that need hospitals.

With the continuous development of information technologies, the Internet has turned into a complex and large-scale infrastructure that profoundly affects people's working and lifestyle [2]. However, the complexity of the traditional network makes the networks difficult to manage. SDN resolves these issues based on the principle of separating the data layers. SDN enables easy management of the traditional network seperating data and control planes. By this way, SDN provides many benefits such as simplifying network management, improve network utilization efficiency, support network innovation, reduce network delays.

*Corresponding Author: Eda Nur Hascokadar, edahsckdr@gmail.com

As the network structure expands, a single controller SDN architecture may be difficult to manage a comprehensive network. To increase the scalability and reliability of the network and to avoid a single point of failure, there is a need for a logically centralized [3], physically distributed multi-controller network architecture. In multi-controller architecture, the Controller Placement Problem (CPP) is encountered. With CPP, the number and location of the controllers have a large impact on end-to-end latency reduction and network performance within the network.

2. Related Works

CPP has been intensely studied in the literature. Papers [4-16] explain the different CPP problems in detail for the proper single controller or multiple distributions of controllers and clarify the benefits of SDN.

In [4], the authors present as a solution to CPP and minimize the response delays for a large-scale network using a hybrid GA-PSO algorithm. Using the hybrid algorithm, the best position is obtained in a shorter time compared to GA.

In [5], the authors solve the distributed controllers' placement problem and apply different scenarios in CPP for the distribution of the network load.

CPP with different aspects such as transmission delay, control plane utilization, and controller workload distribution are studied in [6]. A hybrid gradient descent optimization method with GA is implemented in two different network scenarios and effectively established the balance between the use of the control plane and the network response time.

In [7], the authors examine different controller placement models and their accuracies. Other goals of the study include maximizing the controller capacity, reliability, flexibility while minimizing the network latency, deployment cost, and energy consumption. Concerning CPP, the authors present a feature selection technique adopted in the literature to illustrate the search methods used by the controller placement models.

In [8], the authors present a hierarchical K-median algorithm for CPP. Also, by dividing the large number of single-controller network SDN, they attempt to provide efficient control of the area. They also consider the load balancing.

In [9], the authors consider the packets among the controllers and switches for wide area networks, argue that it is important to reduce the propagation delay. They propose a new approach called the Cluster-Based Network Division Algorithm (CNPA). CNPA divides the network into subnets in steps to shorten the propagation delay between the controller and the controller can place the associated switches to a sub-network. They aim at minimizing the network latency, maximizing reliability and durability, minimizing distribution cost and energy consumption. The solutions are searched for CPP with different purposes. They focus on cooperation between controllers for CPP, cost awareness with the use of multiple controllers, and increasing the number of parameters to be based on optimization.

In [10], the authors review the CPP, and analyze the solution techniques and their limitations. The solution techniques are generally based on objective functions. In obtaining the optimum solutions, various factors such as propagation latency between switches and controllers and constraints such as the capacity of controllers and switches are considered.

In [11], the authors present a cost-effective real-time monitoring and recording system to combat the crime committed on the Internet using SDN architecture. In [12], the authors implement a small scale and low-cost SDN design. The performance of two different load balancing algorithms (Round Robin and Dijkstra) is investigated in [13] using round-trip times for addressing the SDN load balancing problem.

In [14], the authors examine the factor of controllers' load into the problem of the controller placement. For this, they define the Capacitated Controller Placement Problem (CCPP). The evaluation shows that the new strategy can significantly reduce the number of required controllers and reduce the load of the maximum-load controller.

In [15], the authors propose a new mathematical model for the CPP in SDNs. The goal of the model is to minimize the cost of the network while considering different constraints. The simulation results show that the model can be used to plan small scale SDN.

In [16], a new expansion model is presented to introduce the network operators plan and to update the SDN infrastructure. For a given network design and switches, the model finds how many, where and which controllers to install to (re-)design the topology of the new network.

The coronavirus also called COVID-19 or SARS-CoV-2, is one of the worst pandemics in recent history [17]. Especially, in recent months, the financial and the social impacts of the COVID-19 pandemic on our lives and economies are very destructive. To stop the spread of COVID-19, billions of people need a vaccine for coronavirus. Many research teams in companies and researchers in universities try to develop vaccines against COVID-19. But it might not be physically possible to make enough vaccine for everyone and optimistically it could be available in 2 year-time but as scientists say it will be a while before the vaccine is as useful as hoped. To this end, governments take measures such as buy blood test kits, rules for wearing masks and social-distancing, flexible working hours, online schools, severe lockdowns, etc. to slow down the spread of COVID-19. Apart from these and besides, governments need effective resource management and coordination. Especially, the key challenges where to place the pandemic hospitals and what should be the number of pandemic hospitals to be placed in a city should be effectively addressed by governments.

The main contribution of this paper is the introduction and evaluation of an SDN-based approach for the best hospital localization against COVID-19. We propose and explain a GA with the DA method that considers the number of coronavirus patients of cities to reduce the insertion end-to-end delay controller. The proposal uses the ULAKNET data set for Turkey and provides the most appropriate placement for the pandemic hospitals. We experimentally show the controllers are located at the closest distance to all other cities and close to the places where the coronavirus outbreak is common, thus reducing the end-to-end delay significantly. Our approach can be used to analyze, validate, and to evaluate the localization of the countries' pandemic hospitals' demand. If more data support is satisfied especially at the level of villages, towns, cities, and districts better prediction models can be obtained.

3. Technical Concepts

3.1. Software-Defined Network (SDN)

SDN is a new networking paradigm emerging to improve the performance of existing network architecture [18]. SDN, by separating the control plane from the data plane, the switches become simple forwarding devices. Network control of SDN is provided with the help of a controller or controllers with logical management in the control layer. Controller, applications, network policy, network configuration, topology management, finding connections, is responsible for the process flow table entry. SDN architecture data, consisting of control and application layers. The communication between the layers is carried out by North and Southbound interfaces. The basic architecture of SDN is shown in Figure 1.

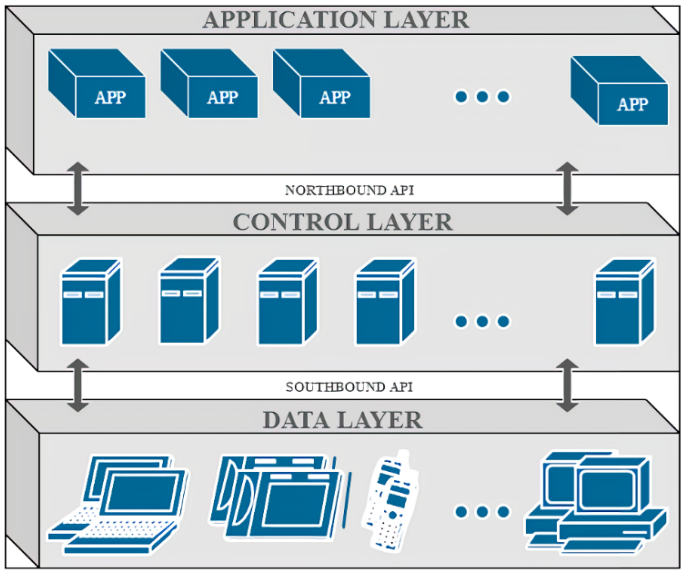


Figure 1: The Basic SDN Architecture

3.2. Multiple Controller Placement Problem

A single controller in the control layer of the SDN architecture may not be sufficient in large-scale networks. There are disadvantages such as a single checkpoint being a single point of failure, bottleneck, and delay in response time. A multi-controller structure is recommended to solve such problems in SDN. After deciding the number of controllers, the significant challenges of multi-controller SDNs to be addressed to increase performance are determining the locations of controllers and determining the number of controllers per module by evaluating the latency.

3.3. Genetic Algorithm (GA)

GA has three steps (selection, crossover, mutation) to have a successful gene. GA process starts with randomly selected individuals. After creating the initial population, it is necessary to determine the fitness function value of each individual. The value of the fitness function is directly proportional to the proximity to the solution according to the type of problem. The higher the fitness function value, the higher the chance of an individual to survive and to reproduce. Individuals that have high fitness function values are selected for crossover. In crossover step, two individuals are selected as a parent and then one or more individuals are produced using different types of crossover operators such as one-point, multi-point, uniform, etc. Then mutation is applied in a probabilistic manner to enrich the gene pool. After these steps, a new population is produced. This procedure is repeated until a specified termination condition is satisfied. The block diagram of GA is given in Figure 2.

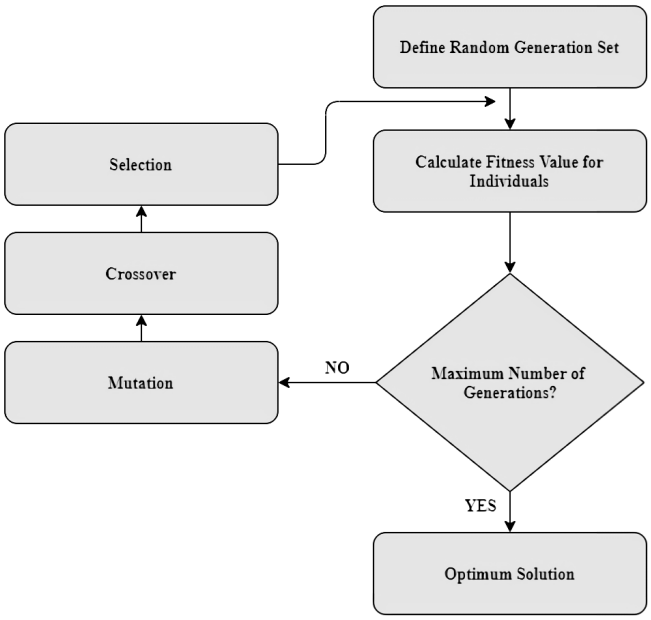


Figure 2: Genetic Algorithm Block Diagram

3.4. Dijkstra Algorithm (DA)

DA is one of the most commonly used algorithms for the shortest path searching. The purpose of DA is to find the shortest distance between nodes in a graph [19]. It uses the greedy search principle as it aims to find the most suitable solution at every step while searching for solutions to the problems it applies to. The input of the DA is the weighted graphs and its output is the shortest path from the origin node to each vertex in the graph.

4. Experiments

In this paper, the hospital placement problem for the coronavirus epidemic affecting the whole world is solved in the SDN infrastructure (see Figure 3) by applying GA and DA approach. In addition to the coordinates of the cities, the coronavirus parameter of each city is added to the ULAKNET database file used.

The ULAKNET data set for Turkey in the Topology Zoo database [20] shown in Figure 4 is used. The study is carried out with Python codes on a machine with Linux based Ubuntu 16.04 operating system, 8 GB RAM, and Core i7, 2.0 GHz processor. The Pycharm platform is used as a compiler for the Python codes.

First, a fixed node map for Turkey using latitude and longitude values in ULAKNET data sets is illustrated in Figure 5. In Figure 5, each city is also represented by a number. For example, 74 number is assigned for İstanbul. The data for İstanbul is as follows:

```
<node id="74">
<data key="d29">1</data>
```

<data key="d30">41.01384</data>
<data key="d31">Turkey</data>
<data key="d32">Red Colour</data>
<data key="d33">74</data>
<data key="d34">28.94966</data>
<data key="d35">Istanbul</data>
</node>

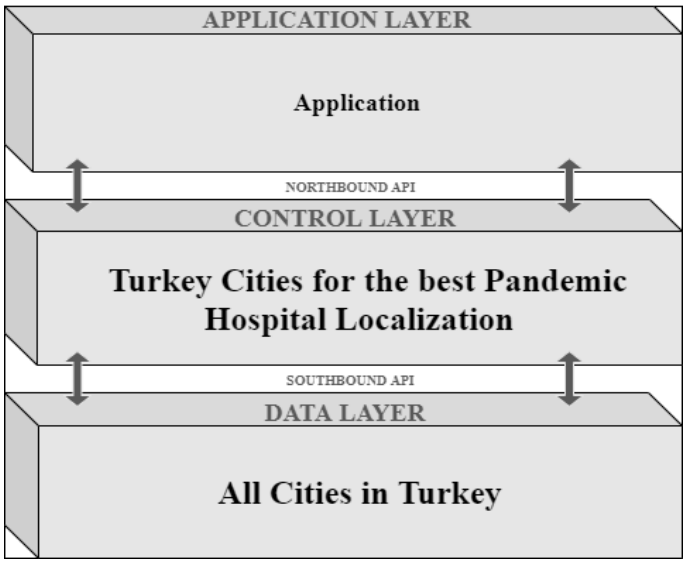


Figure 3: The SDN-based Structure Used for the Best Pandemic Hospital Localization

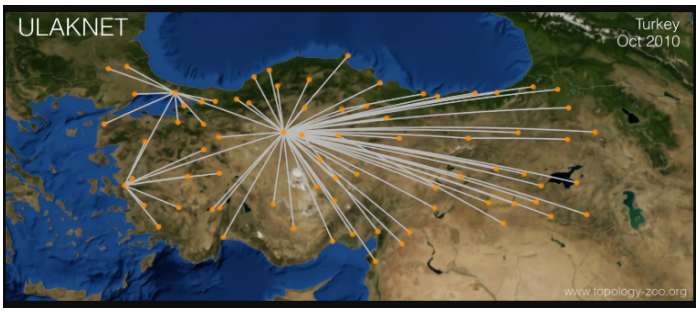


Figure 4: ULAKNET Turkey Map

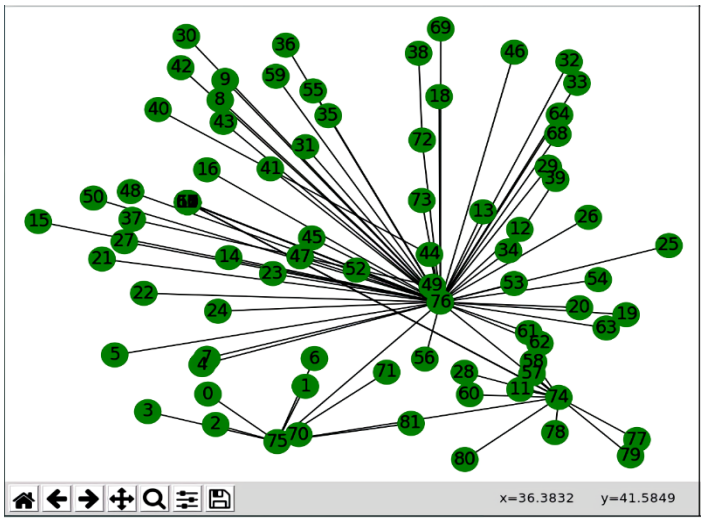


Figure 5: ULAKNET Turkey Map Before the Placement of Multi-Controllers

According to the coordinate data of the cities, the minimum distances of each city to other cities are calculated with DA. A distance matrix obtained for each city is shown in Figure 6.

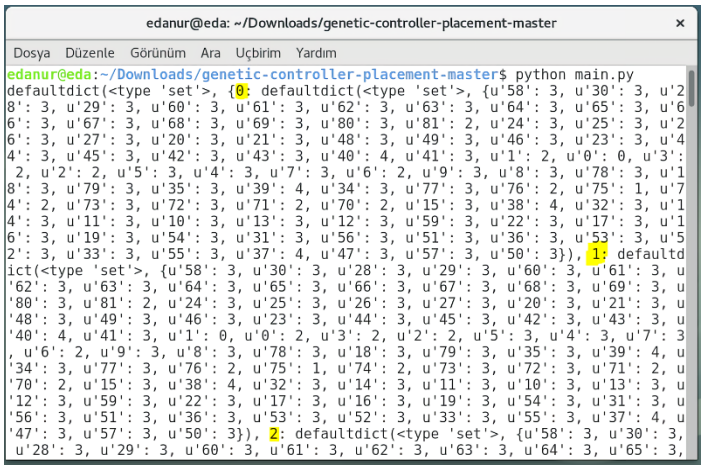


Figure 6: A City Distance Matrix

The number of coronavirus patients is assumed as the coronavirus coefficient of a city. So, the corresponding coronavirus coefficients are added for each city and are shown in Figure 7. The coronavirus patients' data of the cities of Turkey given in Figure 7 are the patients of the date 03/04/2020 [21]. For example, the city number 74 corresponds to Istanbul, and 76 corresponds to Ankara while the coronavirus coefficients of Istanbul and Ankara are 12231 and 860, respectively.

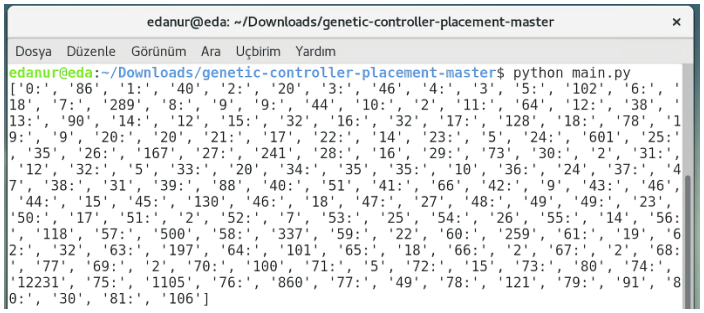


Figure 7: The Coronavirus Matrix of the Cities of Turkey

The parameter setting of GA are listed in Table 1 and the flowchart for the hospital placement problem is presented in Figure 8.

Table 1: GA parameter setting used in the simulations

Parameters	Value
Individuals in the Population	100
The Number of Controllers (Hospitals)	10
Parent Percentage	0.1
Mutation Rate	0.05
Random Selection Rate	0.01
The Number of Generations (Iterations)	1000

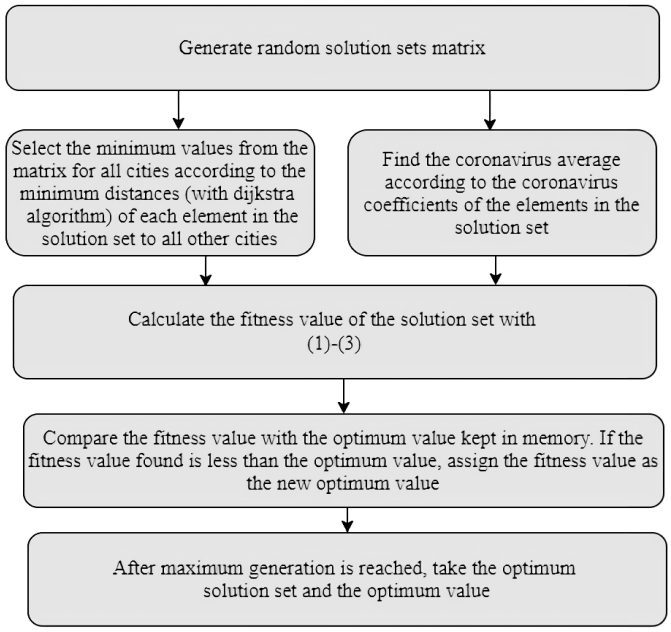


Figure 8: Flowchart for the Hospital Placement Problem

Using the parameters listed in Table 1, 100 random 1x10 solution clusters are created. The randomly generated solutions are illustrated in Figure 9.

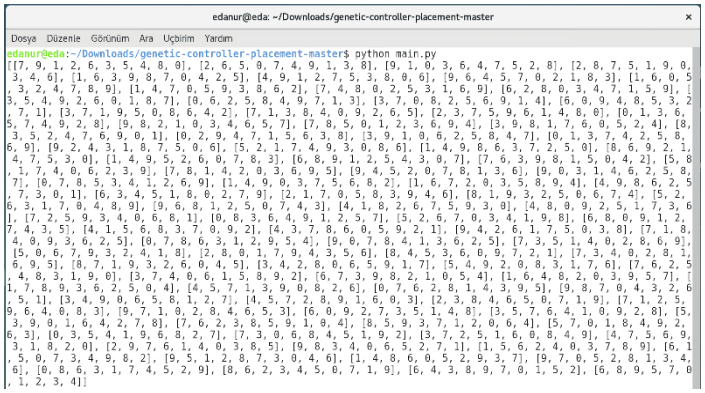


Figure 9: Random Startup Solutions

The fitness of each the random solution sets is calculated using the fitness function. The fitness function has two purposes:

- Choosing solution clusters at minimum distances to cover all cities on the map,
- Choosing solution sets that have higher coronavirus coefficients.

The fitness values of all cities are calculated according to the distance matrix created for each city according to the DA. The city in the random solution set that has a minimum distance from all other cities is selected. After calculating the optimum distance to all cities according to the minimum values in the solution set, the average of the distance from the solution set to all cities is calculated.

For the algorithm used in the simulation: n is the number of the controller (number of pandemic hospitals to be established) and set as 10. x_i is the solution set where $i = 1, 2, 3, \ldots, 10$. h is the weighted graph of cities. $f(x, h)$ is the Dijkstra function. A is the calculated value of the selected solution set with the shortest distances to all other cities. L_Opt and G_Opt are variables for Local Optimum and Global Optimum, respectively.

Equation (1) shows the arithmetic mean of the coronavirus coefficients of the elements in the solution set and (2) takes the average of the values produces according to which of the elements in the selected solution set for each city in the graph of the cities is closest. (2) calculates for the 100 solution sets. After (1) and (2) are calculated, the local optimum value of the selected solution set is obtained by dividing (1) by (2) and (3).

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (1)$$

$$A = mean(min(f(x_i, h), i = 1, \ldots, n)) \quad (2)$$

$$L\_Opt = \frac{A}{\bar{x}} \quad (3)$$

Initially, a global optimum value is defined by (4). In each iteration, the global optimum value is compared with the value in (3). If the value in (3) is better than the global optimum value, the new global optimum value will now be that value (5).

$$G\_Opt = Maxint \quad (4)$$

If 'L_Opt' value is lower than 'G_Opt' then:

$$G\_Opt = L\_Opt \quad (5)$$

The aim of the study is to reach the optimum value when the average distance is minimum and the coronavirus arithmetic average is maximum. Thus, two criteria are taken into consideration for the establishment of pandemic hospitals:

- Solution clusters that are close to the cities on the map and will cover the maximum number of cities
- Solution clusters consisting of cities with a high number of coronavirus and needing hospitals more

Each experiment is run for 1000 iterations. After maximum iteration is reached, the optimum solution set and optimum value are taken. 30 experiments are conducted in the study. The best fitness values obtained of the 30 experiments are shown in Table 2. The number of the experiment and the solution set (cities) obtained during the experiment are also listed in the first and the second column of the Table 2.

The iterations and the convergence graph of the 9th experiment is shown in Figures 10 and 11, respectively. The proposed algorithm is converged faster, as iterations proceed the output gets closer to the optimum value at every iteration.

Each number in the solution set that gives the global optimum value corresponds to a city and represents the most suitable cities for the construction of pandemic hospitals. In Figure 12, the cities obtained from the 9th experiment, which gives the most optimum solution set in Table 2, are shown on the map obtained from the ULAKNET data set within the Topology Zoo database. According to the results we have obtained, the pandemic hospitals that should

be placed in the cities [74, 57, 8, 28, 0, 39, 15, 74, 18, 45]. The names of these cities obtained are as follows:

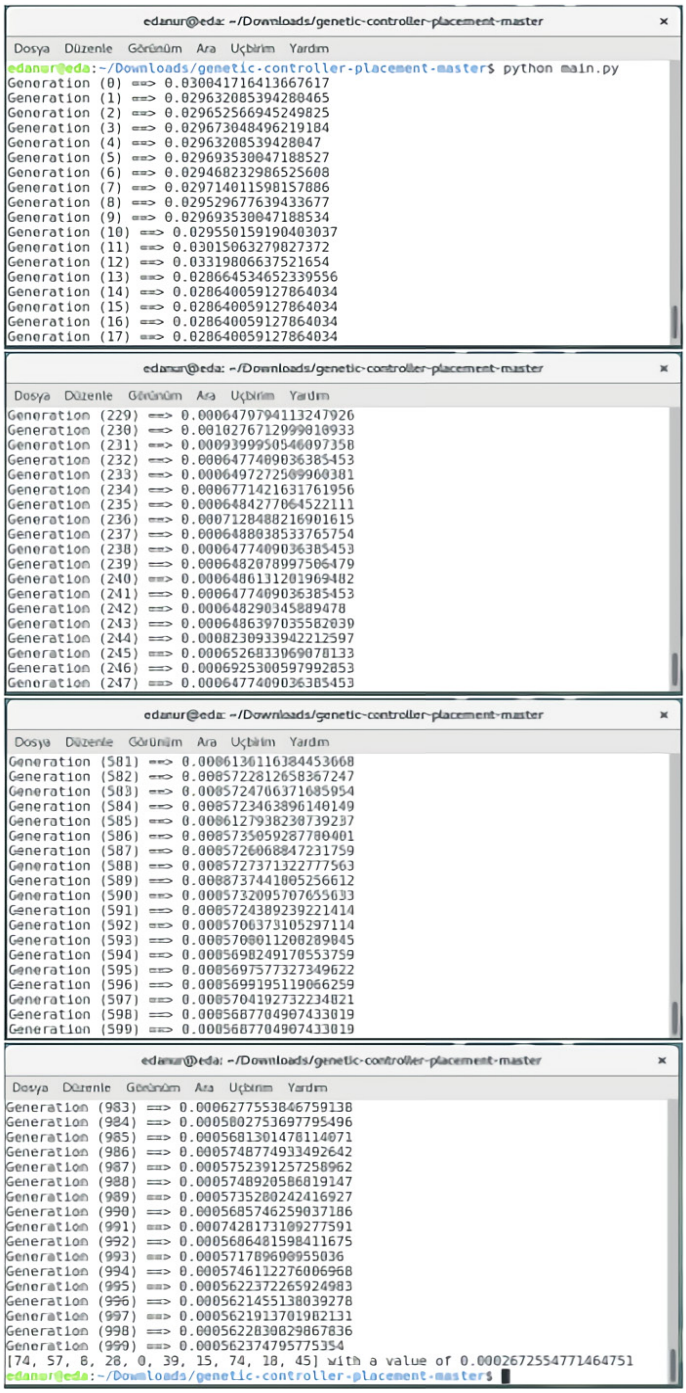


Figure 10: The Optimum Solution for the 9th Experiment

Table 2: The Experimental Results

Exp. No	Cities	Best Fitness Value
9	[74, 57, 8, 28, 0, 39, 15, 74, 18, 45]	0.0002672554771464751
12	[79, 24, 37, 76, 40, 1, 74, 75, 58, 39]	0.0003217251423851137
1	[76, 74, 57, 58, 74, 38, 28, 39, 75, 40]	0.0003219547001035476
14	[60, 42, 39, 49, 12, 40, 74, 24, 76, 33]	0.00032287621623980107
21	[65, 10, 24, 76, 74, 75, 13, 44, 63, 58]	0.0003229917768039527
11	[28, 57, 24, 43, 75, 45, 38, 19, 76, 74]	0.00032310742011823983
5	[51, 74, 52, 62, 53, 57, 58, 38, 7, 76]	0.0003235352480868282
8	[74, 24, 63, 57, 40, 3, 42, 76, 79, 59]	0.0003235708226569537
28	[74, 40, 58, 24, 76, 7, 57, 54, 75, 34]	0.0003238789470047675
17	[24, 58, 18, 75, 76, 37, 40, 7, 74, 39]	0.00032415194841603587
18	[18, 75, 7, 57, 58, 37, 44, 74, 24, 74]	0.00032422337693777505
20	[76, 74, 60, 74, 14, 2, 22, 40, 57, 78]	0.00032422337693777505
10	[44, 18, 76, 34, 58, 5, 74, 74, 30, 34]	0.0003243383497664764
22	[76, 24, 74, 74, 27, 37, 38, 58, 75, 57]	0.0003245685402205335
13	[75, 60, 74, 39, 30, 40, 57, 27, 24, 74]	0.0003247990576496674
29	[18, 40, 67, 63, 24, 81, 64, 25, 74, 76]	0.00032491443919767794
19	[64, 38, 24, 75, 76, 74, 62, 16, 39, 50]	0.0003249690401117191
15	[74, 24, 40, 49, 16, 59, 11, 2, 76, 7]	0.0003250861038869755
7	[75, 76, 39, 74, 24, 57, 46, 49, 7, 63]	0.00032526107622384904
30	[57, 24, 75, 58, 40, 74, 76, 7, 27, 74]	0.0003258893348427155
23	[74, 39, 49, 57, 68, 38, 7, 75, 76, 58]	0.00032590791780073814
2	[76, 58, 35, 74, 37, 57, 75, 60, 17, 27]	0.0003262334629682094 4

24	[7, 1, 27, 74, 76, 75, 39, 24, 58, 40]	0.0003263054392941361
16	[74, 39, 40, 60, 57, 38, 74, 75, 34, 27]	0.00032708917158037114
6	[24, 81, 0, 57, 37, 41, 74, 74, 39, 75]	0.0003281787392685552
3	[58, 75, 67, 76, 55, 31, 74, 7, 39, 60]	0.00032819733296483107
26	[76, 52, 60, 76, 24, 27, 23, 74, 58, 7]	0.00032819733296483107
4	[58, 9, 76, 57, 49, 39, 61, 79, 74, 63]	0.0003696841205081152
27	[27, 37, 39, 57, 39, 80, 74, 76, 51, 7]	0.00037254145084697076
25	[39, 58, 17, 75, 76, 75, 57, 7, 40, 24]	0.0005330035817840696

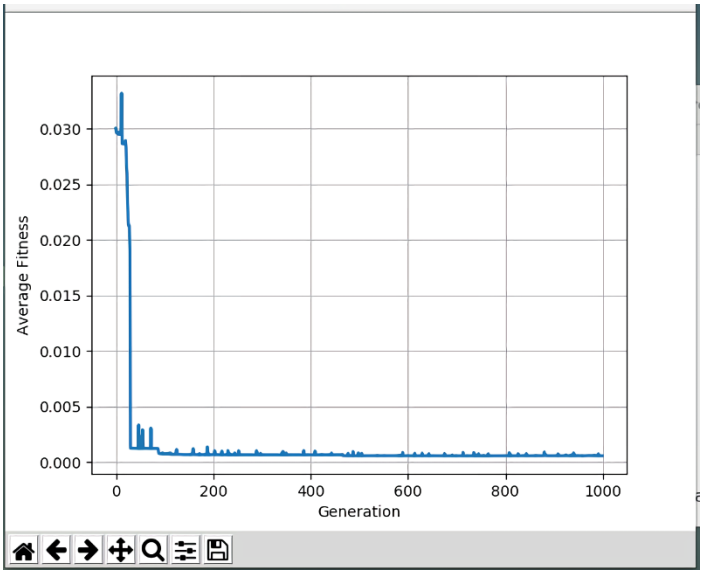


Figure 11: The Convergence Graph of the 9th Experiment

74: İstanbul;
57: Kocaeli;
8: Çanakkale;
28: Bilecik;
0: Denizli;
39: Ordu;
15: Hatay;
74: İstanbul;
18: Erzurum;
45: Kayseri

and demonstrated in the SDN structure in Figure 13. The observation is that in the solution set, there are two Istanbul cities. This corresponds Istanbul needs two pandemic hospitals. However, the capital Ankara is not in the obtained cities. But 24 of 30 experiments obtained Ankara. The cities are not obtained based on their economic, social, and cultural levels. The values obtained from the experiments should be evaluated together with the results obtained from the algorithm of all cities in the solution set, not as a single city.

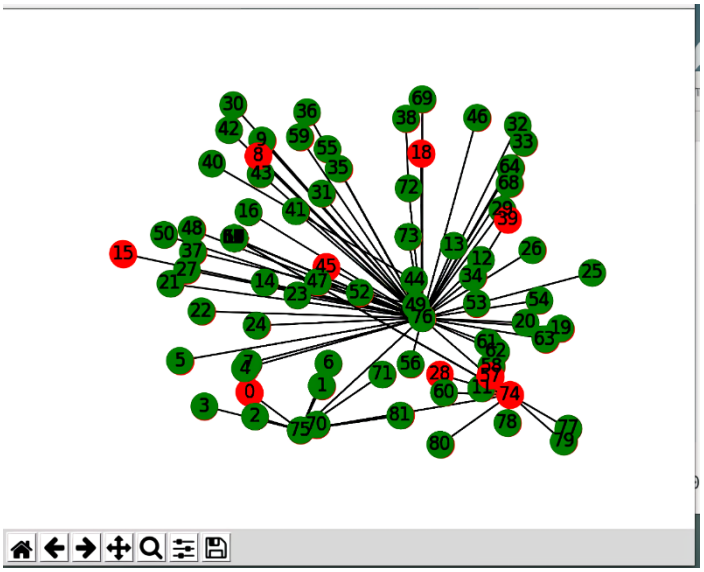


Figure 12: The Best Hospital Localization

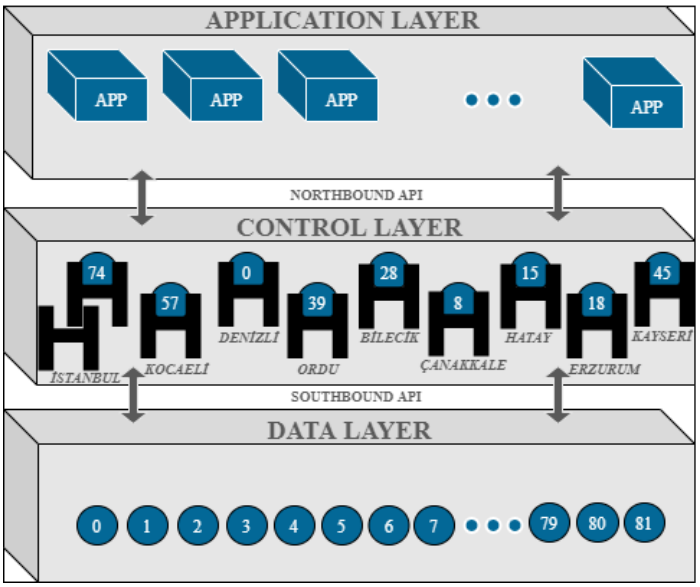


Figure 13: The Obtained Cities for the Pandemic Hospitals in the SDN Structure

5. Conclusion and Future Study

In this paper, we have studied the placement of the pandemic hospitals which should be established against the coronavirus. As it is not possible to establish the pandemic hospitals in every desired location due to the high cost of setting up the pandemic hospitals, the placement of the pandemic hospitals is of great importance. To solve the pandemic hospital placement problem, in this paper, an SDN-based solution has been proposed by applying a GA with DA method according to the distance and coronavirus parameters to the maps obtained from the ULAKNET data set.

From the experimental results, the hospital location found has the closest distance to cover all cities in the selected data set and in places where there is a high coronavirus. Ten pandemic hospitals are placed with maximum coverage.

We need more data at the level of villages, towns, cities, and districts. But now, data for coronavirus in many countries of the World are still limited and even it is very difficult to get the shared data of the countries. Such better data will help us to achieve better prediction models as future work.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank the Erciyes University Scientific Research Projects Coordination Unit (ERU/BAP). This work is supported by the Erciyes University Scientific Research Projects Coordination Unit (ERU/BAP) under project codes FYL-2019-9299.

References

[1] B. Babayiğit, B. Ulu, E.N. Hasçokadar, "Solving multi-controller placement problem in software defined networks with a genetic algorithm," in 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 666–670, 2019, doi: 10.1109/UBMK.2019.8907199.

[2] J. Lu, Z. Zhang, T. Hu, P. Yi, J. Lan, "A survey of controller placement problem in software-defined networking," IEEE Access, **7**, 24290–24307, 2019, doi: 10.1109/ACCESS.2019.2893283.

[3] Z. Guo, R. Liub, Y. Xuc, A. Gushchind, A. Walide, "STAR: Preventing flow-table overflow in software-defined networks," Computer Network, **125**, 15–25, 2017.

[4] L. Liao, V.C.M. Leung, "Genetic algorithms with particle swarm optimization based mutation for distributed controller placement in SDNs," in IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), 1-6, 2017, doi: 10.1109/NFV-SDN.2017.8169836.

[5] N. Mouawad, R. Naja, S. Tohme, "Optimal and dynamic SDN controller placement," in 2018 International Conference on Computer and Applications (ICCA), 1-9, 2018, doi: 10.1109/COMAPP.2018.8460361.

[6] V. Huang, G. Chen, Q. Fu, E. Wen, "Optimizing controller placement for software-defined networks," in International Conference on Cloud Computing and Big Data, Macau, China, 224-232, 2019.

[7] I.O. Adebayo, M.O. Adigun, P. Mudali, "Feature selection strategies for the controller placement problem in SDNs: A Review," in 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 1-5, 2018, doi: 10.1109/ICABCD.2018.8465404.

[8] H. Kuang, Y. Qiu, R. Li , X. Liu, "A hierarchical K-Means algorithm for controller placement in SDN-based WAN architecture," in 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, 263–267, 2018, doi: 10.1109/ICMTMA.2018.00070.

[9] G. Wang, Y. Zhao, J. Huang, W. Wang, "The controller placement problem in software defined networking: a survey," IEEE Network, **31**(5), 21-27, 2017, doi: 10.1109/MNET.2017.1600182.

[10] A.K. Singh,S. Srivastava, "A survey and classification of controller placement problem in SDN," International Journal of Network Management, **28**(3), 1-25, 2018, doi:10.1002/nem.2018.

[11] B. Babayigit, S. Karakaya, "Cost-effective logging using SDN architecture," in International Conference on Advanced Technologies, Computer Engineering and Science , Karabuk, Turkey, 2018.

[12] B. Babayigit, S. Karakaya, B. Ulu, "An implementation of software defined network with Raspberry Pi," in 2018 IEEE 26th Signal Processing and Communications Applications Conference, Izmir, Turkey, 1-4, 2018, doi: 10.1109/SIU.2018.8404546.

[13] B. Babayigit, B. Ulu, "Load balancing on software defined networks," in 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2018 , Kizilcahamam, Ankara, Turkey, 1-4,2018, doi: 10.1109/ISMSIT.2018.8567070.

[14] G. Yao, J. Bi, Y. Li, L. Guo, "On the capacitated controller placement problem in software defined networks," IEEE Communications Letters, **18**(8).1339-1342, 2014, doi: 10.1109/LCOMM.2014.2332341.

[15] A. Sallahi, M. St-Hilaire, "Optimal model for the controller placement problem in software defined networks," IEEE Communications Letters, **19**(1), 30-33, 2015, doi: 10.1109/LCOMM.2014.2371014.

[16] A. Sallahi, M. St-Hilaire, "Expansion model for the controller placement problem in software defined networks," IEEE Communications Letters, **21**(2), 274-277, 2017, doi: 10.1109/LCOMM.2016.2621746.

[17] www.who.int/covid-19, Access Date: June 2020.

[18] Y. Zhang, L. Cui, W. Wang, Y. Zhang, "A survey on software defined networking with multiple controllers," Journal of Network and Computer Applications, **103**, 101-118, 2018.

[19] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numerische Mathematik, **1**(1), 269–271, Dec 1959. [Online]. Available: https://doi.org/10.1007/BF01386390

[20] Topology Zoo. www.topology-zoo.org. Access Date: June 2020.

[21] https://ajansspor.com/haber/hangi-ilde-ne-kadar-corona-virus-vakasi-var-vefat-sayilari-362702, Access Date: June 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 1545-1553 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

NemoSuite: Web-based Network Motif Analytic Suite

Wooyoung Kim*, Yi-Hsin Hsu, Zican Li, Preston Mar, Yangxiao Wang

Computing and Software Systems, School of STEM, University of Washington Bothell, WA 98011, USA.

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 19 December, 2020
Online: 25 December, 2020

Keywords:
Network Motif
Web-GUI
Usability
Efficiency

ABSTRACT

Biological networks represent biological systems, and various graph analysis algorithms have been applied to solve various real-world problems. Network motif analysis, as one of network analyses, is detecting frequently and uniquely over-occurring subgraph patterns in a network. The detection process requires high computational resources, and various tools have been developed to provide efficient solutions. However, they still lack extensible output options and easy accessibility, which restricts substantial scale of experiments for many biological applications. Therefore, we provide NemoSuite (Network Motif in a Suite) as a web-interactive tool for detection and analysis of network motifs. Including both of network-centric and motif-centric approaches, it is a greatly accessible tool emphasizing on efficiency, usability, and extensibility.

1 Introduction

Systems biology focuses more on the interconnections between molecules, such as DNA, RNA, or proteins, than individual components. Biological systems are often represented as biological networks, including protein-protein interaction (PPI), metabolic, or gene-regulatory networks. Biological networks, which model the data in -ome level or in ecosystem level, show individual molecules as graph nodes and their interactions as edges. Researchers applied various graph analysis algorithms to understand biological systems in networks, and we are focusing on network motif analysis as one of graph algorithms. Here, we present an online program extending an original work presented in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [1], which introduces NemoMapPy as an online tool to detect network motifs.

Network motifs are defined as statistically over-recurring subgraph patterns in a network, which are assumed to have biological significance in biological systems. They are defined in line with "sequence motifs" that are playing a key role in analyzing DNA/RNA or protein sequences with the goal of, for example, controlling the translation of the encoding mRNA, characterizing protein families, detecting DNA binding sites, or controlling tumor suppressor gene expression.

Analysis of network motifs have applied to various applications such as predicting protein-protein interactions [2], determining protein functions [3], detecting breast-cancer susceptibility genes [4], investigating evolutionary conservation [5, 6], and discovering essential proteins [7, 8]. Furthermore, a broad spectrum of network motif applications has been explored: "motif clustering" [9], "motif themes" [10], "relative graphlet frequency distances" [11], "motif modes" [12], and "MotifScores" [13].

However, currently existing tools lack the easy accessibility, and flexible scalability, which restrict their range of applications, and thereupon cause some amount of faithlessness on the analysis [14]. Most tools are unable to collect instances of network motifs and lack of functionality, usability, and extensibility. Therefore, we provide NemoSuite (Network Motif in a Suite) to resolve these problems: it includes Nemo and NemoMapPy programs. Nemo program implements a network-centric approach using NemoLib [15, 16]. NemoMapPy implements NemoMap [17] algorithm which is one of the motif-centric methods.

The contributions of NemoSuite include the followings: (1) As a web-interactive graphical user interface (GUI) program, it provides easily accessible tools with great efficiency and usability; (2) It provides various output results based on users' choice, which will explore network motif analysis to various real-world applications; (3) It provides both network-centric and motif-centric tools. As far as the authors' knowledge, NemoSuite is the first program providing both network-centric and motif-centric methods as an web-interactive, and GUI format.

The rest of the paper starts providing the review of various detection algorithms and tools in Section 2. Thereafter, Section

*Corresponding Author, Computing and Software Systems, School of STEM, University of Washington Bothell, WA 98011, USA., kimw6@uw.edu

3 describes NemoSuite, followed by experiments and results in Section 4, and Section 5 respectively, and conclusion in the last section.

2 Related Works

2.1 Network Motif

Recent survey of network motifs showed that there are different types of network motifs [18]. Originally, Milo et al. [19] introduced network motifs as the "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks" in 2002. After discovering the under-represented patterns could also delineate some features of networks, 'anti-motifs' are defined as the subgraphs with significantly less frequencies in the input network compared in a random pool [20]. On the other hand, 'colored-motifs' are introduced as the network motifs in multi-label networks such as metabolic, social or traffic networks [21]. Additionally, Parida proposed 'maximal-motif' as the network motifs with maximal nodes and edges in a given size [22].

However, as we focus on biological networks, we use the original network motifs in this work, whose mathematical definition follows. A network motif M is a connected subgraph of size k which can be found frequently and uniquely in an input network G. To determine the uniqueness, P-value (Eq. (1)) or Z-score (Eq. (2)) of the recurrence of M is calculated within a huge pool of randomly generated graphs.

$$\text{P-value}(M) = \frac{1}{N}\sum_{n=1}^{N} c(n), c(n) = \begin{cases} 1, \text{if } f_R(M) \geq f_G(M). \\ 0, \text{otherwise} \end{cases} \quad (1)$$

$$\text{Z-score}(M) = \frac{f_G(M) - \mu(f_R(M))}{\sigma(f_R(M))} \quad (2)$$

Here, N refers to the number of random graphs, and $f_G(M)$ is the frequency of M in the input G while $f_R(M)$is the frequency of M in the random network R. $\mu(f_R(M))$ is the mean of frequencies of M in the random networks and $\sigma(f_R(M))$ is the standard deviation of frequencies of M in the random networks. For a fair comparison, the random graphs are generated given the same degree sequences as the original graph. In general, a subgraph with P-value less than 0.01 or Z-score greater than 2.0 is considered as a network motif after compared with $1,000$ random networks [23].

2.2 Algorithms

Grow and Kellis classified various algorithms of discovering network motifs into network-centric and motif-centric approaches [24]. The former method first searches all possible patterns from the input graph, then determines which patterns are network motifs through a statistical testing, whereas the latter approach outputs the number of occurrences of the given query pattern in the input network without statistical testing results.

Various algorithms and software programs implement network-centric approach, which include MAVisto [25], MFinder [26], Kavosh [27], NemoFinder [28], NetMODE [29], QuateXelero [30], Motif-Discovery[31], MotifNet [32], ESU algorithm[33], and FANMOD program [34]. Network-centric approach, requiring an input network and the size of motif to search, involves enumerating non-isomorphic k-size subgraph patterns in the original network. It then compares the frequency of each pattern in a number of generated random graphs. The uniqueness of each pattern is determined based on the P-value or Z-score. This approach is specifically useful if the goal is to determine which pattern is a network motif. However, enumeration process limits the size of detectable motifs due to the time-consuming process of scanning for every nodes and edges of the network. Consequently, network-centric algorithms using exhaustive enumeration of k-size subgraphs in the network does not exceed size 8 nodes [24]. Larger size of network motifs can be detected if searched approximately .

Vertices	Number of non-isomorphic subgraphs	
	Undirected	Directed
1	1	1
2	1	2
3	2	13
4	6	199
5	21	9 364
6	112	1 530 843
7	853	880 471 142
8	11 117	1 792 473 955 306
9	261 080	13 026 161 682 466 200
10	11 716 571	341 247 400 399 400 000 000

Table 1: It shows the number of non-isomorphic patterns in undirected or directed graphs increases exponentially with the size of pattern, courtesy of [23].

Focusing on detecting larger size of motifs, Grochow and Kellis introduced a motif-centric approach (GK method), which finds the frequency of each pattern in a given query set by mapping each query graph to all possible location in the input graph [24]. GK method, MODA[35], FASCIA [36], ISMAGS[31], ParaMODA[37], and NemoMap [17] are implementing this motif-centric approach. This approach was able to find motifs of up to 15 nodes with symmetry-breaking technique to significantly reduce isomorphic testing. One disadvantage is that it assumes that the user has pre-knowledge of what patterns are network motifs in the target graph as a set of query patterns have to be provided. Otherwise, this approach is not straightly applicable to finding network motifs. In fact, no algorithms or tools have described how to provide the query set. Therefore, if the goal is to find which pattern is significantly frequent than others, the user has to provide all possible k-size subgraph patterns as a query set to compare, which can suffer heavily since full list of all possible variations of k-size subgraphs grows exponentially with k. Table 1 shows the exponential growth of pattern variations of k-size subgraphs up to 10 nodes, and in the worst-case scenario the application would have to iterate over $341,247,400,399,400,000,000$ of 10-node patterns many of which might not even exist in the target graph.

Table 2 summarizes various aspects of these two approaches. Although both the network-centric and motif-centric approaches have

Table 2: Summary of network-centric and motif-centric approaches.

	Network-centric	Motif-centric
Input	An input network, a motif size k, the number of random graphs to generate	An input network and (a) query pattern(s)
Output	All discovered subgraph pattern's frequency in the input network, Z-score and/or P-values	Each query pattern's frequency in the input network
Processes	Enumerate or sample all possible non-isomorphic subgraphs, and each pattern's frequency is compared in a huge random pool	Search the instances of each query graph in the input network, and its frequency or the set of instances of the pattern is provided
Tools	FANMOD, MAVisto, MFinder, Kavosh, NemoFinder, NetMODE, QuateXelero, MotifNet, Motif Discovery	GK, MODA, FASCIA, ISMAGS, ParaMODA, NemoMap
Benefits	Can determine which patterns are network motifs. Non-existing patterns in the network will not be searched.	Can search large size (more than 8) of a pattern. Relatively fast to collect all instances of a given query graph.
Drawbacks	Exponential computational time growth with increasing size of motifs and networks. Statistical analysis adds more computational burden.	Exponential computing time increase on the number of possible query patterns with increasing size of motifs. It cannot be directly applied for discovering network motifs.

drawbacks, each approach can be respectively applied to different problems with appropriate goals. If the goal is determining which patterns will be network motifs in a target graph, then network-centric approach is appropriate. Motif-centric method would work better for a pattern matching. Therefore, we aimed to provide a tool with both approaches to maximize the usability and efficiency.

2.3 Tools

There are various tools that follow network-centric or motif-centric approaches. After careful review of the tools [18, 38, 39], we found out that most tools are out-dated, dependent on operating systems or other external tools, restricted output options, inaccessible, or lack of usability. As shown in Table 3, we tested the following tools such as mFinder, FANMOD, Kavosh, QuateXelero, Fascia, ISMAGS, MotifNet, and Motif-Discovery, because others are unable to locate, unable to install, out-dated, incompatible with current security setting, or inexecutable.

mFinder, FANMOD, Kavosh, and QuateXelero are network-centric based tools that accept a text file as an input graph, the size of network motifs, and the number of random graphs to generate. While others require a list of edges in the text graph file, Kavosh and QuateXelero additionally require the number of nodes at the first line in the file. Before running these programs, however, since each node should be labelled as an integer, non-integer labels should be converted to integer labels. They provide outputs consisting of each pattern's frequency, and its Z-score or P-value so that users can determine which patterns are network motifs. When we design our program, NemoSuite, we set FANMOD program as an example, since it is a GUI-based program, efficient, with additional filtering options, and can provide the set of instances as well. MotifNet, as a web-program, also succeeds FANMOD, taking two files (edge list and node list), the size of network motif, and the number of random graphs. However, it requires the user to submit the work in a queue, which seemed discontinued the service currently, as we never got the results back. Motif-Discovery, as a cytoscape[31] plug-in, can read string-labelled nodes, visualize the network, and provide the frequency and Z-score for each pattern. Unfortunately, since the Z-scores are significantly different from the ones by other tools, it is unclear how to determine network motifs.

Fascia and ISMAGS are motif-centric based motif discovery tools. Fascia requires two text files. One is the input graph and the other is a query graph. The input graph file should contain the number of nodes, the number of edges, and the list of edges where the node ID's are integers. A query text file is a list of edges. As a command-based program, it provides the number of occurrence of the query pattern in the input network. On the other hand, ISMAGS is cytoscape plug-in, which can visualize the input network. In this work, we excluded other tools including MODA, since we failed to install or run them properly.

Table 3 lists the tools, approach, year of publication, the input options, the output options, interface, and shortcomings. In general, current available tools lack of visualization, usability, easy accessibility, and comprehensive results. To overcome the shortcomings of other tools, we provide NemoSuite as a comprehensive and efficient web-interactive program.

3 Methods: NemoSuite

We developed 'NemoSuite' (Network Motif in a suite) to provide a unified access to network motif analysis including network-centric and motif-centric approaches. NemoSuite consists of two separate programs: 'Nemo' and 'NemoMapPy.' Nemo program employs a network-centric approach using a NemoLib [15, 16] that improves ESU (Enumerate SUbgraphs) algorithm [33]. NemoMapPy program follows a motif-centric method by implementing NemoMap [17] algorithm in Python and Flask framework. Figure 5 shows the NemoSuite homepage with Nemo and NemoMapPy program pages. The program is available at https://bioresearch.css.uwb.edu/biores/NemoSuite/.

Table 3: Network motif detection tools are compared based on their approach (network-centric vs. motif-centric), shortcoming, year, input, output, and interface options.

Tool	Approach	Year	Input	Output	Interface	Shortcoming
mFinder	network	2002	edge list file (integer ID)	text file: freq, z-score, p-value	command-line	restricted input and output options, command-line
FANMOD	network	2006	edge list file (integer ID)	html: freq, z-score, p-value, visual. Instance set	GUI	restricted input and output options
Kavosh	network	2009	#n,edge list file (integer ID)	text file: freq, z-score, motif matrix	command-line	restricted input and output options, command-line
QuateXelero	network	2013	#n,edge list file (integer ID)	text file: freq, z-score, motif matrix	command-line	restricted input and output options, command-line
Fascia	motif	2013	#n,#m,edge list file (integer ID)	count	command-line	restricted input and output options, command-line
ISMAGS	motif	2016	cytoscape input	count	cytoscape plug-in	exception error, plug-in
MotifNet	network	2017	edgelist, nodelist	html: freq, z-score, p-value, visual. Instance set	web-interactive	queue-based, no-prompt result
Motif-Discovery	network	2018	cytoscape input	text file: freq, z-score	cytoscape plug-in	restricted output options, significantly different Z-scores from others, cytoscape-dependent

3.1 *Nemo*

Nemo, as shown in Figure 5(b), follows a network-centric approach which includes the steps of enumeration, random graph generation, and statistical analysis. The user should select an input file (see Fig 1 (a) for the file format), along with the directness, the size of motif and the number of random graphs to generate. Nemo program works on both directed and undirected graphs, and the input graph's node can be represented as string or integer. There are three output options: NemoCount, NemoProfile, and NemoCollection, whose formats are introduced by Kim and Haukap [40], and illustrated in Figure 1 (b).

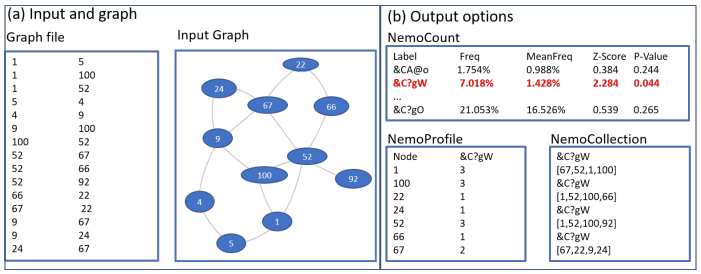


Figure 1: Nemo: (a) shows the input file format with visualized graph on the right; (b) shows the three output options including NemoCount, NemoProfile, and NemoCollection.

'NemoCount' is a default output format which lists the patterns of each detected subgraphs, along with its frequency in the input graph, average frequency in the random graphs, then finally shows its statistical results so that the user can determine which pattern is a network motif. Each pattern is labeled with g6 format which is used to identify isomorphic graphs in the Nauty program [41]. 'NemoProfile' option returns two files. One is a file of *NemoProfile* that contains two dimension of matrix of N rows and M columns, where N is the number of nodes and M is the number of network motifs. For example, from the Figure 1 (b), the node '100' is found in three instances of the motif **&C?gW**, and the node '67' belongs to two instances of the same motif. Wang and Kim showed that NemoProfile can be used as a data feature to detect essential genes [42]. Another file is a *SubgraphProfile*, which is an extended NemoProfile containing not only motifs but also non-motifs in the column. If none of patterns are network motifs, this option provides SubgraphPrifle only. Likewise, 'NemoCollection' also provides two files: *NemoCollection* and *SubgraphCollection*. While NemoCollection provides the set of network motif instances as a file, SubgraphCollection provides instances of all subgraph patterns.

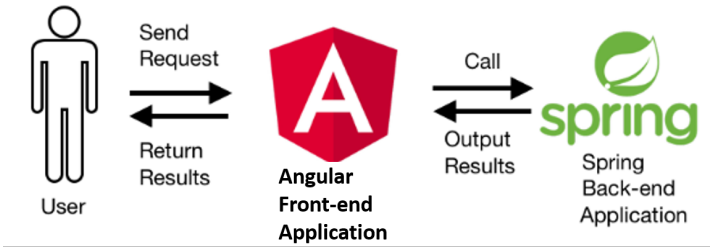


Figure 2: Block diagram for Nemo program.

Nemo program is using an angular application for the front-end, and a spring application for the back-end as shown in Figure 2. The front-end uses the angular framework, which takes advantage of

the Model-View-Controller (MVC) pattern. Developers can easily display the data and the model from the back-end and control the model with the 'appcomponent.' Also, it can be easily deployed with angular. The back-end uses the Spring framework, which produces RESTful APIs. With this simple and intuitive design, developers can easily update modules for additional purposes. Nemo program also has high maintainability with the REST architecture style. Developers can fix any defected part if any, and restart the Spring application on the server while the client application remains untouched. Figure 3 shows the overall architecture of the program.

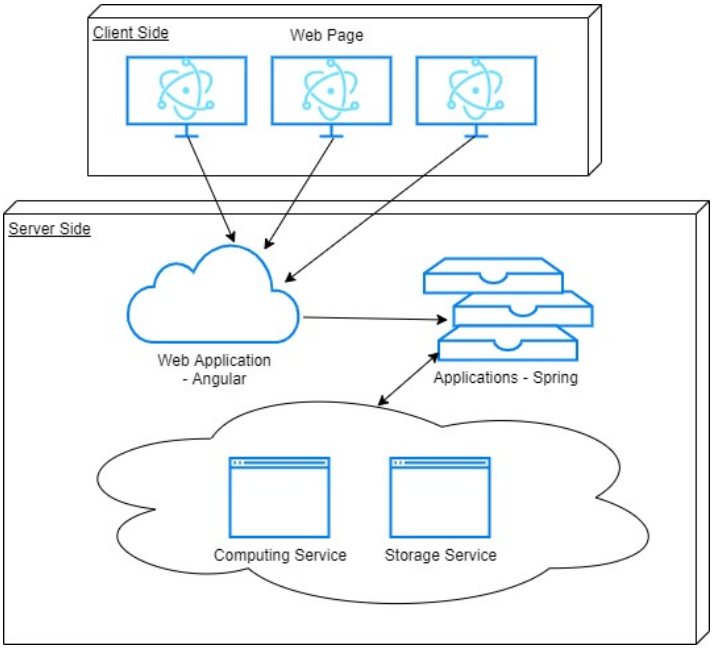


Figure 3: Descriptive architecture of Nemo program.

3.2 *NemoMapPy*

NemoMapPy implements a motif-centric approach which requires an input file, and a query pattern. While other motif-centric tools do not provide statistical results, NemoMapPy provides the frequency and statistical results of the query pattern in the input graph if the number of random graphs are entered, so that whether this pattern is a network motif can be determined with its P-value or Z-score.

NemoMapPy, as an online tool, refined and upgraded the features of NemoMap [17] that improved the performance of ParaMODA [37], which also increased the efficiency of the motif-centric algorithm introduced by Grochow and Kellis (GK) [24], and MODA program [35]. In fact, NemoMap improved the performance of ParaMODA, MODA and GK method by modifying the symmetry-breaking method. It removed the need of maintaining a list of all mappings to prevent duplicate occurrences. It also reduced some of symmetry-breaking conditions, thereafter, contributes on optimizing space and computational time. NemoMap has been proven to be more efficient than ParaMODA, MODA and GK in most cases, exception with some simple query patterns.

NemoMapPy ported NemoMap with additional features on the web [1], and was implemented in Python 3 with the Flask web framework [43]. As a consequence, NemoMapPy can accept streaming text inputs as well as file uploading. The query pattern can be provided as a text input, file uploading, or g6 format. Since Nemo program provides g6 format for detected pattern, the g6 format helps a quick transition from Nemo to NemoMapPy.

Figure 4 illustrates the overall architecture of NemoMapPy. Nginx [44] is used in the first layer, which is a reverse proxy for Gunicorn [45] that redirects incoming requests to Gunicorn with remarkable performance. Gunicorn is a WSGI server which allows multiple applications to run side-by-side in the same process. It handles the load balancing for a program, distributes incoming requests, and communicates with the Flask services accordingly. The back-end of NemoMapPy program is developed by python Flask framework, which takes advantages of simplicity and flexibility. It is convenient for developers to deploy or undeploy it from the server, and the simplified usability make Flask program faster to debug. Here we focused on providing an easy and accessible interface with multiple input options so that users can use NemoMapPy in a variety of situations.

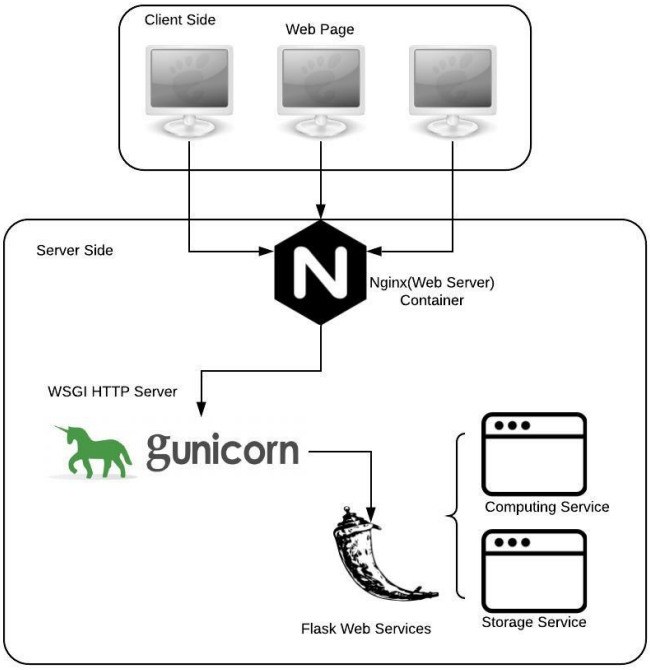


Figure 4: Descriptive architecture of NemoMapPy program

Initially, NemoMapPy only worked on undirected graph and the input file should have integer-labeled vertices [1]. Also, it only provided the frequency of the query pattern as a result. Here, we improved the initial NemoMapPy by adding the following features. (1) NemoMapPy now works both on directed and undirected graphs; (2) the vertices can be integer or string; (3) the query pattern's statistical significance is provided along with its frequency; (4) the input format of query graph pattern has three options. Input text, a file, or g6 format; and (5) the instances of the query pattern is provided as a file.

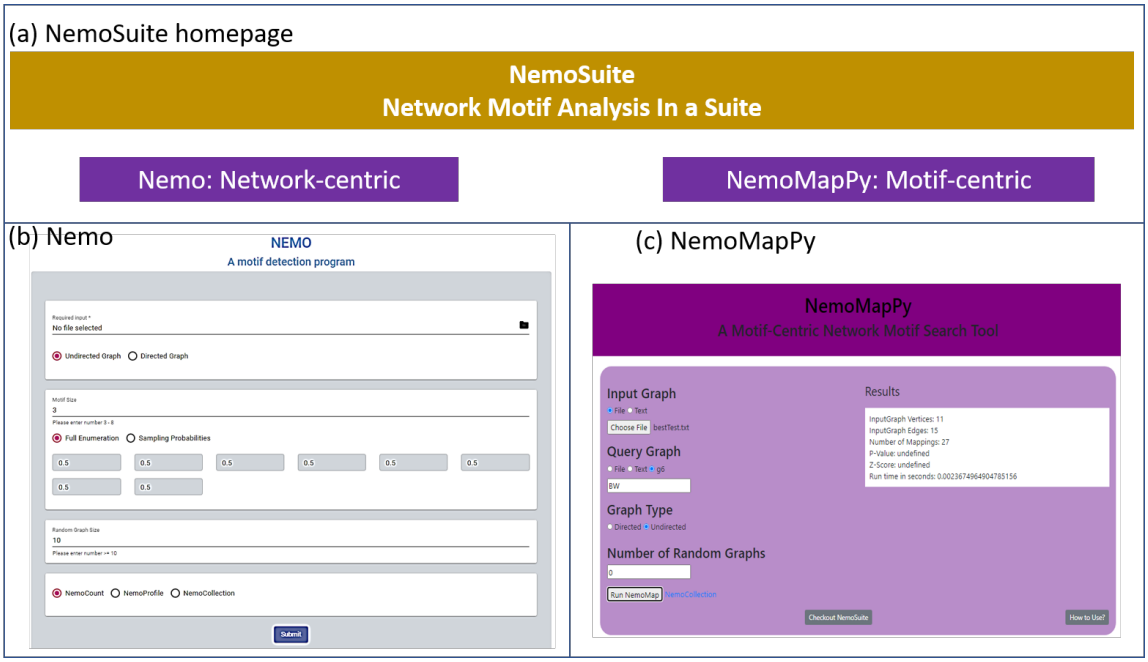


Figure 5: NemoSuite: (a) is the homepage view of NemoSuite program; (b) is the homepage of Nemo program with network-centric method; (c) shows the homepage for NemoMapPy with motif-centric method.

4 Experiments

NemoSuite is an online network motif detection program with both of network-centric and motif-centric methods. It is designed to satisfy the following functional and non-functional requirements.

Functional Requirements

- The user can use the program with any web browser.
- The user can upload the data file, which is an undirected or directed graph.
- The user can use the data file with string or integer type of nodes.
- The user can use both programs to detect the frequency of network motifs (NemoCount).
- The user can use both programs to obtain the instances of network motifs (NemoCollection).
- The user can use the Nemo program to obtain the network motif profile (NemoProfile).

Non-functional Requirements

- The returned results should be clear and easy to understand.
- The running time should be reasonably fast based on the size of the input data.
- The program works for any devices that have access to the internet and have a web browser.

In other words, as a web program, NemoSuite provides easy accessibility for users to detect network motifs, and apply them for various real-world problems. With additional features, the program provides good usability, functionality and computational efficiency. In the following section, we provide results of the comparison of NemoSuite with other programs to demonstrate its computational efficiency and good usability.

We compared Nemo with FANMOD [34] program, since both follow the motif-centric ESU algorithm in essence. Also, after reviewing all other tools, we found FANMOD is efficient and the most usable program among all others. Therefore, we compared Nemo program with FANMOD to evaluate the accessibility, usability, functionality and computational efficiency.

NemoMapPy is implementing a motif-centric approach, which is provided by Grochow and Kellis (GK), called GK method. Unlike network-centric methods, few programs are available that follow GK method, such as, MODA [35], ParaMODA, Fascia [36], ISMAGS[31] and NemoMap. They are available as source codes, cytoscape plug-in, or executable programs. However, the standalone MODA program is out-dated and had many restrictions on running the program, ISMAGS is cytoscape plug-in, and Fascia is a parallel and command-based program. Therefore, we compared NemoMapPy with NemoMap for its efficiency, assuming its easy accessibility and good usability are self-demonstrated with its web-interactive features.

5 Results and Analysis

5.1 Nemo vs. FANMOD

As a web program, Nemo program can be launched with any web-browser. Therefore it is easy to access, and easy to use. FANMOD

program is a stand-alone GUI (graphical-user-interface) program that can be downloaded from a website and it can be installed in Windows, Mac, and Linux operating systems. While both programs have good accessibility, Nemo has a better accessibility as the user can use it anytime and anywhere as long as there is an internet connection.

Nemo and FANMOD programs have good usability too. Usability measures "the capacity of a system to provide an environment for the users to perform the tasks safely, effectively, and efficiently [46]." With Nemo and FANMOD, network motifs are easily detected given an input graph, the size of motif, and the number of random graphs to generate. FANMOD has more options for the random graphs, and the user can see the progress of the task. However, Nemo program performs faster than FANMOD and has more input and output options.

Table 4: Input graph files used for testing

Type	**biological**			**collaboration**		**social**
File	Y2k	Y11k	MIPS	CA-GrQc	CA-HepTH	Face-book
#vertex	988	2401	4545	5241	9875	4545
#edge	2455	11000	12317	14484	25973	12317

We tested the performance of Nemo and FANMOD program with three biological networks, two collaboration networks, and one social network. The detail features of testing networks are described in Table 4. When we tested the run-time for network motif detection with different graphs, the size of network motif is set as four, and the number of random graphs as 100. All tests are performed in Windows OS. Table 5 and Figure 6 provide the comparison results.

Table 5: Nemo vs FANMOD run-time (rounded in seconds) comparison on the size of input graph

	Y2k	**Face-book**	**CA-GrQc**	**CA-HepTH**	**Y11k**	**MIPS**
Nemo	24	54	201	205	517	551
FANMOD	72	248	607	561	2,189	2,859

We can observe that Nemo runs relatively stable as the network size grows, while the run-time of FANMOD program increases abruptly. For the Y11k and MIPS networks, Nemo program runs roughly 5 times faster than FANMOD program.

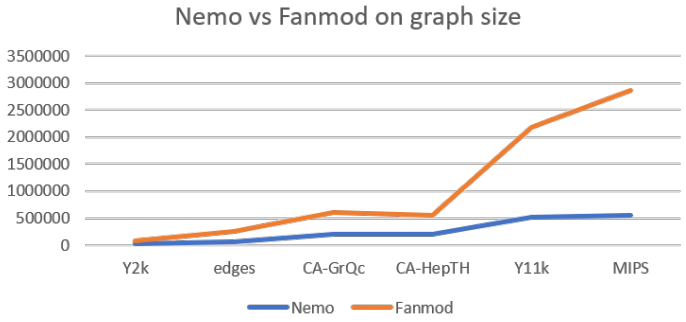


Figure 6: Graph of runtime vs. input graph size between Nemo and FANMOD in millisecond(ms)

We also tested the performance of two programs based on the size of motifs. To save time, we used Y2k data with only 10 number of random graphs, and used Windows OS.

Table 6: Nemo vs FANMOD run-time (in seconds) comparison on the size of motif

	3	**4**	**5**	**6**
Nemo	0.1	2.4	60.8	1,580.5
FANMOD	0.3	7.6	208.5	6,946.9

Table 6 and Figure 7 provide the results. We can observe that Nemo program runs consistently faster than FANMOD program, although the run-time of both programs still increases steeply with the increment of the size of motifs.

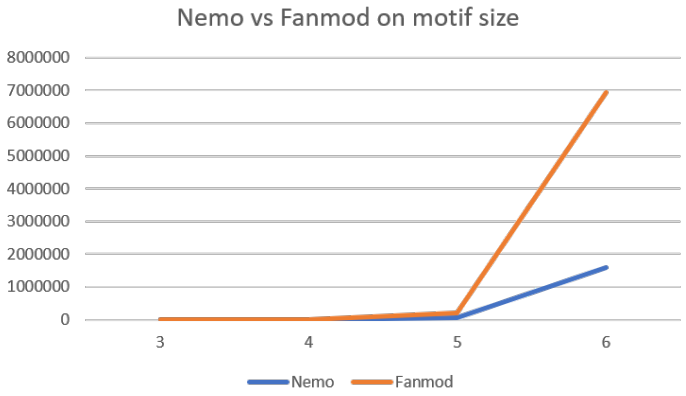


Figure 7: Graph of runtime vs. motif size between Nemo and FANMOD in millisecond(ms)

Through various experiments, we could see that Nemo program provides better accessibility, functionality, usability and computational efficiency than FANMOD program.

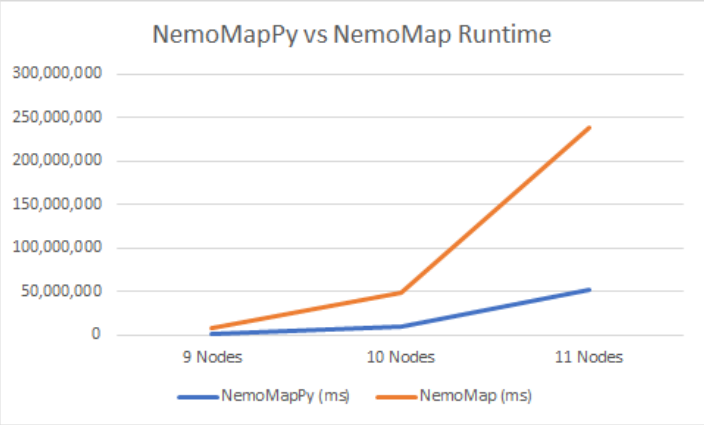


Figure 8: Graph of runtime vs. query graph size between NemoMapPy and NemoMap

5.2 *NemoMapPy vs. NemoMap*

We tested the efficiency of NemoMapPy by comparing with the NemoMap, as NemoMap is known the fastest program with motif-centric approach [17] in most of cases, when not considering parallelization. We tested NemoMapPy with a graph of 1,018 nodes and 1,331 edges, the query graphs that has 9, 10, 11 nodes with average complexity. The complexity of query graph was measured in node-to-edge ratio. We obtained NemoMap source code from the authors of [17] and compiled with Visual Studio 2017. We compared the performance with run-time.

The testing environment was on a desktop computer in Windows OS. In summary, NemoMapPy is approximately 70% faster than NemoMap on the graphs with average complexity and average density, as shown in Table 7, and Fig. 8.

Table 7: NemoMapPy vs NemoMap run-time (in seconds) Comparison

Query Graph	9 Nodes	10 Nodes	11 Nodes
NemoMapPy	1,728,213	9,923,137	51,904,930
NemoMap	8,599,994	48,674,827	239,088,982
% Reduction	79.90%	79.61%	78.29%
# Mappings	13,882,459	251,584,995	1,070,991,776

6 Conclusion

In this paper, we present NemoSuite that provides a combination of easy accessibility, usability, functionality, extensibility, and efficiency that the current network motif detection software programs lack. More specifically, we believe NemoSuite contribute greatly to development of network motif detection programs in the following reasons.

As a web-interactive program, it is easily accessible, and highly usable tool with great efficiency. The various input and output options enable for the users to apply the results to various real-world problems. It is also the first tool that provides both of network-centric and motif-centric approaches. We also demonstrated that NemoSuite is substantially designed to meet the functional and non-functional requirements, and showed its computational efficiency by comparing with other available tools.

Currently, NemoSuite includes Nemo and NemoMapPy programs to represent network-centric and motif-centric methods. Source codes of both programs are available in public github: Nemo is available at https://github.com/lytbfml/Nemolib_App and NemoMapPy at https://github.com/zicanl/NemoMapPy. We plan to add more online tools implementing recent algorithms such as QuateXelero [30], and utilize distributed computing to improve its efficiency further. Additional future works include adding more utility programs such as motif-visualization, graph file conversion, and generating random networks with various methods.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We thank the Computing and Software Systems, School of STEM, at the University of Washington Bothell for the support of this research.

References

[1] P. Mar, W. Kim, "NemoMapPy: Motif-centric network motif search on a web," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1226–1228, 2019, doi:10.1128/mBio.01377-14.

[2] I. Albert, R. Albert, "Conserved network motifs allow protein-protein interaction prediction." Bioinformatics, **20**(18), 3346–3352, 2004.

[3] J. Chen, W. Hsu, M. L. Lee, S.-K. Ng, "Labeling network motifs in protein interactomes for protein function prediction," Data Engineering, International Conference on, **0**, 546–555, 2007.

[4] Y. Zhang, J. Xuan, B. G. de los Reyes, R. Clarke, H. W. Ressom, "Network motif-based identification of breast cancer susceptibility genes," in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 5696–5699, IEEE, 2008.

[5] G. C. Conant, A. Wagner, "Convergent evolution of gene circuits," Nature genetics, **34**(3), 264–266, 2003, doi:http://dx.doi.org/10.1038/ng1181.

[6] S. Wuchty, Z. N. Oltvai, A.-L. Barabasi, "Evolutionary conservation of motif constituents in the yeast protein interaction network," Nature genetics, **35**(2), 176–179, 2003.

[7] W. Kim, "Prediction of essential proteins using topological properties in GO-pruned PPI network based on machine learning methods," Tsinghua Science and Technology, **17**(6), 645–658, 2012.

[8] W. Kim, M. Li, J. Wang, Y. Pan, "Essential Protein Discovery based on Network Motif and Gene Ontology," in Proceedings of IEEE Bioinformatics and Biomedicine, 470–475, 2011.

[9] R. Dobrin, Q. K. Beg, A.-L. Barabasi, Z. N. Oltvai, "Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network," BMC Bioinformatics, **5**, 10, 2004.

[10] L. Zhang, O. King, S. Wong, D. Goldberg, A. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, F. Roth, "Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network," Journal of Biology, **4**(2), 6, 2005.

[11] N. Przulj, D. G. Corneil, I. Jurisica, "Modeling interactome: scale-free or geometric?" Bioinformatics, **20**(18), 3508–3515, 2004.

[12] W.-P. Lee, B.-C. Jeng, T.-W. Pai, C.-P. Tsai, C.-Y. Yu, W.-S. Tzou, "Differential evolutionary conservation of motif modes in the yeast protein interaction network," BMC Genomics, **7**(1), 89, 2006.

[13] Z.-R. Xie, M.-J. Hwang, "An interaction-motif-based scoring function for protein-ligand docking," BMC Bioinformatics, **11**(1), 298, 2010.

[14] A. S. Konagurthu, A. M. Lesk, "On the origin of distribution patterns of motifs in biological networks," BMC Systems Biology, **2**(1), 1–8, 2008.

[15] A. Andersen, W. Kim, NemoLib: A Java Library for Efficient Network Motif Detection, 403–407, Bioinformatics Research and Applications: 13th International Symposium, ISBRA 2017, Honolulu, HI, USA, May 29 – June 2, 2017, Proceedings, Springer International Publishing, Cham, 2017, doi: 10.1007/978-3-319-59575-7_42.

[16] W. Kim, Z. A. Brader, "NemoLib: Network Motif Libraries for Network Motif Detection and Analysis," in Z. Cai, I. Mandoiu, G. Narasimhan, P. Skums, X. Guo, editors, Bioinformatics Research and Applications, 327–334, Springer International Publishing, Cham, 2020.

[17] T. Huynh, S. Mbadiwe, W. Kim, "NemoMap : Improved Motif-centric Network Motif Discovery Algorithm," Advances in Science, Technology and Engineering Systems Journal, **3**(5), 186–199, 2018, doi:10.25046/aj030523.

[18] S. Yu, Y. Feng, D. Zhang, H. D. Bedru, B. Xu, F. Xia, "Motif discovery in networks: A survey," Computer Science Review, **37**, 100267, 2020, doi: https://doi.org/10.1016/j.cosrev.2020.100267.

[19] M. R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," Science, **298**(5594), 824–827, 2002.

[20] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, "Superfamilies of Evolved and Designed Networks," Science, **303**(5663), 1538, 2004, doi:10.1126/science.1089167.

[21] V. Lacroix, C. G. Fernandes, M. F. Sagot, "Motif search in graphs: application to metabolic networks," IEEE/ACM transactions on computational biology and bioinformatics, **3**(4), 360–368, 2006, doi:10.1109/TCBB.2006.55[doi].

[22] L. Parida, "Discovering Topological Motifs Using a Compact Notation," Journal of Computational Biology, **14**(3), 300–323, 2007.

[23] B. H. Junker, F. Schreiber, Analysis of Biological Networks, Wiley, 2008.

[24] J. A. Grochow, M. Kellis, "Network motif discovery using subgraph enumeration and symmetry-breaking," in Proceedings of the 11th annual international conference on Research in computational molecular biology, RECOMB'07, 92–106, Springer-Verlag, Berlin, Heidelberg, 2007.

[25] F. Schreiber, H. Schwobbermeyer, "Mavisto: a tool for the exploration of network motifs," Bioinformatics, **21**, 3572–3574, 2005.

[26] N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, "Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs," Bioinformatics, **20**, 1746–1758, 2004.

[27] Z. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, A. Masoudi-Nejad, "Kavosh: a new algorithm for finding network motifs," BMC Bioinformatics, **10**(1), 318, 2009.

[28] J. Chen, W. Hsu, M. Lee, S. Ng, "NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, 106–115, 2006.

[29] X. Li, D. S. Stones, H. Wang, H. Deng, X. Liu, G. Wang, "NetMODE: Network Motif Detection without Nauty," PLoS ONE, **7**(12), e50093, 2012, doi: 10.1371/journal.pone.0050093.

[30] S. Khakabimamaghani, I. Sharafuddin, N. Dichter, I. Koch, A. Masoudi-Nejad, "QuateXelero: An Accelerated Exact Network Motif Detection Algorithm," PLOS ONE, **8**(7), 1–15, 2013, doi:10.1371/journal.pone.0068073.

[31] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," Genome research, **13**(11), 2498–2504, 2003.

[32] I. Y. Smoly, E. Lerman, M. Ziv-Ukelson, E. Yeger-Lotem, "MotifNet: a web-server for network motif analysis," Bioinformatics, **33**(12), 1907–1909, 2017.

[33] S. Wernicke, "Efficient Detection of Network Motifs," IEEE/ACM Trans.Comput.Biol.Bioinformatics, **3**(4), 347–359, 2006.

[34] S. Wernicke, F. Rasche, "FANMOD: a tool for fast network motif detection," Bioinformatics, **22**, 1152–1153, 2006.

[35] S. Omidi, F. Schreiber, A. Masoudi-Nejad, "MODA: An efficient algorithm for network motif discovery in biological networks," Genes and Genetic Systems, **84**(5), 385–395, 2009.

[36] G. M. Slota, K. Madduri, "Fast Approximate Subgraph Counting and Enumeration," in 2013 42nd International Conference on Parallel Processing, 210–219, 2013, doi:10.1109/ICPP.2013.30.

[37] S. Mbadiwe, W. Kim, "ParaMODA: Improving motif-centric subgraph pattern search in PPI networks," in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1723–1730, 2017, doi:10.1109/BIBM.2017.8217920.

[38] A. Masoudi-Nejad, F. Schreiber, Z. R. Kashani, "Building blocks of biological networks: a review on major network motif discovery algorithms," IET systems biology, **6**(5), 164–174, 2012, doi:10.1049/iet-syb.2011.0011[doi].

[39] D. Jain, R. Patgiri, "Network Motifs: A Survey," in M. Singh, P. Gupta, V. Tyagi, J. Flusser, T. Ören, R. Kashyap, editors, Advances in Computing and Data Sciences, 80–91, Springer Singapore, Singapore, 2019.

[40] W. Kim, L. Haukap, "NemoProfile: Effective representation for network motif and their instances," in 12th International Symposium on Bioinformatics Research and Applications (ISBRA), 2016.

[41] B. McKay, "Practical graph isomorphism," Congr Numer, **30**, 45–87, 1981.

[42] Y. Wang, W. Kim, "Identification of Essential Genes with NemoProfile and Various Machine Learning Models," in Z. Cai, I. Mandoiu, G. Narasimhan, P. Skums, X. Guo, editors, Bioinformatics Research and Applications, 319–326, Springer International Publishing, Cham, 2020.

[43] A. Ronacher, "Flask," https://flask.palletsprojects.com, 2010.

[44] W. Reese, "Nginx: The high-performance web server and reverse proxy," Linux J., **2008**(173), 2008.

[45] B. Chesneau, "gunicorn," https://gunicorn.org/, 2009.

[46] J. Y. Lee, J. Y. Kim, S. J. You, Y. S. Kim, H. Y. Koo, J. H. Kim, S. Kim, J. H. Park, J. S. Han, S. Kil, H. Kim, Y. S. Yang, K. M. Lee, "Development and Usability of a Life-Logging Behavior Monitoring Application for Obese Patients," Journal of Obesity & Metabolic Syndrome, **28**(3), 194–202, 2019, doi:10.7570/jomes.2019.28.3.194.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1634-1642 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

On Design of IoT-based Power Quality Oriented Grids for Industrial Sector

Nesma N. Gomaa[1,2], Khaled Y. Youssef[*,3], Mohamed Abouelatta[1]

[1]*Faculty of Engineering, Electronics and Electrical Communication, Ain Shams University (ASU), Cairo, 11757, Egypt*

[2]*Faculty of Engineering, Electronics and Electrical Communications, October University Modern Sciences and Arts (MSA), Cairo, 11757, Egypt*

[3]*Faculty of Navigation Science and Space Technology, Beni-Suef University (BSU), Beni Suef, 62511, Egypt*

ARTICLE INFO

Article history:
Received: 16 September, 2020
Accepted: 16 December, 2020
Online: 25 December, 2020

Keywords:
Energy Management System
Internet of Things (IoT)
Power system
Power Factor
Harmonics
Power quality

ABSTRACT

Manufacturing industry is facing several challenges in power quality and energy consumptions that play a significant role in the cost of goods sold in addition to the operational efficiency of manufacturing plants. The current techniques are not enough to manage the manufacturing processes from power perspectives as it is focusing only on the monitoring of the power grids using smart meters connected to SCADA systems. In this paper, a novel technique is proposed that is based on energy-aware manufacturing process control model using internet of things technology. The model is applied on cabling industry use case in addition to implementation of a MATLAB model for this purpose whereas the IoT physical layer is collecting , analyzing and communicating the electric power parameters correlated with the manufacturing process parameters received from PLC IoT nodes. Accordingly, the effective power quality is enhanced using manufacturing process control rather than additional correction nodes in traditional techniques. On applying the model on cabling use case, the Total Harmonic Distortion (THD-current) is improved 10 times to be 3.1% against 31.4% and the power factor is improved by 33% from 0.7 to 0.93 without additional correction nodes.

1. Introduction

This paper is an extension of work originally presented in 15th International Computer Engineering Conference (ICENCO) [1]. Power Quality (PQ) has recently become a very important topic because of its effect on energy producers, equipment's, manufacturers and consumers. There are so many industries nowadays that produce and use high technology for production thus high quality and high reliability are needed for these supplying strength [2]. The advantages of gathering real time monitoring is to develop the perception of awareness for power quality in factories.

Despite these circumstances, the executions of energy quality steps in improving energy efficiency have not yet satisfied expectations. The reason for the deficits in energy efficiency realization include lack of time and clarity on energy consumption [3]. The widespread use of non-linear loads leads to a huge distortion of voltage and especially current waveforms that affect any power electronic system [4]. In power systems, harmonic distortion and power factor disturbance increase on daily basis due to an increase in the use of nonlinear loads such as system resonances, voltage distortion, interference with communications networks, transformers overloading, a negative impact on energy metering and a decrease in the overall system efficiency. Thus, these two factors are set to be a serious problem in terms of power quality and this may affect other pieces of equipment connected to a particular system. The optimal performance between increasing in energy demand without severe increases in energy costs come to be our challenge now days. Hence, according to IEEE and International Electro-technical Commission (IEC) standards, industries must devote a greater attention to keeping harmonics and Power Factor (PF) within permissible limits [5, 6]. Furthermore, (IoT) could have significant contribution on energy management, hence several (IoT) applications to reduce the energy consumption have been developed [7, 8].

Power quality (PQ) monitoring methods that are recently produced have the capability to calculate and store data

[*]Corresponding Author: Khaled Y. Youssef, khalid_youssif@yahoo.com

https://dx.doi.org/10.25046/aj0506194

automatically with a high sampling rate. Significant quantities of measured data, such as current signals, voltage signals, total harmonic distortion, and power factor, can be obtained by monitoring devices. Yet, they do not identify distortions automatically, and involve off-line analysis of the data recorded. Consequently, in order to identify the type of end-use equipment, harmonic variations, and power factor detection a real time technique study is provided in this paper to develop the perception of awareness for power quality in factories and examine the impact of energy efficiency in industries. In today's power systems connected to nonlinear loads, voltage and current waveform distortion are more prevalent. The prevalence and interaction of loads of this kind with the power supply system increases the need to use power monitoring equipment to assess the quality of electric power. This paper proposes a new architecture which makes use of real-time data to ensure prediction of machine failure, using power quality data and ways to improve the energy efficient process. The algorithm proposed can be used for automated correction of industrial energy profile using Internet of Things (IoT) technology.

2. Literature review

Systematic approaches to eliminate the consumption of energy has been studied. Hence in order to explain the significance of problems with power quality, a detailed representation of any nonlinear system may be needed in the case of weak systems as interactions may occur between harmonics of different orders in such systems that are not predicted by linear time invariant models. Thus to identify and evaluate energy quality indices for various categories of electricity users, a fuzzy logic-based approach is suggested in [9]. Low energy quality is a serious concern for domestic, commercial and industrial users, in [10] issues of power quality in electrical systems is illustrated briefly.

Harmonics in electricity networks have recently gained even more attention from the power industry as a key cause of PQ problems, in [11], the total harmonic and unbalance distortion is well-defined to help factories to reduce harmonic distortion. In [12], an index is determined to represent the thermal effects on three phase induction motors as a result of harmonic distortions. Researchers in [13,14] illustrate other methods of PQ disturbances such as inter harmonics and voltage fluctuations. Furthermore , a Harmonic Trap Filter in conjunction with a line reactor designed and implemented by [15], using Simulink to demonstrate the principle of harmonic analysis in a three-phase system.

Additionally, the dramatic increase of electricity each year take the initiative on the aspect of energy efficiency in order to address the environmental issues. One of these aspects is the power factor correction. In [16], it is studied the importance of power factor which reflects how much the industries or organizations will consume the electricity efficiently. In [17], the authors investigate a system to carry out the effect of power factor on inductive loads using capacitor bank. Moreover, a system using a thyristor switching capacitor was proposed by [18], which helped to improve the quality of power and regulate the reactive power under the regularization of power factor.

Decreasing energy consumption is important for increasing productivity thus the relationship between electricity and economic growth is vital .To conserve energy and enhance its efficiency, smart grids (SG) were developed to do that [19]. John S. In [20], it is highlighted and clarified the different demand response (DR) programs. A number of DR-based programs are offered to organization managers and consumers, in general, where they choose which program they want to follow in order to decrease the overall energy consumption. Likewise, as energy saving has been one of the critical problems in designing electronic devices, Smart home appliance control system was developed in the IoT environment, a smart control system based on the technologies of internet of things has been proposed in [21].

3. Architecture for power quality management using IoT technology

In this paper a new architecture is proposed to increase the levels of power efficiency in industries. A smart energy management technique is illustrated in Figure 1. Using this block diagram will help factories to reduce their energy consumption. This architecture depends on tracking different loads in the downstream zone.

The architecture shown in Figure 1 is focused on monitoring the different loads attached in a Sensor Area Network (SAN), for the purpose of automated real-time monitoring of transient behaviour and accordingly the detection of anomalies related to power quality issues as voltage dwells, low power factor, and high order harmonics, that might happen as a result of grid's load non-linear behaviour. Most of this phenomenon is currently detected by capture of a measurement sample from the sub-system under test and that is usually reactive rather than proactive.

The concept emphasized by the study is about converting the grid system into an autonomous grid that has the ability to auto detect the phenomenon using a transit-time domain analysis and thus paves the way for artificial intelligence techniques on board to automatically detect the abnormal phenomena even on odd occurrence based on the intra-load communication, source-load communication and load on board processing. The concept also states the edge of machine to machine communications that is the core base of the well-known IoT technology.

According to Figure 1, the central processing server (CPS) is the node that is responsible for orchestration of the correction nodes on the correction plane according to the dynamicity of the load, and thus the CPS calculates and signals the configuration parameters to the corresponding correction nodes based on the measurements done by the IoT layer. The key difference between the existing techniques of correction and the proposed technique is mainly the holistic management of correction for the detected problem according to the multi-factor causal analysis.

One example is the power factor transient's detection that will be captured by IoT layer that is then escalated to CPS layer which in turn take the holistic decision to off –load the power factor capacitor bank and signal the synchronous power factor correction node to be in place in addition to active harmonic cancellation node in place of the passive one. The solution needs a level of intelligence as the classical capacitor bank response time to the change in loads will cause the failure of the classical technique. Moreover, the active cancellation will become effective due to the significant presence of several orders of harmonics rather than the

classical passive filter mode. The edge of the IoT layer in this example is in the consolidated view for the potential sources of problem with the decision of act on the optimum solution of correction.

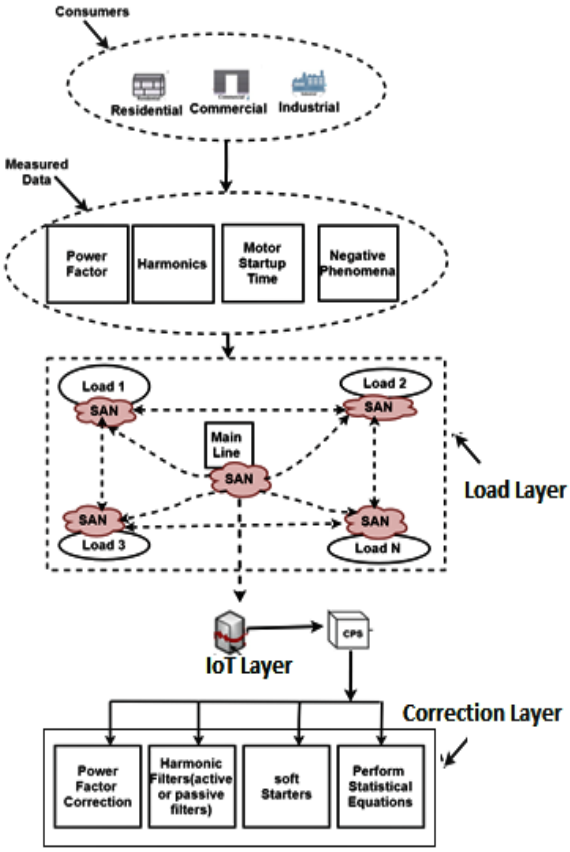


Figure 1: Schematic for the modules combining an intelligent energy management system in factories.

Illustrated in Figure 2 the concept of IoT smart grid which adds intelligence at each step by using online monitoring of transmission lines for disaster prevention and efficient use of power in order to decrease the overall energy consumption inspired by the factory. The three-layer architecture provides a high level framework through which different approaches might be implemented. The first layer consists of Machine drives, which are pumps, electric motors, and fans, account for about half of the manufacturing sector's delivered electricity, layer two is the IoT layer which consists of IoT beacon which are used as linking bridge that helps to provide total tracking and the last layer is named by correction plane layer in which the power factor correction bank and the harmonic filter exists.

The energy efficiency scheme is demonstrated by the flow chart in Figure 3, which begins by monitoring various production lines where each line has different load numbers, X is the order of each load, and N is the total number of loads. Two different parameters are evaluated on each load; power factor and harmonics. The CPS can detect if any of these parameters are not functioning effectively. A signal is sent to the CPS in the event of any errors. Where the problem can be overcome by the CPS,

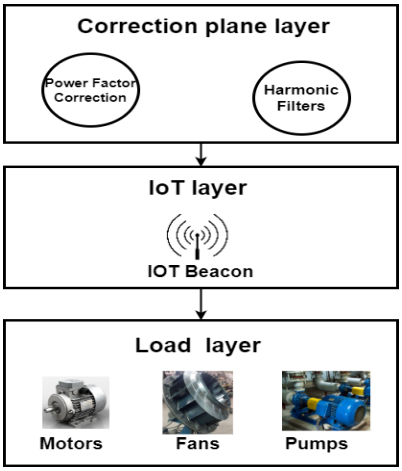


Figure 2: IoT architecture for smart grid.

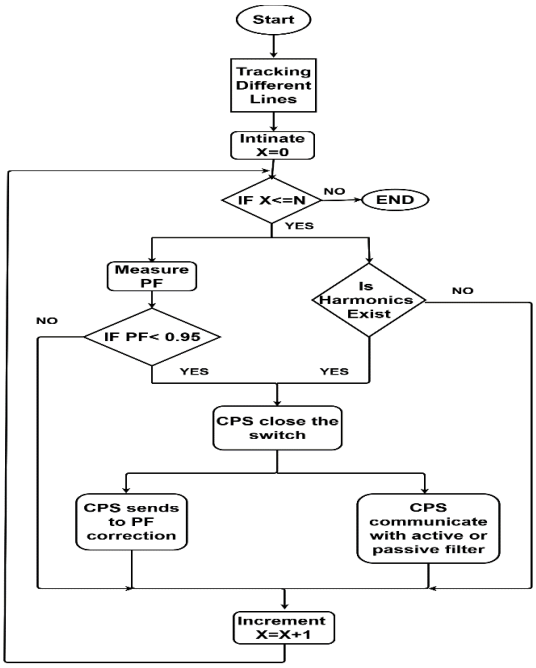


Figure 3: Flow Chart for energy saving analysis.

4. Power Factor correction and Harmonics detection Matlab simulation model results and analysis

A Matlab Model is designed to limit the voltage and current harmonics to be applicable with the IEC standards and to improve the power factor value. A Three-phase harmonic filters with power factor correction model for High-Voltage Direct Current (HVDC) is designed in Figure 4, using Matlab Simulink platform to reduce the voltage and current distortion in the power system. The model is designed using a frequency of 60 HZ and a firing angle α= 19°. One capacitor bank C1 of 150 Mvar designed by a "Thee-Phase Series RLC Load". Three types of different filters combination were also used including one C-type high pass filter tuned to the 3rd F1 of 150 Mvar ,one doubled –tuned filter 11/13th F2 of 150 Mvar and one high pass filter tuned to the 24th F3 of 150 Mvar.

The model used a switch control device for reducing the load imbalance and improving the efficiency of data processing. This

switch control is used to switch on or off when the CPS detect that the data received from the load is not achieving the threshold. The filter can be activated and deactivated using the breaker in the circuit. When the breaker (i.e. CPS) is closed, the filter is activated, the current and voltage will pass through the filter and will successively be filtered. Conversely when the breaker is opened the filter will be deactivated and will have no effect on the output waveforms of the system.

After running the developed Simulink model with different values of α. A various responses are observed. The waveforms in Figure 5, 6 respectively illustrate that when the breaker is opened, the filter is deactivated, a highly distorted waveform was detected because of the nonlinear loads that can be existing in the system. Hence when the breaker is closed, the filter can remove the harmonics given the desired waveform which is free from distortion.

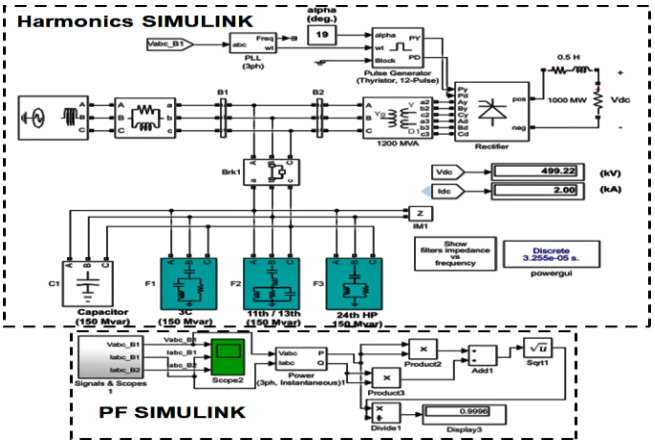


Figure 4: Three Phase Harmonic Filter with aid of PF correction (Simulink Design).

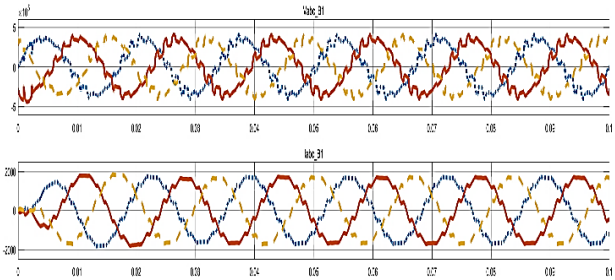


Figure 5: Output Waveforms when filter is disconnected.

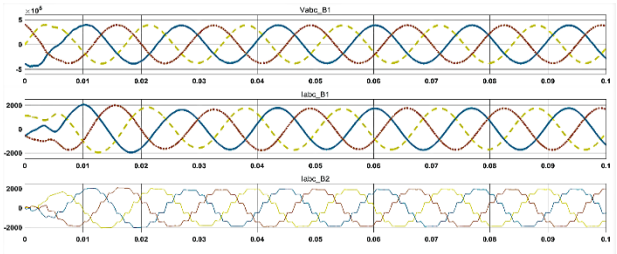


Figure 6: Output Waveforms when filter is connected.

Table 1 and table 2 illustrate that by using a suitable harmonic filter with the aid of a suitable capacitor value, we can control the flow of these fluctuations. From theses measureable results, it can be inferred that the system has worked successfully. As for example, it is clear that with the assistance of the harmonic filters and a capacitor, the THD voltage has fallen dramatically from 17.1 percent to 0.6 percent with an impedance of 407 Likewise, from 7.4 % to 0.3 percent, the new THD has decreased. The PF has risen from 0.3 to almost 0.9, coming close to the ideal 1.0.0. The proposed technique is therefore highly effective. 1.

Table 1: Comparison between voltage THD_v and current THD_i before and after adding the capacitor with the aid of the filter

Z (Ω)	THD_v Without Filter (%)	THD_i Without Filter (%)	Capacitor (F)	THD_v with Filter (%)	THD_i with Filter (%)
407	17.1	7.4	6.2E-06	0.6	0.3
400	16.2	7.1	8.0E-06	0.7	0.4
301	17.0	7.8	9.3E-06	0.71	0.4
261	17.01	7.12	1.0E-05	0.72	0.4
224	17.04	7.11	1.1E-05	0.69	0.5
193	17.08	7.14	1.2E-05	0.73	0.5
169	18	7.51	1.2E-05	0.8	0.7
151	17.53	7.19	1.1E-05	0.98	0.95
338	17.98	7.18	8.4E-06	0.02	0.1
488	16.82	7.00	6.1E-06	0.74	0.4
319	16.80	7.20	9.5E-06	0.7	0.4
335	17	7.31	9.1E-06	0.7	0.4
242	17.01	7.08	1.1E-05	0.7	0.4
187	17.2	7.16	1.3E-05	0.80	0.5
279	17.52	7.3	1.0E-05	0.8	0.4
156	17.69	7.4	1.9E-05	0.8	0.5
138	17.03	7.5	1.39E-05	0.9	0.51
152	17.7	7.22	1.40E-05	0.6	0.49
131	18	7.20	1.41E-05	0.7	0.71

Table 2: Comparison between PF before and after adding the capacitor

Z (Ω)	Distorted P.F	Capacitor (F)	Corrected P.F
406	0.30	6.2E-06	0.8
359	0.34	8.0E-06	0.91
300	0.4	9.4E-06	0.98
260	0.47	1.0E-05	0.99
223	0.55	1.1E-05	0.92
192	0.64	1.2E-05	0.94
168	0.7	1.2E-05	0.95
150	0.82	1.1E-05	0.95
338	0.36	8.5E-06	0.96
487	0.25	6.1E-06	0.97
318	0.15	9.5E-06	0.91
334	0.14	9.1E-06	0.99
241	0.41	1.1E-05	0.99
186	0.53	1.4E-05	0.98
279	0.17	1.0E-05	0.97
156	0.32	1.9E-05	0.95
137	0.72	1.5E-05	0.94
152	0.65	1.5E-05	0.99

5. Existing and future power quality system model for different loads

In the current technological transition scenario, power is very precious. A lot of researches need to improve the power quality system and find out the causes of power loss. In industries the huge increase in the use of inductive loads will leads to deficiency of the power system hence a routinely techniques have been used to detect the failures of their loads and correct it[22]. The classical approach shown in Figure 7 is to connect a direct act device with a current transformer (C.T) across the industrial input power lines. The C.T is directly installed at each load. In the classical scenarios, the technique of switching method (on/off) is used. Each load is supported by a controller device which is responsible for detection or cancellation to any failure on each line.

For example, to correct the power factor, these old techniques make use of a microcontroller to control the switching phase of the linked capacitor bank at the end of the main power supply. Each load has its specific control device connected to the main power supply close to the variable load. The controller's job is to read the power factor, voltage and current of the device. Then the controller determines, according to a proven algorithm, to cut off the power for the specified load and decide how many capacitors should be added to raise the power factor to the value required.

With this classical approach, there are two major problems: (1) the correction is a fixed one, so that if the industrial load differs , the correction becomes either less efficient or on the other hand, potentially overly efficient, resulting in an undesirable low leading power factor , (2) many industrial capacitor banks have inherently unpredictable maintenance problems. It would therefore be highly desirable to provide a device that would automatically correct the power factor of an induction motor in real time, so that as the applied mechanical load is varied, the motor runs at or near optimum efficiency continuously.

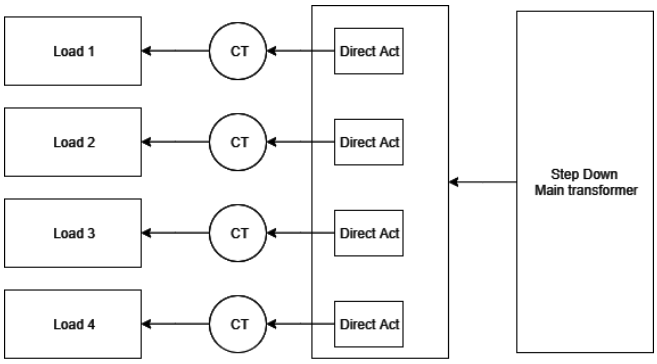


Figure 7: The classical power quality model for different load.

Substantial energy savings could be achieved through the Building Energy Management System (BEMS) within the Smart Electric Grid setting. A new technique based on IoT real time data acquisition using Message Queuing Telemetry Transport (MQTT) protocol which is the most commonly used protocols in IoT projects. The proposed architecture make use of real-time data to ensure prediction of machine failure using power quality data and ways to improve energy efficient process. The development of an energy monitoring system with the ability to acquire, process and store data in real-time is required for improving energy efficiency operations of the manufacturing industries and to enable proactive maintenance .This architecture prove the concept of sense, infer and act. As shown in Figure 8, the IoT beacons can contact with each other through the MQTT protocol. Thus, they can agree together that for example load number (2) will be compensated only. Consequently, tune tone generated by centralized harmonic cancellation to connect central node to a specific load.

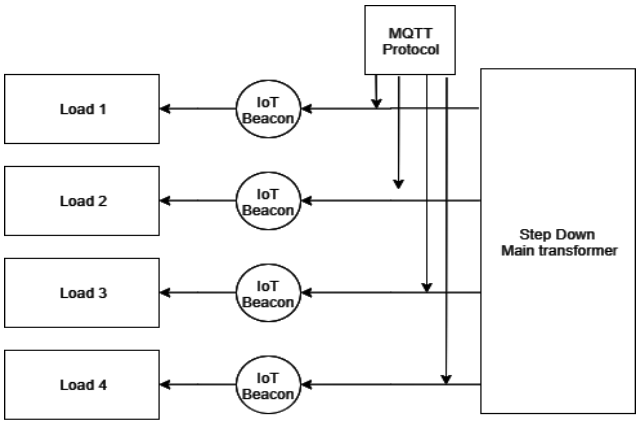


Figure 8: Future power quality system model for different loads.

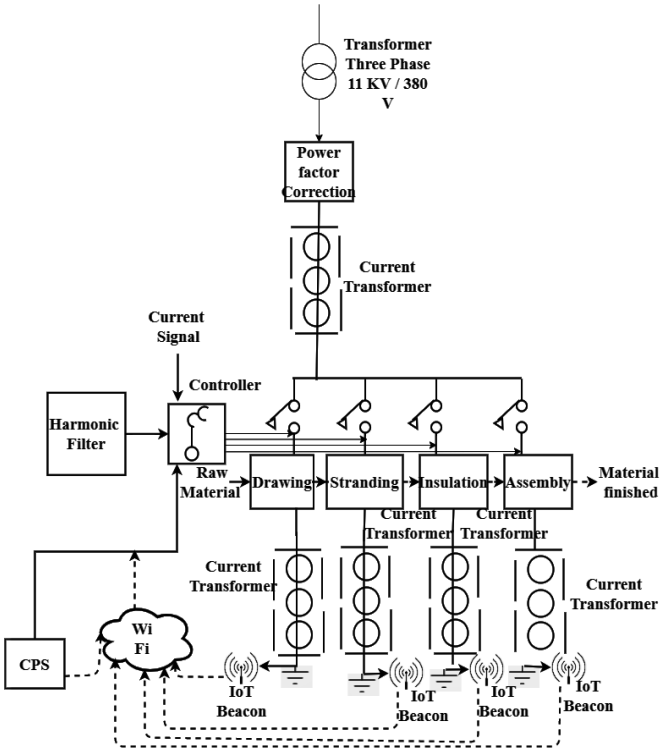


Figure 9: The proposed model for energy management technique.

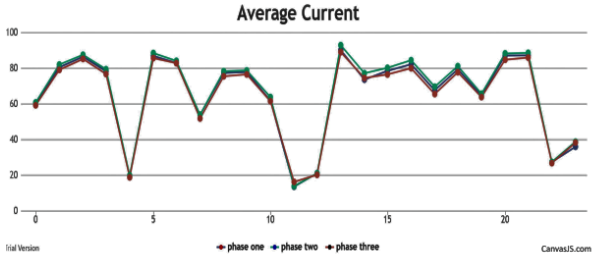
6. Power factor and harmonics improvement: A Case study and Analysis of the processes of the production of a metals in Factories

This work is a case study and review of the technological and energy aspects of the recommended design of the metal factory production lines in Egypt. The study provide the factory with some technologies which can be integrated into the factory production

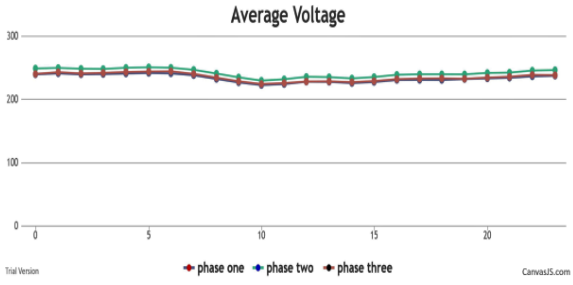
processes in order to improve the product quality, minimize the need for electrical energy, reduce maintenance time and cost, and increase reliability and real-time data accuracy.

This study reviewed the production line of metals in factories. The metal production line consists of four main process illustrated in Figure 9 which are drawing, stranding, insulation and assembly. In this power quality assessment, a power quality analysis of metal production process has been shown where power is given to different loads through transformer. Readings of power quality parameters such as voltage unbalance, current unbalance, power factor, harmonics and voltage fluctuation in each phase of the metal production have been studied.

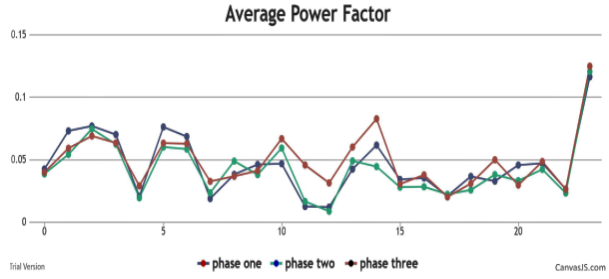
The purpose of the electrical distribution system is to support proper operation of the loads. When any one of the loads fails to meet its purpose, it is time to investigate the problem, find the cause and initiate corrective action. When a load does not operate properly, the other loads agree together through the proposed architecture shown in Fig.9 that this load should be compensated only. Figure 10-13 show the average voltage distortion, average current distortion, average power factor and average harmonic in the drawing phase, stranding phase, insulation phase and assembly phase respectively from the metal factory. Finally all these parameters are measured at the main line, so that the power factor and the harmonics at the main line can be fixed directly. Thus, Figure 14 and Figure 15 illustrated the reading at the main line before and after correction correspondingly.



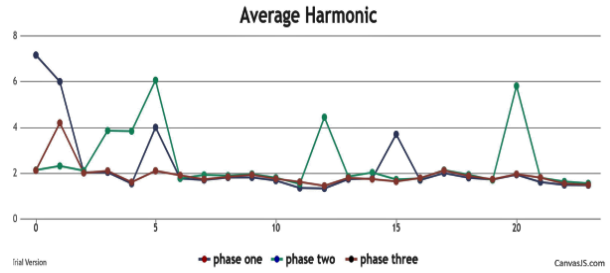
(a)



(b)

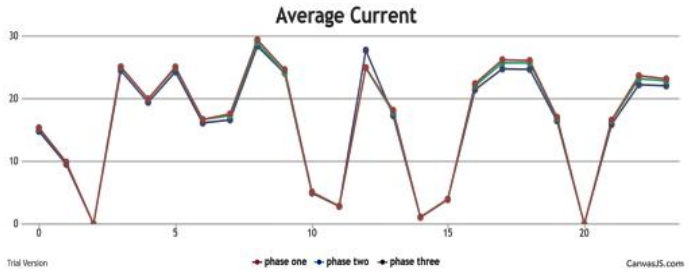


(c)

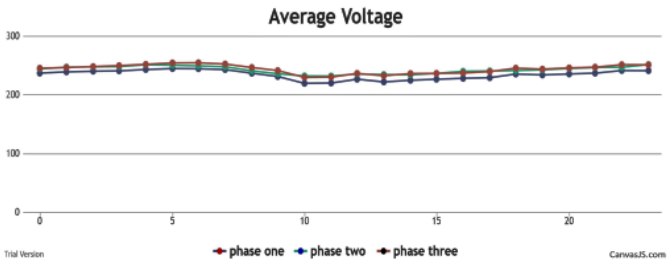


(d)

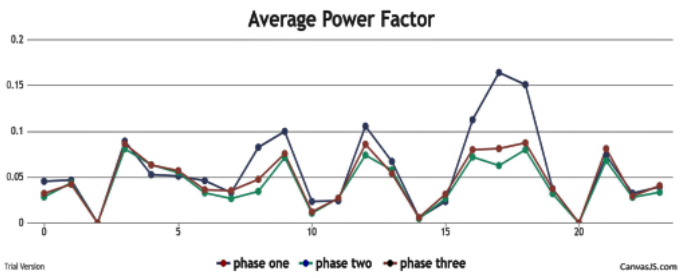
Figure 10: (a) The average current distortion, (b) average voltage distortion, (c) average power factor and (d) average harmonic in the drawing phase respectively.



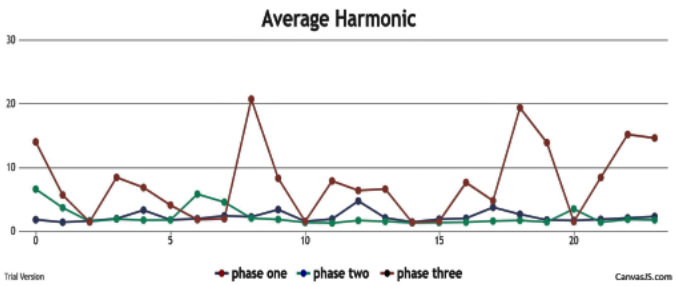
(a)



(b)

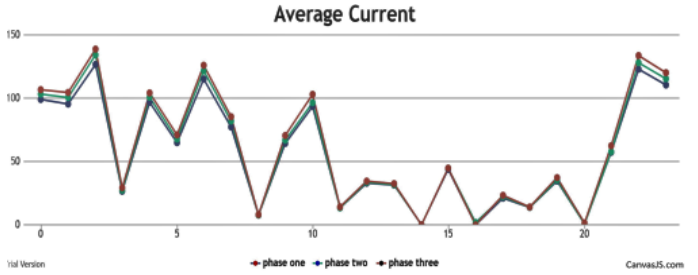


(c)

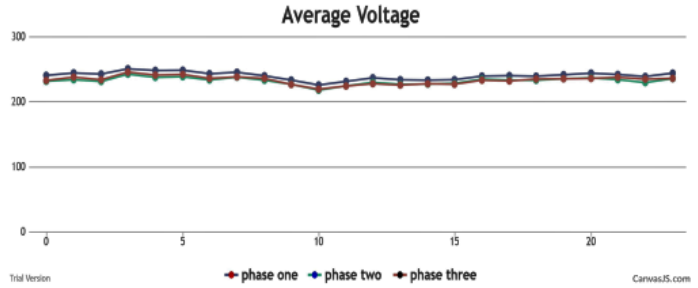


(d)

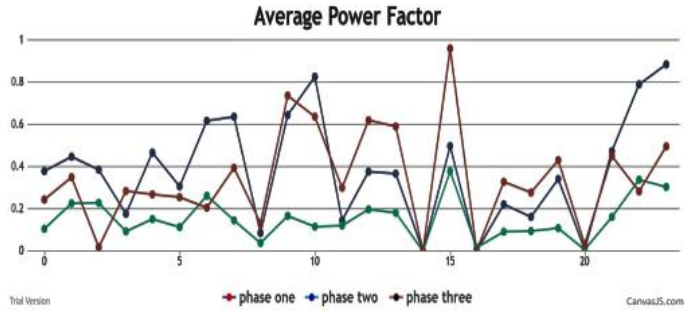
Figure 11: The average current distortion, average voltage distortion, average power factor and average harmonic in the Stranding phase respectively.



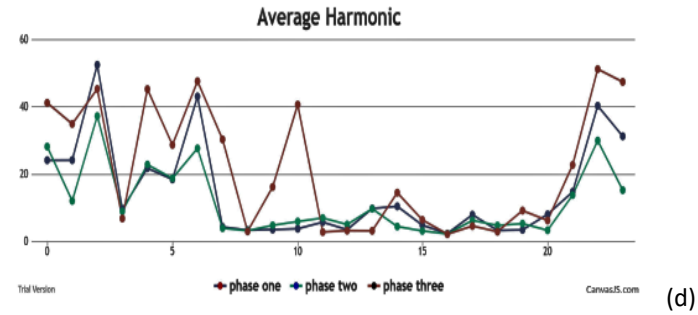
(a)



(b)

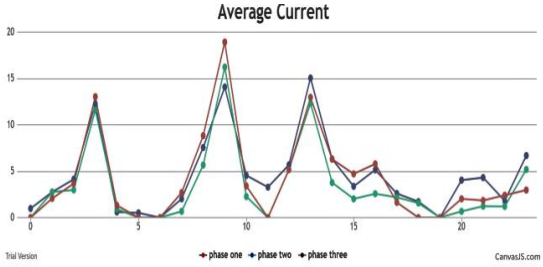


(c)

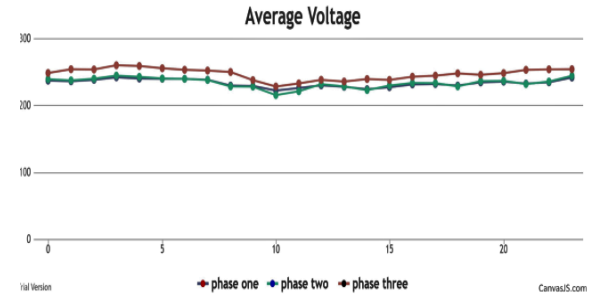


(d)

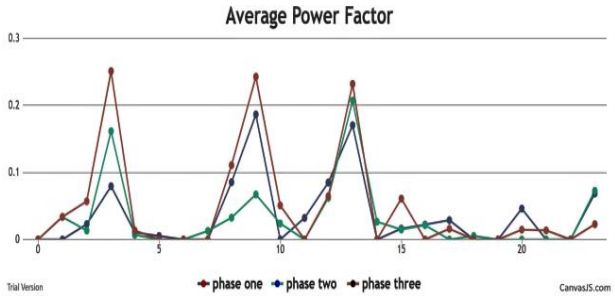
Figure 12: (a) The average current distortion, (b) average voltage distortion, (c) average power factor and (d) average harmonic in the insulation phase respectively.



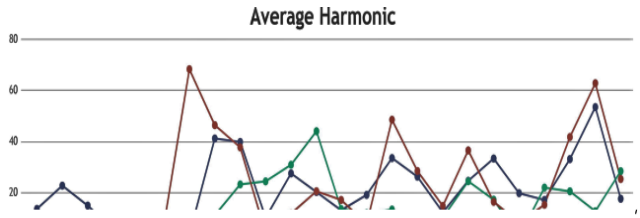
(a)



(b)

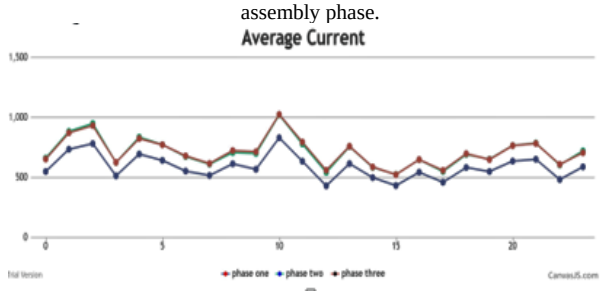


(c)

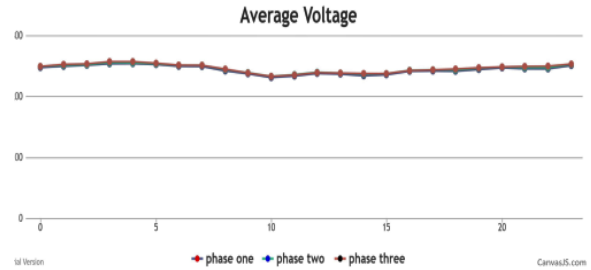


(d)

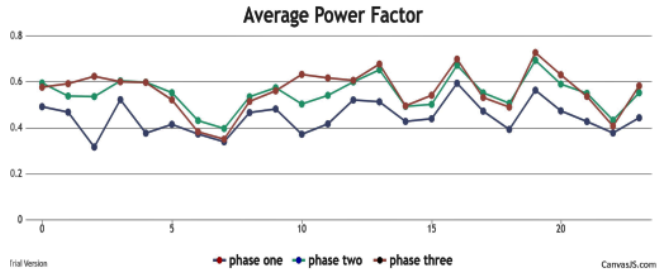
Figure 13: The Assembly phase (a) The average current distortion, (b) average voltage distortion, (c) average power factor and (d) average harmonic in the assembly phase.



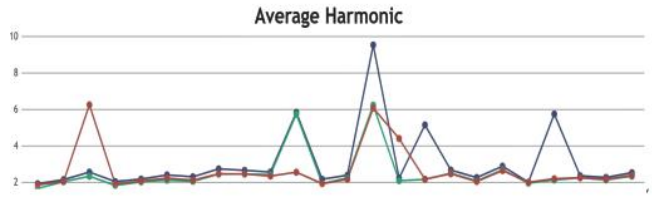
(a)



(b)

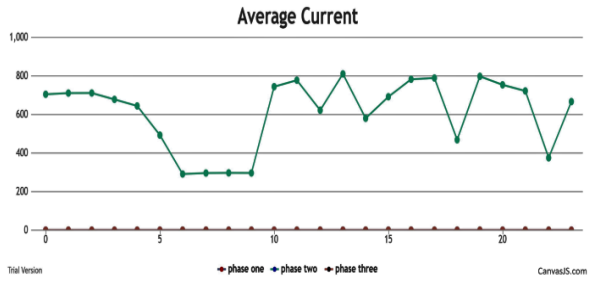


(c)

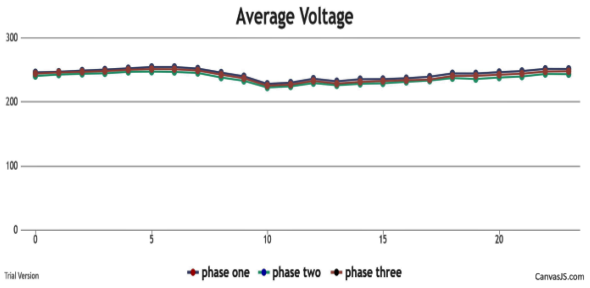


(d)

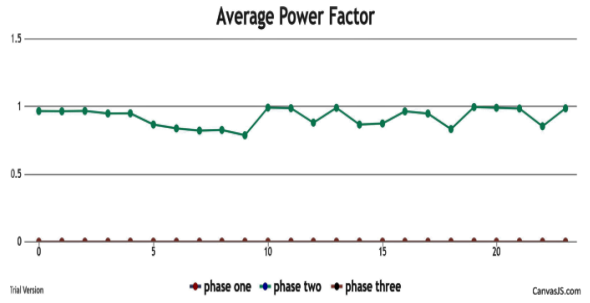
Figure 14: The average current distortion, average voltage distortion, average power factor and average harmonic readings respectively for the main line before correction.



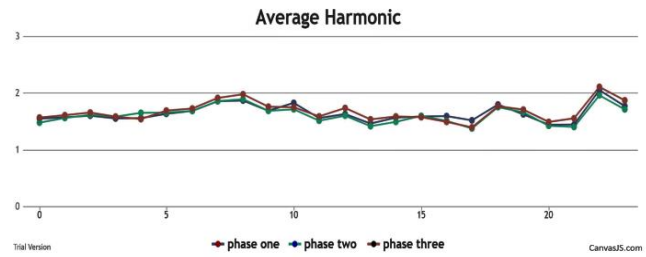
(a)



(b)



(c)



(d)

Figure 15: The average current distortion, average voltage distortion, average power factor and average harmonic readings respectively for the main line after correction.

The proposed system shown in Figure 9 will monitor the load current to detect the harmonics wave up to 49 th order and performs the computing process to analyse and calculate the power factor, then to transfer them to the high current capacity rectifier module along with the module processor in order to generate the compensation current in reverse direction, with overall response time less than 10ms and perform exceeding 97%.

Furthermore, Figure 16 illustrated the power consumption for one of the loads in four days. The measuring is taken by calculating the consumption of power for the three phases of one of the loads over time. Thus, it can be concluded that when using the proposed architecture a daily real time monitoring can be conducted which offers an understanding of energy usage and efficiency across different loads.

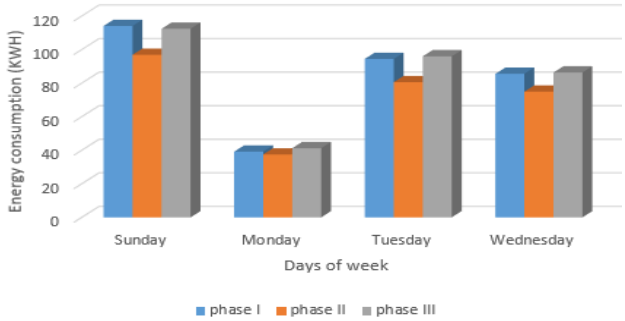


Figure 16: Energy consumption on one of the loads through four days

7. Discussion and conclusion

The manufacturing industry has been challenged with a lot of power quality problems in the overall including power factor, harmonics, unbalance ratios, power swells and dwells. Traditional techniques are moving towards correction steps as power factor correction using capacitor banks, synchronous power factor correction on individual loads, passive harmonic cancellations, active harmonic cancellations, soft switches to absorb spikes, voltage regulators, inverting technologies and others. However, those techniques are more focusing on correction of power rather than correction of manufacturing process that leads to the impact.

In the presented work , energy-aware process control model is proposed that is depended on IoT technology that enables the communication between PLC on machines and the energy management analytic platform on board to pinpoint the process that causes the phenomena to happen rather than the machine. For example, if two different machines working in parallel and doing the same operation type, machine x and machine y, that are installed in a factory. If the two machines are operating during three time shifts daily .Once the proposed model explore that machine x is doing wide spectrum of harmonic compared to machine y. The model will operate machine y for only two time shifts while machine x will operate for only one time shift. Hence the proposed model is able to stop the noisy machine and improve the overall power quality.

The model is simulated using MATLAB and is applied on a use case in factory that is specialized in cable industry. The results shows that the model pinpoints key areas of improvement of reducing the operation hours of a specific malfunctioning machine by one shift (eight hours) and increasing on the other side of operation hours of other two healthy machines doing the same operation type by 3 hours each has led to significant improvement of the power factor by 33% to reaches 0.93 versus 0.7 without impacting average OEE of the operation. On the other hand, the current total harmonic distortion is improved 10 times from 31.4% to 3.1% as shown in Figure 17 and 18 respectively.

The proposed model enables the cost-effective process management concept to achieve the same production results with better power quality and less energy consumption with minimum addition of correction nodes.

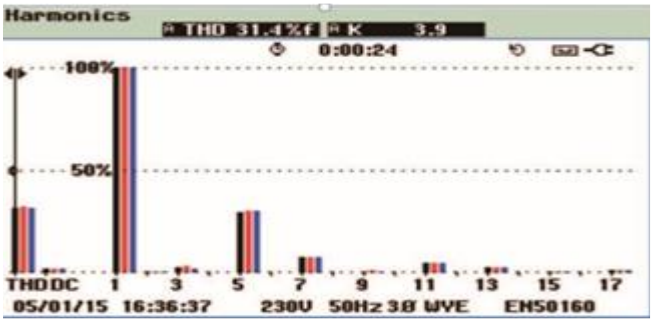


Figure 17: The Total Harmonic distortion (THD) before applying the proposed technique.

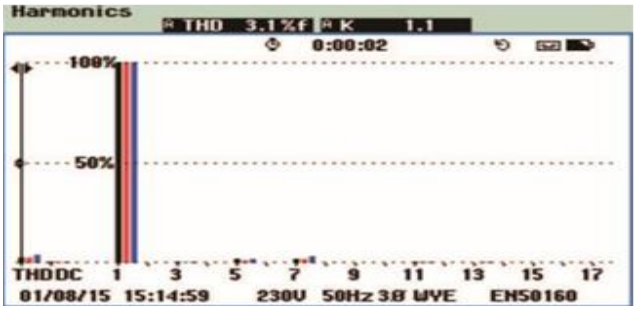


Figure 18: The Total Harmonic distortion (THD) after applying the proposed technique.

8. References

[1] N.N. Gomaa, K.Y. Youssef, M. Abouelatta, "An IoT-based energy efficient system for industrial sector," ICENCO 2019 - 2019 15th International Computer Engineering Conference: Utilizing Machine Intelligence for a Better World, 132–137, 2019, doi:10.1109/ICENCO48310.2019.9027440.

[2] V. Muni, R. Sangepu, "Effect of Power Quality Issues in Power System and Its Mitigation by Power Electronics Devices," Discovery, **28**(105), 72–79, 2015.

[3] M. Krones, E. Müller, "An approach for reducing energy consumption in factories by providing suitable energy efficiency measures," Procedia CIRP, **17**, 505–510, 2014, doi:10.1016/j.procir.2014.01.045.

[4] M.Z. Fortes, L.B. De Sousa, "Analysis of the Effects of Voltage Unbalance on Three-Phase Induction Motors," **3**(2), 29–37, 2018, doi:10.1007/s12053-015-.

[5] S. Rüstemli, E. Okuducu, M.N. Almalı, S.B. Efe, "Reducing the effects of harmonics on the electrical power systems with passive filters," Bitlis Eren University Journal of Science and Technology, **5**(1), 1–10, 2015, doi:10.17678/beujst.57339.

[6] S. A, "Power Quality and Harmonic Analysis in Three Phase Systems," Journal of Electrical & Electronic Systems, **7**(4), 2018, doi:10.4172/2332-0796.1000291.

[7] Z. Cao, Y.W. Chung, Y. Xiong, C.C. Chu, R. Gadh, "IoT based manufacturing system with a focus on energy efficiency," IEEE PES Innovative Smart Grid Technologies Conference Europe, 545–552, 2016, doi:10.1109/ISGT-Asia.2016.7796443.

[8] R.O. Aburukba, A.R. Al-Ali, T. Landolsi, M. Rashid, R. Hassan, "IoT based energy management for residential area," 2016 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2016, 2016, doi:10.1109/ICCE-TW.2016.7521035.

[9] J. Huang, Z. Jiang, "Power Quality Assessment of Different Load Categories," Energy Proceda, **141**, 345–351, 2017, doi:10.1016/j.egypro.2017.11.041.

[10] D. Ogheneovo Johnson, "Issues of Power Quality in Electrical Systems," International Journal of Energy and Power Engineering, **5**(4), 148, 2016, doi:10.11648/j.ijepe.20160504.12.

[11] M.J. Ghorbani, H. Mokhtari, "Impact of harmonics on power quality and losses in power distribution systems," International Journal of Electrical and Computer Engineering, **5**(1), 166–174, 2015, doi:10.11591/ijece.v5i1.pp166-174.

[12] S.X. Duarte, N. Kagan, "A power-quality index to assess the impact of voltage harmonic distortions and unbalance to three-phase induction motors," IEEE Transactions on Power Delivery, **25**(3), 1846–1854, 2010, doi:10.1109/TPWRD.2010.2044665.

[13] M. Mazadi, S.H. Hosseinian, W. Rosehart, D.T. Westwick, "Instantaneous voltage estimation for assessment and monitoring of flicker indices in power systems," IEEE Transactions on Power Delivery, **22**(3), 1841–1846, 2007, doi:10.1109/TPWRD.2007.893188.

[14] Ö. Salor, B. Gültekin, S. Buhan, B. Boyrazoğlu, T. Inan, T. Atalik, A. Açik, A. Terciyanli, Ö. Ünsar, E. Altintaş, Y. Akkaya, E. Özdemirci, I. Çadirci, M. Ermiş, "Electrical power quality of iron and steel industry in Turkey," IEEE Transactions on Industry Applications, **46**(1), 60–80, 2010, doi:10.1109/TIA.2009.2036547.

[15] R. M Patil, M.S. Nagaraj, P.S. Venkataramu, "Design and Implementation of Harmonic Trap Filter in Conjunction with a Line Reactor using Simulink," International Journal of Electrical and Electronics Engineering, **3**(9), 7–16, 2016, doi:10.14445/23488379/ijeee-v3i9p102.

[16] A. Sharma, "Dependence of Power Factor on Inductive Loads for Microcontroller based Power Systems," IOSR Journal of Electrical and Electronics Engineering, **7**(2), 30–35, 2013, doi:10.9790/1676-0723035.

[17] P. Papechen, S. Sebastian, "Automatic Power Factor Correction Using Capacitor Banks," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, **4**(7), 2321–2004, 2016, doi:10.17148/IJIREEICE.

[18] J.D.M. Mohamed Ibrahim A1 , Karthikeyan A2, Aravind K3, Gokulakrishna S4, "Minimizing Industrial Power Consumption Penalty By Using Thyristor Switching Capacitor," International Journal of Recent Trends in Engineering and Research, **4**(3), 741–749, 2018, doi:10.23883/ijrter.2018.4192.qidou.

[19] A.H. Bagdadee, L. Zhang, "A Review of the Smart Grid Concept for Electrical Power System," International Journal of Energy Optimization and Engineering, **8**(4), 105–126, 2019, doi:10.4018/ijeoe.2019100105.

[20] J.S. Vardakas, N. Zorba, C. V. Verikoukis, "A Survey on Demand Response Programs in Smart Grids: Pricing Methods and Optimization Algorithms," IEEE Communications Surveys and Tutorials, **17**(1), 152–178, 2015, doi:10.1109/COMST.2014.2341586.

[21] M.C.L. Padilla syamala, "An IOT Based Appliances Control for Smart Homes," in 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 170–174, 2017, doi:10.1109/ICICIP.2013.6568171.

[22] M.I. Muhamad, N. Mariun, M.A.M. Radzi, "The effects of power quality to the industries," 2007 5th Student Conference on Research and Development, SCORED, (January), 2007, doi:10.1109/SCORED.2007.4451410.

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1657-1665 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Novel Way to Design ADS-B using UML and TLA+ with Security as a Focus

Pranay Bhardwaj*, Carla Purdy, Nawar Obeidat

College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH 45220, USA

ARTICLE INFO

Article history:
Received: 01 September, 2020
Accepted: 04 December, 2020
Online: 28 December, 2020

Keywords:
ADS-B
TLA+
Security

ABSTRACT

Automatic Dependent Surveillance-Broadcast (ADS-B) is the future of aviation. It is a vast system that provides situational awareness for the aviator and regulator at a very low cost and does so with the aid of multiple disparate systems working closely together and communicating with one another. ADS-B uses the Global Navigation Satellite System (GNSS/ GPS) to locate elements. Weather information and ground-based information is also transmitted wirelessly. The system is designed to be open, unencrypted, and accessible to actors throughout the world. However, this leaves it open to attacks. The use of GNSS and other wireless technologies also carries over their security vulnerabilities into ADS-B. Certain issues have arisen due to both component-system failures and malicious attacks. Most obvious solutions impinge on the openness and transparency of the system. Past research has indicated that security must be built into a system design itself and cannot be retrofitted. We want to showcase such a design process for ADS-B. Our pathway to do so is to first create Universal Modeling Language (UML) diagrams to showcase security and safety issues and responses. These UML diagrams will then help us to model state and sequence diagrams. These will then be used to create a TLA+ model of one selected security methodology. We then run the TLC model checking on it to find loopholes and plug gaps in our scheme. We managed to create such models and prove deadlock-free running using only software tools. Our eventual goal is to develop a comprehensive formal specification for ADS-B model-creation and checking.

1. Introduction

This paper is an extension of work originally presented at the 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), Dallas, TX, USA [1].

Under the 2020 ADS-B mandate [2, 3], the FAA has designated that all commercial traffic and most private traffic must make itself compatible with ADS-B by hardware and software upgrades by the date of January 1. 2020.

ADS-B will replace 'hard' sites like primary/secondary surveillance radar (PSR/SSR) stations with satellite-based GNSS. The onus to report position is now on the aircraft. Frequencies of 1090 MHz and 978 MHz will be used for the actual information transmission. Aircraft will also beam down velocity, altitude, and a host of other data. In exchange, weather data will be beamed up to them. Aircraft will share position information with each other as well, to prevent conflicts.

ADS-B is a lot more economical than the current system. In 2007 the cost to monitor 200 nautical miles of air space was estimated to be $10-14 million using PSR, $6 million using SSR, and $380,000 for ADS-B [4]. Cost savings occur due to not having to set up and maintain expensive PSR/SSR RADAR stations plus increased coverage of formerly 'RADAR-dark' areas, resulting in more safety dividends, among other advantages.

The cost factor makes the choice look easy, but the security and safety weaknesses of ADS-B present a conundrum. Security issues are noted in [5-7], and an example of a deliberate system shutdown is given in [8].

Any attempts to gatekeep ADS-B go against its ethos. Security thus becomes a challenge. Attempts are now being made to introduce security measures into an already deployed ADS-B. But in [9], the authors have postulated that security must be considered at the design phase itself and any attempts to fit it in later will result in underestimation of the system capacity required to run security.

The scope of ADS-B is not restricted to 1090/978 MHz. In [10], the authors talk about a future where the Internet Protocol is the communications backbone for this system, even for voice

* Corresponding Author: Pranay Bhardwaj, Email: bhardwpy@mail.uc.edu

communication for aircraft flying. The radio has not been the sole point of communication for an aircraft for a long time. Now component subsystems of an aircraft independently 'talk' to maintenance and route planning departments of the operator via Aircraft Communications Addressing and Reporting System [11]. IP is also slated to take this communication load.

Following [9], the goal of this research is to rethink the ADS-B system design from a security viewpoint. We want to show what models such a hypothetical design process could have made, and what tools they could have used. A similar methodology is adopted by the authors in [12]. Here we are not recommending specific solutions for the numerous safety and security problems, nor are we devising a way to 'fix' ADS-B as it exists. We want to model how the different subsystems that makeup ADS-B interact with each other when under attack.

Our methodology is to pick a particular attack, pick specific solutions to that attack, and then first model the attack and response using use case diagrams. These will be used to create state and sequence diagrams to give a step-by-step description of the security response process to that attack. Once the sequence of events is clear, we can develop a TLA+ model of our system and run it through the TLC model checker to see if there are any deadlocks or failure points.

The specific attack scenarios and response tools are not ours, nor do we want to create any. We just want to observe how these response tools would look if they were all designed into the system, rather than retrofitted and to see if their in-built design into the system will cause any problems with any other system.

To put it simply, we do not want a situation where, for example, a cyberattack is successfully stopped by the security system but with an unintended result that airplanes are left without GNSS coverage. This could happen if the inter-relationships of competing cybersecurity tools are not well- understood at design time. We want to showcase that our methodology can be followed to model all of ADS-B, and perhaps other large systems in this manner at a low cost to spot deadlocks and bugs early in the design process.

So why do we want to use UML?

In [13], the author has used UML, the Unified Modeling Language [14], to model the Controller Area Network Bus (CANBUS). It has also been used to create security parameters [15,16]. UML focuses on whole 'objects' and each component may be a whole system unto itself. Thus hardware, software, and cyber-physical systems can all be modeled in UML.

Looking from a very high level, we can visualize the ADS-B as similar to the CANBUS. Both are unencrypted, both have various 'nodes' in the system, both rely on messages, both use broadcast protocols and both have no 'central brain.' Finally, security was not a consideration in the original design of either.

In ADS-B, the 'nodes' are the aircraft, the GNSS is a 'sensor' and the bus itself could be the pathways of the 978/1080 MHz frequencies. This is quite an interesting comparison and can be made between ADS-B and many other systems of various sizes and scales.

If ADS-B analogies can be found with the CANBUS, then methodologies to secure the CANBUS can theoretically also be proposed for ADS-B. As in [13], we will begin by modeling selected security techniques in UML and create use case diagrams.

Aviation is a heavily regulated sector and the basic requirement in all these systems↓ operational and security↓ is that they all work well together. Thus, the early introduction of security into the design is necessary because a late addition may introduce unintended security flaws. That is exactly the model we suggest in this paper- a redesigned system that does not stray too much from the current design and does not require major changes in protocol or infrastructure.

Our inspiration to use TLA+ is to get a workable model of how a potential 'built-in security' ADS-B system will look and behave. If successful, we will also be able to spot any failure points or illegal states in our models.

We envision that our paper can be used as a guidebook to model any discrete distributed system and test it for security loopholes at a very low cost. This will lead to the introduction of formal methods in civil aviation cybersecurity.

Formal methods have been used for many years in modeling localized systems [17], and we believe large systems consisting of geographically discrete subsystems can also be modeled this way. ADS-B uses independent systems for position, weather, and collision avoidance↓and each one of these is important to ensuring security and safety.

Our end goal is to create an overall super-design of the ADS-B system incorporating all the different security measures we talked about. In the future, we will attempt to create state and sequence charts for all of ADS-B. We will seek to validate the TLA+ based model checking and formal methods for complete ADS-B design.

In section 2, we discuss some weaknesses of ADS-B security and some well-known attacks. We then do the same for its largest sub-system, GNSS in section 3. In section 4 we create UML use-case diagrams to show how some selected methods can protect us from a cyberattack, and we also look at many other methods specified in the literature and select some for the next stage. In Section 5, we create a state diagram and sequence diagram of how ADS-B could respond to an attack using tools described in section 4. This is not a general view↓ just a specific attack scenario↓ and it feeds into the next section. In section 6, we will finally convert the state diagram to an actual TLA+ text specification/model. We will run it through the TLC model checker in section 6.1 to look for deadlocks and failures. We will then suggest how to apply our method more broadly to deal with the safety and security issues we have identified for the ADS-B system in the concluding remarks and the future work described in section 7.

2. Major Weaknesses of ADS-B

Because of its accessible design, ADS-B is missing even minimal security features [5]. This makes attacking easy and defense complicated. Some overarching vulnerabilities are:

- Physical layer: The 1090 MHz channel used for commercial aircraft ADS-B is severely congested, with up to 50% message loss rate measured at 174 miles [5]. This will only get worse with time, and we are in a very early phase yet.
- Unencrypted: Because ADS-B is unencrypted, the positions of aircraft can be tracked by anyone with homemade equipment and a free open-source application like flightradar24.com (FR24). Eavesdropping is the first step for many attacks [5]. FR24 details a step-by-step guide to making an ADS-B 'snooper' and getting real-time aircraft identification, position, and performance data.
- Unauthenticated: ADS-B has no authentication on messages↓no signatures, no handshake protocol. Aircraft identifiers, like the 24-bit ICAO code, are publicly available in a platform like FR24. Thus, there is no way to identify the source of a message [3]. If the constraint of powerful transmission equipment is fulfilled, a well-funded actor can theoretically beam fake data about an existing contact, or a fake contact itself, to air traffic control. Even worse, it can remove data as well.

Some common attacks that can happen in the above environment are:

- Ghost contacts: Two works on cybersecurity [18,19] mention the possibility of fake, or 'ghost' targets in the system. An attacker could use the openness of the system to download authentic aircraft data and then broadcast it over the frequency with sufficient power. This would appear as a legitimate contact on the ATCS screen. Without confirmation of GPS position, it would be impossible to catch such an attack. An example that combined eavesdropping with ghost injection is described. The appearance of multiple ghost aircraft can cause system and sensory overload and can cause pilots and ATC to make dangerous decisions [6].
- Hybrid attacks: Hybrids of various attacks can be made. In [6], the authors talk about an attack that 'deletes' legitimate data from the system and injects false data in its place.
- Legacy systems: It is not uncommon to find aircraft flying today that are 20-30 years old or more. This is especially true in developing countries, cargo airlines, and various armed forces. Legacy aircraft systems were not designed with an interconnected super system like the ADS-B in mind. In the past, each aircraft was an independent entity. Legacy systems have always been a security challenge when hooked up to large networks. A determined adversary with ample support↓like a state-backed actor↓can overcome all compatibility and air gap barriers to infect an 'old' system with a 'new' bug, as was seen in [20].
- Physical access: Physical isolation or 'air- gapping' that ensured the protection of many systems before will cease to exist. In the past, the output of a PSR/SSR was inaccessible to the public. Now, the detection system, position system, and the results therein are all readable on a home computer in real-time. Physical isolation of infrastructure has given way to an open, unencrypted, unauthenticated setup.

Risks of GNSS Use in Aviation

- Survivability: As of today, four nations possess the capability to destroy satellites in orbit the USA, Russia, the PRC, and India [21]. These 4 nations have also all developed their own global or regional GNSS (GPS, GLONASS, BEIDOU, NAVIC respectively). All these nations have also militarized their GNSS systems↓to guide bombs, missiles, warplanes, naval vessels [22] and even to plan large-formation land maneuvers↓as was seen in the 1990 gulf war. Law enforcement and coast guarding also use GNSS. However, this also makes the GNSS a ripe target in the event of great power warfare. Knocking out GNSS satellites using anti-satellite (ASAT) missiles could lead to a mass blackout of ADS-B position reporting for all countries. This could have disastrous unintended consequences for civil aviation within seconds. In a future PSR/SSR-less world, this could mean crippling the global flight infrastructure. Although the probability of this is low, the authors find this to be one of the most disturbing eventualities.
- Deliberate shutdown: Like any other large system, GNSS may be on reduced performance from time to time, as noted, e.g., in [23]. Besides, deliberate shutdowns may be conducted for various purposes, like military training. The degraded signal quality sent an aircraft into high-speed oscillations in 2016 [24]. As the dependence of aircraft on GNSS grows, any shutdown or breakdown will become a safety issue for civilian aircraft. Dependence of autopilot systems on GNSS to navigate means aircraft could go into abnormal flight regimes without even alerting the crew if a sudden 'quiet death' of GNSS occurred locally during normal flight.
- Jamming and spoofing: GNSS is a satellite-based wireless system, and like every wireless system ever made, GNSS is susceptible to jamming of signals and spoofing↓incorrect location reporting. The Russian Federation has already demonstrated a GNSS spoofing/ jamming capability precise enough to make a single human difficult to locate [25]. It can be devastating for unsuspecting civilian aircraft in an ADS-B based system and could very well be the model of a terror attack.

3. Use Case Diagrams for ADS-B Security

For such a large and distributed system as ADS-B, it must be understood that no single solution exists. In [7], the authors tested the viability of many solutions and found none without a loophole. This is a glaring after-effect of not keeping security paramount in the design process. Of course, we cannot talk about a cybersecurity model without talking about the actual 'weapons' that will make it possible. We will now make UML diagrams of how the various cybersecurity tools we are familiar with could be deployed, in isolation or in combination, to ensure protection. Figure 1 provides the convention we use in the use case diagrams that follow.

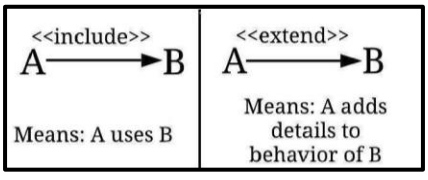


Figure 1: Convention for use case diagrams

In Figure 2, multilateration uses the time delay of signal arrival at multiple (>2) points to estimate the location of the transmitter [7]. Time Delay of Arrival (TDOA) of the signal is

calculated with the use of time stamps stamped on any message. Comparison of timestamps of transmission and reception gives a rough idea of the travel time, from which a rough distance can be calculated. If used at 3 receivers, a close X-Y-Z coordinate can be back-calculated. This is not very different from the SONAR equation.

If this back-calculated position is reasonably close to the reported GNSS coordinate of the transmitter, the allegiance of the message sender can be confirmed. Multilateration can be used to combat ghost aircraft, message modification, and man-in-the-middle attacks, and can provide a backup to GNSS breakdown in the absence of PSR.

Figure 3 is the most important here as it was the most complex one in our original paper. We have thus chosen to extend this diagram to the state diagram, sequence diagram, and later, to the TLA+ specification.

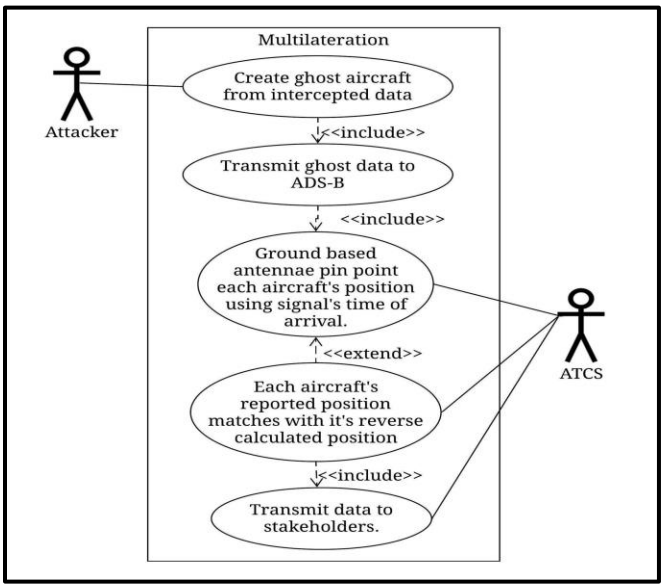


Figure 2: Multilateration and group verification can protect against ghost aircraft injections by matching the physical source of the signal to its reported one.

One of the weaknesses of using static ground posts as receivers for multilateration is that the system may be fooled by the attacker. If the location of these ground antennas is known, the attacker can tailor his transmission time stamps to present a 'false but accurate' picture to the multilateration-systems. A solution to this is to use dynamic receiver posts↓make the other, trusted aircraft in the airspace as multilateration, or triangulation points themselves!

Figure 3 shows multi-point multilateration. Group verification [7] is a kind of multilateration performed in the air. Aircraft already multilaterate each other using TDOA↓it is exactly how the Traffic Collision Avoidance System (TCAS) works. The equipment and algorithms are thus available on board already, and this is in line with our philosophy to cause minimal modification.

Each aircraft locates each member using TDOA of received ADS-B IN signals and estimates the others' positions. If the calculated position differs from the reported one, a 'suspect' airframe is identified, and group members move away from it. This can be successful only when 100% implementation of ADS-B is achieved. Location fixes from different sources are independent of GNSS. The fusion of this data can be an alternate method to locate aircraft. This will counter GNSS spoofing and GNSS jamming attacks.

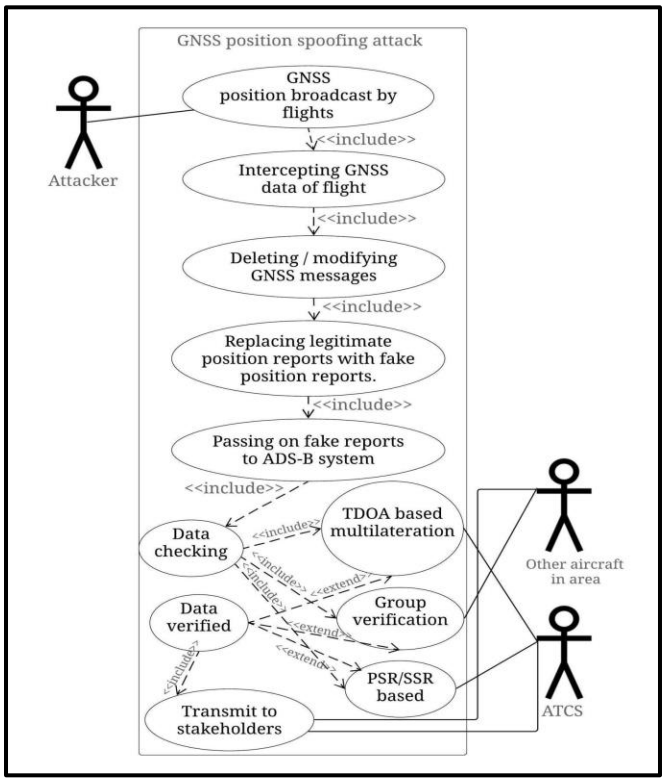


Figure 3: Multilateration may help with GNSS spoofing and unreliable position reports.

Other issues and solutions we will use in our overall model:

- Authentication of messages can be used to identify which messages are from what source. Authentication just confirms that the message is from who you think it to be. This directly allows us to trust or not trust it. Authentication of messages is the closest we come to a silver bullet solution for ADS-B as a single tool protects us from multiple issues↓Denial-of-Service (DOS) attacks and ghost aircraft attacks, just for starters. DOS attacks on airport or airline infrastructure are of concern as these are cheap to wage.
- Kalman filtering has many applications↓the ADS-B system can be 'taught' the typically correct values of airplane parameters [7]. According to their capabilities, scheduling, and operators, most aircraft will fly within a certain envelope of speed, heading, and locations usually. The Kalman filter can be rewarded whenever correct values are noticed, and not when unusual values are noticed. This will protect against ghost injection attacks and false data messages. The system will sense the departure from usual data trends and raise an alarm. If data trends change, it can be fed the new data trend and 'taught' any updates to the system this way. The only disadvantage is that the system is equally likely to be taught the 'wrong' way and thus become a 'bad child.'
- Hardware and software fingerprinting [26] have great potential to introduce economical authentication capability

without making major hardware or software changes. The authors in [26] provide a simple solution to 'fingerprint' the transmitters themselves. Every transmitter in the world is unique, just as every crystal is unique. Individual aircraft-borne ADS-B transmitters can be 'fingerprinted' either via unique hardware or through software properties. This can positively identify the transmitter and provide authentication.

- ADS-B as RADAR? A unique solution of using the ADS-B signal itself like a radar signal, by adding a random bi-phase modulation is proposed in [27]. This system can track other aircraft even with no GNSS. It can provide the much-needed reliability of radar at the low cost of ADS-B. This can counter issues of the survivability of GNSS. This agrees with [4], which talks of a fusion of GNSS and non- GNSS sources for locating aircraft. However, this will require changes to ADS-B transmitting recruitment and is thus low on our list of options.
- Weather RADAR for aircraft tracking: Weather RADAR is ubiquitous and big and small weather radar stations exist all over a developed country like the US. Since this RADAR can track small cloud formation, moisture, and bird/ locust flocks, can it be used to detect hard-skinned objects like airplane skin? Some research has been conducted in the use of weather RADAR to track aircraft and we believe that if it is properly developed, it can act as an emergency backup in the case of a major GNSS failure. For small general aviation (GA) aircraft at least, a method is suggested in [28]. This method can distinguish between small aircraft and large birds based on pressure waves created by aircraft propellers. In [29], the authors managed to detect aircraft with weather RADAR using some signal processing. More than 90% of aircraft in the USA are GA and most of these are propeller equipped [30]. Thus, a solution for GA aircraft as presented in [28] solves a huge chunk of aircraft tracking issues in case of GNSS blackout.
- Identification can be achieved by building a small ADS-B receiver like the ones FR24 uses, to receive aircraft data. Thus, in the event of a GNSS failure, weather RADAR can track the physical location of GA aircraft while an ADS-B receiver can identify them. One only must ignore the ADS-B location data in such a case. The use of a non-ADS-B source as a redundant backup is in the spirit of [7].
- TESLA protocol for encryption: In [31], the authors suggest a method that encrypts only the 24-bit ICAO aircraft identifier. It uses Timed Efficient Stream Loss-tolerant Authentication (TESLA) for authentication. It has performed well under testing without reducing security, performance, or affecting any major changes to ADS-B philosophy.

4. TLA+ conversion

TLA+ stands for 'Temporal Logic of Actions' [32]. It is a logical system or toolbox, developed by Leslie Lamport. which allows one to write formal specifications and to include scalable security protocols. The TLA+ specification can then serve as a formal 'model- checker' for any future realization of the ADS-B system and its components. It is not a classical programming 'language.' TLA+ is based on mathematics and uses set theory to model systems. Each system has states, and each state can be defined by a set of certain values. When these values change, so does the state. Any transition (step) is the change of one state to another or a change of one set of values to another set of values.

An entire system can be modeled as a superset of all possible state sets. We do not go too deep into the specifics of each state↓we are more concerned with the what rather than the how. Thus, abstraction and conciseness are a very important skill to master before modeling the system in TLA+. While economics and computer overhead would be major factors in practical systems design, we drop them when dealing with TLA+.

As a testimonial of its effectiveness, the application of TLA+ by Amazon revealed it to be excellent for bug detection [33,34]. Microsoft also used it to verify the design of a subsystem of the Alpha 21364 Microprocessor [17].

Our work here can be viewed as a guide on how to design large systems in the future↓by comparative analysis and extensive model-driven design.

ADS-B has many subsystems that make up the whole. GNSS is responsible for position reporting to aircraft. Aircraft then report their position to air traffic control.

Aircraft get weather and aerodrome information, and share position reposts with each other for collision avoidance. All this communication happens over the 978/1090 MHz spectrum. So not only are there several independent operations but also there are independent failure points. All these systems must be a part of any formal method examination of the ADS-B system. The specific solutions for responding to attack are not our own↓they are specified in section 4. What we are doing is modeling a security solution for the ADS-B system that will have safety and security included by design, and suggesting a procedure to design such a system and simulate it.

Several works on smart systems in TLA+ have been researched. In [35], the authors design a smart school system in TLA+. In [36], the authors model a sewerage system in UML and TLA+.

The authors in [35] start by making a UML specification, a sequence diagram, and a state diagram. These diagrams aid in making the sequence of events clear. The system's pathways, legal and illegal actions, and outcomes are modeled.

4.1. Why do we make sequence and state diagrams?

TLA+ defines every system as a sum of all the states it can exist in. A system can have multiple states, and there will be a defined 'step', or a sequence of events, to go from one state to another. Thus, we first develop a sequence diagram, which shows all the steps between states, and a state diagram which shows the states themselves. Only then can we draw a true TLA+ model.

Let us take the below example of applying methodologies of [35] and [36] to the ghost aircraft and GNSS spoofing issues as noted in Figure 3.

4.2. ADS-B response to Ghost aircraft and GNSS spoofing

In sections 2 and 4, we talked about security issues caused if an aircraft is either a fake contact or is reporting untruthful position information about itself to ADS-B, via GNSS spoofing. Since ADS-B envisions doing away with PSR/SSR, we had

examined some other backup systems that can physically locate an aircraft if GNSS reports are unreliable. The UML diagram for this was Figure 3. Let us now look at a sequence diagram for the situation, in Figure 4.

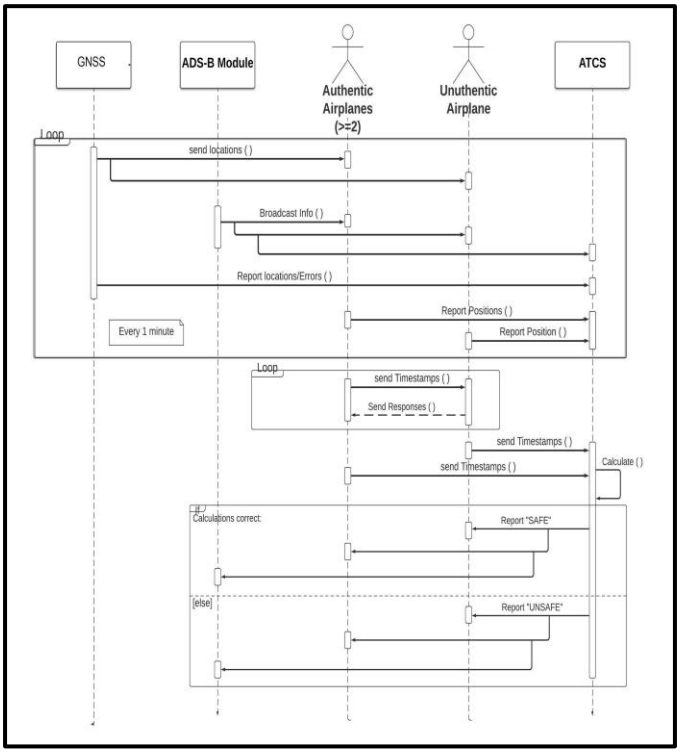


Figure 4: Sequence diagram: Reverse calculation of the aircraft's true position can be done using time-delay analysis of transmitted signals

GNSS reports aircraft position to airplanes, which includes both the 'good' and 'bad' airplane. All aircraft will then report their position to ATCS. A time delay analysis of the signals from all aircraft will be performed, and this will be used to get a location fix of the aircraft in physical space. Aircraft will also be able to perform multilateration and triangulation on each other in airspace↓within a '4-ship-cell' for example↓and get location fixes for each other's locations.

A comparison of the back-calculated position data with the GNSS reported position data will point out which contacts are reporting true data and which are not. Thus, aircraft can be labeled 'SAFE' or 'UNSAFE.' This information can then be provided to other aircraft in the area, and ATCS. Note here that 'ADS-B' stands for the system, but practically it may mean a supervisory body, like the FAA Technical Operations Services, or any other service which needs to be notified of security incidents. Let us now look at a state diagram of the same function in Figure 5.

This specifies an interval of 1 minute after which the whole cycle is repeated. This is arbitrary and will practically be a decision governed by radio traffic and economic factors. Since aircraft move fast and move in all 3 dimensions, this interval cannot be too long. Also, when we say 'ATCS', we mean a local or regional airspace manager. We do not envision a global central ATCS system in this paper.

Let us now use the 3 inputs, Figures 3,4, and 5 to start making the TLA+ formal specification for the above response.

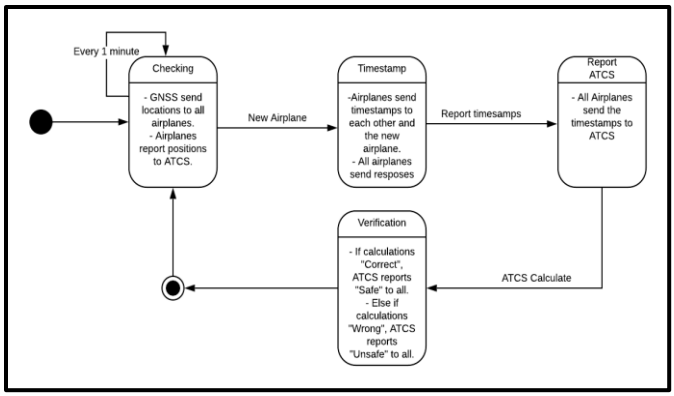
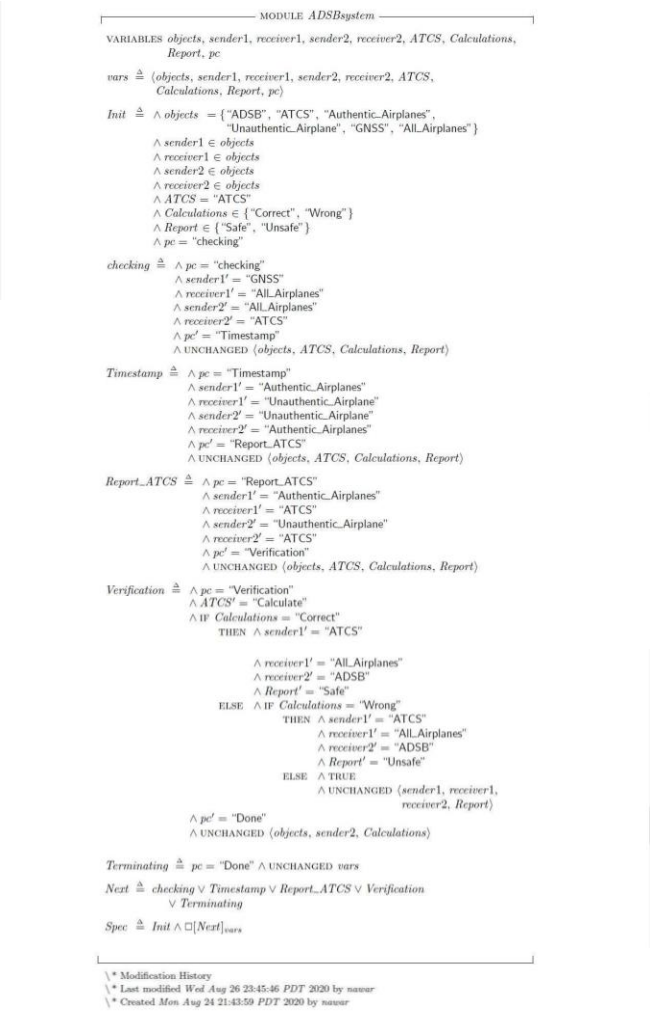


Figure 5: State Diagram of the system under Figure 4 conditions.

5. TLA+ Specification

We present now, the TLA+ specification for the attack response model sequenced in Figure 4 and stated in Figure 5. A block by block explanation follows.



MODULE *ADSBsystem*

VARIABLES *objects*, *sender1*, *receiver1*, *sender2*, *receiver2*, *ATCS*, *Calculations*, *Report*, *pc*

$vars \triangleq \langle objects, sender1, receiver1, sender2, receiver2, ATCS, Calculations, Report, pc \rangle$

$Init \triangleq \land objects = \{\text{"ADSB"}, \text{"ATCS"}, \text{"Authentic_Airplanes"}, \text{"Unauthentic_Airplane"}, \text{"GNSS"}, \text{"All_Airplanes"}\}$
$\land sender1 \in objects$
$\land receiver1 \in objects$
$\land sender2 \in objects$
$\land receiver2 \in objects$
$\land ATCS = \text{"ATCS"}$
$\land Calculations \in \{\text{"Correct"}, \text{"Wrong"}\}$
$\land Report \in \{\text{"Safe"}, \text{"Unsafe"}\}$
$\land pc = \text{"checking"}$

$checking \triangleq \land pc = \text{"checking"}$
$\land sender1' = \text{"GNSS"}$
$\land receiver1' = \text{"All_Airplanes"}$
$\land sender2' = \text{"All_Airplanes"}$
$\land receiver2' = \text{"ATCS"}$
$\land pc' = \text{"Timestamp"}$
$\land \text{UNCHANGED } \langle objects, ATCS, Calculations, Report \rangle$

$Timestamp \triangleq \land pc = \text{"Timestamp"}$
$\land sender1' = \text{"Authentic_Airplanes"}$
$\land receiver1' = \text{"Unauthentic_Airplane"}$
$\land sender2' = \text{"Unauthentic_Airplane"}$
$\land receiver2' = \text{"Authentic_Airplanes"}$
$\land pc' = \text{"Report_ATCS"}$
$\land \text{UNCHANGED } \langle objects, ATCS, Calculations, Report \rangle$

$Report\_ATCS \triangleq \land pc = \text{"Report_ATCS"}$
$\land sender1' = \text{"Authentic_Airplanes"}$
$\land receiver1' = \text{"ATCS"}$
$\land sender2' = \text{"Unauthentic_Airplane"}$
$\land receiver2' = \text{"ATCS"}$
$\land pc' = \text{"Verification"}$
$\land \text{UNCHANGED } \langle objects, ATCS, Calculations, Report \rangle$

$Verification \triangleq \land pc = \text{"Verification"}$
$\land ATCS' = \text{"Calculate"}$
$\land \text{IF } Calculations = \text{"Correct"}$
$\text{THEN } \land sender1' = \text{"ATCS"}$
$\land receiver1' = \text{"All_Airplanes"}$
$\land receiver2' = \text{"ADSB"}$
$\land Report' = \text{"Safe"}$
$\text{ELSE } \land \text{IF } Calculations = \text{"Wrong"}$
$\text{THEN } \land sender1' = \text{"ATCS"}$
$\land receiver1' = \text{"All_Airplanes"}$
$\land receiver2' = \text{"ADSB"}$
$\land Report' = \text{"Unsafe"}$
$\text{ELSE } \land \text{TRUE}$
$\land \text{UNCHANGED } \langle sender1, receiver1, receiver2, Report \rangle$
$\land pc' = \text{"Done"}$
$\land \text{UNCHANGED } \langle objects, sender2, Calculations \rangle$

$Terminating \triangleq pc = \text{"Done"} \land \text{UNCHANGED } vars$

$Next \triangleq checking \lor Timestamp \lor Report\_ATCS \lor Verification \lor Terminating$

$Spec \triangleq Init \land \Box[Next]_{vars}$

* Modification History
* Last modified *Wed Aug 26 23:45:46 PDT 2020* by *nawar*
* Created *Mon Aug 24 21:43:59 PDT 2020* by *nawar*

Figure 6: TLA+ model specification in literal form.

In the module, the keyword "VARIABLES" represents the variables in this module where "vars" is the tuple of variables:

<objects, sender1, receiver1, sender2, receiver2, ATCS, Calculations, Report, pc >

In every TLA+ module, there is an "Init" and a "Next" state. In the Init state, the variables get their initial values, while the next state represents all possible states that come next. In this module, the Next state represents:

Next == checking \/ Timestamp \/ Report_ATCS \/ Verification\/ Terminating

This includes 5 different states. The 'Checking' state occurs when the GNSS sends positioning data to all the airplanes, and then all the airplanes report their locations to ATCS.

In the group verification strategy noted in Figure 3, airplanes were using timestamps on exchanged signals to triangulate each other using time delay of transmission. The 'Timestamp' state represents this communication between airplanes. When a new airplane that possibly might be an attacker appears, it must communicate with other authentic airplanes (even if just for TCAS) and exchange timestamps.

In the 'Report_ATCS' state, all airplanes also report to ATCS with timestamps for routine operations↓and these can be used for position verification from a ground point of view.

The system must then enter the 'verification' state when ATCS does the time delay and radio directional calculations on these signals and compares them to the first GNSS location reports received from all airplanes. Based on these calculations the ATCS will identify which airplanes are reporting their position faithfully and which are not. It will then broadcast the result with a safe/unsafe report to everyone in the airspace.

The termination state allows infinite stuttering to prevent deadlock on the termination (when pc = "Done").

The statement:

Spec == Init /\ [][Next]_vars

represents the specification of the whole system, where Init means the starting state and []Next_vars means Next state must be true for the entire behavior with the option of keeping all variables unchanged.

5.1. TLC Result

The TLC model checker is a tool to validate the TLA+ module and check it for any errors. We run the TLC model checker for our TLA+ ADS-B module and it shows that the module is valid with no errors as well as no deadlock. Also, as appears in the figure below, the TLC model checker shows how many times each state was visited in the module, and what values changed to arrive at the next state.

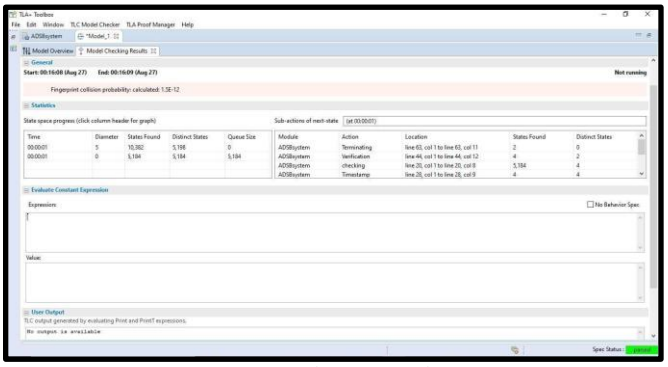


Figure 7: TLC Model check for validity of TLA+ spec in Figure 6.

6. Conclusion and Future Work

At this stage, our model for Figure 3 for verifying physical aircraft locations to combat ghost aircraft and incorrect position reporting issues shows no deadlock. We see that a security system using multilateration and time delay of arrival can identify which aircraft are ghosts and which are not, without getting into a deadlock for resources with the ADS-B system itself.

What this leads us to conclude is that this multi-level modeling of use case-> state diagram-> sequence diagram-> TLA+ specification-> TLC model check can be used to model a large and discrete system like ADS-B to see how it would function and what it would fail at. This is what TLC does it goes through iterations and finds failure points. It throws different numbers at the problem and sees what 'sticks to the wall,' what causes the system to break. We can conclude, at least on this level, that our procedure to model the ADS-B this way and the ability to test the model is valid.

It is always advisable to design security into a system from the beginning of the design process. We hope our method of using UML + state diagrams + sequence diagrams + TLA+ can not only provide the reader with a stable, practical ADS-B model that is secure and realistic but also act as a model of how to design security into other discrete systems in the future.

At the beginning of this paper, we compared the vast ADS-B to a small CANBUS. The roadmap we used here has already been used to formally model smart schools and sewerage systems. When a certain method is successfully used to design various systems of different sizes and scope successfully, it has the potential to develop into a standard. We hope other authors can expand on our work. Our diagrams can be used as an example of how to do this. They also provide a means to examine multiple solutions for strengths, weaknesses, and costs, before any detailed design or implementation takes place.

*6.1. Future work- **An Overall ADS-B** Model in TLA+*

The only weakness of our model is that it is not detailed enough. A true ADS-B system model with all subsystems and security systems and multiple attackers will be much larger in scope. While we model only one attack above, a real-world model will have all attacks and all security systems integrated and will run them in different ways and see when and where two security systems 'collide.'

And that is our future goal.

Following our methodology, the first step towards envisioning the entirety of ADS-B in TLA+ would start with a diagrammatic representation of what we want, as in Figure 8. We believe this will be an extensive task requiring more hands and time than we have right now. We can, however, think about what such a system would look like.

We make some basic assumptions.

No true communication can take place as long as the lines of communication remain unverified. We must find a way to secure the ADS-B physical layer without compromising on openness.

A lot of solutions impinge on trusting information systems like GNSS. Before we trust any data, we must know it came from the right source. Thus, a periodic 'wellness check' on vital infrastructures like transmission systems, radio frequencies, and GNSS should be the second step.

The third step will utilize the data gleaned from steps 1 and 2 to create a model of the system. We would get an idea of how safe or unsafe it is and would be able to deploy mitigation measures accordingly.

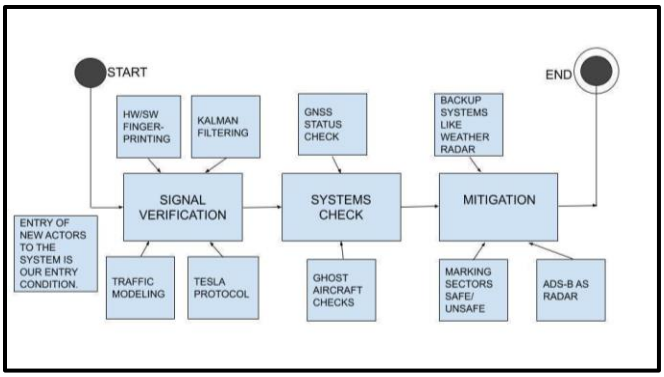


Figure 8: What an overall security solution for ADS-B might look like.

Once again, this would most probably happen on a local level. However, in the spirit of TLA+, we concern ourselves more with the 'what' than the 'how' and do not get too deep into details.

On a higher level, an overall ADS-B specification will consist of multiple figures like Figures 3,4,5, and 6, created for each security challenge, coming together and coalescing into one super-state diagram. This is why we see this paper as a stepping stone to achieving a formally designed ADS-B specification and showing the proper steps to carry out this process.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

All the use case diagrams in section 4 were made using chart-making tools at www.lucidchart.com and the authors would like to thank the website.
The authors would also like to thank Mr. Leslie Lamport for the TLA+ tool which is a fundamental part of this paper.
The authors would also like to thank the University of Cincinnati for its support.

7. References

[1] P. Bhardwaj, C. Purdy, "System design methodologies for safety and security of future wireless technologies in aviation," in 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), 235-238, 2019, doi: 10.1109/MWSCAS.2019.8885347.

[2] FEDERAL AVIATION ADMINISTRATION, May 2010, Federal regulation 14 CFR 91.225S, [Online], Available: http://www.ecfr.gov/cgi-bin/text-idx?node=14:2.0.1.3.10#_top.

[3] FEDERAL AVIATION ADMINISTRATION, May 2010, Federal regulation 14 CFR 91.227, [Online], available: http://www.ecfr.gov/cgi-bin/text-idx?node=14:2.0.1.3.10#se14.2.91_1225.

[4] International Civil Aviation Organization, Guidance material on comparison of surveillance technologies (Ed. 1), [Online], Available: http://www.icao.int/APAC/Documents/edocs/cns/gmst_technology.pdf.

[5] M. Strohmeier, V. Lenders, I. Martinovic, "On the security of the automatic dependent surveillance-broadcast protocol," IEEE Communications Surveys and Tutorials, 2015, doi:10.1109/COMST.2014.2365951.

[6] M. Schäfer, V. Lenders, I. Martinovic, "Experimental analysis of attacks on next generation air traffic communication," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, doi:10.1007/978-3-642-38980-1_16.

[7] M. Riahi Manesh, N. Kaabouch, "Analysis of vulnerabilities, attacks, countermeasures and overall risk of the Automatic Dependent Surveillance-Broadcast (ADS-B) system," International Journal of Critical Infrastructure Protection, **19**, 16-31, 2017, doi:10.1016/j.ijcip.2017.10.002.

[8] FAA Safety Team, 2019, Team notice NOTC8274, [Online], Available: http://www.faasafety.gov/SPANS/noticeView.aspx?nid=8274.

[9] P. Kocher, R. Lee, G. McGraw, A. Raghunathan, S. Ravi, "Security as a new dimension in embedded system design," in Proceedings - Design Automation Conference, 2004, doi:10.1145/996566.996771.

[10] K. Sampigethaya, R. Poovendran, S. Shetty, T. Davis, C. Royalty, "Future E-enabled aircraft communications and security: The next 20 years and beyond," Proceedings of the IEEE, 2011, doi:10.1109/JPROC.2011.2162209.

[11] ICAO International Communications Group, April 2006, Introduction to ACARS messaging services as implemented via Iridium satellite link, http://www.icao.int/safety/acp/inactive%20working%20groups%20library/acp-wg-m-iridium-7/ird-swg07-wp08%20-%20acars%20app%20note.pdf.

[12] C. W. Lin, A. G. Vincentelli, Security-aware design for cyber-physical systems, Book, Springer, 2017, doi: 10.1007/978-3-319-51328-7.

[13] G. Kalakota, Hierarchical partition based design approach for security of CAN bus based automobile embedded system, Electronic Thesis or Dissertation, University of Cincinnati, 2018.

[14] Object Management Group, 2005, Introduction to OMG's Unified Modeling Language (UML®), http://www.uml.org/what-is-uml.htm.

[15] J. Vidal, F. De Lamotte, G. Gogniat, P. Soulard, J.P. Diguet, "A co-design approach for embedded system modeling and code generation with UML and MARTE," in Proceedings -Design, Automation and Test in Europe, DATE, 2009, doi:10.1109/date.2009.5090662.

[16] J. Jürjens, P. Shabalin, "Tools for secure systems development with UML," in International Journal on Software Tools for Technology Transfer, 2007, doi:10.1007/s10009-007-0048-8.

[17] S. Tasiran, Y. Yu, B.Batson, "Using a formal specification and a model checker to monitor and direct simulation," Design Automation Conference, 2003, doi:10.1109/dac.2003.1219024.

[18] Darlene Storm, August 2012, "Curious hackers inject ghost airplanes into radar, track celebrities' flights," Computerworld, [Online], Available: http://www.computerworld.com/article/2472455/curious-hackers-inject-ghost-airplanes-into-radar--track-celebrities--flights.html.

[19] A. Costin and A. Francillon, July 2012, "Ghost in the air (traffic): on insecurity of ADS-B protocol and practical attacks on ADS-B devices," [Online], Available: http://www.researchgate.net/publication/267557712.

[20] S. Collins, S. McCombie, "Stuxnet: the emergence of a new cyber weapon and its implications," Journal of Policing, Intelligence and Counter Terrorism, 2012, doi:10.1080/18335330.2012.653198.

[21] Niall Firth, MIT Technology Review, June 2019, "How to fight a war in space (and get away with it),"http://www.technologyreview.com/2019/06/26/725/satellite-space-wars/.

[22] Defense Intelligence Agency, January 2019, "Challenges to Security in Space,"http://www.dia.mil/Portals/27/Documents/News/Military%20Power%20Publications/Space_Threat_V14_020119_sm.pdf.

[23] European GNSS Service Center, European Global Navigation Satellite Systems Agency, 23 November 2017, Notice advisory to Galileo users (NAGU) 2017045," [Online], Available: http://www.gsc-europa.eu/notice-advisory-to-galileo-users-nagu-2017045.

[24] Federal Aviation Administration, 2016, "EMB-300 Phenom yaw damper failure due to unreliable or unavailable GPS signal," [Online], available: http://www.faa.gov/documentLibrary/media/Notice/GENOT_7110_711_EMB-300.pdf.

[25] C4ADS innovation for peace, 2019, "Above us Only Stars- Exposing GPS Spoofing in Russia and Syria," [Online], Available: http://www.c4reports.org/aboveusonlystars.

[26] M. Strohmeier, I. Martinovic, "On passive data link layer fingerprinting of aircraft transponders," in CPS-SPC 2015 - Proceedings of the 1st ACM Workshop on Cyber-Physical Systems-Security and/or Privacy, co-located with CCS 2015, 2015, doi:10.1145/2808705.2808712.

[27] M.S. Huang, R.M. Narayanan, Y. Zhang, A. Feinberg, "Tracking of noncooperative airborne targets using ADS-B signal and radar sensing," International Journal of Aerospace Engineering, 2013, doi:10.1155/2013/521630.
[28] S. Bachmann, V. DeBrunner, D. Zrnic, "Detection of small aircraft with doppler weather radar," in 2007 IEEE/SP 14th Workshop on Statistical Signal Processing, 443-447, 2007, doi: 10.1109/SSP.2007.4301297.
[29] S. Rzewuski, K. Kulpa, A. Gromek, "Airborne targets detection using weather radar," in 2015 Signal Processing Symposium, SPSympo 2015, 2015, doi:10.1109/SPS.2015.7168305.
[30] Aircraft Owners and Pilots Association, 2019, "2019 State of general aviation,"http://download.aopa.org/hr/Report_on_General_Aviation_Trends.pdf.
[31] H. Yang, Q. Zhou, M. Yao, R. Lu, H. Li, X. Zhang, "A practical and compatible cryptographic solution to ADS-B security," IEEE Internet of Things Journal, 2019, doi:10.1109/JIOT.2018.2882633.
[32] Leslie Lamport, 4 September 2018, "A High-Level View of TLA+," [Online], Available: http://lamport.azurewebsites.net/tla/high-level-view.html.
[33] C. Newcombe, T. Rath, F. Zhang, B. Munteanu, M. Brooker, M. Deardeuff, "How amazon web services uses formal methods," Communications of the ACM, 2015, doi:10.1145/2699417.
[34] C. Newcombe, "Why Amazon chose TLA+," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, doi:10.1007/978-3-662-43652-3_3.
[35] N.H. Obeidat, C. Purdy, "Modeling a smart school building system using UML and TLA+," in Proceedings - 3rd International Conference on Information and Computer Technologies, ICICT 2020, 2020, doi:10.1109/ICICT50521.2020.00028.
[36] S. Latif, A. Rehman, N.A. Zafar, "Modeling of sewerage system linking UML, automata and TLA+," in 2018 International Conference on Computing, Electronic and Electrical Engineering, ICE Cube 2018, 2019, doi:10.1109/ICECUBE.2018.8610971.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1666-1673 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Real-time Gradient-Aware Indigenous AQI Estimation IoT Platform

Hasan Tariq[*,1], Abderrazak Abdaoui[1], Farid Touati[1], Mohammad Abdullah Al Hitmi[1], Damiano Crescini[2], Adel Ben Mnaouer[3]

[1]*Department of Electrical Engineering, College of Engineering, Qatar University, 2713, Qatar*

[2]*Dipartimento di Ingegneria dell'Informazione, Brescia University, 25121, Italy*

[3]*Department of Computer Engineering and Computational Sciences, Faculty of Engineering, Applied Sciences and Technology, Canadian University Dubai, 117781, UAE*

ARTICLE INFO

Article history:
Received: 16 July, 2020
Accepted: 08 October, 2020
Online: 28 December, 2020

Keywords:
Air quality
IoT
Gas sensors node
Multi-variable environmental mapping

ABSTRACT

Environmental monitoring has gained significant importance in outdoor air quality measurement and assessment for fundamental survival as well as ambient assisted living. In real-time outdoor urban scale, instantaneous air quality index estimation, the electrochemical sensors warm-up time, cross-sensitivity computation-error, geo-location typography, instantaneous capacity or back up time; and energy efficiency are the six major challenges. These challenges lead to real-time gradient anomalies that effect the accuracy and pro-longed lags in air quality index mapping campaigns for state and environmental/meteorological agencies. In this work, a gradient-aware, multi-variable air quality sensing node is proposed with event-triggered sensing based on position, gas magnitudes, and cross-sensitivity interpolation. In this approach, temperature, humidity, pressure, geo-position, photovoltaic power, volatile organic compounds, particulate matter (2.5), ozone, Carbon mono-oxide, Nitrogen dioxide, and Sulphur dioxide are the principle variables. Results have shown that the proposed system optimized the real-time air quality monitoring for the chosen geo-spatial cluster (Qatar University).

1. Introduction

Living conditions or probability of life forms are dependent on environmental conditions, i.e. combination of gases, temperature, humidity, pressure, and light [1]. The quality of respiratory life processes is directly related to air quality at a particular geo-location [2]. Gradual deterioration in urban air quality is ambient each year due to the increasing population, chemical emissions from machinery, and depreciation in green ecology [3]. Air quality sustainability is a major concern in the age of imperialism [4] due to exponential development and construction density in the urban landscape.

An increase of 41% in the design and development of gas and pollutant sensing systems was observed in the environmental market since 2001 [5]. The top contributions AirNut, PA-I and PA-II, Egg, PATS+, and S-500, CairClip, Portable ASLUNG, AirSensEUR, Met One, AQY v0.5, Vaisala AQT410, 2B Tech, and AQMesh V3.0 systems had measurement capabilities in specific pollutants and gases presented in [6–8]. FIS SP-61 by FIS, O3-3E1F by CityTechnology, AirSensEUR v.2 by LiberaIntentio, and S-500 by Aeroqual, and AirSensEUR used a built-in AlphaSense OX-A431 limited to O3 [9–11]. The PMS1003 and PMS3003 by Plantower; DC1100 PRO and DC1700 by Dylos; OPC-N2 by AlphaSense had only sensing support for particulate matter (PM) [11–13]. The CO-3E300 by City Technology; CO-B4 by Alphasense, MICS-4515 by and SGX Sensortech, Smart Citizen Kit by Acrobotic and the RAMP had only abilities to measure carbon mono-oxide [4]. It illuminated the requirement of structure real-time sensor data processing framework for multiple gases sensing nodes with warm-up times below 5mins at ambient outdoor environmental conditions.

The top 10 environmental protection agencies (EPAs) unanimously agreed on the standard of four core gases for outdoor air quality [14] i.e. Ozone, Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide, and two prime pollutants particulate matter PM2.5 and PM10. The concentration or ratio of four critical gases and air particles part per million (ppm) constitutes air quality index (AQI) is the fundamental design using Smart Rig [15]. The ratio of

[*] Corresponding Author: Hasan Tariq, Lab106A, B09, Qatar University, P.O. Box 2713, Doha, Qatar +97450419852, hasan.tariq@qu.edu.qa

Ozone (O3), Sulphur Dioxide (SO2), Carbon Monoxide (CO), and Nitrogen Dioxide (NO2) are the standard gases discussed in major environmental standards for AQI [4,14,15]. A standard outdoor air quality node (O-AQN) was needed to measure discrete and cumulative AQI with minimum data processing and power consumption.

Related Works

The existing O-AQNs, Urban AirQ, Smart Citizen Kit, SeReNo V1, AirQ Mesh needed improvement in energy efficiency, GPS, time-series vector structuring, and remote data transmission to ensure dynamic air quality mapping [16–20].

1.1. Urban AirQ

The AirQ platform [17] by Vishal et al was a smart and cost-effective solution for air quality monitoring. It supported wireless data communication technologies including WiFi, 4G, Bluetooth, LoRa, etc. It was supported by non-embedded data management and the implementation of smart data analytics algorithms exhibited at an AirQ Dashboard and Smartphone App. It had a dependency on external hardware and software with more processing needs with only air quality sensors.

1.2. Smart Citizen Kit

The Smart Citizen Kit [18] by Gullien et al was dependent on an additional resource named "Smart Citizen Station". Needed improvement in multi-variable time-series measurements and location accuracy, energy efficiency, and reduced data processing. Multi-variable sensing with simultaneous measurement of gases and environmental variables.

1.3. SeReNo V1

The Sensor Node (SeReNo) V1 [19] by Farid et al was the first node developed with indoor air quality monitoring and energy harvesting capabilities. SeReNo V2 presented in this work [20] is the improvement of [19] that needed to be upgraded with GSM and GPS, more Flash to perform AQI data processing needed for optimized energy efficiency, electrochemical gas sensors warm-up times, and outdoor air quality.

1.4. AirQ Mesh

The networked AirQ Mesh [21] by Nuria et al focused on the review and performance of existing AirQ platforms and their maximum utilization leading to credible results by EPA and Air Quality Directive (AQD). Data payloads were very high due to response time estimation errors leading to high storage requirements with energy challenges in the field were the gaps in the Nuria et al contribution.

Furthermore, the improvement potential in context-aware data processing model and thresholding for detection was discussed by Hasan et al in [22,23] for multi-sensor single-board node. The energy efficiency schemes were one major gap highlighted by [24]. The works Lebilium Waspmote [21] was Zigbee but needed GSM/GPRS support for outdoor sensing and have been more contributory if equipped with geo-spatial traces or cadastral layouts support for all the measured variables like AirBox [22]. An infrastructure embedded system was needed to accommodate the maximum sensing, minimum data processing, AQI estimation, and long-haul deployability [19–23].

2. Materials and Methods

This work proposed a smart AQM IoT system with three major components:

- The Real-time Gradient Aware Multi-Variable Sensing Model (GAM)
- The SeReNo V2 AQM Sensing Platform
- Real-time AQI Estimation Engine (AQM-EE)

First, we proposed the GAM and then the pillars of GAM led to the design and fabrication SeReNo V2 node.

3. The Real-time Gradient Aware Multi-Variable Sensing Model (GAM)

A real-time structured multi-variable structured data time-series vector was needed to proceed with AM. Let us consider an EPA standard outdoor air quality index (O-AQI) real-time variables as temperature T in centigrade, pressure P in pascals, humidity H in %, volatile organic compounds as VoC (ppm), particulate matter as PM (ug/m3), Ozone as $O3$ (ppm), Nitrogen Dioxide as $NO2$ (ppm), Carbon Monoxide as $CO2$ (ppm), and Sulphur Dioxide as $SO2$ (ppm). The real-time O-AQI data was proposed as a commutative time series multi-variable vector $V_{O\text{-}AQI}$ of two non-linear time-series with t_1 and t_2 of environmental E and gas G sensors data at a particular geo-location L given as:

$$V_{O\text{-}AQI}(t) = [E(t_1), G(t_2)] : L(t) \quad \text{where } t = (0, 1, 2, 3, \ldots\} \tag{1}$$

The practicality of response time of heterogeneous sensors was taken in account for non-linear time-series decomposition t, gas sensor response time t_2 is greater than the response time of environmental sensors t_1 with relationship $t_2 > t_1$ given as:

$$t_2 = 3t_1 \qquad where\ [t_1, t_2] \in t \tag{2}$$

The environmental sensor variables function E for sensor array $A_E(T, P, H, VoC, PM)$ as $E(A_E, t_1)$; and for gas sensors array $A_G(O3, NO2, SO2, CO)$ as $G(A_G, t_2)$ and position vector L as reference function GPS using GSM network cell locations (using $AT+CIPGSMLOC=1,1$) for L_{GPRS} and GPS module as L_{GPS} (using AT+CGPSINF). For precise AQM the L_{GPS} must belong to the slope of L_{GPRS1} and L_{GPRS2} in a particular slope format by NEMEA specifier for consecutive cells is given as:

$$L_{GPS}(X, Y) \in [L_{GPRS1}(X2, Y2), L_{GPRS2}(X1, Y1)] \tag{3}$$

The agreed L_{GPS} was termed as $L(t)$ where condition (3) was satisfied. From equations (1), (2), and (3) the finalized AQM vector of $V_{O\text{-}AQI}$ was derived as:

$$V_{O\text{-}AQI}(t) = [E(A_E(T, P, H, VoC, PM), t_1), G(A_G(O3, NO2, SO2, CO), t_2)] : L(t) \tag{4}$$

The transmission of this structured AQM vector can be power leeching. Here we introduce a gradient impact vector as a function of warm-up time t_W, cross-sensitivity classifier C_S, geo-location $L(t)$, and energy efficiency E_E. Three bounded value conditions we applied on GAM programmed in the SeReNo V2 firmware:

a) The mandatory system unit to stay on is the power is the micro-controller with power consumption P_M and at sleep mode, it has $P_{M\text{-}SLEEP}$ power with all the critical sensors active only.

b) For E_E the system has to have instantaneous power P_i varying between $P_{M\text{-}SLEEP} < P_i < P_M$.

c) The t_W had to be in the lower section of the curve of P_i. Likewise, at the constant magnitude of vector A_E at t_n and t_{n+1}, the difference cross-sensitivities had to stay constant at a particular $L(t)$.

4. The SeReNo V2 AQM Sensing Platform

A generic multi-variable AQM node consists of an EPA recommended sensors for $V_{O\text{-}AQI}$ based on equation (4), application-specific signal conditioning, power, and communication sections. The block diagram of a generic AQM node is given below in figure 1.

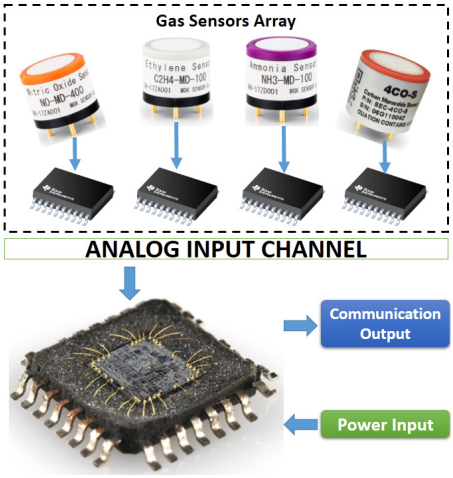
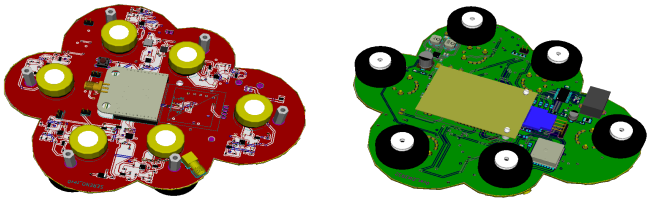


Figure 1: A Generic Multi-Variable AQM Sensing

In our previous works [14,15,19,20], we had filled the design gaps in indoor air quality monitoring and developed a multi-sensor AQM node called Sensors V1 revision 1 and revision 2 presented in figure 1. In this work, work proposed a novel O-AQM node by improving the system architecture of SeReNo V1 as SeReNo V2. In our recent work [24] we had developed SeReNo V2 nodes tested for long-haul autonomous operations of 1.5 years that were improved in this work. The existing SeReNoV2 node was improved with GAM and AQI-EE to meet the state of the art needs.



a. Top View b. Bottom View

Figure 2: Two 3D Layouts of SeReNo V1 AQM Node

For outdoor AQM SeReNo V1 was improved to V2 as per global EPAs standards recommended sensors, energy harvesting long-range communication support. The comparison of major capabilities of SeReNo V1 and V2 is exhibited in the table below.

Table 1: SeReNo V1 and V2 Characteristics

Blocks	SeReNo V1	SeReNo V2
Energy Block	PV and Manual Charging + LiPo Battery 3.7V / 1200mAh + 6 Thermo-Electric Generators	Dual PV + Manual Charging
Communication Block	RF 433Mhz	GSM/GPRS
Sensors Block	6 Indoor Sensors Only interfaced with Analog Input	10 Outdoor Sensors Networked with I2C (4 Gas Sensors + 2 Pollutant Sensors + 4 Environmental Sensors)
Environmental Sensors Block	Absent	4 Sensors
Controller Block	RFC200	ATMega328P
Location Block	NIL	GPS
Local Display	NIL	Present
Energy Optimization	Partial	Full
Sensors Control	NIL	Present
Outdoor Cross Sensitivity	NIL	Present

GAM was a core engine driving the multi-sensor O-AQM IoT system. Mathematically modeled design and equations (1~4) parameters are addressed in the hardware blocks. A 1-1 correspondence electronics and instrumentation system was designed SeReNo V2 presented in figure 3 presented below.

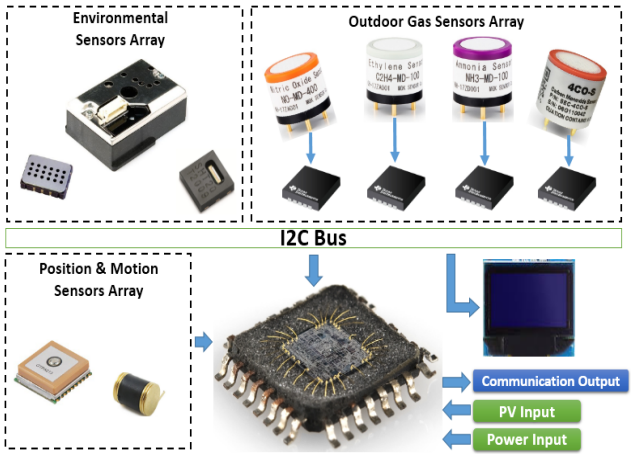
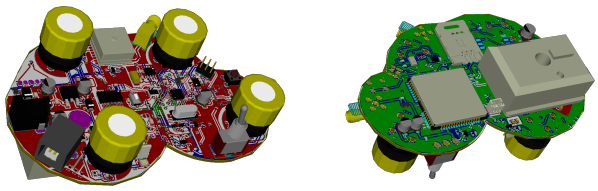


Figure 3: Block Diagram of SeReNo V2 AQM Node

Fig 3 contains three sets of sensors The PCB design and completely assemble functional O-AQM node is presented in figure 4 given below.



a. Top View b. Bottom View

Figure 4: Two 3D Layouts of SeReNo V2 AQM Node

Three SeReNo V2 nodes with GAM and AQI-EE as improved firmware were deployed in QU for outdoor testing. The deployment was focused on the resolution of six key improvements mentioned in the abstract and literature review through proposed GAM and AQI-EE for six months, i.e. QU Greenhouse exhibited in figure 5 below.



Figure 5: SeReNo V2 Deployment in QU to utilize the GAM

This deployment was for six months and all the data was collected from Ubidots IoT platform as CSV files for further processing. The first month was stability assessment of GAM and AQI-EE firmware and the remaining 5 months were the measurement campaign.

5. Real-time AQI Estimation Engine (AQM-EE)

An AQI magnitude is estimated from 5 critical air variables or termed as pollutants, i.e. PM2.5, O3, NO2, SO2, and CO. The critical variable with the maximum AQI magnitude is termed as the "principle pollutant" and entire computation revolves around its significant figures among all the critical air variables. Let I_p be the index of the principal pollutant; C_p as the rounded-off concentration of pollutant p; $B_{P\text{-}high}$ as the breakpoint greater or equal to C_p; $B_{P\text{-}low}$ as the breakpoint less than or equal to C_p; I_{high} is the AQI corresponding to $B_{P\text{-}high}$; I_{low} is the AQI corresponding to $B_{P\text{-}low}$. The AQI is generically estimated as:

$$I_P = [(I_{high} - I_{low})/(B_{P-high} - B_{P-low})] \times (C_P - B_{P-low}) + I_{low} \quad (5)$$

Every pollutant was formulated in using equation (5) and given by in equations (6 to 10).

$$I_{O3} = [(I_{high} - I_{low})/(B_{O3-high} - B_{O3-low})] \times (C_{O3} - B_{O3-low}) + I_{low} \quad (6)$$

$$I_{SO2} = [(I_{high} - I_{low})/(B_{SO2-high} - B_{SO2-low})] \times (C_{SO2} - B_{SO2-low}) + I_{low} \quad (7)$$

$$I_{CO} = [(I_{high} - I_{low})/(B_{CO-high} - B_{CO-low})] \times (C_{CO} - B_{O3-low}) + I_{low} \quad (8)$$

$$I_{NO2} = [(I_{high} - I_{low})/(B_{NO2-high} - B_{NO2-low})] \times (C_{NO2} - B_{NO2-low}) + I_{low} \quad (9)$$

$$I_{PM} = [(I_{high} - I_{low})/(B_{PM-high} - B_{PM-low})] \times (C_{PM} - B_{PM-low}) + I_{low} \quad (10)$$

Furthermore, the standard AQI evaluation template for a specific region provided by EEA as Common Air Quality Index (CAQI) and US EPA as AQI were adapted in this AQM-EE and given as a cumulative state chart given below in figure 6 and 7.

Pollutant	Index level (based on pollutant concentrations in µg/m3)				
	Good	Fair	Moderate	Poor	Very poor
Particles less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-800
Particles less than 10 µm (PM_{10})	0-20	20-35	35-50	50-100	100-1200
Nitrogen dioxide (NO_2)	0-40	40-100	100-200	200-400	400-1000
Ozone (O_3)	0-80	80-120	120-180	180-240	240-600
Sulphur dioxide (SO_2)	0-100	100-200	200-350	350-500	500-1250

Figure 6: CAQI EEA AQI Cumulative State Chart

The SeReNo V2 was successfully able to give real-time AQI-EE estimations that have plenty of environmental, metrological, and health applications and significance.

Breakpoints							AQI	Category
O_3 (ppm) 8-hour	O_3 (ppm) 8-hour[1]	PM_{10} (µg/m³)	$PM_{2.5}$ (µg/m³)	CO (ppm)	SO_2 (ppm)	NO_2 (ppm)		
0-0.064	–	0-54	0-15.4	0-4.4	0-0.034	(2)	0-50	Good
0.065 0.084	–	55 154	15.5 40.4	4.5 9.4	0.035 0.144	(2)	51 100	Moderate
0.085-0.104	0.125-0.164	155-254	40.5-65.4	9.5-12.4	0.145-0.224	(2)	101-150	Unhealthy for sensitive groups
0.105-0.124	0.165-204	255-354	65.5-150.4	12.5-15.4	0.225-0.304	(2)	151-200	Unhealthy
0.125-0.374 (0.155-0.404)[4]	0.205-0.404	355-424	150.5-250.4	15.5-30.4	0.305-0.604	0.65-1.24	201-300	Very unhealthy
(3)	0.405-504	425-504	250.5-350.4	30.5-40.4	0.605-0.804	1.25-1.64	301-400	Hazardous
(3)	0.505-0.604	505-604	350.5-500.4	40.5-50.4	0.805-1.004	1.65-2.04	401-500	Hazardous

Figure 7: US EPA AQI Cumulative State Chart

The EPA rated sensors produced the results in thresholds of the chart in figure 6.

The AQM-EE uses 5 KPIs to evaluate the performance of GAM given as:

5.1. Energy Efficiency:

The instantaneous energy consumed by node E_N for a minute is the sum of 60 secs for the product of supply voltage V_{DC}, instantaneous current I_i, and time T_i given in equation 11.

$$E_N = \sum_{i=0}^{n} V_{DC} * I_i * T_i \quad (11)$$

The difference between the maximum rated power consumption of node and minimum instantaneous power consumption in terms of watt-hour is termed as energy efficiency or E-Efficiency symbolized as E_e and given by:

$$E_e\ (Wh) \approx \left(\frac{N-M}{N}\right)(E_N) \quad (12)$$

5.2. Cross-Sensitivity Computation Error:

The cross-sensitivity computation error is the average of the total difference in a sensor's reaction to other gases being measured.

$$C_s\ (\%) \approx \left(\sum_{i=0}^{n} \Delta G_i\right)/n \quad (13)$$

It is used as the correction factor and must be in the range defined by the vendor.

5.3. Typo-graphic Error:

The difference in the GPS coordinates by GPS and AGPS was termed as typo-graphic or GPS error, presented in equation below:

$$T_E = |L_{GPS}(X, Y) - L_{AGPS}(X, Y)| \quad (14)$$

The acceptable value for this error was marked as 20 meters for precise AQM.

5.4. The ratio of Magnitude E_{BAT}/E_N or Instantaneous Capacity:

The ratio of instantaneous system battery energy E_{BAT} in watt-hour (Wh) and instantaneous node energy E_N in watt-hour (Wh) was termed the instantaneous capacity magnitude C_{MAG} of the node with current operations given as:

$$C_{MAG} = E_{BAT} / E_N \quad (15)$$

For long haul AQM, the C_{MAG} must be greater than 1 to ensure at least 1-hour autonomous operation.

5.5. Improved Warm-Up Time of Gas Sensors:

The start-up warm-up time is mandatory for the gas sensors to acquire chemical equilibrium with the atmosphere in which they are used. The warm-up time was improved by putting all the micro-controller, GPS, and GSM modules to sleep and then get active based on the gradient as mentioned in section 3.

6. Results and Discussion

After the long haul deployment of six months, the data results obtained were displayed on the Ubidots IoT platform as exhibited in figure 7.

The eleven real-time variables were exhibited in Figure 8 sending data through GSM model QuecTel M10. The application of GAM enabled only meaningful data to be sent to the cloud that made time-series more non-linear as only gradients impacted values were being transmitted. The following plots of individual variables give more insights of GAM in SeReNo V2.

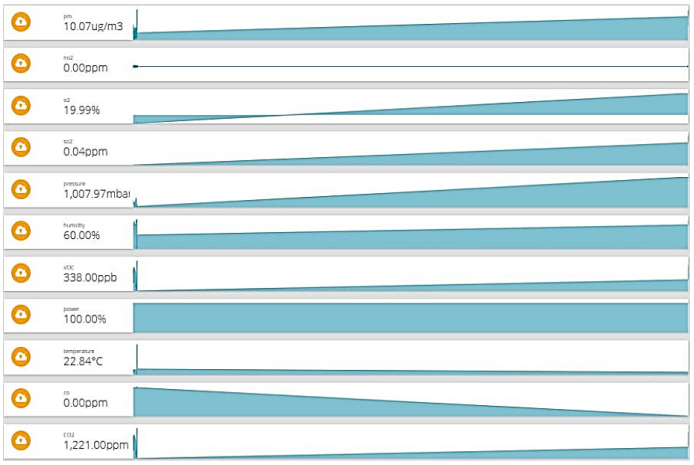


Figure 8: The SeReNo2 IoT Dashboard at Ubidots IoT

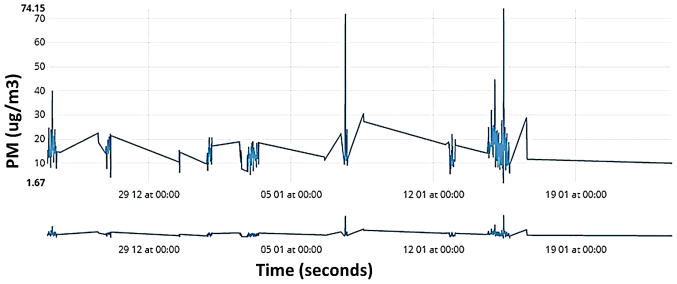


Figure 9: Illustration of Particulate Matter data due to GAM

Figure 9 explained the real-time measurement of PM from December 22, 2019, to January 22, 2020. The max of 74.15 µg/m^3 was observed on January 17, 2020, and a minimum of 1.67 µg/m^3 on December 27, 2019.

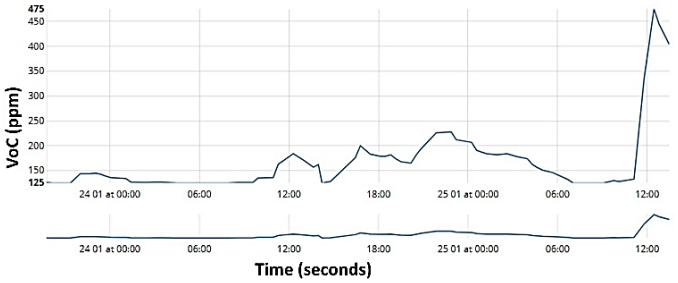


Figure 10: Illustration of VoC and Impact of GAM at 12:00 PM

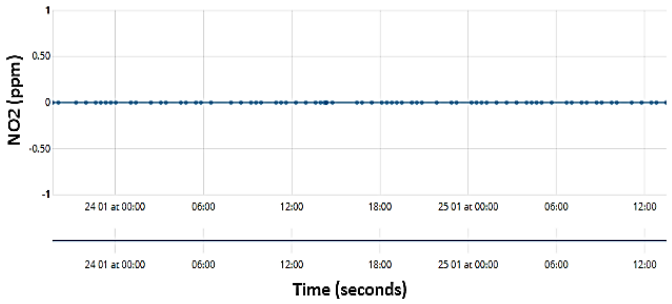


Figure 11: Illustration of Nitrogen Dioxide data due to GAM

In figure 10, it is ambient that amount of VoCs increased significantly due to day time traffic and air pollution in the atmosphere. This the special behavior common in all the days for 6 months.

In figure 11, it is eloquent that NO2 being extremely hazardous stayed zero and was transmitted to cloud whenever there was a signature of cross-sensitivity and time-series for NO2 was also non-linear.

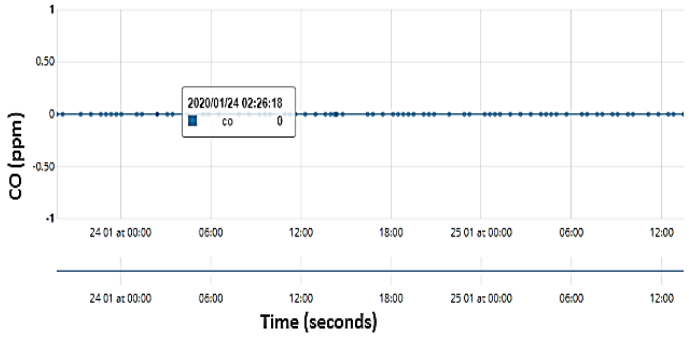


Figure 12: Illustration of Carbon Monoxide data due to GAM

In figure 12, it was observed that CO being hazardous was also observed as a non-linear time-series and the same as NO2 had cross-sensitivity signatures.

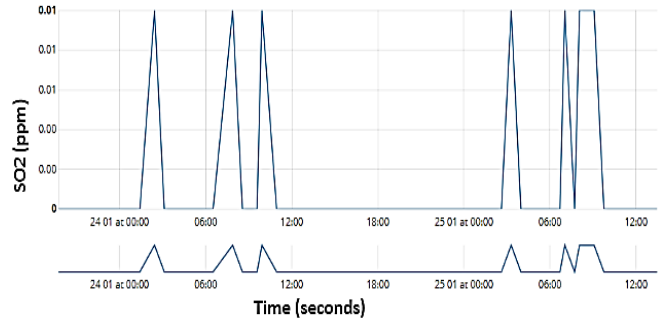


Figure 13: Sulphur Dioxide Signatures in Atmosphere

In figure 13, nominal magnitudes of SO2 were observed in the atmosphere due to plenty of gas processing plants that flare H2S and rest is preserved by Sulphur recovery units (SRUs).

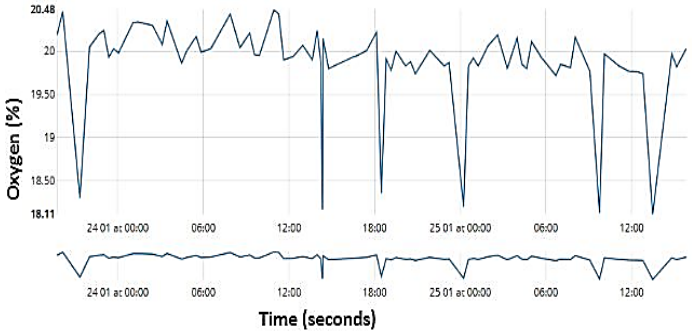


Figure 14: Presence of Oxygen Traces in Air

The SeReNo V2 node was initially equipped with an O2 sensor that was later on replaced by the O3 sensor for 2 years. The breathing conditions were favorable due to the measured average of 19.3 O2 in the air as obvious from figure 14.

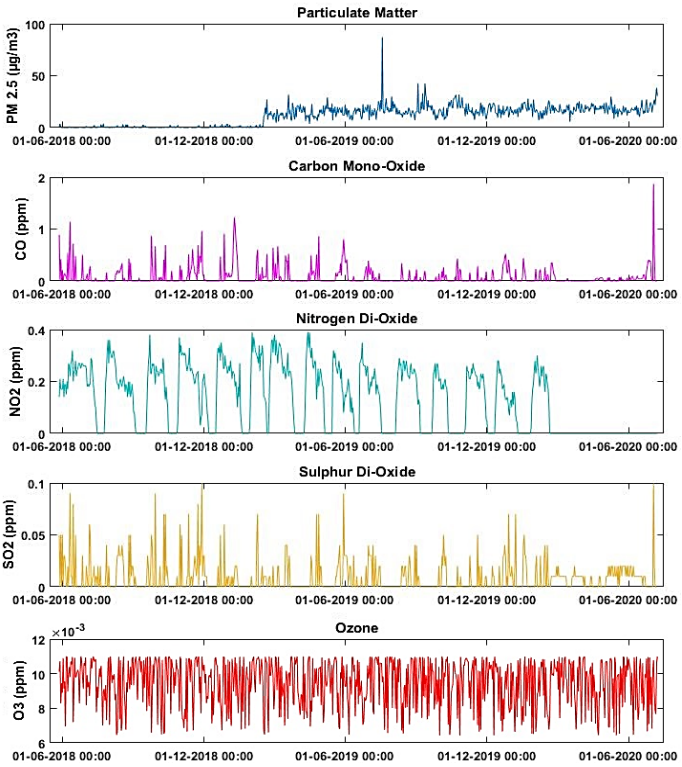


Figure 15: SeReNoV2 Total Deployment Data for EPAs recommended measurement

The first deployment of the SeReNoV2 node was 1.5 years to validate the autonomous operational capabilities as presented in figure 15 at the maximum sampling rate for pollutants and outdoor gases. This deployment was from June 1, 2018, to January 1, 2020.

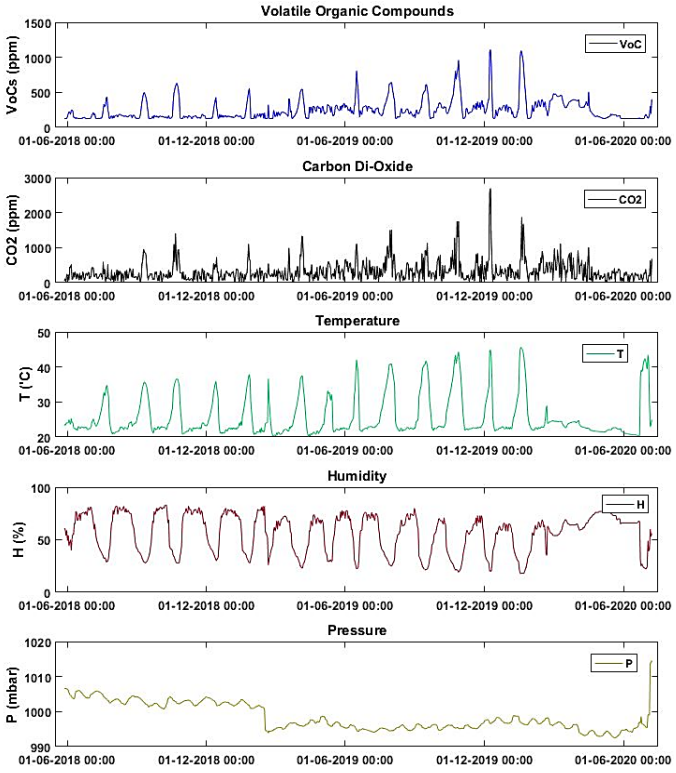


Figure 16: SeReNoV2 Total Deployment Data for EPAs Environmental Variables

In figure 16, the core 5 environmental variables, VoC, CO2, temperature, humidity, and pressure can be observed during the entire deployment tenure.

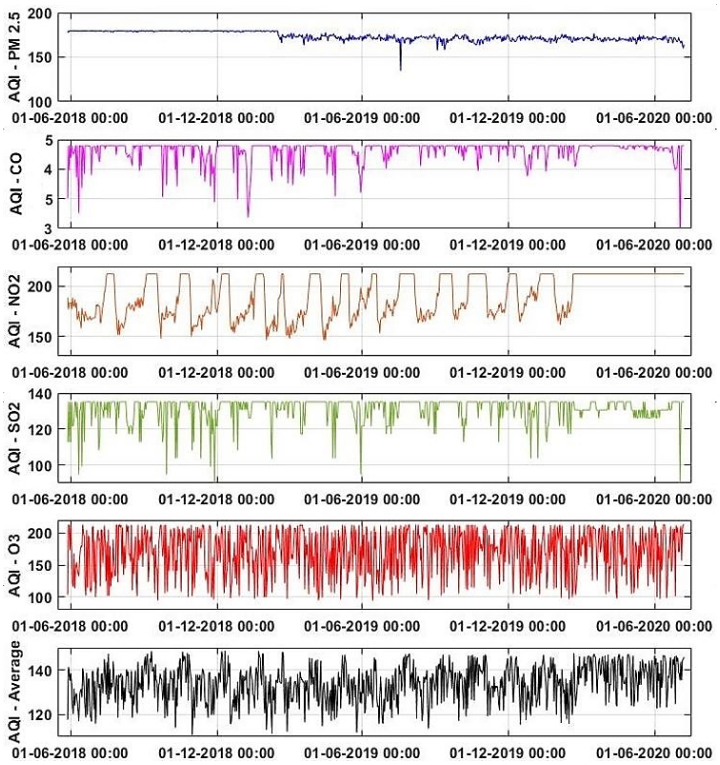


Figure 17: SeReNoV2 Estimation of real-time Discrete and Cumulative AQI

The AQI-EE key performance indicators (KPIs) were estimated for the programmed GAM on SeReNo V2 were the major contributions that enabled all the outcomes presented in figures 17 to 19. 80 meters typographic error can be observed in figure 18.

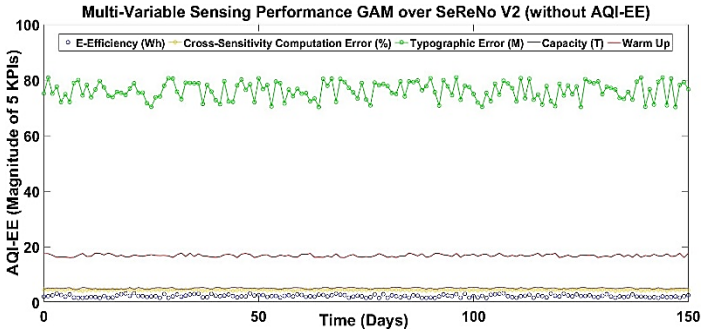


Figure 18: The GAM without AQI-EE KPIs for 5 months

The warm-up time of 18 minutes of electrochemical gas sensors was due to the average humidity of 63%. The warm-up time of electrochemical gas sensors is ideally 2 to 7. The cross-sensitivity error of 5% is normal. The E_e is near 0.8~1% is due to the extremely low power design on SeReNoV2.

The AQI-EE KPIs for GAM efficiency on SeReNo V2 can be observed in these recent 6 months deployment with 5 months of KPI assessment. The correlation of GPS and AGPS minimized typographic error to 3.5 meters. The cross-sensitivities error was also reduced to 2.7~4.8%. The E_e is aligned with the electrochemical gas sensors' operational instances. The capacity has improved to 1.4%. A significant reduction in warm-up times was observed due to the $P_{M\text{-}SLEEP}$ variable. Lastly, the measure data size with respect to power energy efficiency was mapped and significant amounts of data were observed even at very low power consumption as a landmark for future research.

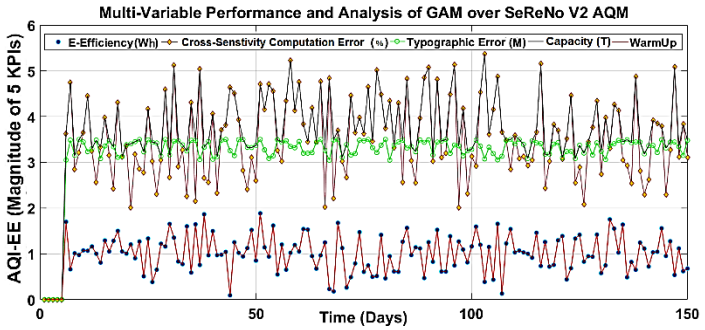


Figure 19: The GAM KPIs for SeReNo V2

7. Conclusion

A gradient-ware urban air quality index estimation multi-variable sensing IoT system was developed. The existing SeReNoV2 nodes were improved with the proposed system. The gradient-aware model and air quality estimation engine contributed to a significant and noticeable reduction in data processing for real-time AQI computation, electrochemical sensors warm-up times, cross-sensitivity errors, geo-location typographic-errors, and improved energy efficiency. The results can be summarized in four key milestones for accurate estimation of urban AQI estimation: 1) implemented gradient-aware sensing model as non-linear time series vector compensating the commutative anomalies in gases and environmental variables measurements; 2) the energy efficiency mechanism was able to keep the node up for 1.5 years constantly without any external charging source; 3) reduced warm times due to the proposed instantaneous power condition, and 4) precise estimation of AQI recommended air quality chart principal parameters. These novel air quality system results demonstrated the potential to handle redundancy in sensing time-triggers and non-linear timers as well.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This publication was made possible by NPRP grant # 10-0102-170094 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "A survey on sensor networks," IEEE Communications Magazine, **40**(8), 102–114, 2002, doi:10.1109/MCOM.2002.1024422.

[2] M. Abderrahim, H. Hakim, H. Boujemaa, F. Touati, "A Clustering Routing based on Dijkstra Algorithm for WSNs," in 2019 19th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), 605–610, 2019, doi:10.1109/STA.2019.8717279.

[3] M. Vinyals, J.A. Rodriguez-Aguilar, J. Cerquides, "A Survey on Sensor Networks from a Multiagent Perspective," The Computer Journal, **54**(3), 455–470, 2011, doi:10.1093/comjnl/bxq018.

[4] F. Touati, H. Tariq, M.A. Al-Hitmi, A. Ben Mnaouer, A. Tahir, D. Crescini, "IoT and IoE prototype for scalable infrastructures, architectures and platforms," International Robotics & Automation Journal, **4**(5), 319–327, 2018, doi:10.15406/iratj.2018.04.00144.

[5] M. Abderrahim, H. Hakim, H. Boujemaa, F. Touati, "Energy-Efficient Transmission Technique based on Dijkstra Algorithm for decreasing energy consumption in WSNs," in 2019 19th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), 599–604, 2019, doi:10.1109/STA.2019.8717210.

[6] P. Temdee, R. Prasad, Communications for Context-Aware Applications, 65–96, 2018, doi:10.1007/978-3-319-59035-6_4.

[7] F. Touati, H. Tariq, D. Crescini, A.B. Manouer, "Design and Simulation of a Green Bi-Variable Mono-Parametric SHM Node and Early Seismic Warning Algorithm for Wave Identification and Scattering," in 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), 1459–1464, 2018, doi:10.1109/IWCMC.2018.8450277.

[8] N. Zaghdoud, W.H. Alouane, H. Boujemaa, F. Touati, "Secure Performance of AF and DF Relaying in Cooperative Noma Systems," in 2019 19th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), 614–619, 2019, doi:10.1109/STA.2019.8717242.

[9] H. Taheri, P. Neamatollahi, O.M. Younis, S. Naghibzadeh, M.H. Yaghmaee, "An energy-aware distributed clustering protocol in wireless sensor networks using fuzzy logic," Ad Hoc Networks, **10**(7), 1469–1481, 2012, doi:https://doi.org/10.1016/j.adhoc.2012.04.004.

[10] H. Lin, D. Bai, D. Gao, Y. Liu, "Maximum data collection rate routing protocol based on topology control for rechargeable wireless sensor networks," Sensors, **16**(8), 2016, doi:10.3390/s16081201.

[11] M. Abderrahim, H. Hakim, H. Boujemaa, F. Touati, "Multihop transmission strategy using dijkstra algorithm to improve energy efficiency in WSNs," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer International Publishing: 97–107, 2018, doi:10.1007/978-3-030-02849-7_9.

[12] T. Khan, M.A. Shah, K. M., "Communications, Applications, Technology and Beyond: A Surveyon Smart Grid," Sindh University Research Journal-SURJ, **48**(3), 651–656, 2016.

[13] H. Tariq, A. Tahir, F. Touati, M.A.E. Al-Hitmi, D. Crescini, A. Ben Manouer, "Geographical area network-structural health monitoring utility computing model," ISPRS International Journal of Geo-Information, **8**(3), 2019, doi:10.3390/ijgi8030154.

[14] U. Asghar, F. Touati, D. Crescini, A. Galli, A. Ben Mnaouer, "Development of Highly Efficient Multi-invariable Wireless Sensor System Design for Energy Harvesting," 2018.

[15] A. Abdaoui, F. Touati, H. Tariq, A. Ben Mnouar, "A smart rig for calibration of gas sensor nodes: Test and deployment," DTS 2020 - IEEE International Conference on Design and Test of Integrated Micro and Nano-Systems, (i), 18–20, 2020, doi:10.1109/DTS48731.2020.9196182.

[16] H. Luo, W. Li, X. Wu, "Design of indoor air quality monitoring system based on wireless sensor network," IOP Conference Series: Earth and Environmental Science, **208**, 12070, 2018, doi:10.1088/1755-1315/208/1/012070.

[17] V. Choudhary, J.H. Teh, V. Beltran, H.B. Lim, "AirQ: A Smart IoT Platform for Air Quality Monitoring," in 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), 1–2, 2020, doi:10.1109/CCNC46108.2020.9045550.

[18] G. Camprodon, Ó. González, V. Barberán, M. Pérez, V. Smári, M.Á. de Heras, A. Bizzotto, "Smart Citizen Kit and Station: An open environmental monitoring system for citizen participation and scientific experimentation," HardwareX, **6**, e00070, 2019, doi:https://doi.org/10.1016/j.ohx.2019.e00070.

[19] F. Touati, C. Legena, A. Galli, D. Crescini, P. Crescini, A. Ben Mnaouer, "Environmentally powered multiparametric wireless sensor node for air quality diagnostic," Sensors and Materials, **27**(2), 177–189, 2015, doi:10.18494/sam.2015.1056.

[20] H. Tariq, A. Abdaoui, F. Touati, M.A.E. Al-Hitmi, D. Crescini, A.B. Mnaouer, "An Autonomous Multi-Variable Outdoor Air Quality Mapping Wireless Sensors IoT Node for Qatar," in 2020 International Wireless Communications and Mobile Computing (IWCMC), 2164–2169, 2020, doi:10.1109/IWCMC48107.2020.9148392.

[21] N. Castell, F.R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?," Environment International, **99**, 293–302, 2017, doi:https://doi.org/10.1016/j.envint.2016.12.007.

[22] H. Tariq, F. Touati, M.A.E. Al-Hitmi, D. Crescini, A.B. Manouer, "Design and Implementation of Information Centered Protocol for Long Haul SHM Monitoring," in 2019 IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS), 1–7, 2019, doi:10.1109/DTSS.2019.8915286.

[23] H. Tariq, F. Touati, M.A.E. Al-Hitmi, D. Crescini, A. Ben Mnaouer, "A real-time early warning seismic event detection algorithm using smart geo-spatial bi-axial inclinometer nodes for Industry 4.0 applications," Applied Sciences, **9**(18), 1–25, 2019, doi:10.3390/app9183650.

[24] A. Abdaoui, S.H.M. Ahmad, H. Tariq, F. Touati, A.B. Mnaouer, M. Al-Hitmi, "Energy Efficient Real time Outdoor Air Quality Monitoring System," in 2020 International Wireless Communications and Mobile Computing (IWCMC), 2170–2176, 2020, doi:10.1109/IWCMC48107.2020.9148229.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 1674-1682 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Japanese Abstractive Text Summarization using BERT

Yuuki Iwasaki [*1], Akihiro Yamashita[1], Yoko Konno[2], Katsushi Matsubayashi[1]

[1]*National Institute of Technology, Tokyo College, 193-0997, Japan*

[2]*CHOWA GIKEN Corporation, 001-0021, Japan*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 12 December, 2020
Online: 28 December, 2020

Keywords:
Abstractive text summarization
BERT
Japanese text summarization

ABSTRACT

In this study, we developed and evaluated an automatic abstractive summarization algorithm in Japanese using a neural network. We used a sequence-to-sequence encoder-decoder model for practical purposes. The encoder obtained a feature-based input vector of sentences using the bidirectional encoder representations from transformers (BERT) technique. A transformer-based decoder returned the summary sentence from the output as generated by the encoder. This experiment was conducted using the Livedoor news corpus with the above model. However, two problems were revealed. One is the repetition of a specific phrase while the model is generating text. The other is that the model can not handle out-of-vocabulary words. As solutions, we use repeat block in n-gram words and WordPiece. In addition, to evaluate the performance of the model, we compared the summarization accuracy between our model and a long short term memory based pointer-generator network. As revealed by the results, our model comprehends the meanings of sentences better than a pointer-generator network but makes more word-based mistakes.

1 Introduction

This paper is an extension of work originally presented in 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) [1].

Societies with a highly developed information system make it easy for citizens to obtain the information they need. However, for this reason, the information obtained often has redundancy. Therefore, skills are needed to extract only the necessary information. In other words, a summarization skill is important. Thus, we need a system that automatically takes the necessary points of a text and applies deep learning with a neural network. At the present time, models using bidirectional encoder representations from transformers (BERT) [2] have achieved the highest scores in abstractive summarization tasks [3]. Although there are already many models using BERT in English and some other languages, a model in Japanese has yet to be found. We therefore developed models using BERT in Japanese and evaluated them through a comparison with another model. The code is available at `https://github.com/IwasakiYuuki/Bert-abstractive-text-summarization`.

Text summarization is one of the natural language processing that effectively summarizing long sentences. The Algorithms of text summarization used in machine learning are mainly divided into two types: extractive and abstractive summaries. In the extractive text summarization, a summary sentence is generated by combining important sentences in the source text. On the other hand, the latter, abstract type, understands the input sentences and generates the corresponding summary sentences by itself. Although both types are similar in that they summarize the main points of input sentences, they are more flexible in this respect because abstract abstract summarization generates the corresponding summary sentence by itself, while extractive summarization can only process sentences collected from input sentences. In this study, we focused on abstract summarization.

In recent years, various models have been proposed for abstractive summarization. An author proposed an abstractive summarization model using Bidirectional Encoder Representations from Transformers (BERT) [3]. Experimentation results as reported in [3] revealed that the developed model achieved a new state-of-the-art performance on both CNN/Daily Mail and New York Times datasets. Viswani et al. proposed an abstractive summarization model called a pointer-generator network based CopyNet [4]. This pointer-generator network model has advantages in terms of both abstractive and extractive summarizations.

The model developed in this study was built with reference to a text summarization model using BERT and has two stages. In the first stage, the input text was encoded into a contextual representation using BERT, and after processing the input text with BERT, a

[*]Corresponding Author: Yuuki Iwasaki, Tokyo College, s20607@tokyo.kosen-ac.jp

summary draft was generated using a transformer-based decoder. In the second stage, the draft summary text was re-validated using BERT to generate a crisper summary text, which was then processed with BERT. In this experiment, only the first stage was employed.

In section 2, we introduce the several well-know neural network models and layers used in this study, i.e., BERT and multi-head attention (the layers in BERT), and pointer-generator network used for comparison with our model. In section 3, we describe an overview of the proposed model along with its structure. In section 4, we detail the settings and dataset applied while training our model. In section 5, the results of the experiments described in section 4, are evaluated, along with those of another model, i.e., a pointer-generator network. Based on the evaluations, we compare our model with a pointer-generator network based on scores from the Extracted places and ROUGE-N [5]. In section 6, we provide some concluding remarks regarding our experiments and discuss areas of future study.

2 Related Studies

2.1 Pointer-Generator Networks

A pointer-generator network [6] is an LSTM-based model for an abstractive summarization. Long short term memory(LSTM) [7] is a neural network model that is often used for natural language processing. Figure 1 shows a model overview of a pointer-generator network. The most significant differences between this model and our approach are the structures, i.e., the “copy mechanism” and “coverage mechanism”. A pointer-generator network can-not compute in parallel owing to the structure of the model. In addition, it requires a long time and a large amount of data for training to obtain the meanings of both words and sentences at the same time.

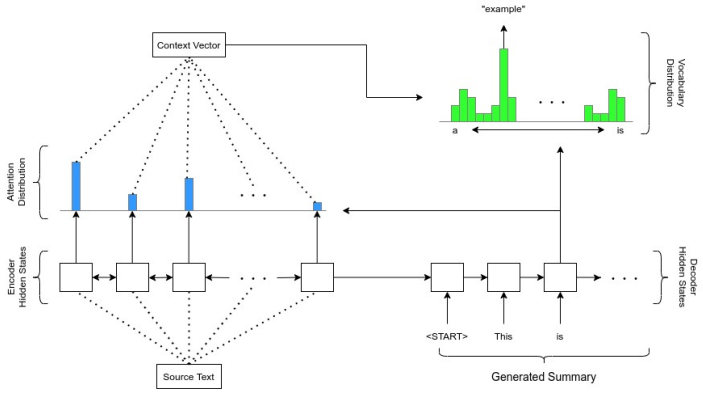


Figure 1: Overview of pointer-generator network. As shown on the left side of the figure, an ’Attention Distribution’ helps handle out-of-vocabulary words. In addition, as shown on the right side, the ’Vocabulary Distribution’ resolves the repeating problem. Referred to [6]

Copy mechanism A neural network outputs the probability of a predetermined group of words when generating them. Therefore, it cannot output proper nouns. However in sentence summaries, proper nouns are often important. Thus, a pointer-generator network dynamically allocates the words in the input text to the IDs using a copy mechanism and thus the model can handle out-of-vocabulary words.

Coverage mechanism Sentence summarization tasks occasionally have a problem of generating the same words repeatedly. A pointer-generator network incorporates a mechanism called a coverage mechanism to solve this problem. The coverage mechanism keeps the distribution of the words generated up to the t-step, and then penalizes them for producing the same word. The penalty is achieved by adding a new term to the loss function.

2.2 BERT

In recent years, pre-training models, such as BERT, have been widely incorporated into neural network models. In particular, models trained with BERT have achieved state-of-the-art performance in natural language processing tasks; BERT has been pre-trained with a large unlabeled corpus and can be fine-tuned with another corpus to achieve better performance. Figure 2 shows the process of pre-training and fine-tuning.

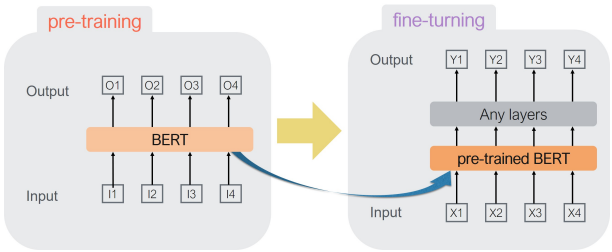


Figure 2: BERT pre-training and fine-tuning processes. A pre-training of the model is shown on the left side of this figure, and a fine-tuning is shown on the right side.

We briefly describe the structure of the BERT model with reference to [2]. BERT has several layers; each layer has a Multi-Head Attention and a linear affine with a residual connection. In our experiment, we utilize the BERT-based model which has 12-layers and 768-hidden sizes.

Pre-training In a general neural network model, the model is trained at once for the tasks. Then, the model is trained for each task at the same time as the meanings of the words and sentences are obtained. For this reason, a large amount of data is needed to train the model. However, there are two processes applied during BERT training, pre-training and fine-tuning. Pre-training tasks of BERT are called masked language model (MLM) and next sentence prediction (NSP). An MLM is a task for predicting masked words in an input text, and an NSP is a task for predicting the next sentence of an input sentence. Through a pre-training, the model obtains the meanings of the words and sentences, as well as the coherence of the texts.

Fine-turning After pre-training of the model, we re-train the model for a specified task, in this case, a summarization. A pre-trained model already has the obtained meanings of words and sentences, and thus we simply tune the model toward our task. This means it can be trained with a low data quantity. At the fine-tuning

stage, we use the pre-trained BERT model and some layers, as shown in Figure 2.

2.3 Multi-Head Attention

BERT and a transformer primarily comprise multi-head attention layers. The multi-head attention divides the attention input into multiple parts and concatenates them with multiple outputs. Multi-head attention is more accurate than attention and can be calculated through (1), (2), and (3) as described in reference [4].

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2}$$

$$MultiHead\,(Q,K,V) = Concat\,(head_i,\ldots,head_n)\cdot W^O \tag{3}$$

Equation (1) represents a formula for a single input query in multi-head attention, in which the input is split into units of a head as shown in (2), and each output is concatenated as shown in (3) to generate the overall output. In practice, we compute the attention function on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V, and W_i^Q, and W_i^K, and W_i^V are the parameter matrices. Here, d_k used in (1) is the number of dimensions of key K.

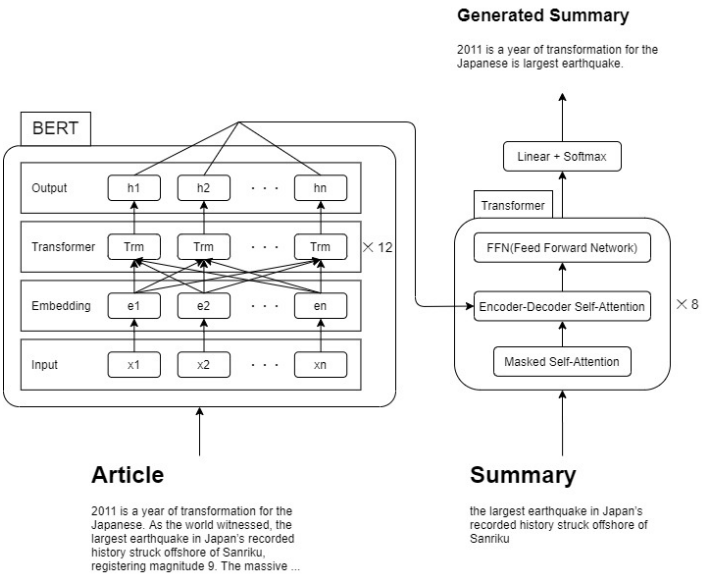


Figure 3: Overview of our text summarization model. Our model includes BERT in the encoder and a transformer in the decoder. During the training phase, we set the article body and summary into the encoder and decoder inputs, respectively.

3 Model

Figure 3 shows an overview of the model used in our experiment.

3.1 Encoder

Several pre-learning models, such as BERT, have been widely utilized in encoder-decoder models. Because BERT is efficient for fast training with high precision, it ensures a higher accuracy in existing models. In a study conducted by Zhang et al. [3], BERT is used as the encoder to achieve state-of-the-art performance for an abstract text summarization task. In this experiment we applied a pre-learning model BERT as the encoder.

3.2 Decoder

In our model, we configured a transformer-based decoder as an encoder. It was not the proposed decoder choice for BERT to send the output generated by the encoder to the input of the decoder at the same time.

Furthermore, we opted for a transformer-based decoder as opposed to a recurrent neural network (RNN) such as an LSTM [7] and GRU [8] for the following reasons: (1) Time for training - Transformer-based decoders are built using multi-head attention, which can perform parallel computations, and are, hence, faster. (2) Accuracy - Transformers are far more accurate than RNNs on machine translation tasks. (3) Long-range dependency - The attention used in the transformers makes it easier to learn long-range dependencies compared to RNNs such as an LSTM.

3.3 Abstractive summarization model

The input is denoted as $X = \{x_1, x_2, \ldots, x_n\}$, sequence representing sentence breaks is denoted as $S = \{s_1, s_2, \ldots, s_n\}$, and the corresponding summary is denoted as $A = \{a_1, a_2, \ldots, a_n\}$. We started by entering X and S into BERT.

If $f_{sen}(x)$ is assumed to be the number of sentence of x, sequence S is computed as $S = f_{sen}(X)\ mod\ 2$. The resulting BERT encoder output is denoted as H. Next, we input H and the output of the decoder at the t-th time step.

The probability of the vocabulary at the t-th time step can be obtained as shown in (4). This probability was conditioned on the decoder output until the t-th time step and the output of the encoder H.

$$P_t(w) = f_{decoder}(w \mid H, Y_{<t}) \tag{4}$$

The loss of training L, is calculated as shown in (4) using the probability of vocabulary $P_t(w)$.

$$L = -\sum_{i=0}^{n} \log P(y_t \mid H, a_{t-1}) \tag{5}$$

4 Experiments

4.1 Setting

During this experiment, we used a pre-trained model with BERT that was developed at the Kurohara and Kawahara laboratories of Kyoto College [9]. Most of the BERT hyperparameters were same as those of BERT-Base (i.e., 12-layers, 768-hidden, and 12heads) in [2]. The model was trained for 30 epochs with 1.8 billion Japanese Wikipedia corpuses. The input text was divided into sub-words with byte pair encoding (BPE [10]) using the morphological analysis system Juman++ [11].

The vocabulary size stood at 32,000 words. The decoder in our model comprises eight multi-head attention layers; the division of hidden size as 3,072; and the embedding vector as 768, similar to the encoder.

We used Adam as the model optimizer, and set the parameters as $\beta_1 = 0.9, \beta_2 = 0.999, and \epsilon = 1.0 \times 10^{-9}$. The maximum learning rate was set at 1.0×10^{-9}. In addition, the dynamic learning rate was adopted as the model learning rate in [4].

The learning rate is computed as shown in (6),

$$lr = max\_learning\_rate * \frac{min\left(cs^{-0.5}, ws^{-1.5} * cs\right)}{ws^{-0.5}} \quad (6)$$

In principle, the learning rate increases linearly up to the warmup step (ws). If the current step (cs) exceeds the warmup step, the learning rate gradually decreases. The learning rate peaks when the current step and warmup step are equal. At this point, the learning rate is denoted as $lr = max\_learning\_rate$. For our experiment, we set $warmup\_step\,(ws) = 4000$ and $max\_learning\_rate = 0.0001$ for training. The model includes BERT with 12 multi-head attention layers. The batch size is set to 4 for GPU memory because the maximum input sequence was set to 512.



Figure 4: An example of Livedoor news corpus. Every article has three sections, a title, a summary and an article body. Each section is labeled in the above figure.

4.2 Dataset

The Livedoor News corpus contains 130,000 Japanese news articles from Livedoor News, each news article is accompanied by a three-line summary. The article text and summary are set as input and output of the experiment, respectively. From the dataset, 100,000 data points were used for training and 30,000 data points were used for validation. The maximum length of the input sequence was set to 512 tokens, but some data values exceeded this limit. In such cases, only the first 512 tokens were entered into the model.

Every article on the Livedoor news website has text and a three-line summary. Figure 4 shows an example of an article body and summary. In this example, the article body at the bottom of the figure is input into the model, and the summary at the middle of the figure is output to the model.

4.3 Training

If 100,000 live news corpus data points in 1 epoch were trained in 15 epochs, the loss of each lexicon in the validation data was about 6.5 words, and the accuracy of identifying the correct lexicon was about 67%. The graphics card used for training was a Titan X (Pascal) with 12 GB of memory. The training took about three days.

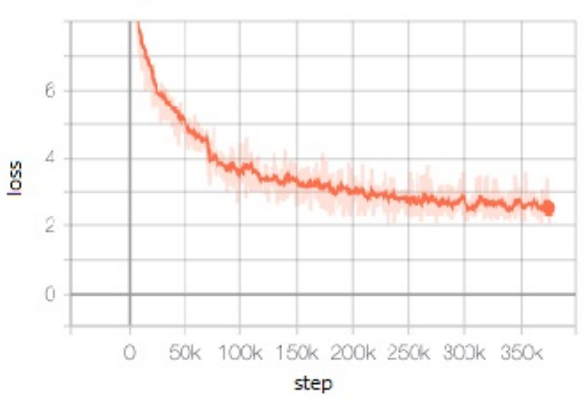


Figure 5: Outputs of loss function for every training step. A gradual decrease in the numbers of loss function outputs is implied in this figure.

5 Results and Discussions

In this section, we can see the evaluations of the generated summaries with two models, i.e., our model (BERT+Transformer) and a pointer-generator network. From the generated summaries, we evaluated the models on the following criteria:

- Grammatical Accuracy
- Vocabulary mistake
- importance level

To check the validity of the generated sentences, we focused on grammatical and vocabulary errors. We also extracted and checked which parts of the original text were important for the generation of the summary text. The above three points are evaluated subjectively. In addition, we calculated and compared their respective ROUGE-N scores.

5.1 Generated summaries

The texts in the appendix are the article bodies and their summaries that generated by the two trained models, our model and the pointer-generator network (translated from Japanese into English). The texts under "Input" are article bodies in the Livedoor news corpus, and the texts under "Output" are summaries generated by our model and the

pointer-generator network from each article body. In the appendix, we provide three article bodies and generated their summaries as examples. When generating the text, a beam search with a width of 4 was used. Prior to the training process, WordPiece [12] was used to further divide out-of-vocabulary words into multiple words. For example, the word "smoke" which is not in the vocabulary of the model, is able to represent the combination of "smo" and "ke" in the vocabulary. For the quantitative effect of WordPiece, refer to an ablation study [13].

One problem that arises during the summary sentence generation is the "repetition problem". This means that an arbitrary phrase will repeatedly appear in the summary sentence. We used a REPEAT block to get around this. The repeat block detects the repetition of a sentence in n-grams increments and returns to the (n-1)-steps before the repetition and reapplies the word generation. In this experiment, we used the repeat block of the tri-gram that had the highest ROUGE-N score.

For example, you can see the *t*-th step generated summaries by the model as follows:

summary : It is going to rain in **the(1) the(2) the(3)**

In the above example, the same three words are generated from the end of the word, and thus the approach returns to the two steps prior to "the(1)" and regenerates words other than the those previously generated.

summary : It is going to rain in **the(1) evening**

Input 1 in the appendix demonstrates that both models were able to learn correctly as, it can be seen, to some extent, that the summary text retained the key points of the input text. In addition, there are no errors in grammar or vocabulary. However, if we look at the summary of a pointer-generator network, the sentence "The Berlin International Film Festival Silver Bear Award is one of the three biggest film festivals in the world for "smoke". " is not extremely important. Therefore, in the above example, we thought the summary generated by our model was better than that of the pointer-generator network owing to the importance.

Input 2 in the appendix is regarding the risks to the human body from sitting for long periods of time. At first glance, both summaries appear to be correct, but the word "skillfully" is included in the summary of our model. This is not the word that should be used. Therefore, we felt that the pointer-generator network was able to produce a more accurate summary in this example than our approach.

Input 3 in the appendix deals with the response of a real estate company when a house is an accidental property. No erroneous words appear in either summary. However, the sentence "It is said that the building should be disliked because it is a place to live continuously." in the summary of our model, is not grammatically correct. In terms of content, however, we believe that our model is able to extract the more important parts. For example, the content regarding payment is fairly important but is not included in the summary of the pointer-generator network.

In the present example, our model extracts the more important parts, but the grammar was correct for the pointer-generator network. From the three examples provided in the appendix, it appears that the pointer-generator network has fewer vocabulary and grammatical errors, although our model is able to extract more of the important content. In addition, you can see that the locations that the pointer-generator network extract is mostly in the first half of the input text. By contrast, our model extracts from all throughout the input text. In other words, our model is more expressive, whereas the pointer-generator network is less prone to grammatical and vocabulary errors. In addition, because the pointer-generator network has a copy mechanism, we thought it might be related to the fact that the model is somewhere between abstract and extractive text.

5.2 *Extracted places*

We will now see which parts of the input text are extracted by the summary statement generated in the previous section. Figure 6, 7 and 8 shows the places that the model extracts from the input texts. Each horizontal axis represents the number of words, and the bar graph shows the word positions referenced by the models.

Figures 6, 7 and 8, indicate that the locations extracted by the pointer-generator network are mostly in the first half of the input texts. By contrast, those extracted by our model(BERT+Trasnformer) are scattered throughout the input text. This leads us to believe that our model may be able to understand the entire content of the input text better than the pointer-generator network.



Figure 6: Input 1 highlighting the extracted locations

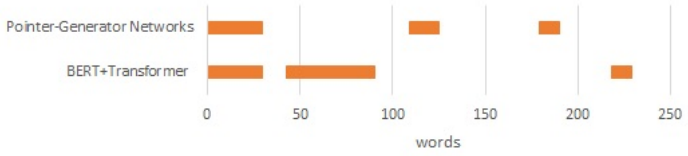


Figure 7: Input 2 highlighting the extracted locations

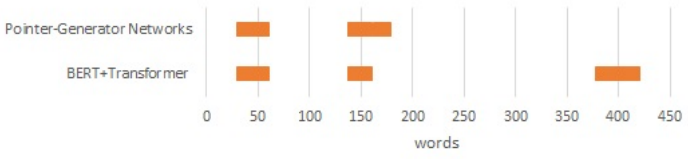


Figure 8: Input 3 highlighting the extracted locations

5.3 ROUGE-N socre

Next, we compare the pointer-generator network with our model using the ROUGE-N [5], which is an index frequently used in the evaluation of summary sentences, which represents the percentage of agreement between words in N-gram units.

As an example, let us try to calculate the ROUGE-N score of the following sentence.

summary : It is going to rain in the evening today
references : It is going to rain today

Referring to [5], the equation for obtaining the ROUGE-N score is shown below.

$$ROUGE_N = \frac{\sum_{C \in Reference} \sum_{Gram_n \in C} Count_{match}(Gram_n)}{\sum_{C \in Reference} \sum_{Gram_n \in C} Count(Gram_n)} \quad (7)$$

Considering the above example, we have the same six pairs on uni-gram, and four pairs on a bi-gram. From this, the ROUGE-1,2 score of the example could be derived as follows:

$$\begin{aligned} ROUGE_1 &= \frac{6}{9} \\ &= 0.667 \\ ROUGE_2 &= \frac{4}{8} \\ &\fallingdotseq 0.500 \end{aligned}$$

In our experiment, we calculated the ROUGE-1,2 scores with 1,000 Japanese news articles (sample size $n = 1,000$) chosen at random from the verification datasets. The standard error, *SE*, calculated using the unbiased standard deviation, u, and the sample size, n, is expressed as follows:

$$SE = \frac{u}{\sqrt{n}} \quad (8)$$

Using the above equation, the margin of error, e, for the 95% confidence interval (95% CI) is expressed as in Equation 9.

$$e = \pm 1.96 \times SE \quad (9)$$

The ROUGE-1,2 scores, the standard errors, *SE*, and the unbiased standard deviations, u, for both the models are presented in Table 1.

Table 1: ROUGE-1, 2 scores for pointer-generator networks and our model (BERT + transformer)

Model		Score	u	e
Pointer-Generator Networks	ROUGE-1	0.463	0.123	±0.008
	ROUGE-2	0.221	0.142	±0.009
BERT + Transformer	ROUGE-1	0.470	0.107	±0.007
	ROUGE-2	0.215	0.126	±0.008

Table 1 indicates that the ROUGE-1,2 scores of our model and the pointer-generator network were similar. However, our model's structure was designed for multiple tasks, whereas the pointer-generator network was a model designed for text summarization tasks because of the copy mechanism. In addition, our model generated all the words in the output sentences. Thus, we hypothesized that our model achieved a better comprehension of the text.

5.4 Word Mover's Distance

Next, we compared the pointer-generator network with our model using word mover's distance (WMD) [14] to evaluate their semantic similarity. The WMD is a novel distance function between text documents. The ROUGE scores are based on literal word overlap; hence, it is necessary to evaluate a score that is not based on literal word overlap. Among a number of recently proposed semantic similarity metrics, WMD is shown to be the most reasonable solution to measure semantic similarity in reformulated texts [15]. We evaluated the WMD for both models with 1,000 Japanese news articles chosen at random from the verification datasets and used the Japanese word embedding vectors [16] to calculate the WMD. Table 2 presents the WMD scores, unbiased standard deviation, and margin of error for the pointer-generator networks and our model.

Table 2: WMD of pointer-generator networks and our model (BERT + transformer)

Model	WMD	u	e
Pointer-Generator Networks	0.533	0.122	±0.008
BERT + Transformer	**0.521**	0.106	±0.007

As shown in Table 2, our model has a slightly lower WMD score than the pointer-generator network. This implies that the summaries generated by our model were more similar to the article's summaries than those generated by the pointer-generator network. Thus, the hypothecate that our model achieved a better comprehension of the text than the pointer-generator network was strengthened.

6 Conclusion and Future works

We conducted an experiment to demonstrate an abstractive summarization of Japanese text with a neural network model using BERT. In addition, we conducted a comparison between the qualitative and quantitative aspects of our model and a model frequently used for text summarization, i.e., a pointer-generator network. Our model is composed of a BERT encoder and a transformer-based decoder. The dataset used in this paper was the Livedoor news corpus consisting of 130,000 datapoints, of which 100,000 were used for training.

The results of the experiment revealed that the model was able to learn correctly as the summary sentence captured the key points of the text to a certain extent. However, the contents of the summary sentence were repeated, and the model could not handle unknown words. As a solution, we applied two mechanisms, a repeat block and WordPiece. The repeat block detects repeated n-grams in the generated summaries and regenerates summaries at the (t-n-1)-th step excluding the repeat word. To handle this, WordPiece further divides the out-of-vocabulary words into sub-words.

The results of evaluations of our model and the pointer-generator network revealed that their ROUGE-1,2 scores were similar. However, our model's structure was designed for multiple tasks, whereas the pointer-generator network was a model designed for text summarization tasks because of the copy mechanism. Therefore, we hypothesized that our model achieved a better comprehension of the text. In addition, the comparison of the two models to evaluate their semantic similarity, that is, the WMD, strengthened our hypothecate. While extracting phrases, the pointer-generator network extracted phrases from the first half of the input text, whereas our model extracted phrases from throughout the text. However, a qualitative evaluation of our model revealed that it made numerous grammatical and vocabulary mistakes. We believe that this problem can be solved by improving the model.

In the future, we will explore these recommendations through additional experiments and compare the results.

7 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19K12906.

References

[1] Y. Iwasaki, A. Yamashita, Y. Konno, K. Matsubayashi, "Japanese abstractive text summarization using BERT," in 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 1–5, 2019, doi:10.1109/TAAI48200.2019.8959920.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018, doi:10.18653/v1/N19-1423.

[3] H. Zhang, J. Cai, J. Xu, J. Wang, "Pretraining-Based Natural Language Generation for Text Summarization," 789–797, 2019, doi:10.18653/v1/K19-1074.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, I. Polosukhin, "Attention is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 6000–6010, Curran Associates Inc., Red Hook, NY, USA, 2017, doi:10.5555/3295222.3295349.

[5] C.-Y. Lin, E. Hovy, "Automatic Evaluation of Summaries Using N-Gram Co-Occurrence Statistics," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, 71–78, Association for Computational Linguistics, USA, 2003, doi:10.3115/1073445.1073465.

[6] A. See, P. J. Liu, C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1073–1083, Association for Computational Linguistics, Vancouver, Canada, 2017, doi:10.18653/v1/P17-1099.

[7] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," Neural Comput., **9**(8), 1735–1780, 1997, doi:10.1162/neco.1997.9.8.1735.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014, doi:10.3115/v1/D14-1179.

[9] S. Kurohashi, Y. Murawaki, "BERT pretrained model with Japanese corpus," http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB (accessed 2019-07-16).

[10] R. Sennrich, B. Haddow, A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1715–1725, Association for Computational Linguistics, Berlin, Germany, 2016, doi:10.18653/v1/P16-1162.

[11] T. Kudo, K. Yamamoto, Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphologiaical Analysis," IPSJ-NL, **161**, 89–96, 2004.

[12] M. Schuster, K. Nakajima, "Japanese and korean voice search," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5149–5152, IEEE, 2012, doi:10.1109/ICASSP.2012.6289079.

[13] K. Bostrom, G. Durrett, "Byte Pair Encoding is Suboptimal for Language Model Pretraining," ArXiv, 2020, doi:abs/2004.03720.

[14] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, "From word embeddings to document distances," in International conference on machine learning, 957–966, 2015, doi:10.5555/3045118.3045221.

[15] I. Yamshchikov, V. Shibaev, N. Khlebnikov, A. Tikhonov, "Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric," arXiv, 2020, doi:abs/2004.05001.

[16] M. Suzuki, "Japanese Wikipedia entity vector," http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/ (accessed 2020-10-30).

Appendix

The following are three articles from Livedoor news and a summary of our model and the pointer-generator network generated as examples. The examples below were used for the discussions provided in section 5.

Input 1 Director Wayne Wang from Hong Kong, who has won the Silver Bear Award at the Berlin International Film Festival, one of the three biggest film festivals in the world in the movie "Smoke" (1995), and Japan's world-famous Beat Takeshi and Nishijima Hidetoshi. The movie "When a Woman Sleeps" (published on February 27). It was the first time in 12 years since "Blood and Bone" starring in a film other than his own work in 2004, and said "I was so worried about my acting that I held my head," he said unexpectedly. Is it because of the mysterious style that the boundaries between dreams and delusions and reality are vague? Three people talked to each other. This film is a movie based on a short story by Javier Marias, a Spanish writer, and is set in a quiet resort hotel, and a mysterious couple (Takeshi and Shiori Nana Shiori), two years older than their parents and children. It is a mystery of love and hate that stares at a writer (Nishijima) who is obsessed with a relationship that has become obsessed with and has run out of curiosity. "I thought that the main character was Nishijima," said the smiling Takeshi, and read the script, "I thought I was foolish. A normal script would tell the story and the ending." He said that he had a hard time understanding a complicated story. Nishijima, who praises himself for being more acquainted with movie expressions, relieved, "If I say "good", I'm sure there will be no mistake." I watched all the acting all the time. When I finished watching, I didn't even remember what kind of movie it was," said an unknown side of Takeshi called "Kitano of the World." Nishijima, who immediately decided to appear, "Wayne Wang is a work that shoots Beat Takeshi, so it will come out (laughs)," said "This time especially Beat Takeshi is obsessed with love, even at the shooting site. I was shocked a lot and was really moved. It was wonderful." Director Wang, who worked on Japanese movies for the first time, also said, "I wanted to work with Takeshi rather than want to shoot in Japan." Takeshi played an unusual role in keeping a video recording of a young beautiful woman (Kana) sleeping. When asked, "What do

you want to leave your dear ones?", "I've been in the entertainment world for a long time, so I thought, "I thought that the gag of Takeshi at that time, that laugh was amazing." It's okay if you can leave one," he said. (Interview and text: Megumi Shibata)

Output 1 (Pointer-generator network) The Berlin International Film Festival Silver Bear Award is one of the three biggest film festivals in the world for "smoke." For the first time in 12 years, Takeshi said, "I was so worried about my performance that I held my head." "The boundaries between dreams and delusions and reality are vague and mysterious.

Output 1 (BERT+Transformer) A movie "When a Woman Sleeps", in which Takeshi Beat and Hidetoshi Nishijima formed a tag. Beat Takeshi was said to have been attached to love. "I'm worried about my performance, and I'm worried about the finished product, too," he said.

Input 2 "Mentai Wide-Fukuoka Broadcasting" At the corner of "Special News THE Slide Show" broadcasted on February 9, 2016, we covered the risks to your body from sitting for a long time. The image is an image (taken by Andrea Arbogast, from Flickr). A study of 220,000 men and women over the age of 45 in Australia found that the risk of death was 1.4 times higher for people sitting 11 hours or more a day than for those sitting less than 4 hours a day I got the result. It seems that the risk of obesity, diabetes, cancer, etc. increased, causing death. The detailed cause is unknown, but it is also said that "myokine" that suppresses heart disease is not secreted unless the muscles are contracted. This excessive sitting puts a heavy burden on the lower back, with 1.4 times being normally seated and 1.85 times being leaned forward. Even in Fukuoka, more and more offices are incorporating "sometimes standing work". More than 90% of the desks sold in Sweden and Denmark are of the type that can be moved up and down, and by using this you can do "occasionally standing work". Even in Japan, Rakuten has changed the desks of all employees so that they can be moved up and down, to avoid sitting down. Even in Fukuoka, some companies are making their desks movable up and down and changing the height of the desks themselves. When I asked when to change the height of the desk, this was the most common opinion. Because I may feel drowsy after lunch, switch to standing work at such times. When you are tired from working on a computer or lose your concentration, start a standing job. Stand up when your shoulders, waist, or neck hurts. It seemed that it was good to stand up, and there were interesting opinions. I have less stiff shoulders, lower back pain, and neck pain. Increased communication from the surroundings. When I sat down, I was staring at my computer, but by standing up, I was able to see my surroundings better and start talking. Since I also work as a deskwork worker, I sometimes feel pain in my lower back and shoulders after sitting for a long time. For this reason, I stand up and relax once every hour, but I found it useful to have a desk like this. If you keep sitting down, you may not be able to see the surroundings because you are staring at a desk or a personal computer, so I would like you to incorporate such desks in various offices. (Writer: Piyoko).

Output 2 (Pointer-generator network) At the "Special News THE Slide Show" on the 9th, I heard about the risks to the body from sitting for a long time. It is said that "myokine" that suppresses heart disease is not secreted unless the muscles are contracted. It is said that more and more offices in Fukuoka have adopted "sometimes standing work".

Output 2 (BERT+Transformer) In the program on the 9th, we discussed the risks to the body from sitting for a long time. The risk of obesity, diabetes, cancer, etc. is high, and the risk of death is 14 times higher. In addition, he changed the man's desk to something that can be skillfully moved up and down to avoid sitting.

Input 3 Last year, an incident that involved the illegal construction of a large-scale condominium construction project in Yokohama City reminded me of the difficulty of choosing a property, but I finally got the feeling that I could get a house. If it's a so-called "accident property", can you ask the real estate agent for a refund? Lawyer Masami Takeshita answers. (Consultation). At the end of last year, I bought a second-hand house, but when I looked up the reason because the neighborhood was strange, the previous owner was at home It seems that he tried to hang himself. The real estate company that introduced this property did not tell me that it was an accident property. In this case, is it possible to request the real estate agent to refund the purchase price in full? (Answer). The fact that a building purchased for residential use commits suicide and lacks "comfortable living" means that "a building is a place to live continuously, so there are also psychological flaws due to the historical background of the building that should be disliked. (Defect)" is understood. If the buyer makes a contract without knowing that there is a defect, the seller will be liable for the defect guarantee, and if the defect does not meet the contract purpose, the contract will be canceled, and if it does not reach that level, You are obliged to compensate the buyer for damages. In addition, there are court cases that the contract can be canceled if there is suicide and it is judged that it is reasonable to feel that it is not suitable for living because it lacks comfort to live in. .. It is the case that the cancellation of the sale of condominiums, which was found to have committed suicide six years ago when the deposit was paid, was approved. On the other hand, there are some cases in which the suicide was canceled 7 years ago, and there are also cases in the sale and purchase of a plan to demolish an old house in which the degree of dislike of suicide two years ago due to the removal of a building was not a defect. In your case, depending on the time of suicide and the degree of topicality, you may be able to cancel the contract and request a full refund of the price, and even if you do not go that far, the objective evaluation of the suicide building and the purchase price You can consider the difference claim for damages. The important things that the Homebuilding Act requires an intermediary to explain are not the history of the building, such as suicide, but if the trader knows suicide, which is an important factor in determining the buyer's contract decision, There is a statutory obligation to explain and if you remain silent, you will be liable. However, it is questionable to impose accountability forever, and there seems to be a guidance that sales are 5 years and rent is 3 years. In any case, when you buy a used building, you should ask about its provenance, etc., and you can pursue the default liability

of breach of accountability to the lying seller. (Attorney Profile). Masami Takeshita: Born in Osaka in 1946. Graduated from the Faculty of Law, the University of Tokyo. Registered as a lawyer in 1971. * Weekly post February 26, 2016 issue.

Output 3 (Pointer-generator network) Attorney Masami Takeshita explained why a real estate company can request a refund if the property is an ”accident.” It seems that the building purchased for residential use may lack ”comfortable living” due to suicide. ”There are also psychological flaws due to the disgusting historical background of buildings.

Output 3 (BERT+Transformer) A lawyer explains whether you can request a refund when you purchase the accident property in 2017. It is said that the building should be disliked because it is a place to live continuously. Regarding loans, it seems that there is a possibility that you can cancel the contract and request full refund.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1710-1718 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Efficient and Scalable Ant Colony Optimization based WSN Routing Protocol for IoT

Afsah Sharmin*, Farhat Anwar, S M A Motakabber

Faculty of Engineering, Department of Electrical and Computer Engineering, International Islamic University Malaysia, Kuala Lumpur, 53100, Malaysia

ARTICLE INFO

Article history:
Received: 31 August, 2020
Accepted: 21 December, 2020
Online: 28 December, 2020

Keywords:
Wireless Sensor Networks
IoT
Routing algorithms
ACO
Energy consumption

ABSTRACT

IoT integrates and connects intelligent devices or objects with varied architectures and resources. The number of IoT devices is growing exponentially. Due to the massive wave of IoT objects, their diversity and heterogeneity among their architectures, the existing communication protocols for wireless networks become ineffective in the context of IoT. Wireless Sensor Network (WSN) has the potential to be integrated to the internet of things (IoT). The issues of the routing of WSNs impose nearly similar prerequisites for IoT routing technique. Most of the traditional routing protocols are not appropriate for WSNs and IoT because of resource constraints, computational overhead and environmental interference and do not take into account the different factors affecting energy parameter and do not accommodate node mobility. Routing algorithms must ensure the data transmission in an efficient way, having proper knowledge of the IoT system. For this reason, many intelligent systems have been utilized to design routing algorithms to handle the network's dynamic state. In this paper, an ant colony optimization (ACO) based WSN routing algorithm for IoT has been proposed and analyzed to enhance scalability, to accommodate node mobility and to minimize initialization delay for time critical applications in the context of IoT to find the optimal path of data transmission, improvising efficient IoT communications. The proposed routing algorithm is simulated using MATLAB for performance evaluations. The evaluation results have recorded an improvement in conservation of energy, of almost 50% less consumed energy even with an increase in the number of nodes, by comparing with an existing routing technique based on ant system, a current routing protocol for IoT and the conventional ACO algorithm.

1. Introduction

With the widespread use of IoT devices, the issue of designing efficient routing protocol has attracted more attention in networking research. This paper is an extension of work originally presented in ICOM'19 [1].

According to the prediction of Cisco™ IBSG [2], by 2020, 50 billion devices will be connected to the Internet. IoT provides network connectivity between these smart devices everywhere and all times. The emergence of a new ubiquitous computing era has been created due to the evolution of wireless networks and sensor technologies, allied to the increasing demand for new IoT applications for the provision of smart services [3]. In this context, WSNs play an essential role to the expansion of IoT while providing ubiquity of networks with smart and low-cost devices that are easy to deploy. In an IoT system, a large collection of autonomous and dynamic sensor nodes are used to gather information by detecting physical parameters, communicate and cooperate with their environment and send their data to the internet. Sensor nodes have the ability of self-organization and sensor networks function in a distributed way [4]. IoT with the integration of WSNs, have a wide range of applications spaces that shape human life and also have impact on economic benefits. The physical domain of IoT is presented through the connected networks of objects and nodes utilizing wireless sensors. A large number of small devices surround the environment help to manage the physical world by sensing, processing, communicating and analyzing the data in the IoT network system [5]. The development of new applications and technologies are drawing attention from the research perspective, such as smart home, environmental monitoring, healthcare, transport, agriculture, offices, buildings and smart cities etc.

*Corresponding Author: Afsah Sharmin, Faculty of Engineering, Department of Electrical and Computer Engineering, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia, Tel: +60167101538, afsahsharmin@gmail.com

Designing efficient routing protocol is the key factor to improve the energy efficiency, data transmissions, scalability and prolong network lifetime in WSNs and IoT. However, several considerations are required for resource-constrained network system, such as energy efficiency, scalability, security, autonomy, computational complexity, environmental constraints for wireless link, node mobility, the QoS (quality of service) requirement for a particular application, during IoT routing. The sensed information ought to be sent to the base station for further operations in various IoT applications through the competent forwarding mechanism while keeping in consideration different functionalities of IoT objects states.

Until now, there is a large number of routing algorithms have been proposed by the research community, considering the energy parameter, yet its exploitation is not well thought-out [6]. The existing node energy and the distinctive components affecting this energy parameter need to be considered to spare node energy and enhance the communication quality of the network. The node's lifetime relies upon the battery-life to a huge range, and the irrational energy utilization will effectuate the system to expire early and decrease the lifetime of the network [7]. Hence the key research issue is the designing of efficient route calculation algorithm that can ensure efficient information communication within IoT while maintaining scalability. Most of the traditional routing protocols are not adaptable other than energy efficient, if the varied difficulties in various applications are considered or due to the dense and complex conditions and a wide range of radio obstruction. Many intelligent systems including the working mechanism of the biological systems have been employed for designing routing algorithms to handle the network's dynamic state while keeping pace with energy efficiency during information communication [8]. Ant colony optimization (ACO) based algorithms emphasize the design of routing protocols that are robust, adaptable and scalable [9]. The coordination of ants depends on the ability of self-organization that ant colony optimization based swarm intelligence techniques possess [10]. The probabilistic approach of ACO is used to determine the routes, and the pheromone update formulation is used for further updating of the pheromone trail [11].

An ant colony based routing technique, named EICAntS, for effective communications within IoT is presented in [12]. This algorithm considers the energetic parameters. The pheromone estimations in the ant system is related here by the calculated global efficacy factor. An improvement with regard to network lifetime and conservation of energy is shown from the evaluation results. The algorithm does not include the heuristic information and pheromone update strategy consisting of pheromone evaporation rate and/or the amount of pheromone deposited. The energy effect precisely addresses the data class handled by the node and does not specify the different factors affecting energy parameter such as the energy consumption in free-space and multi-path fading standard of wireless communication as well as no particulars are provided on how to compute the nodes' energy level. The REL routing technique for IoT applications, which is focused on energy and link quality information, is proposed in [13]. Testbed experiments are used besides simulation for the evaluation where REL increases the system and network lifetime, reliability, energy-efficiency, QoS of IoT applications and reduces packet loss rate. In addition, it offers a path determination system which is basically an end to end route. For this purpose, it depends on cross-layer data with insignificant overhead. Furthermore, it permits information transmission with a reasonable appropriation of wireless and remote assets. A piggyback and on-demand system guides nodes to become energy proficient where the residual energy is sent to their neighboring nodes. Nonetheless, no enhancement system is utilized in contradiction of ant based protocol, which utilizes the ant colony system.

In [14], LEACH-MA protocol is proposed which is based on modified ACO. The residual energy parameter is used along with LEACH (Low-energy adaptive cluster hierarchy routing) protocol for the selection of current cluster head. The measure of energy consumption is minimized by this algorithm. The energy and distances are combined in this approach for choosing the cluster head. But no facts have been provided on the threshold value for selecting the cluster head. In [15], a hybrid tree-based search approach, called ANT-BFS, is proposed to discover the optimum route for information communication. This scheme integrates breadth first search with ACO to minimize the amount of energy consumption. However, the memory and computational time requirements may arise some issues for this technique in case the cluster head and the sink are located far from each other, so it might not be suitable for large scale. Content based routing (CCR) protocol is presented in [16] that ensures reliability of information transmission for IoT applications. It utilizes the process of data aggregation and ensures good load balancing. This method transmits data based on the message content and incorporates link quality information. The traffic reduction gain is achieved by this technique, where an objective function is used by the nodes for routing of the heterogeneous sorts of content. Each node builds a distinct routing entry by using the content to select the next node for transferring the data via the certain reliable communication link. The energy consumption has been conserved by this method by forwarding aggregated data to selective nodes. However, no details were provided on the estimation of the reliability while taking into consideration of the parameters that are also unspecified for this reliability.

In [17], an ant-based routing algorithm considering the energy supervision is presented. Reward and punishment technique are adopted with the pheromone update rules. The energetic parameter is considered and this technique increases network lifetime and energy-efficiency. However, various factors of energy utilization are required to be addressed. The node delay parameter is used here, but no details were found on the delay factor. An improvised energy saving ant colony based routing protocol for sensor networks is proposed in [18]. Three phases are introduced in this multipath protocol, such as discovery of neighbor through link information, transmission of packets through EWMA (exponentially weighted moving average) technique and efficient & reliable end to end delivery. The simulation results indicate that the proposed routing protocol for dynamic networks stands efficient in general performance, particularly regarding energy efficiency and throughput when contrasted with standard and other novel proposed routing algorithms. A multi-constrained technique that ensures QoS, known as IAMQER focused on ant colony, is proposed in [19]. The simulation results show that average energy consumption is minimized and packet delivery ratio is increased by utilizing this technique. Also, a route evaluating function is presented here. But IAMQER strategy might suffer longer processing delay as it is based on traditional ACO method.

This paper is an extension of work initially reported in [20], where the necessary parameters for communication process have been taken into account by the proposed method, such as mobility and energy parameters. Most of the proposed routing algorithms have addressed merely some communication parameters that are not adequate enough to enhance the communication quality needed by the energy constrained IoT applications. The important communication parameters must be considered by the mechanisms that mediate in the correspondence procedure, for instance routing, for resolving the issues of routing and for efficient communications in the IoT network system. The proposed protocol has utilized the advantages of the ant system to discover the optimum path for information communication and to improve communication quality within an IoT system.

2. System Model

2.1. The Network Model

The considered network model consists of *M* sensor nodes with random distribution in an $L \times L$ rectangular area. The proposed algorithm comprises the system model for IoT communications which is designed with several nodes or sensors M_i. The adopted graph G with the nodes and connecting links has constructed the network. It is assumed that the sensor nodes have the same computation and preliminary energy. The nodes can update the information about its neighbors. The received signal strength indication (RSSI) is used to compute the approximate distance of the senders by the nodes. For that calculation, the transmitted power of the objective is acknowledged.

2.2. Proposed ACO Algorithm

The most pheromone path is chosen as the shortest and optimal path in the traditional ACO algorithm [11] and the energy parameter is not considered. Hence the network's node energy on that path reduces abruptly and lessens the entire network's lifespan. Furthermore, novel routing protocols are required to manage the overhead of mobility. As the topological changes for the mobility of the sensor nodes and the sink nodes generate frequent updates in the network, which may drain the node energy extremely in an energy-constrained system. This paper expands our research works presented in [1] in terms of network performance to improve communications within IoT. The proposed system of ours [21] has been investigated more here to enhance scalability, to accommodate node mobility, and to minimize initialization and processing delay for time critical applications in the context of IoT. An improved network routing algorithm based on ACO is proposed here by analyzing the nodes' balanced consumption of energy.

At the place of node *i* and time *t*, each ant *m* will adhere to the following probabilistic formula to select the next node *j* being the forwarding node of the subsequent route for the enhancement that is proposed of the ACO algorithm of ours [20] for the next hop routing:

$$P_{ij}^{m} = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}[\vartheta_{ij}(t)]^{\gamma}E_j}{\sum_{s\subset allowed_m}[\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}[\vartheta_{is}(t)]^{\gamma}E_s}, & j \subset allowed_m \\ 0, \quad others \end{cases} \quad (1)$$

here,
$$\eta_{ij}(t) = \frac{1}{d_{ij}} \quad (2)$$

where $\tau_{ij}(t)$ and $\eta_{ij}(t)$ are the amount of pheromone and heuristic information on edge (*i, j*) respectively and $\eta_{ij}(t)$ is typically $1/d_{ij}$. α and β are two parameters that control the influence of the pheromone intensity and heuristic information respectively. d_{ij} is the distance between *i* and *j*. The average mobility parameter is used to calculate the stability factor, $\vartheta_{ij}(t)$, where γ is the mobility constant. E_j is the node residual energy that ant *m* will visit.

To avoid faster local convergence in case large amount of pheromone deposition happens on the routes, the pheromone update is required to improve as well. So, the amount of pheromone is updated and limited by incorporating a threshold value and can be obtained as follows:

$$\tau_{ij}(t+1) = \begin{cases} T, & \tau_{ij}(t+1) > T \\ (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) & else \end{cases} \quad (3)$$

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \quad (4)$$

where ρ is the pheromone coefficient for evaporation, $\rho \in (0,1)$, T represents the threshold value used to limit excessive pheromone deposition, $\Delta\tau_{ij}(t)$ is the increased pheromone concentration of edge (*i, j*), usually given by,

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{R}{L_k} & if\ kth\ ant\ uses\ the\ edge\ (i,j) \\ 0 & otherwise \end{cases} \quad (5)$$

where *R* represents strength of pheromone, L_k denotes the path length of the kth ant.

2.3. Energy Model

The energy model of wireless communications that is presented in [22] is incorporated by the proposed system. The free-space and multi-path fading model are utilized here depending upon the distance *d* in between the sending and receiving nodes and a threshold value, d_0. The following equations for energy consumption for the transmission (E_{tran}) of an *S*-bit data by the sensors are used:

$$E_{tran}(S,d) = \begin{cases} SE_{elec} + S\epsilon_{fs}d^2 & if\ d < d_0 \\ SE_{elec} + S\epsilon_{mp}d^4 & if\ d \geq d_0 \end{cases} \quad (6)$$

where E_{elec} indicates the energy dissipated to run electronic devise circuitry. The energy consumption in free-space and multi-path fading model are given by ϵ_{fs} and ϵ_{mp} respectively. *d* denotes distance and d_0 is the threshold value. The receiving energy, $E_{Rx}(S)$, for an *S*-bit data for a node is provided as follows:

$$E_{Rx}(S) = SE_{elec} \quad (7)$$

To calculate the residual energy of a node n_i, the following equation is used:

$$E_{res_i} = E_{tot_i} - E_{tran_i} \quad (8)$$

where E_{res_i} is the residual energy, E_{tot_i} is the total initial energy and E_{tran_i} is the transmission energy.

2.4. Fitness Function

A function for the evaluation of the routes, path assessing index, is provided here considering the current energy of the nodes, and the path of routing. If remaining energy is not assessed as in the conventional ACO algorithm, which causes early demise of some nodes and ultimately affects the whole network lifetime. The routing path is related to each particular ant after all the ants get to the destination node. The fitness value for the path can be computed as follows:

$$f_{(fitness)}{}_m^k = \frac{E_{res_i}}{L_m^k} \quad (9)$$

where the residual energy level of a sensor node n_i is E_{res_i}. L_m^k is the length of the route for m^{th} ant and k^{th} iteration. The pheromone is updated on the optimum path, having the highest fitness value.

2.5. Planning of Route Phases

Step 1. The route's arrangement phase. At first an initialization signaling is broadcasted by the sink node. Each node acquires its own neighboring node, then updates and adds it to its own routing table. The adjacent node link pheromone value is set to 1. The current node's ant packet is generated by each node that contains the number of nodes and the routing table. N_{max} is set to the maximum number of iterations, and the initial number of iterations is set to 1.

Step 2. The route's organization phase. The next node will be selected by the ant that locates in node *i* according to (1)-(2). Upon receiving the ant package, a node forwards ant package in accordance with the probability P_{ij}^m .

Step 3. The route's optimization phase. Once all ants get to the destination node, i.e. the sink node, the fitness function is used for optimal route selection where the fitness value is computed for the route according to (9) and the route with highest fitness value is chosen for optimal route for data transmission. The concentration of pheromones on this path is then updated according to (3)-(5).

In this route's optimization phase, once the ant package is received by the sink node, it will count how many ants packets each node sends. Let's assume that the total node number is *n* and the number of ants package sent by, for instance, node *i* is X_i (i =1,……., n).

Then the total sum of the network ant package can be expressed by means of:

$$T = \sum_{i=1}^{n} X_i \quad (10)$$

Back ant package is generated by each node. When back ant package is received by the nodes, the back ant package's adjacency linkage information is updated by (3)-(5). The node is selected as the next hop which resides in the back ant package. Following the information in the back ant packets, it is determined whether to send a new packet of ants at that same instance. When new ant package need to be sent, the route is established.

2.6. Proposed Improvement

Forming clusters associated with the group of nodes is deliberated on achieving scalability and robustness. For this reason, the routing protocol presented in [20] can be integrated with a clustering routing technique, such as LEACH (Low-energy adaptive clustering hierarchy) protocol [22]. It improves the selection strategy of optimal cluster head (CH), which is based on probability, node residual energy, and the distance of a node from the base station (BS). For selecting the cluster head node *j*, if it is assumed that node *i* is the current cluster head and next node is *j*, an ant *m* will use the probability calculation given by:

$$Prob_i(\mathrm{t}) = \frac{dist_i{}^{*\alpha} + [P_{ij}^m(t)]^{*\beta}}{\sum_1^{N_i} dist_i{}^{*\alpha} + [P_{ij}^m(t)]^{*\beta}} \quad (11)$$

Where $Prob_i$ provides the probability of each node to be selected as a CH (cluster head), $dist_i$ is the distance of node, α and β are two control parameters and N_i denotes the set of cluster nodes. P_{ij}^m is found from (1). Figure 1 shows the flow diagram of this suggested enhancement.

Using the proposed ACO based routing algorithm and the proposed improvement given above the initial optimal CH (cluster head) is selected. Then the final optimal CH is elected using the maximum fitness function value provided in (9). The optimal initial cluster head sends data to the optimal final cluster head. The optimal final cluster head transmitted data to the sink node.

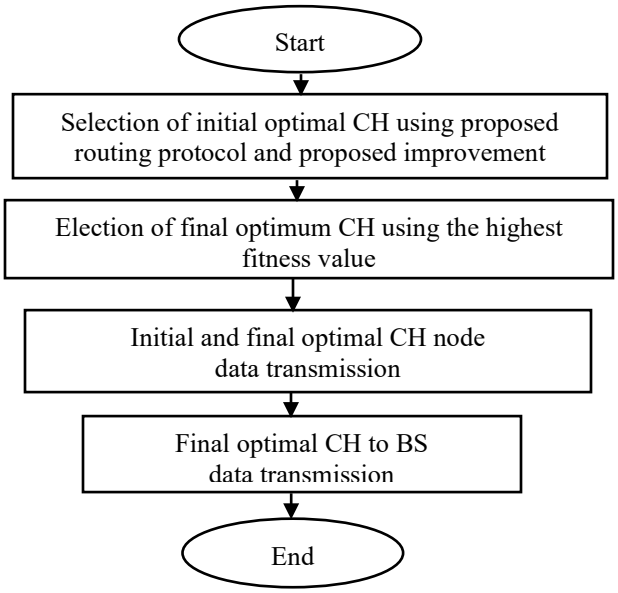


Figure 1: The flow diagram of the proposed improvement

Figure 2 shows the flow diagram of the proposed routing protocol operation.

3. Result and Discussion

The simulation is performed utilizing MATLAB where the proposed routing protocol is evaluated according to some important parameters, i.e., the consumed energy, the lifetimes of nodes, best cost, end-to-end delay, throughput and packet delivery

rate. The method represented here is evaluated against the conventional ACO, a current routing protocol for IoT, e.g., RPL (routing protocol for low power lossy network) [23] and an existing ant colony based algorithm, EICAntS algorithm [12] as benchmark protocols.

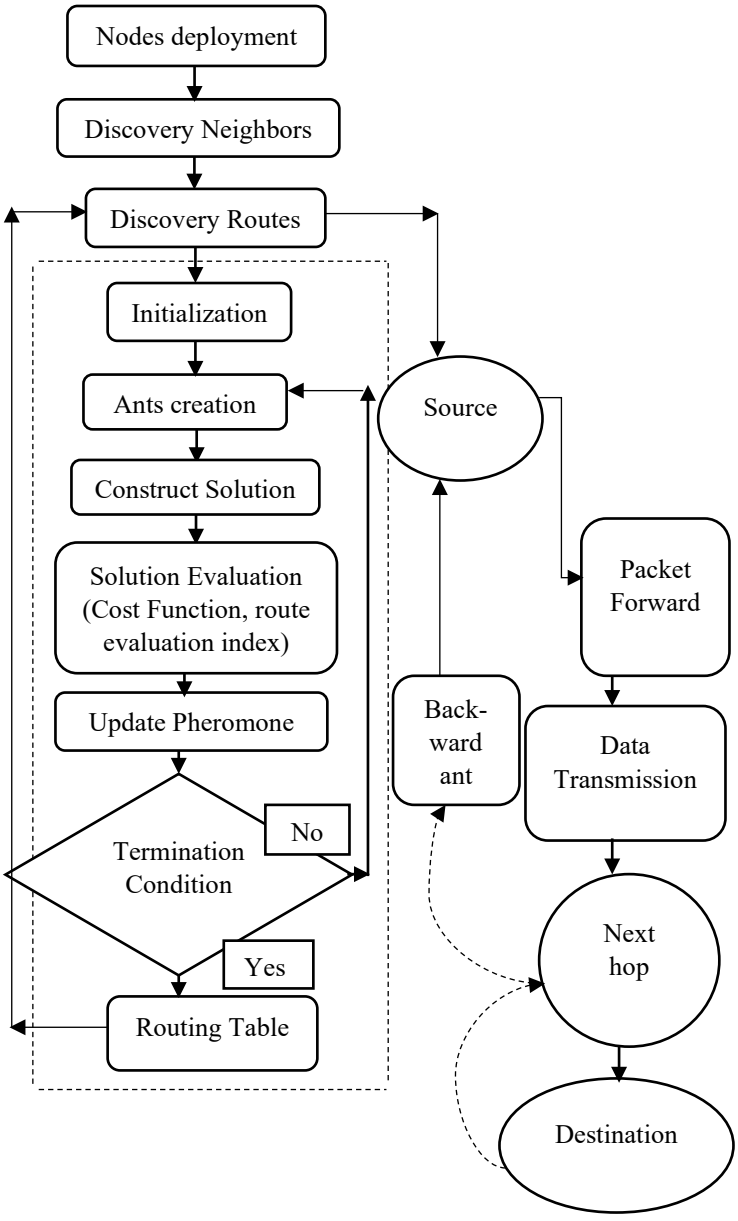


Figure 2: Overview Block Diagram demonstrating the Proposed ACO based Routing Algorithm Operation

The simulation parameters are based on the repeated tests and on the basis of simulation analysis of various parameters and set as: $\alpha = 1$, $\beta = 1$, $\gamma = 1$, $\rho = 0.05$. More parameters are provided in the Table 1.

The proposed system is supporting a mechanism permitting the selection of the best cost route for information communication from the source node to the destination, the base station. The network's parameters are dynamically adjusted to the network evolution, the growing number of nodes, the changed environment issues due to mobility, and outstanding node energy.

Table 1: Simulation Parameters

Parameters	Values
Simulation Region	100m×100m
Number of Nodes	40,50,60,70,80,90,100
T	100
No. of ants	40
Initial energy per node	0.5 joule
Node speed	2 m/s ~ 5 m/s
Message bits transmitted	4000 bits
Transmission distance	50 m

In Figure 3, for the proposed ACO based routing protocol and for the traditional ACO based protocol, the comparison of the node energy consumption per transmission is shown. The considerable improvement is shown for the proposed method. For majority of the nodes, contrasted with the conventional protocol, the proposed method has smaller energy consumption, almost 50% less. The proposed calculation has presented noticeable refinement for the total energy consumed of each node for individual discovery for the routes, contrasted with the benchmark protocol.

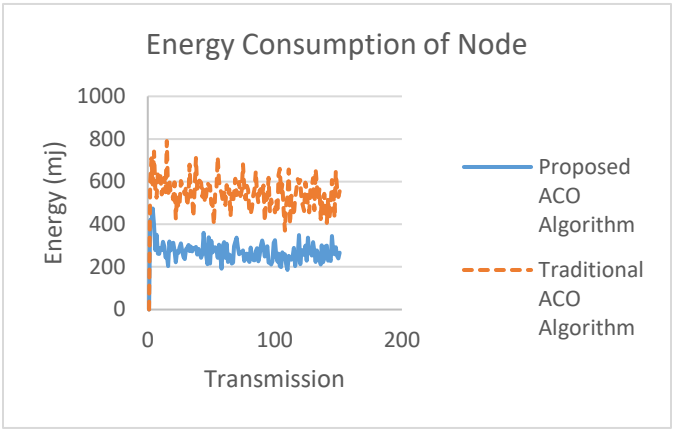


Figure 3: Comparison of node energy consumption per transmission

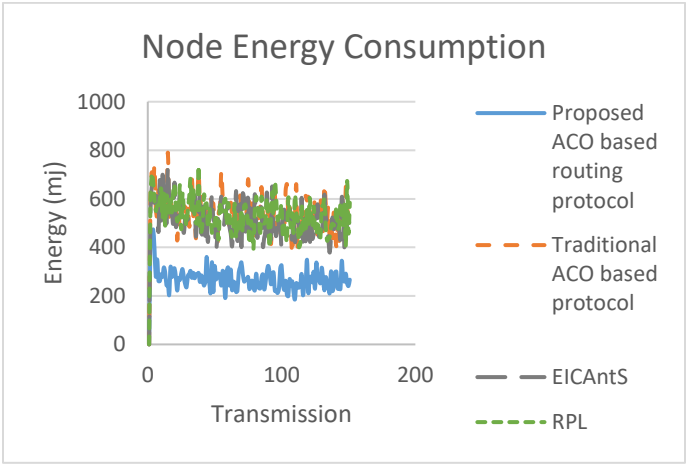


Figure 4: Comparison of node energy consumption per transmission

From Figure 4, we can see the node energy consumption per transmission based on improved method proposed here and other benchmark protocols, based on original ACO, RPL and EICAntS protocol. It is demonstrated from the evaluation results that

compared with the other two algorithms, the proposed ACO algorithm has attained a refinement, a much smaller energy consumption, almost 50% less.

The average energy consumption for the increasing quantity of nodes is represented in Figure 5 and Table 2. Figure 5 shows that the system using the proposed ACO based routing protocol as a resolution has minimized the average energy consumption, almost 50% less, by comparing with the original ant colony-based system, RPL and EICAntS technique. Since the proposed solution allows the determination of the optimal path for packet transmissions, and the retransmissions are avoided as well. Also, this presented method optimizes path selection strategy and minimizes updating phases. The scalability is maintained by the proposed system that has attained lower energy utilization contrasted with the benchmark protocols even when the number of nodes increases in the network.

Table 2: Average Energy Consumption (mj) with the Number of Nodes for 150 transmissions

No. of Nodes	Proposed ACO based routing protocol	Traditional ACO based protocol	EICAntS	RPL
40	270.49	545.89	516.65	532.65
50	281.67	570.07	566.42	574.78
60	303.76	620.61	606.97	632.29
70	330.97	650.51	660.42	679.46
80	350.65	686.47	695.87	692.01
90	372.42	739.92	744.55	735.55
100	401.21	791.33	799.01	768.49

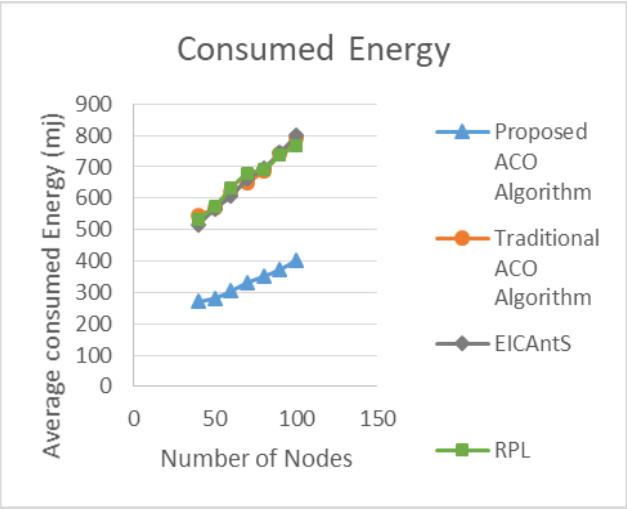


Figure 5: Comparison of the average energy consumption

The effective use of energy lead to increase of network lifetime. The percentage of survival nodes based on improved protocol proposed here and the traditional ACO based protocol are shown in Figure 6 and 7. The issues of ineffectiveness and uncertainty of the routes are alleviated as the protocol proposed here takes into account the energetic parameters. The energy utilization needs to be balanced and the present node energy needs to be deliberated; otherwise, the node energy is depleted and some nodes expire. The network lifetime is affected by this. The proposed approach deliberates the existing energy of the nodes besides the route path. Employing the suggested enhancement according to (11), the longevity, i.e. the nodes' lifetime and lifetime of the network, is increased by the offered ACO based technique.

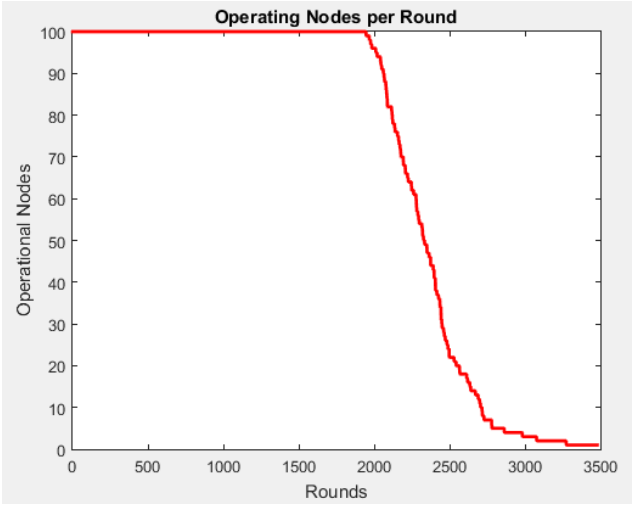


Figure 6: The percentage of survival nodes for the proposed method

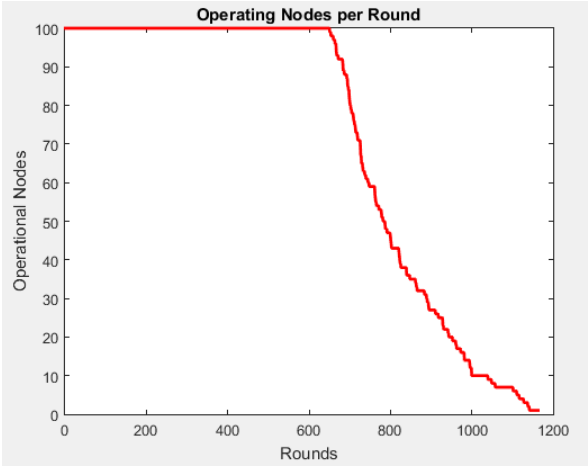


Figure 7: The percentage of survival nodes for the traditional ACO based protocol

We can see from the Table 3 and in the Figure 8 that the average end-to-end delay increases with an increase in the number of nodes. Route discovery and path selection stages affect this parameter, which are instated all the more regularly with an imperative number of nodes and with the topology evolution because of mobility of the nodes. Using the proposed algorithm the route discovery and selection stages are improved and the repetition problem is minimized when many packets should be resent to the destination in case the route might not be the best

always. The proposed system achieved better results with regard to average end to end delay, almost 40% reduced end to end delay, compared to the other benchmark techniques.

Table 3: Average End-to-End Delay (ms) with the Number of Nodes using Fitness Function

No. of Nodes	Proposed ACO based routing protocol	Traditional ACO based protocol	EICAntS	RPL
40	159.78	367.41	364.42	370.67
50	203.59	397.57	394.64	382.81
60	250.81	413.22	412.96	410.75
70	263.37	427.28	414.09	425.13
80	299.25	464.68	460.82	455.42
90	312.43	478.52	465.4	466.35
100	327.46	485.86	479.72	488.96

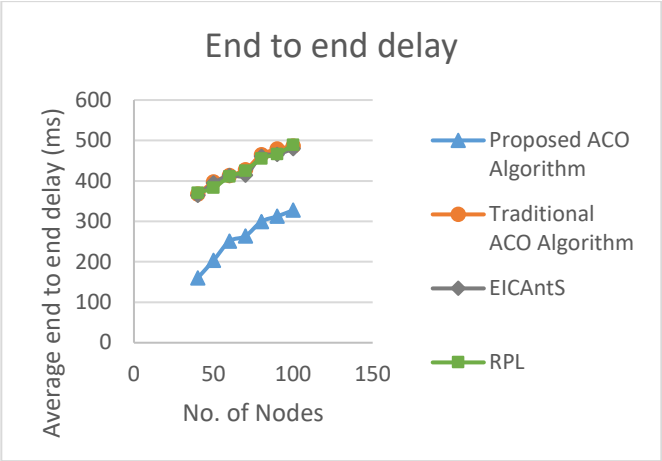


Figure 8: Evolution of average end to end delay with the number of nodes using fitness function

Throughput results are presented in Table 4 and depicted in Figure 9. The quantity of digital data, over a physical or logical connection per time unit, is the measure of the throughput in sensor networks. It is measured in bits/s or bps (bits per second), occasionally in per-second data packets or in per time-slot data packets. It can be defined mathematically, as the total number of packets delivered over the total simulation time:

$$\text{Throughput} = N / \text{Total simulation time} \quad (12)$$

where N is the number of bits received successfully by all destinations.

We can clearly see from the contrast that the results obtained are better than the results recorded by the other network.

Table 4: Throughput with the Number of Nodes over the Total Simulation Time

No. of Nodes	Simulation Time (Proposed Algorithm)	Simulation Time (EICAntS)	Proposed Algorithm	EICAntS
40	42 sec.	42 sec.	3809.52 (bps)	975.24 (bps)
50	45 sec.	45 sec.	4445.45 (bps)	1137.78 (bps)
60	48 sec.	49 sec.	5000 (bps)	1253.88 (bps)
70	51 sec.	51 sec.	5490.196 (bps)	1405.49 (bps)
80	54 sec.	53 sec.	5925.93 (bps)	1545.66 (bps)
90	55 sec.	55 sec.	6545.45 (bps)	1675.63 (bps)
100	57 sec.	57 sec.	7017.54 (bps)	1796.49 (bps)

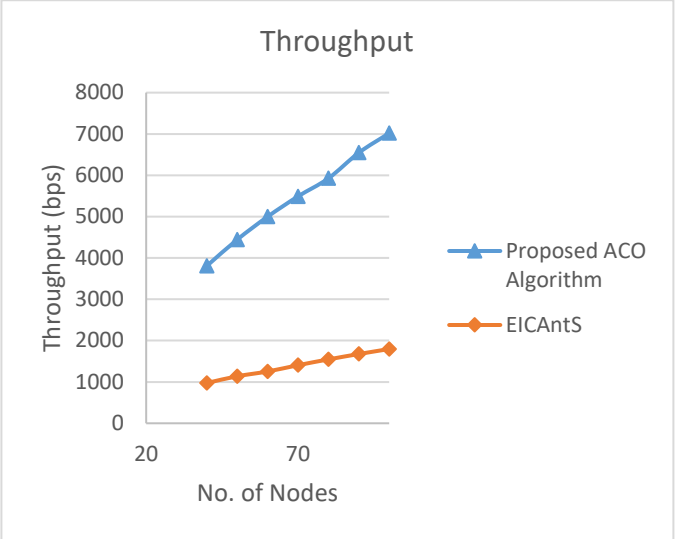


Figure 9: Throughput for the Proposed ACO and EICAntS algorithm

The calculation of the proposed algorithm in accordance with the delivery rates of the packets is provided in the Figure 10. The good results found with the proposed approach, even with an imperative number of nodes because the system has the capability that let many packets to send to get to the destination node in short time.

We can see from the comparison of the best cost as presented in Figure 11 that the network using proposed protocol has minimized initialization delay compared to the benchmark protocols. The network system achieved faster convergence for the initialization delay to discover the best cost route, which is about 30% improvement in convergence. Current node energy, mobility, route path, and cost function including route assessment index are

considered by the proposed method, making it preferable as a solution for the network system to improve communication quality. Figure 12 shows that utilizing the route evaluation calculation, about 60% enhancement can be recorded to find the optimum path with faster convergence in initialization delay.

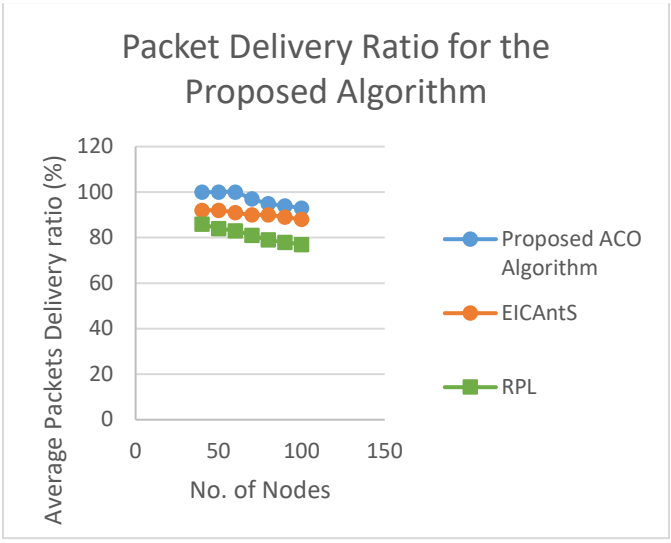


Figure 10: Packet Delivery Rates with the Number of Nodes

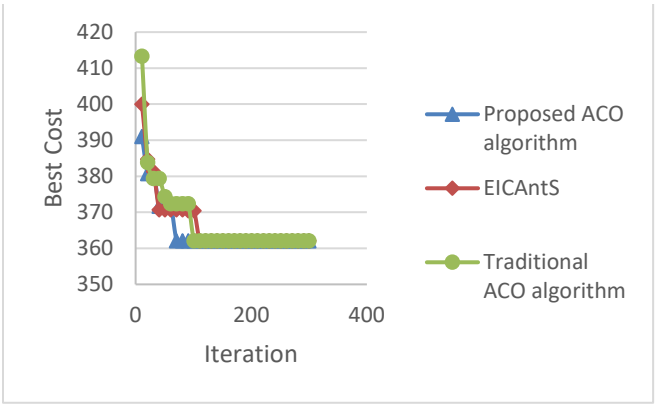


Figure 11: Comparison of the best cost with respect to the number of iterations (not utilizing fitness function)

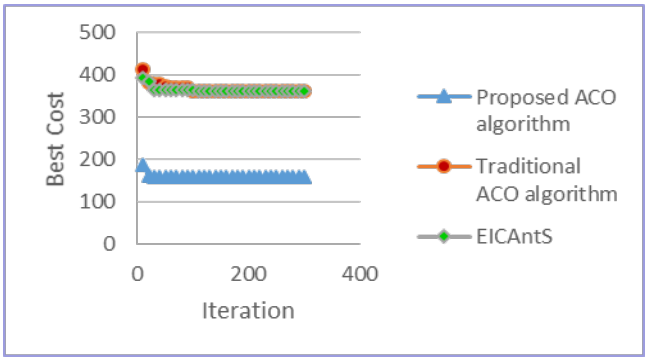


Figure 12: Comparison of the best cost with respect to the number of iterations using fitness function

4. Conclusions

The popularization of IoT-connected devices enabling development of IoT applications possess multiple and complex aspects of IoT while designing efficient communication protocols. The proposed system has utilized the beneficial features of an ant colony system to improve the route determination process and communication quality and mechanism in the context of IoT and to make IoT network system more scalable under varying load conditions. The parameters used in the selection of the routes are dynamically attuned to manage the network's dynamic condition. Current node energy, mobility, route path, and cost function including route assessment index are considered by the proposed method to enable efficient information communications within IoT. In the probability formula of ACO algorithm, the energy factor and the average mobility of the nodes are incorporated as well. It is demonstrated from the evaluation results that the proposed ACO based routing algorithm attained faster convergence for the initialization delay, about 30% improvement to discover the optimum route, and reduced energy consumption, of almost 50% less consumed energy even with the increasing number of nodes, compared with the benchmark algorithms. The proposed system achieved better results in terms of average end to end delay, almost 40% reduced end to end delay, compared to the other standard protocols. Also, it is confirmed that the proposed protocol as energy efficient method even for scalable networks, where the performance does not deteriorate abruptly as the number of nodes increases. Rather, it maintained improved network performance comparing with the other algorithms from the benchmark papers.

References

[1] A. Sharmin, F. Anwar, S.M.A. Motakabber, "Energy-Efficient Scalable Routing Protocol Based on ACO for WSNs," 2019 7th International Conference on Mechatronics Engineering, ICOM 2019, 1–6, 2019, doi:10.1109/ICOM47790.2019.8952053.

[2] Cisco, How the Next Evolution of the Internet Is Changing Everything, White paper, 2011, CISCO White Paper, 2011.

[3] Y.A. Malkani, L. Das Dhomeja, "Secure device association for ad hoc and ubiquitous computing environments," in 2009 International Conference on Emerging Technologies, ICET 2009, 2009, doi:10.1109/ICET.2009.5353132.

[4] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless sensor networks: A survey," Computer Networks, 2002, doi:10.1016/S1389-1286(01)00302-4.

[5] K. Sohraby, D. Minoli, T. Znati, Wireless Sensor Networks: Technology, Protocols, and Applications, 2006, doi:10.1002/047011276X.

[6] G. Anastasi, M. Conti, M. Di Francesco, A. Passarella, "Energy conservation in wireless sensor networks: A survey," Ad Hoc Networks, 2009, doi:10.1016/j.adhoc.2008.06.003.

[7] W. Dargie, C. Poellabauer, Fundamentals of Wireless Sensor Networks: Theory and Practice, 2011, doi:10.1002/9780470666388.

[8] M. Saleem, G.A. Di Caro, M. Farooq, "Swarm intelligence based routing protocol for wireless sensor networks: Survey and future directions," Information Sciences, 2011, doi:10.1016/j.ins.2010.07.005.

[9] T. Gui, C. Ma, F. Wang, D.E. Wilkins, "Survey on swarm intelligence based routing protocols for wireless sensor networks: An extensive study," in Proceedings of the IEEE International Conference on Industrial Technology, 2016, doi:10.1109/ICIT.2016.7475064.

[10] R.A.L. Rabelo, J.V.V. Sobral, H.S. Araujo, R.A.R.S. Baluz, R.H. Filho, "An approach based on fuzzy inference system and ant colony optimization for improving the performance of routing protocols in Wireless Sensor Networks," in 2013 IEEE Congress on Evolutionary Computation, CEC 2013, 2013, doi:10.1109/CEC.2013.6557967.

[11] M. Dorigo, V. Maniezzo, A. Colorni, "Ant system: Optimization by a colony of cooperating agents," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 1996, doi:10.1109/3477.484436.

[12] S. Hamrioui, P. Lorenz, "Bio inspired routing algorithm and efficient communications within IoT," IEEE Network, 2017, doi:10.1109/MNET.2017.1600282.

[13] K. Machado, D. Rosário, E. Cerqueira, A.A.F. Loureiro, A. Neto, J.N. de

Souza, "A routing protocol based on energy and link quality for internet of things applications," Sensors (Switzerland), 2013, doi:10.3390/s130201942.

[14] V. Gupta, S.K. Sharma, "Cluster head selection using modified ACO," Advances in Intelligent Systems and Computing, 2015, doi:10.1007/978-81-322-2217-0_2.

[15] R. Khoshkangini, S. Zaboli, "Efficient Routing Protocol via Ant Colony Optimization (ACO) and Breadth First Search (BFS)," International Conference on Internet of Things (IThings 2014), (March), 375–381, 2014, doi:10.1109/iThings.2014.69.

[16] M.D. Devi, K. Geetha, K. Saranyadevi, "Content Based Routing Using Information Centric Network for IoT," in Procedia Computer Science, 2017, doi:10.1016/j.procs.2017.09.145.

[17] R. Bao, H. Pan, Q. Dong, L. Yu, L. Shao, "Ant colony-based routing algorithm for wireless sensor networks," Chinese Journal of Sensors and Actuators, **24**(11), 1644–1648, 2011, doi:10.3969/j.issn.1004-1699.2011.11.025.

[18] A. Nayyar, R. Singh, "IEEMARP- a novel energy efficient multipath routing protocol based on ant Colony optimization (ACO) for dynamic sensor networks," Multimedia Tools and Applications, 2019, doi:10.1007/s11042-019-7627-z.

[19] Y.L. Wang, M. Song, Y.F. Wei, Y.H. Wang, X.J. Wang, "Improved ant colony-based multi-constrained QoS energy-saving routing and throughput optimization in wireless Ad-hoc networks," Journal of China Universities of Posts and Telecommunications, 2014, doi:10.1016/S1005-8885(14)60267-3.

[20] A. Sharmin, F. Anwar, S.M.A. Motakabber, "A Noble Approach of ACO Algorithm for WSN," Proceedings of the 2018 7th International Conference on Computer and Communication Engineering, ICCCE 2018, 152–156, 2018, doi:10.1109/ICCCE.2018.8539295.

[21] A. Sharmin, F. Anwar, S.M.A. Motakabber, "A novel bio-inspired routing algorithm based on ACO for WSNs," Bulletin of Electrical Engineering and Informatics, 2019, doi:10.11591/eei.v8i2.1492.

[22] W.B. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," IEEE Transactions on Wireless Communications, 2002, doi:10.1109/TWC.2002.804190.

[23] S.S. Solapure, H.H. Kenchannavar, "Design and analysis of RPL objective functions using variant routing metrics for IoT applications," Wireless Networks, **26**(6), 4637–4656, 2020, doi:10.1007/s11276-020-02348-6.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1719-1726 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Design of Power Efficient Routing Protocol for Smart Livestock Farm Applications

Shahenda S. Abou Emira[1,2], Khaled Y. Youssef [3,*], Mohamed Abouelatta[1]

[1]*Faculty of Engineering, Electronics and Electrical Communication, Ain Shams University (ASU), Cairo, 39827, Egypt*

[2]*Faculty of Engineering, Electronics and Electrical Communications, October University Modern Sciences and Arts (MSA), Cairo, 39827, Egypt*

[3]*Faculty of Navigation Science and Space Technology, Beni-Suef University (BSU), Beni-suef, 62521, Egypt*

ARTICLE INFO

Article history:
Received: 16 September, 2020
Accepted: 27 December, 2020
Online: 28 December, 2020

Keywords:
AODV
IoT
Livestock
MANET
Routing
Sustainability

ABSTRACT

The demand on livestock as a source of protein is increasing with the dramatic increase of world population which approaches 8 billion. As a result, the monitoring of the performance of livestock breeding is becoming essential. The technique which could increase livestock production and improve the efficiency of operation is digitalization of livestock management .This could be achieved by models' realization that relates the rates of productivity and the parameters of operation which count on collecting big volume of data that results from digitalization process. In fact, the adoption of IoT technology in livestock environments is usually challenged by energy problems and power efficient communication technologies. In this paper a modified AODV protocol is proposed to enhance the traditional AODV protocol used in MANET networks that was mainly concerned with optimal route selection based on shortest path to send the data packets through it. The proposed technique models and simulates operational parameters of intermediate nodes as a factor to decide on the optimal path selection in IoT based MANET networks in addition to the shortest path that leads to transmission power gain of the source node. The results show a significant power performance enhancement that reaches average power saving of 33.3% compared to the traditional AODV protocol.

1. Introduction

This paper is an extension of work originally presented in 14th International Conference on Computer Engineering and Systems (ICCES) [1]. The increase in world population that reaches 8 billion increases the demand for livestock production. As a result, it is necessary to avoid livestock losses and protect it from diseases as this could threaten the countries' economy and food security. The process of observing livestock was difficult as it took a lot of time to detect if they had medical issues or not. Accordingly, researchers illustrated nowadays different advanced monitoring systems for livestock, these systems depend on remote sensors. Different types of sensors were used as: accelerometer, rumination, body temperature and humidity, and sound sensors. They were placed on the cattle to collect data and send it to the processing unit to take the appropriate action [2].

Today, Different technologies play an important role to increase livestock production and to save their losses as well as improving the operation efficiency in large farms. One of the recent technologies that offers the measure, actuate, sense and infer of physical and physiological parameters is the Internet of Things (IoT) technology. Livestock performance monitoring systems using IoT technology require special type of networks. Adhoc network is one of these special networks as it is energy efficient network and elongates the devices' lifetime. In the Adhoc network, different sensors are used to monitor behavior, stress, diseases and temperature. These sensors provide the IoT network with the collected data which allow the early detection of diseases. As a result, early treatment is offered which saves money and time.

The main problem in IoT technology is the energy source as it needs to work for a long time to reduce the need for replacement of sources or human collaboration. This paper concentrates on the study of suitable network topology that provides sensor network operation using energy efficiently. Accordingly, different

*Corresponding Author: Khaled Y. Youssef, Khaled.youssef@nsst.bsu.edu.eg

techniques against farms' environment are illustrated and simulated. Additionally, a Mobile Adhoc network (MANET) algorithm is introduced to decrease the source energy drainage and assure energy sustainability.

2. Literature Review

Several literatures are presented in previous work that proved the significance of IoT and sensor technologies to improve the livestock farm performance especially the large animal farming. In [3], the author presented a group of micro-sensors which were non-invasive sensors. They were mounted on the livestock using a collar that sent data via the internet wirelessly. As a result, a continuous connection between animals and veterinarian was established.

In [2], the author proposed a system using various sensors which were mounted on the cattle body. These sensors were used to monitor different parameters which allowed to detect the cattle body issues constantly. In [4], the author developed a system using misting systems for heat stress reduction. It was also developed to monitor humidity and temperature ranges. The system assured the avoidance of cattle death and the increase of producing high quality milk by preserving the ranges of the monitored parameters constant.

Although the literature above highlighted the use of rechargeable batteries as a main power source of the sensor nodes used in their studies, the researchers did not consider the battery recharging methodology especially for Adhoc wireless sensor networks. On the other side, the following literature developed techniques that enhance wireless sensor networks' performance.

In [5], the author focused on wireless sensor networks and proposed a gateway based Geographical Energy Aware Routing (M-Gear) protocol. In the system, the sensor nodes were distributed according to their location into four regions. Outside the sensing area, the base station was placed and at the center of the sensing area, the gateway node was located. Their developed protocol performance was compared to the performance of Low Energy Adaptive Clustering Hierarchy (LEACH) protocol. The compared performance was in terms of energy consumption and network lifetime. The comparison proved that M-Gear protocol was better than LEACH protocol.

In [6], the author developed a Modified Sensor Protocol for Information via Negotiation which is called M-SPIN. They made a comparison between the performance of M-SPIN and the traditional SPIN protocol using the broadcast communication. In the developed protocol, the hop distance between nodes and sink node was measured using a distance discovery phase. Additionally, the total number of transmitted packets was reduced as a result of transmitting data only to the sink node. As a result the consumption of total energy was minimized.

In [7], the author proposed an algorithm to modify the Dynamic Source Routing (DSR) protocol. The number of Route Reply (RREP) packets was decreased using the proposed algorithm and it also decreased the size of the header for the data packets which resulted in decreasing the overhead. They also developed an algorithm to reduce the energy consumed in data transmission.

In [8], the author presented a Dynamic MANET On-demand (DYMO) protocol and ant colony optimization algorithm which was capable of managing environmental changes. Graphs were used in ant colony optimization algorithm which was based on probabilities' method to find routes. Two factors were measured to evaluate the discovered routes. The measured factors were the route reliability and delay time. The software used to implement the algorithm was NS-2.

Accordingly, the proposed modified AODV technique capitalize on extending the enhancements in the literature to develop IoT operation aware Adhoc wireless sensor network that is suitable for use in livestock farm applications.

3. Overview on Energy-Efficient Routing Protocols

Mobile Adhoc network (MANET) is a group of self - configuring mobile nodes. These nodes can be used at the same time as router and end node, therefore a fail in the operation of some nodes affects the performance of the network. Energy deficiency of nodes is one of the main problems which affects the performance of network in MANET. As a result, it is important to use energy efficiently in MANET in order to increase the lifetime of node's battery [9]. There are several routing protocols in MANET such as Destination Sequenced Vector (DSDV), Dynamic Source Routing (DSR) and Adhoc On-demand Distance Vector (AODV). These protocols are used to select the optimal route between source and destination in order to send data through it.

3.1. Destination Sequenced Distance Vector

DSDV is one of the proactive routing protocols in MANET. It is a table driven routing protocol and it is based on Bellman-Ford algorithm. In this protocol, a routing table is used at each node to store all available destinations. Additionally, advertisements are made as a severe change may exist in the table's data. In this protocol, a sequence number is also sent to the destination in order to discriminate the old routes from the new ones which avoids the formation of loops. It isn't recommended for highly dynamic networks because it needs regular updates in its routing tables which lead to battery drainage [10].

3.2. Dynamic Source Routing

DSR protocol creates routes only when requested, therefore it is one of the reactive routing protocols. It doesn't use routing tables at the intermediate nodes as it uses source routing. In this protocol, the intermediate nodes send a Route Request (RREQ) packet with its own address identifier. The available routes in the DSR protocol are being tracked by a route cache. Two main mechanisms are used in DSR: route discovery and route maintenance. They are used to discover routes and to notice if any change occurs in the network topology [11].

3.3. Adhoc On-demand Distance Vector

AODV is one of the most known reactive routing protocols in MANET. It creates only a route when a source node needs to send data to a destination node therefore it is better than DSDV as it minimizes the number of broadcasts. It has a good performance in dealing with large number of mobile nodes as it creates dynamically at each intermediate node a routing table. It also has

a sequence number counter which prevents the formation of loops in its routing tables. This sequence number also keeps the nodes updated by comparing their routing tables to check how updated information they have about other nodes [12].

4. Performance Evaluation of Adhoc networks

There are different parameters measured to assess the performance of MANET protocols as summarized in Table 1.

4.1. Packet Delivery Fraction

The ratio between the number of broadcasted packets by the source node and the number of received packets at the destination node is known as PDF. In case of low number of nodes, DSR performs better in this metric but the performance declines with the increase of the number of nodes. On the contrary, DSDV performs better in case of large number of nodes but AODV showed that its performance is uniform in both cases [13].

4.2. Average end to end delay

It is the time it takes from the packet being created at the source node to the packet being delivered at the destination node. In this metric, DSDV performance declines with the increase of the number of nodes. However, the DSR and AODV performance is uniform [13].

4.3. Number of Packets Dropped

It is the failed data packets that couldn't reach the destination node. In case of large number of nodes, DSR and AODV have a good performance. On the other hand, DSDV has the worst performance as it declines with the increase of the number of nodes. The performance of AODV in this metric is the best as it exchanges information periodically between nodes which ensures connection [13].

Table 1: Performance of MANET Protocols

Metrics	AODV	DSDV	DSR
Average end to end delay	Uniform performance	Performance declines when number of nodes increase	Uniform Performance
PDF	Uniform Performance	Good performance when number of nodes increase	Performance declines when number of nodes increase
Number of packets dropped	Good performance when number of nodes increase	Performance declines when number of nodes increase	Good performance when number of nodes increase

5. Proposed Architecture for Livestock monitoring IoT networks

Figure 1 shows the architecture of a group of mobile nodes connected in MANET network. Each node in this network has different sensors such as accelerometer, skin temperature, humidity and ambient temperature, and respiratory rate. The data sensed from these sensors will be transmitted to a central processing server (CPS) to perform the proper action. In such a network, Power should be used efficiently to elongate the battery lifetime of each node. Accordingly, AODV protocol which is an energy efficient routing protocol is used to select the optimal route to send the collected data to the destination node. As a result, the total transmitted power of source node is reduced.

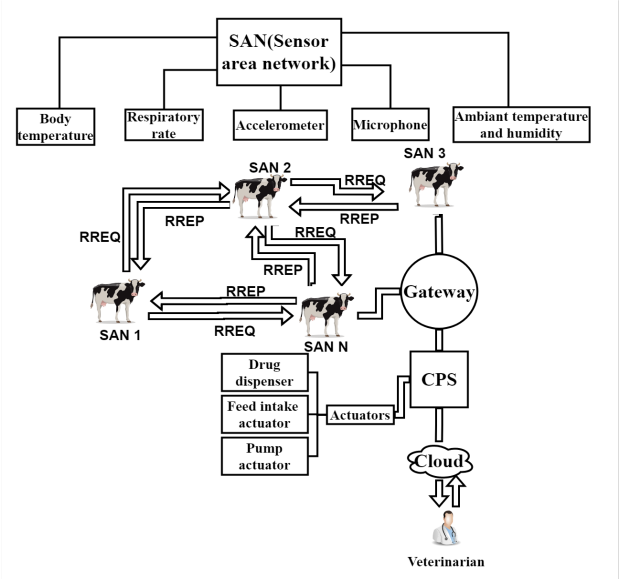


Figure 1: System Architecture for Monitoring livestock Performance

Figure 2 shows the flowchart of how AODV protocol is implemented. When a source node starts to search for a route to send its data through it to the destination node, it searches its routing table. If there is a route in its routing table to destination, it sends its data packets through this route. If there is no route to the destination node, it broadcasts a Route Request (RREQ) packet to all the neighbor nodes. If the node that received the RREQ packet is the destination node or it has in its routing table a route to the destination node, it sends a Route Reply (RREP) packet to the source node. If the receiving node isn't the destination node or there is no route for the destination in its routing table, it broadcasts a RREQ packet to all its neighbors. When the source node receives a RREP packet, it transmits its data packets through the same route where it received the RREP packet. If several RREP packets are received by the source node, it selects the RREP packet received from the shortest route and discard other packets. As a result, the power for transmitting the data packets is reduced which increases the source node battery lifetime.

A modified AODV protocol is proposed to enhance the overall network power performance. The optimal route in the modified protocol is selected depending on the index weight of the path rather than depending only on the shortest path. The index weight is a new factor measured to indicate the operation time of each node in a certain path. Each node is given a certain index

which represents its operation time. The index weight of the path is measured as the total index of nodes in a certain path divided by the number of nodes in this path. The path with the highest index weight has the lowest priority to be selected while the path with the lowest index weight has the highest priority to be selected.

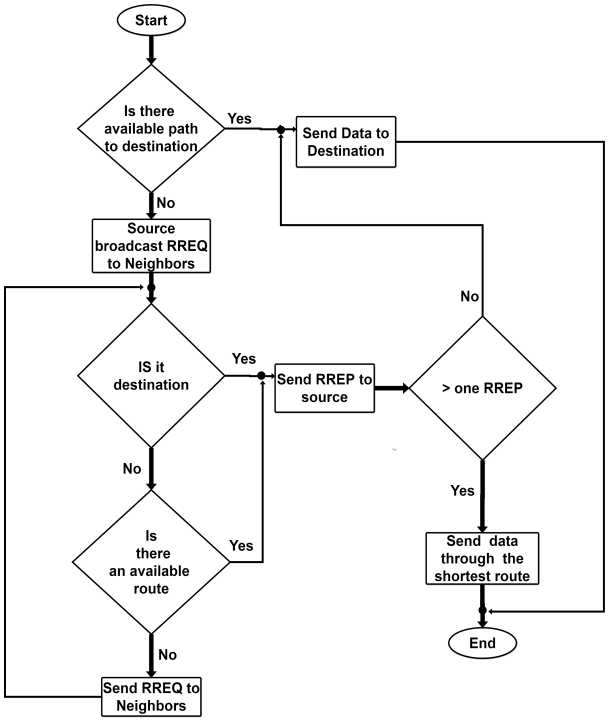


Figure 2: AODV Flowchart

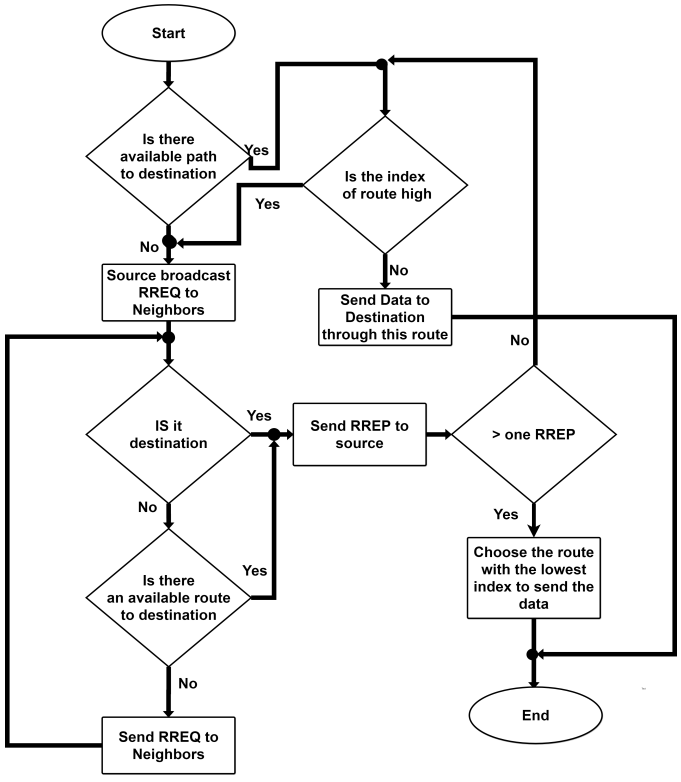


Figure 3: Modified AODV Flowchart

The flowchart in Figure 3 shows how the modified AODV protocol is executed. When a source node starts to search for a route to send its data through it to the destination node, it searches its routing table. If there is a route in its routing table to destination, it checks its index weight. If it is low, it sends its data packets through this route. If there is no route in its routing table or the index weight of the available route is high, it broadcasts a Route Request (RREQ) packet to all the neighbor nodes.

If the node that received the RREQ packet is the destination node or it has in its routing table a route to the destination node, it sends a Route Reply (RREP) packet to the source node. If the receiving node isn't the destination node or there is no route for the destination in its routing table, it broadcasts a RREQ packet to all its neighbors. When a RREP packet reaches the source node, it checks the index weight of the route. If it is high, it sends RREQ packet to all its neighbors searching for other routes with lower index weight. If it is low, it starts to send data packets to the destination through the same route where it received the RREP packet. If the source node receives more than one RREP packet. It selects the route with the lowest index weight to send data packets. As a result, it increases the lifetime of the serving nodes which enhances the overall network power performance.

6. Simulation Model Results

In this simulation, MATLAB software is used to simulate AODV protocol. In Figure 4, it is shown how the source node, destination node and the number of nodes are entered in this model. It is assumed in this simulation that the network has 20 mobile nodes randomly distributed in a farm size of 100 m^2 and a 20 m transmission range between nodes. The source node and the destination node are assumed to be node 2 and node 10 respectively. AODV protocol is implemented in this model to enable the source node to request different available paths to the destination node. The source node will receive RREP packet from the intermediate nodes if they have an available route to the destination node. In Figure 4, the available routes are shown.

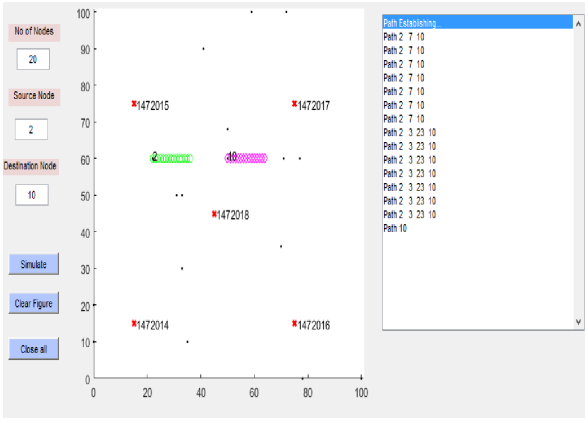


Figure 4: Simulation Model

In our work, two main performance indicators are used to measure the impact of the proposed advanced AODV protocol on the IoT network performance mainly the energy consumption and network lifetime.

i) Energy consumption is based on MATLAB model that is governed by the following equation

Energy Consumption (Joule) = Energy Consumption Rate (Joule/bit) * Transmission rate (packets per sec) * Packet size (bits) (1)

The multipath fading is also accounted in the simulator model given that the amount of Energy consumption rate considered in the model for the transmitter or receiver circuitry is 50 x 10^{-9} and amount of energy consumption for multipath fading is 0.0015 x 10^{-12}.

ii) Network lifetime that is defined as the ratio between its residual energy and the energy consumed to transmit data packets (i.e. the time spent until the node run out of energy). It is the third parameter calculated at each route as shown in Figure 7. In this case, the route with the longest network lifetime is the optimal route as it has the lowest energy consumption and the shortest distance. As a result, route 12 is the optimal route in terms of distance, energy consumption, and network lifetime.

The simulation results for applying AODV protocol in cows' performance monitoring system are shown in this section. The first parameter measured as shown in Figure 5 is the total distance between the source node and the destination node at each route. The optimal route is found based on this parameter as the optimal route is the shortest route which is selected to transmit data through it. In this case, the route with the shortest distance is route 12 so it is the optimal route.

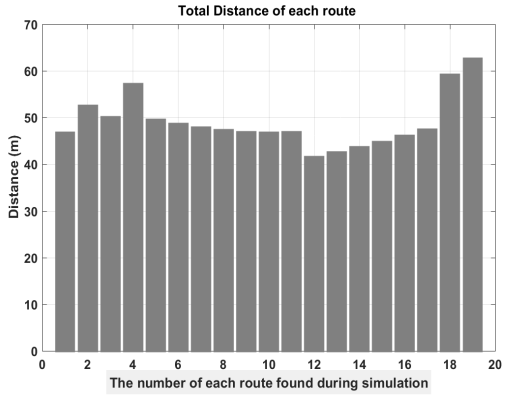


Figure 5: Total Distance versus The Number of Each Route Found During Simulation

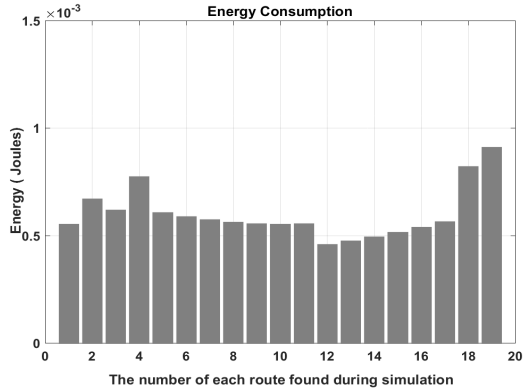


Figure 6: Energy Consumption versus The Number of Each Route Found During Simulation

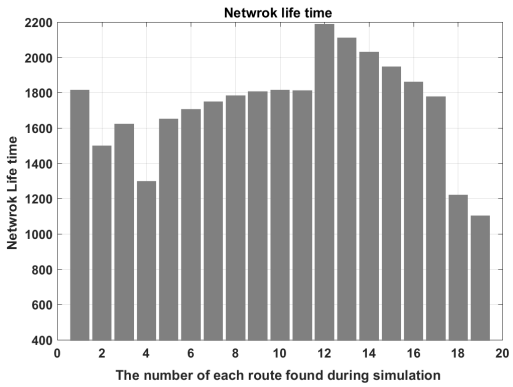


Figure 7: Network Life Time versus The Number of Each Route Found During Simulation

In this simulation, the energy consumption is the second parameter calculated at each route as shown in Figure 6. Equation 1 represents the energy consumption model. The route with the shortest distance has the lowest energy consumption. In this case, route 12 is the optimal route in terms of energy consumption and distance.

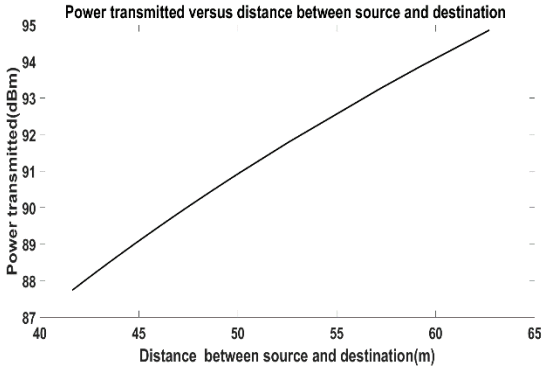


Figure 8: Transmitted Power versus Distance between Source and Destination During Simulation

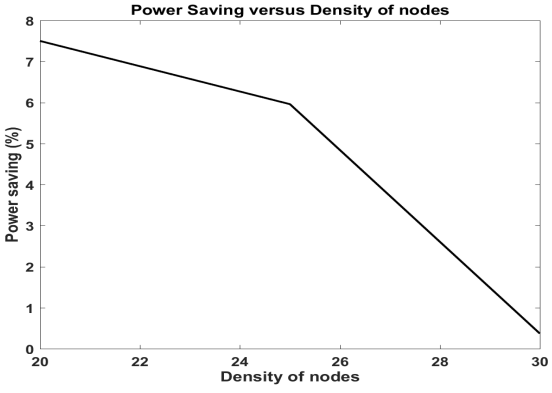


Figure 9: Power Saving versus Node Density in Mobile Adhoc Network

The fourth parameter calculated in this simulation as shown in Figure 8 is the transmitted power at each route. The data are transmitted with minimum power when transmitted through the shortest optimal route to the destination node. In this case, route 12 is the optimal route in terms of distance, energy consumption, network lifetime, and transmitted power.

The fifth parameter measured is the power saving as shown in Figure 9. It indicates when selecting the shortest optimal route in MANET how much transmitted power is saved. It is measured at different number of nodes as shown in Figure 9 to indicate that when the node density increases the power saving decreases. Figure 10 shows the power saving which indicates how much transmitted power is saved when selecting the shortest route in MANET compared to the transmitted power in fixed network. It is measured at different number of nodes to indicate that when the node density increases the power saving decreases as shown in Figure 10.

Table 2: Simulation Results

Route Number	Distance (meters)	Transmitted Power (dBm)	Power Gain
1	46.84294241	89.78211633	
2	52.58886799	91.79210254	
3	50.14213562	90.96446273	
4	57.192105	93.24979289	
5	49.60330529	90.7767743	
6	48.66124402	90.44367794	
7	47.90305917	90.1708796	
8	47.33955199	89.96531538	
9	46.97970233	89.83276007	
10	46.83013443	89.77736582	
11	46.894632	89.80127492	0.924
12 Optimal Route	41.64331698	87.73816223	
13	42.60107524	88.13317207	
14	43.66889973	88.56323963	
15	44.83914468	89.02264243	
16	46.10374679	89.50579849	
17	47.45489265	90.00758945	
18	59.20525888	93.85076102	
19 Worst Route	62.72062331	94.85276426	

Figure 11 shows different paths between source and destination which are path1 (P1), path2 (P2), path3 (P3), path4 (P4) and path5 (P5). It is assumed for each path that the distance between nodes is equal and the total transmitted power of the source node is 30 dBm. It also shows the transmitted power of each intermediate node in each path. There are some nodes which are common in more than one path such as nodes: 3, 4, 7 and 9. These nodes are called busy intermediate nodes as they could be used several times in the network which will lead to the drainage of their batteries. As a result, as mentioned earlier in section 5 a modified AODV protocol is proposed to save the battery of these busy nodes. There are different scenarios for node process shown in Table 3 as each intermediate node will have an index to indicate how busy this node in order to select the optimal path in terms of its index. In Table 4, it shows the transmitted power of each node and its index of processing. It also shows their index weight which is the total index of nodes divided by the number of nodes in each path. The selection of the optimal route will be based on its index weight, the route with lowest index weight is the optimal route to send data through it.

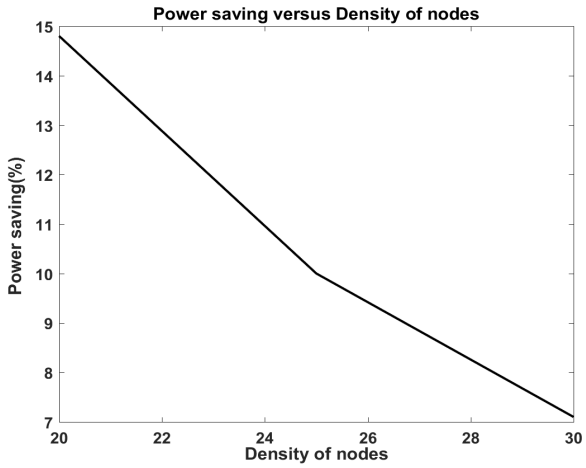


Figure 10: Power Saving versus Node Density in Fixed Network

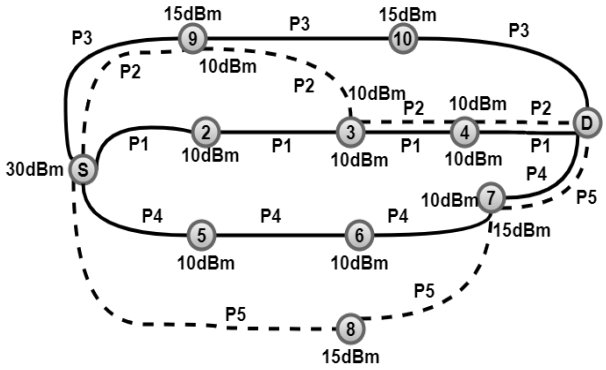


Figure 11: Different Paths Available to Send Data Packets from Source Node (S) to Destination Node (D)

Table 3: Different Scenarios for Node Process and Its Index

Node process	Duration of not processing node (T in hours)	Index(i)
Event Based	T >= 12	0
	6 =< T <12	1
	3 =< T < 6	2
	1 =< T < 3	3
	T < 1 hour	4
Periodic	Periodic Time (Tp in hours)	Index(i)
	Tp >= 12	5
	6 =< Tp <12	4
	3 =< Tp< 6	3
	1 =< Tp < 3	2
	Tp< 1 hour	1
	Data size (D in kilobyte(KB))	Index(i)
	D>=200	5
	100=<D<200	4
	50=<D<100	3
	10=<D<50	2
	D<10	1

Table 4: Parameters of Intermediate Nodes

Path	Intermediate Nodes	Power(dBm)	Index	Index Weight
P1	2	10	1	2.33
	3	10	3	
	4	10	3	
P2	9	10	5	3.67
	3	10	3	
	4	10	3	
P3	9	15	5	4
	10	15	3	
P4	5	10	1	2
	6	10	0	
	7	10	5	
P5	8	15	2	3.5
	7	15	5	

Figure 12 shows the index weight of each path which indicates the status of the intermediate nodes. If the index weight of a path is high, it indicates that its intermediate nodes were processing for long time therefore this path can't be selected to transmit data through it. As a result, this will save the lifetime of their batteries.

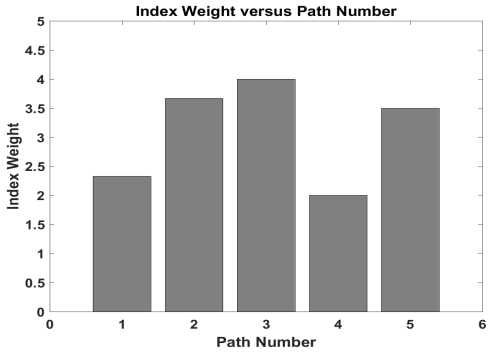


Figure 12: Index Weight versus Path Number

Figure 13 shows the power gain of each path which indicates how much power is saved by selecting a certain path. The optimal path shown in Figure 13 is path 4 as it has the highest power gain.

Figure 14 shows the power gain versus the index weight to indicate how much power is saved when selecting the path with the lowest index weight. As shown in Figure 14, the path with the lowest index weight has the highest power gain.

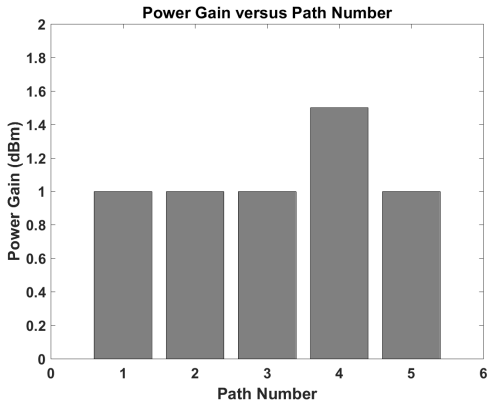


Figure 13: Power Gain versus Path Number

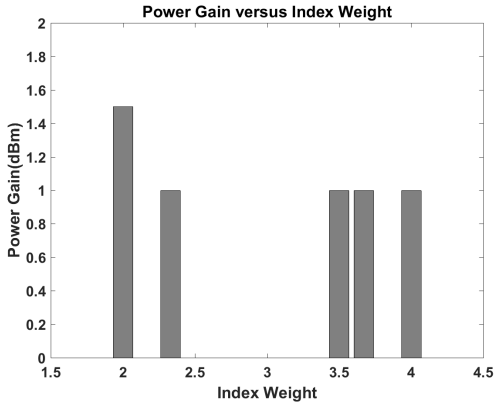


Figure 14: Power Gain versus Index Weight

7. Discussion and Conclusion

The results from applying the AODV routing protocol to select the shortest route showed that when the total distance between source and destination decreases, the energy consumption decreases which increases the network lifetime. It also showed that a decrease in the transmitted power occurs when the total distance decreases. For MANET and classical fixed networks, the power saving is calculated for each route found at different node densities. It showed that when the node density increases, the power saving decreases.

Table 5: Power Gain for Different Path Selection Scenarios

Busy intermediate nodes	Common in Path	Total Power(dBm) on busy node	Scenario	Power gain (dBm)
3	P1	20	1 (P1 selected)	10
	P2		2 (P2 selected)	10
4	P1	20	3 (P1 selected)	10
	P2		4 (P2 selected)	10
7	P4	25	5 (P4 selected)	15
	P5		6 (P5 selected)	10
9	P2	25	7(P2 selected)	15
	P3		8 (P3 selected)	10

Table 6: Traditional AODV versus Proposed Modified AODV Protocol

Routing Protocol	Optimal Path	Hop Count	Optimal Path Selection	Power Gain (dBm)	
Traditional AODV	3	3	Shortest Path	10	By comparing the traditional AODV and the proposed modified AODV, the results show a significant power performance enhancement that reaches average power saving of 33.3%
Proposed Modified AODV	4	4	Lowest Index Weight	15	

The traditional AODV protocol focused only on how to enhance the battery lifetime of the source node depending on selecting shortest optimal route. It didn't highlight the importance of the intermediate nodes which are serving nodes and their batteries could be drained quickly due to their operation time in the network. As a result a modified AODV protocol is proposed, this protocol added a new feature to select the optimal route. The selection of the optimal route in the modified AODV protocol is depending on the operation time of each node in the path. Each node is given an index which specifies its operation time as shown in Table 3. An index weight which is the total index of the nodes in a certain path divided by the number of nodes in this path is measured for each path. This factor indicates if the available path could be used or not. The path with a high index weight has the lowest priority to be selected because it indicates that the operation time of the intermediate nodes in this path is very high and thus the probability to have lower battery state of charge is higher.

Accordingly, the optimal path is the path that has the lowest index weight.

By applying the modified AODV protocol to select the optimal route depending on the index weight of the route as a priority to selection of the shortest route (i.e. if two routes have the same index, then the shortest one is selected). The protocol supports to give the intermediate node with longer operation time the opportunity to focus on sensed data processing. The networking processing function is then off loaded from this intermediate node to other nodes with lower operational time and accordingly higher probability of battery charge lifetime.

As observed in Table 5 the optimal path in terms of power gain is path2 and path4. It is also observed in Table 4 that the best path in terms of index weight is path 4. By comparing the power gain when selecting the shortest route (measured in hop counts) which is path 3 in Figure 11 to the power gain when selecting path4 which has the lowest index weight, it proved that the power is enhanced by 33.3% as shown in Table 6.

As a summary, a convenient network topology is studied in this paper to provide a network of sensors which has an energy efficient operation. Accordingly, in order to decrease the source energy drainage which assures energy sustainability, MANET algorithm is proposed. In MANET, AODV is the most common known routing protocol as its performance is the best compared to other MANET protocols. In this simulation, AODV was applied to allow the selection of the optimal route between source node and destination node. The optimal route is selected as it has the shortest distance, the lowest energy consumption and the longest lifetime.

A modified AODV algorithm is proposed in order to enhance the overall network power performance. The protocol depends on giving intermediate nodes with lower operation time a higher priority to share in the network routing even if the path could be longer which could lead to a power saving that reaches 33.3%. As a result, the paths in the modified AODV are selected based on the index weight of the path. The selected path is the path with the lowest index weight.

References

[1] S.S. Abou Emira, K.Y. Youssef, M. Abouelatta, "Power Efficient Routing Protocol (AODV) for IoT Based Livestock Applications," in 2019 14th International Conference on Computer Engineering and Systems (ICCES), 407–412, 2019, doi:10.1109/ICCES48960.2019.9068149.

[2] M. Meenakshi, S.S. Kharde, "Advance Cattle Health Monitoring System Using Arduino and IOT," **5**(4), 3365–3370, 2016, doi:10.15662/IJAREEIE.2016.0504162.

[3] J. Edwards, D. Gibson, "Novel Technology for the Remote Monitoring of Animals," Companion Animal Society Newsletter, **23**(2), 56–59, 2012.

[4] M. M, S. A, "IOT-BASED MEASURING TEMPERATURE AND HUMIDITY IN THE CATTLE DIARY," International Journal of Sustainable Development, **7**(11), 7358–7360, 2015, doi:10.24327/IJRSR.

[5] Q. Nadeem, M.B. Rasheed, N. Javaid, Z.A. Khan, Y. Maqsood, A. Din, "M-GEAR: Gateway-based energy-aware multi-hop routing protocol for WSNs," in Proceedings - 2013 8th International Conference on Broadband, Wireless Computing, Communication and Applications, BWCCA 2013, 164–169, 2013, doi:10.1109/BWCCA.2013.35.

[6] Z. Rehena, S. Roy, N. Mukherjee, "A modified SPIN for wireless sensor networks," in 2011 3rd International Conference on Communication Systems and Networks, COMSNETS 2011, 2011, doi:10.1109/COMSNETS.2011.5716469.

[7] M. Tamilarasi, S. Chandramathi, T. Palanivelu, "Efficient energy management for mobile ad hoc networks," Ubiquitous Computing and Communication Journal, **3**(5), 12–19, 2001.

[8] A.M. Oranj, R.M. Alguliev, F. Yusifov, S. Jamali, "Routing Algorithm for Vehicular Ad Hoc Network Based on Dynamic Ant Colony Optimization," International Journal of Electronics and Electrical Engineering, **4**(1), 79–83, 2016, doi:10.18178/ijeee.4.1.79-83.

[9] M. Tarique, K.E. Tepe, M. Naserian, "Energy saving dynamic source routing for ad hoc wireless networks," in Proceedings - WiOpt 2005: Third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, 305–310, 2005, doi:10.1109/WIOPT.2005.21.

[10] P. Manickam, T. Guru Baskar, M. Girija, D. Manimegalai, "Performance Comparisons of Routing Protocols in Mobile Ad Hoc Networks," International Journal of Wireless & Mobile Networks, **3**(1), 98–106, 2011, doi:10.5121/ijwmn.2011.3109.

[11] V. Denkovski, B. Stojcevska, T. Dovenski, V. Pachovski, A. Bozinovski, "Performance Evaluation of Routing Protocols in a Wireless Sensor Network for Targeted Environment," JITA - Journal of Information Technology and Applications (Banja Luka) - APEIRON, **10**(2), 2016, doi:10.7251/jit1502110d.

[12] I.D. Chakeres, E.M. Belding-Royer, "AODV routing protocol implementation design," in Proceedings - International Conference on Distributed Computing Systems, 698–703, 2004, doi:10.1109/icdcsw.2004.1284108.

[13] S. Ade, P.. Tijare, "Performance Comparison of Routing Protocols in Mobile Ad Hoc Networks," International Journal of Information Technology and Knowledge Management, **2**(2), 545–548, 2010, doi:10.1007/978-981-10-6890-4_4.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1727-1733 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

New Properties of Crimes in Virtual Environments

Roman Dremliuga[1,*], Natalia Prisekina[1], Andrei Yakovenko[2]

[1]*School of Law, Far Eastern Federal University, Vladivostok, 690922, Russia*

[2]*National Technology Initiative Center for Neurotechnology, Virtual and Augmented Reality Technologies, Far Eastern Federal University, Vladivostok, 690922, Russia*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 04 December, 2020
Online: 28 December, 2020

Keywords:
Virtual reality
Criminal law
Virtual crime

ABSTRACT

Virtual reality is a technology that literally allows constructing of a new reality for its users. VR has huge potential, it is able to change social life, communications, and the academic sphere, but VR may also be applied by criminals. In the present paper properties of potentially committed crimes in VR are analyzed. It is concluded that VR provides criminals with some advantages. The first advantage is the multijurisdictional problem in prosecuting the offender that is usual for most of IT. The second advantage is that a virtual environment gives a realistic experience to the user and literally replaces the real world. The third advantage is an integration of virtual-world environments with haptic devices that may factually affect the user in the real world. All of these allow the commission of new kinds of crimes where properties of conventional and cybercrimes are combined. Among such kinds of crimes analyzed in the paper are cybercrimes that affect the body and mental state of victims. VR is researched as a first tool that in combination with haptic devices gives criminals the opportunity to commit crimes against sexual freedom and to cause physical harm. Thus, VR crimes need a special legal framework that will take into consideration this kind of crime.

1. Introduction

This paper is an extension of work originally presented in 2019 International Conference on Cybersecurity (ICoCSec) [1]. Virtual reality (VR) is one of the most promising breakthroughs. It literally replaces the real world. VR dips humans into an alternative reality where different aspects of human life are imitated. The tasks of introduction of any information technology may vary, for instance, technology could be applied to make the world around us better. At the time of the pandemic, it is the only way to visit new spaces that are so similar to the real world [2]. Virtual reality is not a single technology but a set of technologies that permit a developer to construct an environment that would be recognized by our body and mind as real. VR is a tool to create new realistic worlds. Sometimes this environment is used for work, very often for gaming and social interaction. There are some forecasts that soon people will prefer virtual worlds over reality.

Virtual reality is a widely used technology. The market of VR is anticipated to grow from US$27 billion in 2018 to US$209.2 billion in 2022 [3]. Some researchers noted that VR attracts a lot of investors' attention and large companies invest in the technology [4].

Virtual reality is technology that literally allows constructing a new reality for its users. VR has great potential, it is able to change our everyday life, communications, and academic sphere. Nowadays virtual reality is actively applied in different spheres. The biggest driver is the entrainment and computer games. Also, a VR application is noticeably beneficial in teaching [5,6]. Sometimes virtual reality helps to simulate real situations [7]. VR is widely applied in medicine [8]. Some cases of rehabilitation of felons in VR exist [9,10]. In addition, a virtual world can be tuned in accordance with the needs of the user that makes it more attractive than real.

Virtual reality technology is a variety of hardware and software tools, which help to imitate a real or fictional world for a user. It interacts with our senses (vision, hearing, touch, smell, sense of balance, etc.) and misleads the body and mind that they are in the real world [11]. Virtual reality is an environment that is imitated through technology. The main virtual reality device is a special

*Corresponding Author: Roman Dremliuga, Email: dremliuga.ri@dvfu.ru

helmet, that may mimic the outside world through rendering video and audio. There are also numerous additional devices such as controllers, joysticks, steering wheels, haptic interfaces which increase a realism of VR illusion. All devices are designed to make an imitated world as real as possible.

"Virtual reality" has a common-used definition of the term. It is considered that VR is an information technology with special characteristics. For example, the following definition is suitable for the goals of the research. Virtual reality term is interpreted as "a technically constructed interactive environment with the help of computer facilities; as the generation and operation of objects, real or imaginary, on the basis of their graphic representation, simulation of their physical properties and ability to influence and independent presence in space, as well as the creation of such objects by means of special computer equipment" [12].

Development and introduction of any technology depend on a legal framework that is applied. Investors, developers, and users have to understand current and future legal rules that concern application of technology. In terms of business, the legal expertise is necessary on the step of making the initial decision. Wrong assessment of legal risks could lead to a total misapplication of funds, time, and efforts. Thus, to understand development of VR it is necessary to understand the current and future legal frameworks in this sphere.

The question is whether VR is another technology to which existing cyberspace laws apply or whether entirely new rules need to be developed. In order to form the correct legal framework, it is necessary to understand the conceptual differences between this technology and other technologies.

The sufficient legal risk is to face criminal cases with new properties. Some technologies provide tremendous opportunities for commitment of crimes. If a developer understands which challenges would exist in terms of criminal application, he will design his products by taking into consideration potential risks. The key is understanding how VR could be used by criminals and how it could be treated from a legal point of view.

The idea of this paper is not directly to enumerate possibilities of illegal use of VR but assess future prevalence and social consequences. To achieve this goal, it is necessary to utilize some assumptions that research will rely on.

- If VR maintains to achieve the same illegal result as some other technology, this automation will be replaced by VR in case of higher effectiveness of VR.
- If some property bolsters to spread VR use in society, this property will be the key characteristic that makes VR useful for criminals.
- Enforcement is always behind leaders of the crime 'industry' since it is bound by strict regulation and bureaucracy.

Thus, three aspects were carried out in the present paper. Properties of VR as a tool of crime were analyzed; challenges for the development of regulation that is caused by the characteristics of VR were reviewed. Some recommendations for the development of a legal framework were provided.

2. Research methodology

Research methodology for studying legal challenges of VR technology that goes beyond both dogmatic analysis has to be applied in the research. Analysis in the research follows positions developed within the philosophy of law and legal theory. Among relevant methodologies are the hermeneutic and argumentative approaches.

As the research is interdisciplinary, there is a synergy between methodlogies of disciplines. Interdisciplinary research must extend beyond mere multiplication of disciplinary methods and perspectives, to incorporate some element of integration or, mutual dialogue between disciplines. Interdisciplinarity is absolutely essential for the research as its object lies on the intersection of disciplines. As fairly commented some authors "it is important to have a multidisciplinary approach inconstructing such a methodology as the tools analysed stem from different disciplines and domain of expertise" [13].

Factually the research is based on a comparison of current and future properties of VR technology with current and potential legal framework. The focus on the legal and societal impacts of VR use requires us to observe the results of researches in the sphere of influence of virtual environment on humans. It is clear that such an important question cannot be accepted only on the basis of one or several studies, therefore, it is necessary to use the results of a whole set of studies to draw conclusions about the characteristics of VR, important for the formation of legal norms for its regulation.

The expected results of the study are directions where the legal framework applied to VR has to be upgraded. Recommendations with a high level of abstract should be developed. In future, de lege ferenda proposals could be based on our results.

Despite analysis and conclusions of the paper could be applied in different legal systems, the study is mostly focused on legal practice and conceptions of the United States and EU states as leaders in introduction of VR. The reseacrh is less applicable to asian technological leaders due to sufficicent differences in legal frameworks. Some ways of VR applications that are usual in the United States and EU are prohibited in other countries. For instance, virtual sex or VR pornograhpy are illegal in China, Indonesia and Malaysia.

Thus, the methodology can be shortly described as the application of a legal case study to the research of technology with an enormous social impact. The paper as some others [14] utilized a case study methodology. It is manifested in gathering and analyzing proven effects of VR on a user in terms of social, ethical, and legal risks. As the area of risk is quite wide, the paper focuses on risks related to the commission of criminal offences in VR.

3. Legal challenges

3.1. Increasing significance of VR

The main purpose of VR is to provide an environment for spending time. The application of VR for gaming dominates nowadays but VR is also used for workshops and business meetings. It is actively applied for educational purposes and simulation of difficult tasks. The advantage of VR is clear: a

customized reality that provides a high level of similarity of human reactions with the real nature. It is effective for games where people pass through all spectrum of emotions as well as for training where people become ready for the stress of a real situation. Most VR applications in business communications or games imply audio and video information exchange in some virtual space that increases the level of participation.

The level of realism could be even improved when real appearance is applied. In virtual reality, a user is frequently seen by others as an anthropomorphic avatar whose appearance may be similar to the user's appearance in the real world or may be completely different. Despite the voice and personal appearance of the virtual world can be altered, this does not exclude the possibility that voice and visual appearance of the user would be the same as in the real world.

The big advantage of VR is that people are able to communicate in a realistic way even if they are separated in the real space. Thus, people, in the form of relatively realistic avatars, communicate as they would be face-to-face. VR provides the possibility to imitate eye contact and to manipulate virtual objects that participants see together. It is similar to a video call, but participants of VR meetings have a higher level of involvement.

VR technology is far from perfection. Headsets cannot track eye movement with high accuracy. VR Graphics is far from a full imitation of the representation of the real world and is rather applied to construct some simple spaces with a low quantity of details. Facial expressions are still rendered in virtual reality quite poorly.

Despite the above, virtual reality is a very popular environment for spending time and human interaction. Under different studies, the amount of VR's application for spending time and social interaction is fastly growing [15]. There are approximately 170 million users of virtual reality worldwide. A double increase in the number of VR consumers is evidenced in recent years.

The growth of legal applications of VR implies an increase in criminal use. It is anticipated that the number of virtual reality users should correlate with the number of crimes committed in VR, as this was observed with the growth of the number of Internet users. As more people use VR worldwide, there are more potential victims and criminals in VR. Some researchers found the proliferation of the use of the global network (the Internet) as a major factor in the growth of cybercrime. Thus, the number of criminal acts in VR is expected to increase due to the growth of the number of its users [16].

Wide use of VR leads to a classic dilemma. A legislator has to choose between the need for a free exchange of computer information, as a factor of digital economy growth and the need to defend state and social security by restrictions on IT use. To find the balance between public security and the necessary freedom of technology application for its development, a lot of different factors have to be taken into consideration.

As with any technology, VR may facilitate criminal activity. Virtual environment may be a tool or instrument to commit a crime as well as an additional factor that facilitates it. Recently, the use of high technology to commit crimes increases, and it concerns virtual reality. An opportunity how criminal law is applicable in case of crime in virtual environment should be studied.

3.2. VR as a potential tool to commit a crime

VR is in some sense a unique technology that makes crimes with the application of that technology dissimilar with others. It may constitute difficulties for the investigations of such offences. VR as a tool confers crimes by new properties. Virtual reality is not simply another new technology, which requires to design technical standards. Rather VR application requires to introduce a new legal framework. There are several reasons why a traditional legal framework does not qualify. Basically, due to its special characteristics, VR requires fundamentally new legal frameworks and solutions.

The first, commonly known problem is cross border jurisdiction. It is difficult to identify a person with whom the user interacts in VR. Virtual environment and its users are often in different jurisdictions. A world that is undivided in VR may exist regardless of borders of the concrete state in terms of its users. Millions VR users have no idea where the virtual reality platform is physically located, where their data are processed and stored. Thus, virtual space may be outside the jurisdiction of the state where the victim applies to law enforcement.

Some authors commented that: “It will take a lot to get domestic police interested in investigating a crime where the criminal is in a foreign country. … Still, the greater difficulties of extradition are likely to exceed the greater ease of proof. And many VR street crimes might thus be practically ignored by traditional police departments” [17]. Thus, the multijurisdictional challenge is the multidimensional problem. There are a lot of questions as from legal assessment of committed offence as well from criminal procedure rules.

Any digital cross border space poses a challenge to law enforcement and the legislator. People from states with different legal systems meet in a virtual environment. Cross border character of VR spaces creates a situation where the crime has anonymous, remote in space, and global character. When virtual reality is applied to commit a crime, deanonymization and pursuing criminals become very difficult and even impossible.

The second feature of virtual reality as a place, tool, or instrument of crime is the capability of a virtual environment to create an illusion of reality. Sometimes this property forces real consequences for the potential victims of crime in VR.

Studies demonstrate that people uploaded at VR may be afraid of walking across a wooden plank that is on the ground in the real world but placed in virtual reality between roofs of skyscrapers. The experimental environment forces similar psycho-emotional reactions of a probationer as if it happens in the real world [18]. Dizziness, nausea, stomachache, and shaking body are just a few of the reactions to virtual heights [17]. Thus, experience in a virtual environment, in turn, may bring real consequences for body and mind. The VR customer may be literally frightened to a state of heart attack by a game in which he (she) feels quite real [19]. In case of the absence of consequences for the body, it may be a difficult psychologically traumatic experience. Thus, a virtual experience could cause permanent physical or psychological harm.

The realism of virtual environments has also positive effects, for example, VR is successfully applied for autism control [20] or in treatment of phobias [21]. But the side effect of VR realism is

that it is a perfect tool for hackers or other cybercriminals to "hack" into the victim's heads. Earlier, cybercriminals only had distant access to property, business reputation, and other values, but now they can literally get inside the brain. Harm also could be caused by the application of VR in medical treatment without direct intention [22].

In comparison with the real world, VR provides certain advantages for criminals in terms of the range of psycho-emotional reactions that could be provoked. In the physical world, it is very difficult to place a victim into the "Grand Canyon" and is even impossible to organize a meeting with a monster from science fiction. Hence, it is possible to distantly manipulate consciousness and emotions through VR by imitating any real or fictional environment. [23]. The character of such intervention is limited only by the imagination of the criminals. Current studies demonstrate that with the help of VR any emotion for up to 8 minutes could be forced [24].

Virtual worlds become more realistic due to the development of VR software and hardware. Objects are rendered with a higher degree of detail. More sophisticated models are applied to imitate interaction between objects in virtual reality. The link between events in VR and the body of the user gets more channels. Some researchers underline that higher realism implies a more significant psycho-emotional response of the customer on events in the computer environment [25].

As some researchers claim “Multimodal inputs will be combined in unprecedented patterns of stimulation that will ultimately make a VR experience more real than reality. Strong social emotions, in turn, will likely reinforce the sense of presence and make it easier to ascribe mental states to other virtual characters” [26]. Thus, intervention into human mental state is conceivably even more aggressive in future because new developments in sphere of VR technology.

The next challenge is development of haptic devices or other devices that are capable to affect the customer in the reality. Haptic technology enhances, new ones appear each year [27]. For example, there are enough investments in technologies for remote sexual intercourse. In the case of virtual environment application, skin sensations are networked between the users. The widespread use of such technologies implies that a criminal is potentially able to commit illegal intimate contact.

In some studies, conducted in the United States, it is underlined that hacker’s replacement of another user in a sexual act or unauthorized access to the user’s body in VR with the use of a haptic interface has similar characteristics with crimes against sexual freedom. In the legal research papers, such act committed with intention is considered to be either sexual harassment or rape [28]. Before haptic technology, it would be impossible to commit a rape distantly. Thus, haptic devices and similar equipment provide the capability to intervene in body inviolability.

The opinion exists that the above challenge could be overcome by a technical solution. Some authors state that “the ability to define default consent in software can make VR safer than the real world—for instance, well-designed software may let me consent in advance to certain types of touching but not others, or touching by some people but not others, and touching that isn’t consented to won’t even be felt. The question that remains is to what extent, if at all, it should be viewed as the victim’s job to set software consent boundaries; more on that below” [17]. Thus, the questions here: who would be responsible for the introduction of defense from body encroachments, what would be the liability of the developer or the seller for negligent informing about such kind of risk? And the main question is whether it is still crime or not?

All discussed above presuppose that VR allows to committing crimes which are impossible to commit with traditional IT. For instance, homicide could be committed on the Internet without VR, but not rape. A victim who has been raped in VR may not immediately realize what happens. Thus, a victim may be in a helpless state. As the victim is unable to understand that some illegal actions happen, he (she) is unable to take measures against the criminal. In some sense, it implies that rape or harassment in VR is probably even more dangerous because it is latent even for the victim.

Moreover, software and hardware for distant sex have low security against cyberattacks [29]. Often developers in this sphere are focused on new user engagement than on fixing vulnerabilities. It indicates that such kind of products could be easily hacked. Potentially criminals even could provoke VR users to move outside their secure boundaries in the reality through application of malware [30].

Due to the innovations in the VR sphere criminals are able to encroach not only on the mind but also on the body of victims. It is a complicated question whether the sexual assault in VR is comparable with real-world crime in terms of consequences for a victim. Of note, in the case of VR, there is a great number of difficulties in the prosecution of the offender. In addition, criminal significantly reduces the sum of risks related to the act (risk to be caught at the crime scene, risk of suffering if the victim resists, risk of public censure even if the guilt is not proven, and some others).

Development of haptic suits with feedback poses a risk of increasing harm-related crimes. The suit by electrical impulses informs the body of what happens with it in VR. The haptic device stimulates skin when some object touches the user in virtual world. Currently, there are no regulations on the permissible power of impact on the body. It is not legally defined what the degree of pain and harm is possible in VR. Of note, such a rule has to depend on the individual characteristics and other parameters. Due to the absence of regulation a risk of inflicting real suffering from haptic suits in virtual reality exists.

For instance, «new peripherals and technologies enter the market supporting a variety of interaction fidelities, such as data glove controllers, skeletal motion capture suits, tracking systems, and haptic devices» [31].

3.3. Immoral acts

As described above some manifestations of VR are not covered by modern legal regulations but there are some difficulties in addressing the mentioned challenges even in terms of moral rules. As law and morality are both means of conducting control, they are strongly linked. It does not imply that both social regulators have to correlate each other, but at least legislator has to take into consideration the norms of molality. And if a gap in the legal framework exists, VR has to be assessed from a moral point of view. A moral system of norms is clearly often very dynamic and opaque, but it is often helpful to know what is good or what is bad from a moral point of view. Moreover, the last social initiative in countries of Western Europe and Northern America as MeToo and

BLM demonstrate that moral norms could have even bigger impact on society than legal rules.

The first question is arising from identity flexibility in a virtual environment. In VR any user can choose any appearance that is provided by a software, risk exists that a person will apply the wrong appearance. At the same time, avatars and names influence on the perception of the user by other persons. These characteristics that are very significant for the attitude of other persons could be easily changed. It raises some ethical questions that are significant from ethical and sometimes legal point of view.

Imagine a teenager who uses VR poses as an adult and engages in an intimate conversation or even virtual sexual contact with an adult person. Who would be liable for such actions? Who has to take precautionary measures? Who must be in charge of the customer's appearance? What would the legal consequences of similar action be? Probably the legislator has to oblige VR developers to incorporate technical solutions for this kind of case. If one person misleads another person by his/her appearance in VR, how does it correspond with moral or even legal rules?

Even if users represent themselves in VR without misleading other persons, appearance also could be an issue. A demonstrative example is an avatar that looks like a naked body. Public nudity is banned by ethical and legal norms in most of the states, but a video demonstration of a naked body is rather not. Whether the naked body in VR is regarded as pubic nudity or is it a demonstration of video content from an ethical and legal perspective? Does the legislator have to oblige the developer to add measures against the use of such avatars?

A prosecution of public nudity in VR is harder from a legal and ethical point of view. What is a nature of this act? Factually it is a rendering of video content. However, a demonstration is accompanied by emulating personal presence. Thus, nudity in VR is not the same as nudity in games or movies. At the same time in terms of the intent of a person, it could be considered as public nudity. Virtual avatar is able to be real human size, with real human movements, seemingly standing close to other users. Mentioned avatars are perceived much more like a real person than a picture on a screen. Such actions may offend someone's feelings or cause psychological trauma in a demonstration for minors.

The issue of a proper avatar in VR is not limited to the nudity problem. For instance, an avatar may offend someone's feelings in other ways. It may be intolerant of gender, race, religion, political beliefs. More ways to offend or to cause suffering exist in VR. Imagine that someone's deceased relative's face is used in VR. Some users are able to use Hitler or famous serial killer's face for their avatars.

It is able to harm, disgust, or provoke users which see the avatar in VR. Appearance could be chosen as intentionally as well as recklessly. Developers can clearly provide technical measures to filter or block graphical representation of a user in virtual reality. But as it is linked with additional costs of software development, special regulation obliging VR companies has to be enacted.

Another problem is compliance with ethical norms applied in reality. As users all come to VR from the real world, some expectation exists on the conduct of the user in a virtual environment. There is an enormous temptation to transpose the moral norms of the ordinary world into the virtual world. Is it appropriate from a moral point of view to scream, create a mess, etc. in the VR art gallery?

There is a suggestion that "VR environment operators can easily implement code that can deal with the screamers. The operator could, for instance, allow each user to control the perceived volume, for that user, of any other user. That is good not just to silence the screamers, but also to quiet down acquaintances who are a bit too loud, or to amplify acquaintances who mutter. And this should be technically trivial to code" [17]. But it provokes the same question who will oblige the operator or the developer of the application to bear additional expenses on the implementation of these rules. Any state that implements such rules will automatically decrease chances in the race for technological dominance in the sphere of VR.

Of note, the issue of content filtering in VR is within the scope of the free speech agenda. When the legislator imposes additional legal requirements on VR developers, it implies reducing–or elimination–of the ability of individuals to speak freely in VR [32]. It is difficult to differentiate harmful behavior from an expression of someone's freedom of speech right on the level of development, as any case depends on context, viewer, or environment. Thus, a restriction for the developer or the operator is a threat to the freedom of speech in a virtual environment.

Another issue related to legal restrictions in VR is enforcement. The developer that designs or supports VR software or hardware is able to incorporate restrictions in its products; however, it is inevitably linked with real-time tracking of the user's conduct. It is presumably a high price for secure VR.

Development of restrictive measures will face also technical issues. Experience of companies which control and process of big data shows that "training moderators to overcome cultural biases or emotional reactions in the application of rules to facts can be analogized to training lawyers or judges. Regardless of the label, training content moderators involves a repetitive process to override cultural or emotional reactions and replace them with rational valid resolutions" [33]. So, an automatic realization of rules could produce unjust decisions, discriminative ruling, and other negative effects. It is probably more difficult to automatically make just decisions about the character of VR scenes than to create a transparent judicial system. This part of companies' work is not limited by the assertion that technical decisions have to be developed, the company has to create a division to manage disputes with customers which are not enjoyed by rule application.

Big IT companies are often criticized for a lot of power, but the capabilities of a virtual environment will allow private companies to have almost unlimited power over customers and their data. Factually, companies get control over all channels of human perception in different spheres from gaming to sexual relations. "VR settings mimicking as close as possible the multisensory nature of real-world human interactions would be a welcome and powerful addition to the field. Experiments aligning auditory, vestibular, and tactile signals with the visual domain are already improving VR immersion and realism" [26]. Thus, some private companies will control our reality, minds, and bodies.

4. Discussion

The study is just the beginning. This paper describes the key challenges related to managing risks of virtual environment use. VR is in the spotlight of social attention and attracts more attention from researchers. The study could be developed in different directions. There are several topics for further research. First of all,

it is necessary to systematize theoretical and practical legal knowledge related to the regulation of VR. It could include but not be limited to the research literature, legal practice, and international public agreements, as well as case-study.

Also, it is necessary to define the term "VR" and differentiate its categories and types, depending on the depth of immersion. It is helpful in terms of systematization and unification of results of current and future studies. It is necessary to categorize types of crimes that could be committed in VR. As some authors fairly comment on significance of legal definition and categorization of VR "it will serve as a basis for choosing the optimum regime of their legal protection" [34].

In terms of multidisciplinary dialogue, it is necessary to find out categories of the permissible technical parameters of VR functioning. It is very supportive in terms of preventing the destructive impact of technology on the psychophysiological state of the subject. For instance, there is a need to regulate the frame rate to avoid: headache, nausea, dizziness, epilepsy, and other disorders. It is especially critical for the application of VR by minors and juveniles.

Some authors claim that it is necessary "to add additional investigation and analysis testing stages to the development of virtual reality technologies in efforts to protect the public. These tests might not focus on physical health and safety concerns, but rather on physiological and social influences" [35]. However, in accordance with the analysis, the main concerns are mental and physical health. Moreover, any measures that take into consideration social influences would be very local and depend on the volatile social environment.

There is a need to define and regulate the allowable limits of the subject's functional permissibility in virtual reality (virtual ethical standard). Moreover, it is necessary to study the potential consequences of non-compliance with the prescribed rules of conduct, in order to prevent the direct projection of actions in VR in real life, such as the loss of boundaries of objective reality and the unconscious killing of a person in the real world.

In some papers, authors conclude that current law is very flexible and could be also applied to VR [22]. These studies do not take into consideration the dual nature of VR, i.e. this cybertechnology is rapidly merged with real world. The question arises in this study that current law is not applicable in cases where VR manifests itself in cyberspace and space of the real world. For instance, illegal access to haptic devices is not the same as usual "hacking".

There is research on restrictions of speech in VR in terms of US legislation. However, there is no conclusion how it corresponds to the First amendment and leaves a lot of space for further discussion [36]. In accordance with the analysis of current research in the sphere of medicine and psychology, it is difficult to conclude whether it is appropriate to apply the law of real world or virtual worlds for VR. The main reason is that virtual space and behavior of users there are often perceived as reality by other users.

In addition, there are several problems related to VR regulation that have to fall within the scope of criminal law. For instance, rules for the regulation of rendering in VR, application of real-world standards in VR, total control of users' conduct by IT giants, and appearance that misleads other users.

5. Conclusion

In summary, our study demonstrates that the offender gets new opportunities in VR. It literally helps criminals to hack the head or body of the victim. VR allows to provoke the emotions and to manipulate the consciousness of the victim at a completely new level. It is proved in the research that VR use could not be regulated as any other IT technology as the effect of perception by the victim is comparable in strength with the effect of events in the real world. In the paper characteristics of the virtual environment that have to be taken into consideration during a process of VR, regulation development is underlined.

Moreover, VR combined with haptic devices requires even more sophisticated regulation as such environment merges with the real world. Due to the integration of skin-feedback devices into the virtual environment, the consequences of VR's actions also have an influence on the user in the real world that could not be considered as just illegal manipulation with computer information. As many types of criminal acts with harm to the victim's body can now be committed remotely, a new kind of crime with dual nature is originated.

The article shows that there are not enough criminal law measures to combat crimes in VR. First of all, the legislation should provide for the obligations of the developer of VR technologies in order to minimize the possibility of criminals. Development and adoption of such measures is not an easy task, it is necessary to minimize economic costs in the implementation of these measures, and to prevent serious restriction of freedom of speech.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

The study was carried out with the financial support of the Russian Foundation for Basic Research in the framework of research project No. 18-29-16129

References

[1] R. Dremliuga, R. Iakovenko, N. Prisekina "Crime in virtual reality: Discussion", 2019 International Conference on Cybersecurity, ICoCSec 2019, 8970947, 81-85, 2019.

[2] A.O.J. Kwok, S.G.M. Koh "COVID-19 and Extended Reality (XR)", Current Issues in Tourism, 2020. https://www.tandfonline.com/doi/full/10.1080/13683500.2020.1798896

[3] Forecast augmented (AR) and virtual reality (VR) market size worldwide from 2016 to 2022 (in billion U.S. dollars). Statista, 2018. Retrieved October 26, 2020 from https://www.statista.com/statistics/591181/global-augmented-virtual-reality-market-size/.

[4] M. Meißner, J. Pfeiffer, C. Peukert, H. Dietrich, T. Pfeiffer "How virtual reality affects consumer choice", Journal of Business Research, 117, 219-231, 2020.

[5] M. Figueiredo, R. Mafalda, A. Kamensky "Virtual reality as an educational tool for elementary school", Smart Innovation, Systems and Technologies, 198 SIST, 261-267, 2021.

[6] Y.P. Zinchenko, P.P. Khoroshikh, A.A. Sergievich, S.A. Gutnikov, K.S. Golokhvast "Virtual reality is more efficient in learning human heart anatomy especially for subjects with low baseline knowledge", New Ideas in

Psychology, 59,100786, 2020. DOI: 10.1016/j.newideapsych.2020.100786
[7] S. Ahn, J. Bostick, E. Ogle, K.L. Nowak, K. McGillicuddy, J. Bailenson "Experiencing Nature: Embodying Animals in Immersive Virtual Environments Increases Inclusion of Nature in Self and Involvement with Nature", Journal of Computer-Mediated Communication, **21**(6), 399-419, 2016. https://doi.org/10.1111/jcc4.12173
[8] M.C. Howard "A meta-analysis and systematic literature review of virtual reality rehabilitation programs", Computers in Human Behavior, **70(**1), 317-327, 2017.
[9] B. Ticknor Virtual Reality and Correctional Rehabilitation: A Game Changer", Criminal Justice and Behavior, **46**(9), 1319-1336, 2019.
[10] B. Ticknor "Using virtual reality to treat offenders: An examination", International Journal of Criminal Justice Sciences, **13**(2), 316-325, 2018.
[11] G.P. Otjutsky "Understanding the virtual reality: methodological difficulties", Historical, philosophical, political and legal sciences, cultural studies and art criticism. Issues of theory and practice, **2-2** (52), 161-170, 2015.
[12] E.A. Solodkina "To the definition of the concept of virtual reality", Bulletin of the Peoples' Friendship University of Russia. Series: Philosophy, (1), 190-194, 2004.
[13] V. Jaillot, M. Istasse, S. Servigne, M. Rautenberg, I. Lefort, "Describing, comparing and analysing digital urban heritage tools: A methodology designed with a multidisciplinary approach", Digital Applications in Archaeology and Cultural Heritage, 17,e00135, 2020.
[14] P.P. Shrestha, B. Davis, G.M. Gad, "Investigation of Legal Issues in Construction-Manager-at-Risk Projects: Case Study of Airport Projects", Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, **2**(3),04520022, 2020.
[15] E. Fritsch "Growth in wearable technologies", Screen Printing, 110(5), 22-26, 2019.
[16] R. Dremliuga "Subculture of Hackers in Russia", Asian Social Science, **10**(18), 158-162, 2014. DOI: 10.5539/ASS.V10N18P158
[17] Mark A .Lemley, E. Volokh "Law, Virtual Reality, and Augmented Reality", University of Pennsylvania Law Review, **166**, 1051-1138, 2018. Available at SSRN: https://ssrn.com/abstract=2933867 or http://dx.doi.org/10.2139/ssrn.2933867
[18] Q . Wang, F. Hu, H.Wang "Are you afraid of heights and suitable for working at height?", Biomedical Signal Processing and Control, 52, 23-31, 2019.
[19] T.W. Cleworth, R. Chua, J.T. Inglis, M.G. Carpenter "Influence of virtual height exposure on postural reactions to support surface translations", Gait and Posture, 47, 96-102, 2016.
[20] S.N.V. Yuan, H.H.S. Ip "Using virtual reality to train emotional and social skills in children with autism spectrum disorder", London Journal of Primary Care, **10**(4), 110-112, 2018.
[21] J.-M. Park, G.-Y. Noh "The effects of vr simulation game on driving fear and efficacy", JP Journal of Heat and Mass Transfer, **15**(Special Issue 2), 161-169, 2018.
[22] Y.Y. Lim, P. Poronnik, T. Usherwood, B. Reeve "Medical negligence laws and virtual reality in healthcare", Australian journal of general practice, **49**(8), 525-529, 2020.
[23] R. Ueoka, A. AlMutawa, "Emotion hacking VR: Amplifying scary VR experience by accelerating actual heart rate", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10904 LNCS, 436-445, 2018.
[24] B. Meuleman, D. Rudrauf "Induction and profiling of strong multi-componential emotions in virtual reality", IEEE Transactions on Affective Computing, 2018, Available at: https://ieeexplore.ieee.org/document/8432120/
[25] J. Hvass, O. Larsen, K. Vendelbo "Visual realism and presence in a virtual reality game", 3DTV-Conference 2017-June, 1-4, 2018.
[26] A. Monti, S.M. Aglioti "Flesh and bone digital sociality: On how humans may go virtual", British Journal of Psychology, **109**(3), 418-420, 2018.
[27] A. Leleve, T. McDaniel, C. Rossa "Haptic Training Simulation", Frontiers in virtual reality, **1**(3), July 2020. https://doi.org/10.3389/frvir.2020.00003
[28] L. Strikwerda "Present and Future Instances of Virtual Rape in Light of Three Categories of Legal Philosophical Theories on Rape", Philosophy and Technology, **28**(4), 491-510, 2015.
[29] M.J. Faustino "Rebooting an Old Script by New Means: Teledildonics—The Technological Return to the 'Coital Imperative'", Sexuality and Culture, **22**(1), 243-257, 2018.
[30] M.U. Rafique, S.-C.S. Cheung "Tracking Attacks on Virtual Reality Systems", IEEE Consumer Electronics Magazine, **9**(2), 8977814, 41-46, 2020
[31] D. Duggan, C. Kingsley, M. Mazzeo, M. Jenkins "Exploring Extended Reality as a Simulation Training Tool Through Naturalistic Interactions and Enhanced Immersion", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11575 LNCS, 272-282, 2019.
[32] Jack M. Balkin, "Information Fiduciaries and the First Amendment", UC Davis Law Review, **49**(4), 183-1234, 2016. Available at SSRN: https://ssrn.com/abstract=2675270.
[33] K. Klonick. "The new governors: The people, rules, and processes governing online speech", Harvard Law Review, **131**(6), 1599-1670, 2018.
[34] T.A. Kochubey, "Legal protection of virtual and augmented reality technologies", Advances in Intelligent Systems and Computing, 1100 AISC, 184-191, 2020.
[35] B. Kenwright "Virtual Reality: Ethical Challenges and Dangers [Opinion]" IEEE Technology and Society Magazine, **37**(4),8558774, 20-25, 2018.
[36] M.J. Blitz "The First Amendment, video games, and virtual reality training" in "Research Handbook on the Law of Virtual and Augmented Reality", Edited by W. Barfield and M.J. Blitz, Edward Elgar Publishing, 712 p.2018.